

# A Bayesian Framework for Learning Shared and Individual Subspaces from Multiple Data Sources

Sunil Kumar Gupta, Dinh Phung, Brett Adams, and Svetha Venkatesh

Department of Computing  
Curtin University, Perth, Australia  
sunil.gupta@postgrad.curtin.edu.au,  
{d.phung,b.adams,s.venkatesh}@curtin.edu.au

**Abstract.** This paper presents a novel Bayesian formulation to exploit shared structures across multiple data sources, constructing foundations for effective mining and retrieval across disparate domains. We jointly analyze diverse data sources using a unifying piece of metadata (textual tags). We propose a method based on Bayesian Probabilistic Matrix Factorization (BPMF) which is able to explicitly model the partial knowledge common to the datasets using shared subspaces and the knowledge specific to each dataset using individual subspaces. For the proposed model, we derive an efficient algorithm for learning the joint factorization based on Gibbs sampling. The effectiveness of the model is demonstrated by social media retrieval tasks across single and multiple media. The proposed solution is applicable to a wider context, providing a formal framework suitable for exploiting individual as well as mutual knowledge present across heterogeneous data sources of many kinds.

## 1 Introduction

Recent developments in computing and information technology have enabled us to jointly analyze data from multiple sources such as multiple news feeds, social media streams and so on. Discovering structures and patterns from multiple data sources helps us to unravel certain commonalities and differences which otherwise is not possible when analyzing each data source separately. This information provides valuable inputs for various data mining and representation tasks. Whilst the data mining community has developed techniques to analyze a single data source, there is a need to develop formal frameworks for analyzing multiple data sources exploiting their common strengths.

However, modeling the data across multiple heterogeneous and disparate sources is a challenging task. For example, how do we model *text*, *image* and *video* together? At the semantic level, they provide much richer information together, and the question is how to exploit these strengths at lower levels? One solution is to exploit textual metadata for each data source in the form of *tags*. These tags are rich metadata sources, freely available across disparate data sources (images, videos, blogs etc.) and at topical or conceptual levels, they are often more meaningful than what current content processing methods extract [2]. However, tags can be ambiguous, incomplete and subjective [7] due to a lack of constraints during their creation. Due to this problem, performance of any data mining task using tags suffers significantly. Therefore, it is imperative to model the uncertainties of tag data to improve the performance of several data mining

tasks. Work on tag denoising has been, broadly speaking, aimed at determining tag relevance through modification or the recommendation of additional tags [11,7]. But these approaches typically focus solely within the internal structure of a given tagging source, and are thus bounded by the information content and noise characteristics of the tagging source. Moreover, these methods, working individually for each data source, lack in exploiting the collective strengths of all data sources.

Addressing the problem of constructing a unified framework for disparate sources, we develop a Bayesian framework which can model the uncertainties of multiple data sources jointly using the textual tags from each source as a unified piece of metadata. Our method allows multiple data sources to exploit collective strength by learning probabilistic shared subspaces and, at the same time, *crucially* retains the differences of each data source by learning probabilistic individual subspaces. Retaining the differences between data sources is very important; ignoring this aspect often leads to “negative knowledge transfer”. Another strength of the proposed framework is that both the shared and individual subspaces are probabilistic in nature, which helps in modeling the uncertainties involved in real world applications. Similar work by Gupta et al [5] also models the shared and individual subspaces but has certain limitations. First, their framework is restrictive in part as it can model only nonnegative data sources. Second, their model supports only two data sources, which renders the model unusable when working with more than two data sources. Third, the subspaces learnt are not probabilistic and do not cater for uncertainties such as missing tags and tag ambiguity.

Previous works on shared subspace learning are mainly focused on supervised or semi-supervised learning. Ji et al [6] and Yan et al [12] provide frameworks for extracting shared structures in multi-label classification by learning a common subspace which is assumed to be shared among multiple labels. Si et al [10] propose a family of transfer subspace learning algorithms by minimizing Bregman divergence between the distributions of the training and test samples. This approach, while being fairly generic for transfer learning, is not appropriate for multi-task learning and can not exploit the knowledge strengths from the multiple data sources. In another work, Gu and Zhou [4] propose multi-task clustering by learning a shared subspace for all tasks. This approach provides a way to learn shared subspaces, but has no way of controlling the sharing level, a crucial aspect when dealing with heterogeneous data sources. Moreover, the sharing is imposed among all tasks which is unrealistic in many scenarios.

Our framework is based on the state-of-the-art Bayesian probabilistic matrix factorization (BPMF) model, recently proposed in [9]. We extend BPMF to enable joint modeling of multiple data sources deriving common and individual subspaces, and derive inference algorithms using Rao-Blackwellized Gibbs Sampling (RBGS). To demonstrate the usefulness of our approach, we examine two applications – improving social media retrieval using auxiliary sources, and cross-social media retrieval. We use three disparate data sources – Flickr, YouTube and Blogspot – to show the effectiveness of shared subspace learning frameworks.

Our main contributions are :

- The construction of a novel Bayesian shared subspace learning framework for extraction of shared and individual subspaces across an arbitrary number of data sources using joint matrix factorization.
- Efficient inference based on RBGS (Gibbs sampling) for joint factorization.

- Algorithms for retrieval within one social medium or across disparate social media using Bayesian shared subspace learning framework.
- Two real-world applications using three popular social media sources: Blogspot, Flickr and YouTube. We demonstrate (1) improvement in social media retrieval by leveraging auxiliary media, and (2) effective cross-social media retrieval.

