Unsupervised Modeling of Multiple Data Sources : A Latent Shared Subspace Approach

Sunil Kumar Gupta

December 2011

Department of Computing

Unsupervised Modeling of Multiple Data Sources : A Latent Shared Subspace Approach

Sunil Kumar Gupta

This thesis is presented for the Degree of
Doctor of Philosophy
of

Curtin University

December 2011

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledge has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature: ....
Date: ....

Contents

List of Figures

List of Tables

List of Algorithms

Abstract

The growing number of information sources has given rise to joint analysis. While the research community has mainly focused on analyzing data from a single source, there has been relatively few attempts on *jointly* analyzing multiple data sources exploiting their statistical sharing strengths. In general, the data from these sources emerge without labeling information and thus it is imperative to perform the joint analysis in an *unsupervised* manner.

This thesis addresses the above problem and presents a general shared subspace learning framework for jointly modeling multiple related data sources. Since the data sources are related, there exist common structures across these sources, which can be captured through a *shared* subspace. However, each source also has some individual structures, which can be captured through an *individual* subspace. Incorporating these concepts in nonnegative matrix factorization (NMF) based subspace learning, we develop a nonnegative shared subspace learning model for two data sources and demonstrate its application to tag based social media retrieval. Extending this model, we impose additional regularization constraints of mutual orthogonality on the shared and individual subspaces and show that, compared to its unregularized counterpart, the new regularized model effectively deals with the problem of negative knowledge transfer – a key issue faced by transfer learning methods. The effectiveness of the regularized model is demonstrated through retrieval and clustering applications for a variety of data sets. To take advantage from more than one auxiliary source, we extend above models generalizing two sources to multiple sources with an added flexibility of allowing sources having arbitrary sharing configurations. The usefulness of this model is demonstrated through improved performance, achieved with multiple auxiliary sources. In addition, this model is used to relate the items from disparate media types allowing us to perform cross-media retrieval using tags.

Departing from the nonnegative models, we use a linear-Gaussian framework and develop Bayesian shared subspace learning, which not only models the mixed-sign data but also learns probabilistic subspaces. Learning the subspace dimensionalities for the shared subspace models has an important role in optimum knowledge transfer but requires model selection – a task that is computationally intensive and time consuming. To this end, we propose a nonparametric Bayesian joint factor analysis model that circumvents the problem of model selection by using a hierarchical beta process prior, inferring subspace dimensionalities automatically from the data. The effectiveness of this model is shown on both synthetic and real data sets. For synthetic data set, successful recovery of both shared and individual subspace dimensionalities is demonstrated, whilst for real data set, the model outperforms recent state-of-the-art techniques for text modeling and image retrieval.

Acknowledgements

This thesis marks the culmination of an intense yet gratifying journey at Curtin that started nearly 2.5 years ago. During this period, I was fortunate to have worked with some great people who had been inspirational and of great importance for shaping the research contained in this thesis.

I am particularly grateful to my supervisor, Prof. Svetha Venkatesh, for the tremendous freedom I have enjoyed during my time at IMPCA, and for her support and advice when came time to make important decisions, in research as well as in life. I am grateful to my co-supervisor, Dr. Dinh Phung, for providing me a warm and outstanding intellectual environment at IMPCA, for having countless passionate discussions on almost everything that came on my way. Working with Dinh has provided me an unending supply of inspiration, guidance and feedback. Things would have been much more difficult for me without his support. I would like to express my sincere gratitude to my other co-supervisor Dr. Brett Adams for his invaluable ideas, independent view points, useful discussions and feedbacks, which have been of great importance for this thesis.

In addition to my supervisors, the other close colleagues and friends at IMPCA – Sonny, Truyen, Kurt, Thin, Buddha and Santu – have played an important role towards this thesis by providing stimulating conversations, helping in programming related issues and tolerating my never-ending interruptions. We have had many interesting discussions regarding work and other useful matters and I look forward to their continued collaborations. My special thanks go to the folks who enlivened my life on meetings over coffee/meals or get-togethers!

Other staff members – Mary Simpson, Wanquan and Patrick – were also helpful. I especially thank Ms Mary Mulligan and Mr. Jeff Bilman for their administrative support and invaluable feedbacks while proofreading this thesis.

On a more personal note, I would like to thank my parents for their continuing support 5,000 miles away from home during my whole adventure. I dedicate this thesis to them. Finally and more than all, I am grateful to my wife Shweta for her steady support and love during good and bad times. Without her, much of this endeavor would have been meaningless.

Relevant Publications

Part of this thesis has been published or documented elsewhere. The details of these publications are as follows

- Chapter 3:
  - S. K. Gupta, D. Phung, B. Adams, T. Truyen and S. Venkatesh. Nonnegative shared subspace learning and its application to social media retrieval. In *Proceedings of the 16th ACM International Conference on Knowledge Discovery and Data Mining* (KDD), pages 1169–1178, ACM 2010.
- Chapter 4:
  - S. K. Gupta, D. Phung, B. Adams, S. Venkatesh. A Matrix Factorization Framework for Jointly Analyzing Multiple Nonnegative Data Sources. In *Proceedings of the 9th Text Mining Workshop, held in conjunction with the 11th SIAM International Conference on Data Mining* (SDM), 30th Apr, Arizona, USA, 2011.
- Chapter 5:
  - S. K. Gupta, D. Phung, B. Adams, S. Venkatesh. Regularized Nonnegative Shared Subspace Learning. *Data Mining and Knowledge Discovery* (DAMI), Springerlink:10.1007/s10618-011-0244-8, Springer 2011.
- Chapter 6:
  - S. K. Gupta, D. Phung, B. Adams and S. Venkatesh. A Bayesian Framework for Learning Shared and Individual Subspaces from Multiple Data Sources. In *Proceedings of the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (PAKDD), pages 136–147, Springer 2011.