The novelty of our approach lies in the framework and algorithms for learning *across* diverse data sources. Our probabilistic shared and individual subspaces not only exploit the mutual strengths of multiple data sources but also handle the uncertainties.

The significance of our work lies in the fact that theoretical extensions made to BPMF [9] for multiple data sources allow for the flexible transfer of knowledge across disparate media. In addition, RBGS sampling derived for our model achieves better inference (i.e. Markov chain mixes better) compared to [9]. Our work brings a broad scope of open opportunities and applications—it is appropriate wherever one needs to exploit multiple and heterogeneous datasets for knowledge transfer among them, such as collaborative filtering or sentiment analysis.

The rest of the paper is organized as follows. Section 2 presents the Bayesian shared subspace learning formulation and describes Gibbs sampling based inference procedure. Section 3 and 4 demonstrate the applicability of the proposed framework to social media retrieval applications. Conclusions are drawn in Section 5.

## 2 Bayesian Shared Subspace Learning (BSSL)

We introduce a framework for learning individual and shared subspaces across an arbitrary number of data sources. Let a set of  $n$  data sources be represented by data matrices  $\mathbf{X}_1, \dots, \mathbf{X}_n$ , which, for example, can be term-document matrices (each row a word and each column a document with *tf-idf* features [1]) for retrieval applications or user rating matrices (each row a user and each column an item with ratings as features) for collaborative filtering applications. We assume that matrices  $\mathbf{X}_i$  have the same number of rows. Whenever it is not the case, one can always merge the vocabularies of each data source to form a common vocabulary. Our goal is to factorize each matrix  $\mathbf{X}_i$  as  $\mathbf{X}_i = \mathbf{W}_i \cdot \mathbf{H}_i + \mathbf{E}_i$  such that the decomposition captures arbitrary *sharing* of basis vectors among data sources whilst preserving their *individual* bases. For example, when  $n = 2$ , we create three subspaces: a shared subspace matrix  $W_{12}$  and two individual subspaces  $W_1, W_2$ . We thus write  $\mathbf{X}_1$  and  $\mathbf{X}_2$  as

$$\mathbf{X}_1 = \underbrace{[W_{12} \mid W_1]}_{\mathbf{W}_1} \underbrace{\begin{bmatrix} H_{1,12} \\ H_{1,1} \end{bmatrix}}_{\mathbf{H}_1} + \mathbf{E}_1 \text{ and } \mathbf{X}_2 = \underbrace{[W_{12} \mid W_2]}_{\mathbf{W}_2} \underbrace{\begin{bmatrix} H_{2,12} \\ H_{2,2} \end{bmatrix}}_{\mathbf{H}_2} + \mathbf{E}_2 \quad (1)$$

To define subspaces at data source level, we define  $\mathbf{W}_1 = [W_{12} \mid W_1]$  and  $\mathbf{W}_2 = [W_{12} \mid W_2]$  for the two data sources. Note however that the encoding coefficients corresponding to the shared subspace  $W_{12}$  are different, and thus, an extra subscript is used to make it explicit in  $H_{1,12}$  and  $H_{2,12}$ . *Notation-wise, we use bold symbols  $\mathbf{W}, \mathbf{H}$  to denote the entire decomposition at a data source level and normal capital letters  $W, H$  at the shared level.*  $\mathbf{E}_1$  and  $\mathbf{E}_2$  denote the residual factorization error.

To see how we can generalize these expressions for  $n$  datasets, we continue with this example by constructing the power set over  $\{1, 2\}$  as  $S(2) = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$ . Our intention is to create an index set over the subscripts  $\{1, 2, 12\}$  used in matrices presented in Eq (1) so that a summation can be conveniently written. To do so, we further use  $S(2, i)$  to denote the subset of  $S(2)$  in which only elements involving  $i$  are retained, i.e.  $S(2, 1) = \{\{1\}, \{1, 2\}\}$  and  $S(2, 2) = \{\{2\}, \{1, 2\}\}$ . With a slight relaxation in the set notation, we rewrite them as  $S(2, 1) = \{1, 12\}$  and  $S(2, 2) = \{2, 12\}$ . Thus, Eq (1) can be re-written as

$$\mathbf{X}_1 = \sum_{v \in \{1, 12\}} W_v \cdot H_{1,v} + \mathbf{E}_1 \text{ and } \mathbf{X}_2 = \sum_{v \in \{2, 12\}} W_v \cdot H_{2,v} + \mathbf{E}_2$$

For a set of  $n$  datasets, let  $S(n)$  denote the set of all subsets over  $\{1, 2, \dots, n\}$  and for each  $i = 1, \dots, n$ , denote by  $S(n, i) = \{v \in S(n) \mid i \in v\}$  the index set associated with the  $i$ -th data source. Our proposed shared subspace learning framework seeks a set of expression in the following forms for  $i = 1, \dots, n$

$$\mathbf{X}_i = \mathbf{W}_i \cdot \mathbf{H}_i + \mathbf{E}_i = \sum_{v \in S(n, i)} W_v \cdot H_{i,v} + \mathbf{E}_i, \quad i = 1, \dots, n \quad (2)$$

It is also clear from Eq (2) that the total subspace  $\mathbf{W}_i$  and its corresponding encoding matrix  $\mathbf{H}_i$  for the  $i$ -th data matrix are horizontally augmented matrices over all  $W_v$  and vertically augmented over all  $H_{i,v}$  for  $v \in S(n, i)$  respectively. That is, if we explicitly list the elements of  $S(n, i)$  as  $S(n, i) = \{v_1, v_2, \dots, v_Z\}$  then  $\mathbf{W}_i, \mathbf{H}_i$  are

$$\mathbf{W}_i = [W_{v_1} \mid W_{v_2} \mid \dots \mid W_{v_Z}] \text{ and } \mathbf{H}_i = \begin{bmatrix} H_{i,v_1} \\ \vdots \\ H_{i,v_Z} \end{bmatrix} \quad (3)$$

## 2.1 Bayesian Representation

We treat the residual errors ( $\mathbf{E}_i$ ) probabilistically and model each  $\mathbf{E}_i, \forall i$  as i.i.d. and normally distributed with mean zero and precisions  $\Lambda_{\mathbf{X}_i}$ . Although we consider Bayesian shared subspace learning for arbitrary number of data sources, for simplicity, we show the graphical model for the case of two data sources in Figure 1. For each  $i \in \{1, 2, \dots, n\}$  and  $v \in S(n, i)$ , our probabilistic model is then given as

$$\begin{aligned} p(\mathbf{X}_i(m, l) \mid \mathbf{W}_i, \mathbf{H}_i, \Lambda_{\mathbf{X}_i}) &= \mathcal{N}(\mathbf{X}_i(m, l) \mid \mathbf{W}_i^{(m)} \mathbf{H}_i^{[l]}, \Lambda_{\mathbf{X}_i}^{-1}) \\ p(W_v \mid \mu_{W_v}, \Lambda_{W_v}) &= \prod_{m=1}^M \mathcal{N}(W_v^{(m)} \mid \mu_{W_v}, \Lambda_{W_v}^{-1}) \\ p(\mathbf{H}_i \mid \mu_{\mathbf{H}_i}, \Lambda_{\mathbf{H}_i}) &= \prod_{l=1}^{N_i} \mathcal{N}(\mathbf{H}_i^{[l]} \mid \mu_{\mathbf{H}_i}, \Lambda_{\mathbf{H}_i}^{-1}) \end{aligned}$$

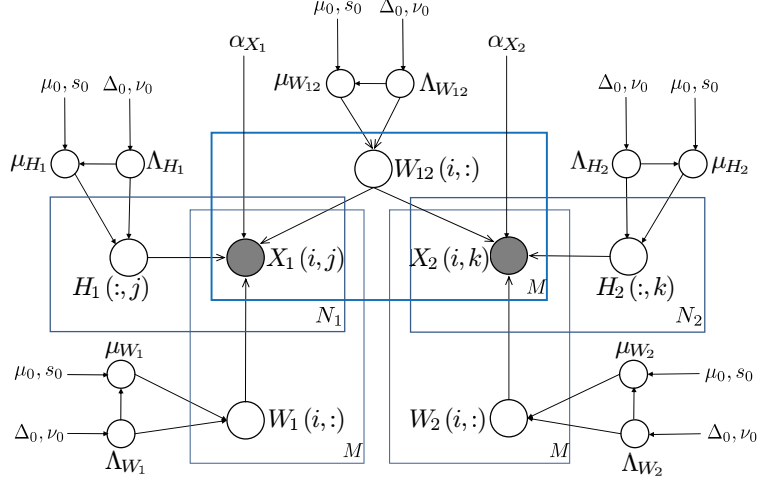


Fig. 1: Graphical Model for BSSL (a special case for two data sources, i.e.  $n = 2$ )

where  $\mathbf{B}^{(m)}$  denotes the  $m$ -th row, i.e.  $\mathbf{B}(m, :)$  while  $\mathbf{B}^{[l]}$  denotes the  $l$ -th column, i.e.  $\mathbf{B}(:, l)$  of a matrix  $\mathbf{B}$ . Since  $\mathbf{E}_i$ 's are i.i.d., we set  $\Lambda_{\mathbf{X}_i} = \alpha_{\mathbf{X}_i} \mathbf{I}$  for each  $i$ . Going fully Bayesian, we further use a normal-Wishart prior on the parameters  $\{\mu_{W_v}, \Lambda_{W_v}\}$ . The normal-Wishart prior is given by

$$p(\mu_{W_v}, \Lambda_{W_v} | \Psi_0) = \mathcal{N}(\mu_{W_v} | \mu_0, (s_0 \Lambda_{W_v})^{-1}) \mathcal{W}(\Lambda_{W_v} | \Delta_0, \nu_0)$$

where  $\mathcal{W}(\cdot | \Delta_0, \nu_0)$  is Wishart distribution with  $K_v \times K_v$  scale matrix  $\Delta_0$  and  $\nu_0$  degree of freedom. Similar priors are placed on the parameters  $\{\mu_{\mathbf{H}_i}, \Lambda_{\mathbf{H}_i}\}$ . For future reference, we define all the hyperparameters as  $\Psi_0 \triangleq \{\mu_0, s_0, \Delta_0, \nu_0, \alpha_{\mathbf{X}_1}, \dots, \alpha_{\mathbf{X}_n}\}$ .

## 2.2 Gibbs Inference

Given data matrices  $\{\mathbf{X}_i\}_{i=1}^n$ , the goal of BSSL is to learn the factor matrices  $W_v$  and  $\mathbf{H}_i$  for all  $i \in \{1, 2, \dots, n\}$  and  $v \in S(n, i)$ . In our Bayesian setting, this translates to performing posterior inference on the distribution of random (row or column) vectors from  $W_v$  and  $\mathbf{H}_i$ . Since we are treating these vectors as Gaussian with proper conjugate prior normal-Wishart distribution, posterior inference can be conveniently carried out using Gibbs sampling, which is guaranteed to converge asymptotically.

In a typical Gibbs sampling setting, our state space for sampling is  $\{W_v, \mu_{W_v}, \Lambda_{W_v}\}$  and  $\{\mathbf{H}_i, \mu_{\mathbf{H}_i}, \Lambda_{\mathbf{H}_i}\}$  conditioned on the hyperparameters  $\Psi_0$  and data  $\{\mathbf{X}_i\}_{i=1}^n$ . However,  $\{\mu_{W_v}, \Lambda_{W_v}\}$  and  $\{\mu_{\mathbf{H}_i}, \Lambda_{\mathbf{H}_i}\}$  are nuisance parameters which can be integrated out to reduce the variance of the Gibbs samples (for better mixing of Markov chain) – a scheme which is known as Rao-Blackwellized Gibbs Sampling (RBGS).

After integrating out these nuisance parameters, our state space reduces to only the factor matrices  $\{W_v, \mathbf{H}_i\}$  for all  $i \in \{1, 2, \dots, n\}$  and  $v \in S(n, i)$ . Our Gibbs sampler then iteratively samples each row of  $W_v$  and column of  $\mathbf{H}_i$  conditioned on the observed data and the remaining set of variables in the state space from the previous

Gibbs iteration. Algorithm 1 outlines these sampling steps while the rest of this section shall briefly explain how to obtain the Gibbs conditional distributions as in Eqs (6-9).

Recalling  $\mathbf{W}_i$  and  $\mathbf{H}_i$  from Eq. (3), the conditional distribution over  $W_v^{(m)}$ , conditioned on matrices  $W_u$  for all  $u \in S(n, i)$  except  $v$ , all the rows of matrix  $W_v$  except  $m$ -th row (denoted by  $W_v^{(\setminus m)}$ ), the coefficient matrices  $\mathbf{H}_i$  for all  $i$ , observed data  $\mathbf{X}_i$  for all  $i$  and the hyperparameters  $\Psi_0$ , is given by

$$p\left(W_v^{(m)} \mid \mathbf{X}_{1:n}, W_v^{(\setminus m)}, W_{\{u \in S(n, i)\} \setminus v}, \mathbf{H}_{1:n}, \Psi_0\right) \propto \left[ \prod_{i=1}^n \prod_{l=1}^{N_i} p\left(\mathbf{X}_i(m, l) \mid \mathbf{W}_i^{(m)} \mathbf{H}_i^{[l]}, \Lambda_{\mathbf{X}_i}^{-1}\right) \right] \times p\left(W_v^{(m)} \mid W_v^{(\setminus m)}, \Psi_0\right) \quad (4)$$

---

**Algorithm 1** Rao-Blackwellized Gibbs Sampling (RBGS) for BSSL.

---

- 1: **Input:** Hyperparameters  $\Psi_0$ , number of samples  $L$ .
  - 2: For each  $i$ , initialize matrices  $\mathbf{W}_i$ ,  $\mathbf{H}_i$  randomly.
  - 3: **for**  $r = 1, \dots, L$  **do**
  - 4:   For each  $v$ , draw  $r$ -th sample  $[W_v]^r$  from normal distribution parameterized by Eqs.(6–7).
  - 5:   For each  $i$ , draw  $r$ -th sample  $[\mathbf{H}_i]^r$  from normal distribution parameterized by Eqs.(8–9).
  - 6: **end for**
  - 7: For each  $v$  and  $i$ , get an estimate of  $W_v$  and  $\mathbf{H}_i$  using the Gibbs samples as  $W_v \approx \frac{1}{L} \sum_{r=1}^L [W_v]^r$ ,  $\mathbf{H}_i \approx \frac{1}{L} \sum_{r=1}^L [\mathbf{H}_i]^r$ .
  - 8: **Output:** Samples  $\{[W_v]^r\}_{r=1}^L$ ,  $\{[\mathbf{H}_i]^r\}_{r=1}^L$  and estimates  $W_v$ ,  $\mathbf{H}_i$  for each  $v$  and  $i$ .
- 

Note that the above posterior is proportional to the data-likelihood as a function of  $W_v^{(m)}$  and  $\mathbf{H}_i$  for each  $i$  and the predictive distribution of  $W_v^{(m)}$  given  $W_v^{(\setminus m)}$ . The predictive distribution of  $W_v^{(m)}$  conditioned on  $W_v^{(\setminus m)}$  and  $\Psi_0$  is obtained by integrating over the parameters of the normal–inverse–Wishart posterior distribution and is *multivariate Student-t* [3]. Assuming  $\nu_l > K_v + 1$ , this predictive density has finite covariance and is known to be approximated well by a normal distribution through matching the first two moments [3]. Thus, the predictive distribution is given as

$$p\left(W_v^{(m)} \mid W_v^{(\setminus m)}, \Psi_0\right) \approx \mathcal{N}\left(W_v^{(m)} \mid \mu_{W_v^{(m)}}^{pred}, \Lambda_{W_v^{(m)}}^{pred}\right) \quad (5)$$

$$\text{where } \mu_{W_v^{(m)}}^{pred} = \frac{s_0 \mu_0 + \sum_{\substack{l=1 \\ l \neq m}}^M W_v^{(l)}}{s_0 + (M - 1)}, \Lambda_{W_v^{(m)}}^{pred} = \frac{(s_m + 1) \Delta_m^{-1}}{s_m (\nu_m - K_{W_v} - 1)},$$

$$\Delta_m^{-1} = \Delta_0^{-1} + \sum_{\substack{l=1 \\ l \neq m}}^M W_v^{(l)} \left(W_v^{(l)}\right)^\top + \frac{s_0 (M - 1)}{s_0 + M - 1} (\mu_0 - \bar{\mu}_{W_v \setminus m}) (\mu_0 - \bar{\mu}_{W_v \setminus m})^\top,$$

$$\bar{\mu}_{W_v \setminus m} \triangleq \frac{1}{(M-1)} \sum_{\substack{l=1 \\ l \neq m}}^M W_v^{(l)}, \quad s_m = s_0 + M - 1, \quad \nu_m = \nu_0 + M - 1, \quad W_v \in \mathbb{R}^{M \times K_v}$$

Using Eqs (4) and (5), the posterior distribution can be written as

$$p\left(W_v^{(m)} \mid \mathbf{X}_{1:n}, W_v^{(\setminus m)}, W_{\{u:u \neq v\}}, \mathbf{H}_{1:n}, \Psi_0\right) = \mathcal{N}\left(W_v^{(m)} \mid, \mu_{W_v^{(m)}}^{post}, \Lambda_{W_v^{(m)}}^{post}\right) \text{ where}$$

$$\Lambda_{W_v^{(m)}}^{post} = \Lambda_{W_v^{(m)}}^{pred} + \sum_{i=1}^n H_{i,v} \Lambda_{\mathbf{X}_i} H_{i,v}^\top \quad (6)$$

$$\left(\mu_{W_v^{(m)}}^{post}\right)^\top = \left[\Lambda_{W_v^{(m)}}^{post}\right]^{-1} \left[\Lambda_{W_v^{(m)}}^{pred} \left(\mu_{W_v^{(m)}}^{pred}\right)^\top + \sum_{i=1}^n H_{i,v} \Lambda_{\mathbf{X}_i} \left(\mathbf{A}_{i,v}^{(m)}\right)^\top\right] \quad (7)$$

and  $\mathbf{A}_{i,v}^{(m)} \triangleq \mathbf{X}_i^{(m)} - \sum_{\{u:u \neq v\}} W_u^{(m)} H_{i,u}$ . Similar to  $W_v^{(m)}$ , the posterior distribution over the  $l$ -th column of matrix  $\mathbf{H}_i$  conditioned on its remaining columns is normally distributed with mean vector and precision matrix as

$$\Lambda_{\mathbf{H}_i^{[l]}}^{post} = \Lambda_{\mathbf{H}_i^{[l]}}^{pred} + \mathbf{W}_i^\top \Lambda_{\mathbf{X}_i} \mathbf{W}_i \quad (8)$$

$$\mu_{\mathbf{H}_i^{[l]}}^{post} = \left[\Lambda_{\mathbf{H}_i^{[l]}}^{post}\right]^{-1} \left[\Lambda_{\mathbf{H}_i^{[l]}}^{pred} \mu_{\mathbf{H}_i^{[l]}}^{pred} + \mathbf{W}_i^\top \Lambda_{\mathbf{X}_i} \mathbf{X}_i^{[l]}\right] \quad (9)$$

### 2.3 Subspace Dimensionality and Complexity Analysis

Let the number of rows in  $\mathbf{X}_1, \dots, \mathbf{X}_n$  be  $M$  and the number of columns be  $N_i$ , giving  $M \times N_i$  dimension for  $\mathbf{X}_i$ . Since each  $\mathbf{W}_i$  consists of an augmentation of individual and shared subspaces  $W_v$ , we use  $K_v$  to denote the number of basis vectors in  $W_v$ . Assuming  $R_i$  to be the total number of basis vectors in  $\mathbf{W}_i$ , we have  $\sum_{v \in \mathcal{S}(n,i)} K_v = R_i$ . Determining the value of  $K_v$  is a model selection problem and depends upon the common features among various data sources. According to a heuristic proposed in [8], a rule of thumb is to use  $K_v \approx \sqrt{M_v/2}$  where  $M_v$  is the number of common features in sharing configuration  $v$ . Following the above notation for the dimensionalities of matrices, for each  $i \in \{1, 2, \dots, n\}$ , assuming  $R_i < M$  (generally the case for real world data), the complexity of sampling  $\mathbf{H}_i$  matrices from its posterior distributions is  $\mathcal{O}(M \times N_i \times R_i)$  whereas the complexity involved in sampling  $\mathbf{W}_i$  matrices from its posterior distributions is  $\mathcal{O}(M \times N_i \times R_i^2)$ . Thus the computation complexity of BSSL remains similar to BPFM model [9] and does not grow any further.

## 3 Social Media Applications

We demonstrate the usefulness of BSSL in two real world applications. (1) Improving social media retrieval in one medium by transferring knowledge from auxiliary media sources. (2) Performing retrieval across multiple media. The first application can be seen as an multi-task learning application, whereas the second application is a direct manifestation of mining from multiple data sources.

### 3.1 BSSL based Social Media Retrieval

Let the tag-item matrix of the target medium (from which retrieval is to be performed) be denoted as  $\mathbf{X}_k$ . Further, let us assume that we have many other auxiliary media sources which share some features with the target medium. Let the tag-item matrices of these auxiliary media be denoted by  $\mathbf{X}_j$ ,  $j \neq k$ . In a multi-task learning setting, we leverage these auxiliary sources to improve the retrieval precision for the target medium, and given a set of query keywords  $S_Q$ , a vector  $q$  of length  $M$  (vocabulary size) is constructed by putting *tf-idf* values at each index where the vocabulary contains a word from the keywords set or else setting it to zero. Next, we follow Algorithm 2 for BSSL based retrieval.

### 3.2 BSSL based Cross-Social Media Retrieval

To retrieve items across media, we use the common subspace among them along with the corresponding coefficient matrices for each medium. As an example, for  $n = 3$  (three media sources), we use the common subspace matrix  $W_{123}$  and coefficient matrices  $H_{1,123}$ ,  $H_{2,123}$  and  $H_{3,123}$  for first, second and third medium respectively.

Similar to subsection 3.1, we construct a vector  $q$  of length  $M$  using a set of query keywords  $S_Q$ . We proceed similar to Algorithm 2 with the following differences. Given  $q$ , we wish to retrieve relevant items from each domain, which is performed by projecting  $q$  onto the augmented common subspace matrix ( $W_{123}$  for the case when  $n = 3$  media sources) to get its representation  $h$  in the common subspace. Next, we compute similarity between  $h$  and the columns of matrices  $H_{1,123}$ ,  $H_{2,123}$  and  $H_{3,123}$  (the representation of media items in the common subspace spanned by columns of  $W_{123}$ ) to find similar items from medium 1, 2 and 3 respectively. The results are ranked based on these similarity scores either individually or jointly.

For both retrieval applications, we use *cosine-similarity* as it seems to be more robust than Euclidean distance based similarity measures in high-dimensional spaces. As we are dealing with distributions, we also tried out *KL-divergence* based similarity measures, but *cosine-similarity* gives better results.

---

**Algorithm 2** Social Media Retrieval using BSSL.

---

- 1: **Input:** Target  $\mathbf{X}_j$ , auxiliaries  $\mathbf{X}_k$ ,  $k \neq j$ , query  $q$ , set of items from medium  $j$ , denoted as  $\mathcal{I} = \{I_1, I_2, \dots, I_{N_j}\}$ , number of items to be retrieved  $N$ .
  - 2: Get Gibbs estimates of  $\mathbf{W}_j$  and  $\mathbf{H}_j$  using Algorithm 1.
  - 3: Project  $q$  onto the subspace spanned by  $\mathbf{W}_j$  to get  $h$  as  $h = \mathbf{W}_j^\dagger q$  where  $\dagger$  is Moore-Penrose pseudoinverse of a matrix.
  - 4: For each item (indexed by  $m$ ) in  $\mathbf{X}_j$ , with its subspace representation  $h_m = m$ -th column of  $\mathbf{H}_j$ , compute its cosine similarity with query projection  $h$ :  $\text{sim}(h, h_m) = \frac{h^\top h_m}{\|h\|_2 \|h_m\|_2}$
  - 5: **Output:** Return the top  $N$  items in decreasing order of similarity.
-

Table 1: Description of the YouTube-Flickr-Blogspot data set.

<i>Media</i>	<i>Dataset Size</i>	<i>Concepts Used for Creating Dataset</i>	<i>Avg. Tags/Item (rounded)</i>
Blogspot	10000	'Academy Awards', 'Australian Open', 'Olympic Games', 'US Election', 'Cricket World Cup', 'Christmas', 'Earthquake'	6
Flickr	20000	'Academy Awards', 'Australian Open', 'Olympic Games', 'US Election', 'Holi', 'Terror Attacks', 'Christmas'	8
YouTube	7000	'Academy Awards', 'Australian Open', 'Olympic Games', 'US Election', 'Global Warming', 'Terror Attacks', 'Earthquake'	7

## 4 Experiments

### 4.1 Dataset

To conduct the experiments, we created a social media dataset using three different sources : YouTube<sup>1</sup>, Flickr<sup>2</sup> and Blogspot<sup>3</sup>. To obtain the data, we first chose common concepts – ‘Academy Awards’, ‘Australian Open’, ‘Olympic Games’, ‘US Election’ – and queried all three websites using their service APIs. To have some pairwise sharing, we additionally used the concept ‘Christmas’ to query Blogspot and Flickr, ‘Terror Attacks’ to query YouTube and Flickr, and ‘Earthquake’ to query Blogspot and YouTube. Lastly, to retain some differences between each medium, we used the concepts ‘Cricket World Cup’, ‘Holi’ and ‘Global Warming’ to query Blogspot.com, Flickr and YouTube respectively. Table 1 provides further details of the dataset size and average tag counts for each data source. The total number of tags from all three sources combined is 3740.

### 4.2 Subspace Learning and Parameter Setting

For clarity, let us denote YouTube, Flickr and Blogspot tag-item matrices as  $\mathbf{X}_1$ ,  $\mathbf{X}_2$  and  $\mathbf{X}_3$  respectively. To learn BSSL based factorization, we use Eqs (6)–(9) to sample  $W$  and  $H$  matrices. Recalling the notation  $K_v$  (dimensionality of subspace spanned by  $W_v$ ), for learning factorization, we set the individual subspace dimensions as  $K_1 = K_2 = K_3 = 10$ , pairwise shared subspace dimensions as  $K_{12} = K_{23} = K_{13} = 15$ , and the common to all subspace dimension as  $K_{123} = 25$ . To obtain these parameters, we first initialize them using the heuristic described in [8] and then do cross-validation based on retrieval precision. In addition, we also set the error precisions  $\alpha_{\mathbf{X}_1} = \alpha_{\mathbf{X}_2} = \alpha_{\mathbf{X}_3} = 2$ , hyperparameters  $\mu_0 = [0, \dots, 0]^T$ ,  $s_0 = 1$ ,  $\Delta_0 = \mathbf{I}$  and  $\nu_0 = K_v$  for corresponding  $W_v$ ,  $H_{i,v}$ . The values of  $\alpha_{\mathbf{X}_i}$  depend upon the quality of the tags and a small value implies high tag uncertainty. For the dataset described

<sup>1</sup> <http://code.google.com/apis/youtube/overview.html>

<sup>2</sup> <http://www.flickr.com/services/api/>

<sup>3</sup> <http://www.blogger.com/>

above, Gibbs sampling usually takes around 50 iterations to converge (convergence plots are omitted due to space restrictions), however, we collect 100 samples to ensure convergence. The first 20 samples are rejected for “burn-in” and the remaining 80 Gibbs samples are averaged to get an estimate of  $W_v, H_{i,v}$  matrices.

### 4.3 Experiment 1 : Social Media Retrieval using Auxiliary Sources

To carry out our experiments, we choose YouTube as the target dataset and Blogspot and Flickr as auxiliary datasets. To perform BSSL based retrieval from YouTube, we first generate samples of basis matrix  $\mathbf{W}_1 \triangleq [W_1 | W_{12} | W_{13} | W_{123}]$  and representation matrix  $\mathbf{H}_1 \triangleq [H_{1,1} | H_{1,12} | H_{1,13} | H_{1,123}]$  according to Eqs (6)–(9) and then get an estimate of  $\mathbf{W}_1$  and  $\mathbf{H}_1$  following Algorithm 2.

To compare the performance with other methods, we choose three baselines. The first baseline performs retrieval by matching the tag-lists of videos without any subspace learning. To get the similarity with other items, Jaccard coefficient<sup>4</sup> is used and the results are ranked based on the similarity scores. The second baseline is retrieval work based on subspace learning using Principle Component Analysis (PCA). For the third baseline, we use a recent state-of-the-art BPMF model proposed in [9] for Bayesian matrix factorization. For both second and third baselines, we do not use any auxiliary data (e.g. tags of Flickr or Blogspot) and use the tags of YouTube only, but increase the YouTube datsize so as to keep the total datsize equal to the combined data (target + auxiliary) used for the first baseline to make the comparison fair.

To evaluate our retrieval algorithm, we use a query set of 20 concepts defined as  $\mathbb{Q} = \{ \text{‘beach’, ‘america’, ‘bomb’, ‘animal’, ‘bank’, ‘movie’, ‘river’, ‘cable’, ‘climate’, ‘federer’, ‘disaster’, ‘elephant’, ‘europe’, ‘fire’, ‘festival’, ‘ice’, ‘obama’, ‘phone’, ‘santa’, ‘tsunami’} \}$ . Since there is no public groundtruth available, we manually go through the set of retrieved items and evaluate the results.

Figure 2 compares the retrieval performance of BSSL with all the baselines in terms of *precision-scope (P@N) curve*<sup>5</sup>, *mean average precision (MAP)* and *11-point precision-recall curve* [1]. Figure 2 clearly shows that BSSL outperforms the baselines in terms of all three evaluation criteria. Although, BPF performs better than PCA due to its ability to handle uncertainties well, it can not surpass BSSL as it is confined to the tag data of YouTube only. Intuitively, BSSL is able to utilize the related data from auxiliary sources and resolve the tag ambiguities by reducing the subjectivity and incompleteness of YouTube tags. In essence, the use of multiple sources in subspace learning helps discover improved tag co-occurrences and gives better results.

### 4.4 Experiment 2 : Cross Media Retrieval

The effectiveness of BSSL for cross-media retrieval is demonstrated using the YouTube-Flickr-Blogspot dataset with the subspace learning parameters as in subsection 4.2. For evaluation, we again use *precision-scope (P@N)*, *mean average precision (MAP)* and *11-point interpolated precision-recall curves*. Let  $q$  be a query term,  $G_i$  be the ground

<sup>4</sup> Jaccard  $(A, B) = |A \cap B| / |A \cup B|$ .

<sup>5</sup> For example, P@10 is the retrieval precision when considering the top 10 retrieved items.

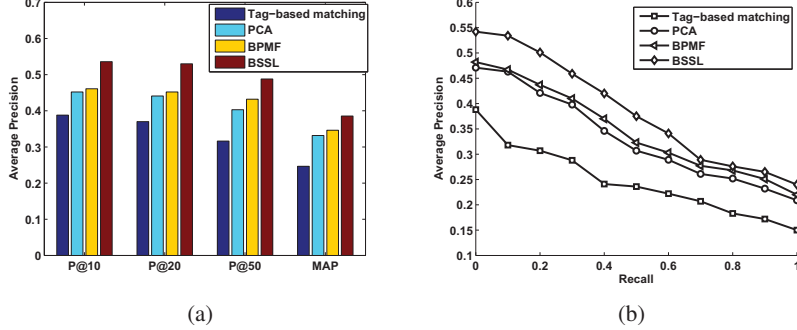


Fig. 2: YouTube retrieval results using auxiliary sources Flickr and Blogspot (a) Precision-Scope, MAP (b) 11-point interpolated Precision-Recall; for tag-based matching (baseline 1), PCA (baseline 2), BPMF [9] (baseline 3) and proposed BSSL.

truth set for the  $i$ -th medium and  $A_i$  be the answer set for query  $q$  using a retrieval method for the  $i$ -th medium. Then, the precision and recall measures for cross-media retrieval are

$$Precision = \frac{\sum_{i=1}^n |A_i \cap G_i|}{\sum_{i=1}^n |A_i|}, \quad Recall = \frac{\sum_{i=1}^n |A_i \cap G_i|}{\sum_{i=1}^n |G_i|}$$

As far as baselines are concerned, we note that both BPMF and PCA are not applicable for cross-media retrieval as they do not support analysis of multiple data sources in their standard form. Therefore, we compare the performance of BSSL against tag-based matching (based on Jaccard coefficient without any subspace learning) only. Other works on cross-media retrieval [13,14] use the concept of a Multimedia Document (MMD), which requires *co-occurring* multimedia objects on the same webpage which is not available in our case. Therefore, these methods can not be applied directly.

Figure 3 depicts the cross-media retrieval results across all three media - Blogspot, Flickr and YouTube. To generate the graphs, we again average the retrieval results over the query set  $\mathbb{Q}$  defined in subsection 4.3. It can be seen from Figure 3 that BSSL significantly outperforms tag based matching in terms of all three evaluation criteria. Improvement in terms of *MAP* criteria is around 13%. This improvement in performance is due to the learning of shared subspaces which not only handle the problem of ‘synonymy’ and ‘polysemy’ in tag-space, but also the uncertainties probabilistically.

## 5 Conclusion

We have presented a Bayesian framework to learn individual and shared subspaces from multiple data sources (BSSL) and demonstrated its application to social media retrieval across single and multiple media. Our framework, being based on the principle of Bayesian probabilistic matrix factorization (BPMF) [9], provides an efficient algorithm to learn the subspace. Our Gibbs sampler (RBGS) provides better Markov chain mixing than BPMF without increasing the complexity of the model. Our experiments have demonstrated that BSSL significantly outperforms the baseline methods for both retrieval tasks on Blogspot, Flickr and YouTube datasets. More importantly, our solution

provides a generic framework to exploit collective strengths from heterogeneous data sources and we foresee its wider adoption in cross-domain data mining and beyond.

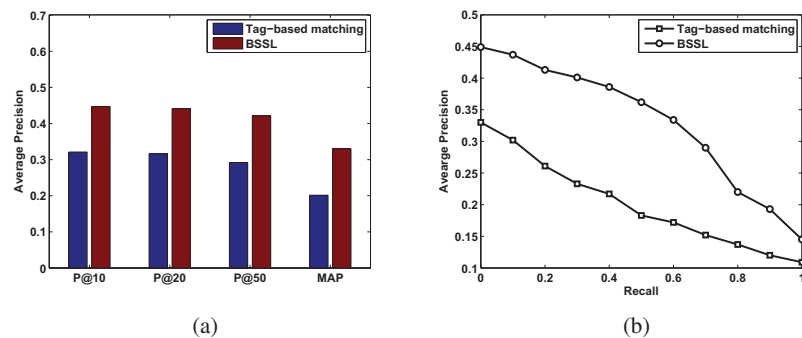


Fig. 3: Cross-media retrieval results: (a) precision-scope, MAP (b) 11-point interpolated precision-recall; for tag-based matching and proposed BSSL.

## References

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern information retrieval. Addison-Wesley (1999)
2. Datta, R., Joshi, D., Li, J., Wang, J.: Image retrieval: ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)* 40(2), 1–60 (2008)
3. Gelman, A.: Bayesian data analysis. CRC press (2004)
4. Gu, Q., Zhou, J.: Learning the shared subspace for multi-task clustering and transductive transfer classification. *ICDM* pp. 159–168 (2009)
5. Gupta, S.K., Phung, D., Adams, B., Tran, T., Venkatesh, S.: Nonnegative shared subspace learning and its application to social media retrieval. *SIGKDD* pp. 1169–1178 (2010)
6. Ji, S., Tang, L., Yu, S., Ye, J.: Extracting shared subspace for multi-label classification. *SIGKDD* pp. 381–389 (2008)
7. Li, X., Snoek, C.G.M., Worring, M.: Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia* 11(7), 1310–1322 (2009)
8. Mardia, K.V., Bibby, J.M., Kent, J.T.: Multivariate analysis. Academic Press, NY (1979)
9. Salakhutdinov, R., Mnih, A.: Bayesian probabilistic matrix factorization using markov chain monte carlo. *ICML* pp. 880–887 (2008)
10. Si, S., Tao, D., Geng, B.: Bregman divergence based regularization for transfer subspace learning. *IEEE Transactions on Knowledge and Data Engineering* 22(7), 929–942 (2009)
11. Sigurbjörnsson, B., Van Zwol, R.: Flickr tag recommendation based on collective knowledge. *WWW* pp. 327–336 (2008)
12. Yan, R., Tesic, J., Smith, J.: Model-shared subspace boosting for multi-label classification. *SIGKDD* pp. 834–843 (2007)
13. Yang, Y., Xu, D., Nie, F., Luo, J., Zhuang, Y.: Ranking with local regression and global alignment for cross media retrieval. *MM* pp. 175–184 (2009)
14. Yi, Y., Zhuang, Y., Wu, F., Pan, Y.: Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. *IEEE Transactions on Multimedia* 10(3), 437–446 (2008)