Abbreviations

BP	Beta Process
BeP	Bernoulli Process
BSSL	Bayesian Shared Subspace Learning
CCA	Canonical Correlation Analysis
CRP	Chinese Restaurant Process
DBSCAN	Density Based Spatial Clustering of Applications with Noise
DP	Dirichlet Process
EM	Expectation-Maximization
FA	Factor Analysis
HBP	Hierarchical Beta Process
HDP	Hierarchical Dirichlet Process
IBP	Indian Buffet Process
i.i.d.	independent and identically distributed
MAP	Mean Average Precision
MCMC	Markov Chain Monte Carlo
ML	Maximum Likelihood
MS-NMF	Multiple Shared Nonnegative Matrix Factorization
NGD	Normalized Google Distance
NJFA	Nonparametric Joint Factor Analysis
NMF	Nonnegative Matrix Factorization

NMI	Normalized Mutual Information
PCA	Principle Component Analysis
RBGS	Rao-Blackwellized Gibbs Sampling
RMS-NMF	Regularized Multiple Shared Nonnegative Matrix Factorization
RS-NMF	Regularized Shared Nonnegative Matrix Factorization
SIFT	Scale-Invariant Feature Transform
S-NMF	Shared Nonnegative Matrix Factorization
SVD	Singular Value Decomposition

Notations

The following table lists some of the common notations used in this dissertation.

Notation	*Meaning*
∇	Gradient
$\mathbb{R}$	Set of real numbers
$\mathcal{V}$	Vocabulary of words
$\emptyset$	Null set
$\lVert x \rVert_2$	Euclidean or L_2-norm of vector x
$\lVert \mathbf{A} \rVert_F$	Frobenius norm of matrix $\mathbf{A}$
$\mathrm{Tr}\left(\mathbf{A}\right)$	Trace of matrix $\mathbf{A}$
$\lvert \mathbf{A} \rvert$	Determinant of matrix $\mathbf{A}$
$\mathbf{A}^{\mathsf{T}}$	Transpose of matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	Inverse of matrix $\mathbf{A}$
$\mathbf{A}^{\dagger}$	Moore-Penrose pseudo inverse of matrix $\mathbf{A}$
x	Data vector
$\mathbf{X}$	Data matrix
$\mathbf{A} \odot \mathbf{B}$	Element-wise multiplication of matrices $\mathbf{A}$ and $\mathbf{B}$
$\mathbf{A} \oslash \mathbf{B}$	Element-wise division of matrices $\mathbf{A}$ and $\mathbf{B}$
$\mathcal{O}$	Order of complexity
$\mathcal{N}$	Normal distribution
$\mathcal{W}$	Wishart distribution
$\Gamma\left(.\right)$	Gamma function
$\cup$	Union operation of sets
$\cap$	Intersection operation of sets
$\mathcal{F}$	Sigma field
$\triangleq$	Equal by definition
$\square$	End of a proof
$\delta_{\phi}\left(\theta\right)$	Dirac delta function; returns 1 if $\phi = \theta$ and 0 otherwise
$\delta_{\phi}\left(S\right)$	Dirac measure; returns 1 if $\phi \in S$ and 0 otherwise

Chapter 1

Introduction

Feeding our insatiable appetite for content, disparate data sources surround us. Data from a single source is often not rich enough and users look for information across multiple sources and modalities. The research community has focused on data mining and analysis of a single data source, but limited work addresses issues arising from the joint understanding of multiple data sources. This has created open opportunities to develop formal frameworks for analyzing multiple related data sources, exploiting joint properties to strengthen analysis and data mining. Discovering structures and patterns from multiple data sources often unravels commonalities and differences, otherwise not possible with isolated data source analysis. This information is valuable for various data mining and representation tasks.

As an example, consider social media. Entirely new genres of media have been created around the idea of participation, including wikis (e.g. Wikipedia), social networks (e.g. Facebook), media communities (e.g. YouTube), news aggregators (e.g. Digg), social bookmarks (e.g. Del.icio.us), blogs and micro-blogs (e.g. Blogger, Twitter). These applications are significant because they are ranked highest by traffic volume and attention. Modeling the content of unstructured data across heterogeneous yet similar data sources is critical for social media mining and retrieval tasks. Open questions are: how can we effectively analyze disparate data sources with greater precision and at deeper conceptual levels? Can we establish the correspondence or similarity of items in one data source with items in another data source?

Similar requirement arises in transfer learning applications where the aim is to improve performance on a target source by transferring knowledge from other related auxiliary sources. The motivation for using the auxiliary sources may be driven by different causes in different scenarios. For example, there might be a lack of data in the source of interest (target source) whereas plenty of data may be available in other related sources. Although the distribution of these related sources is not identical to that of the target source, combining data from these sources in an appropriate manner has been shown to improve performance. In other applications, it may be that the plenty of data from a particular source are available but these examples are very noisy. Further assume that there exist other related but less noisy data sources which could be used as auxiliary sources and might be useful towards learning accurate common structures. For example, tags associated to Flickr images are noisy, subjective and incomplete whereas similar sources such as LabelMe are relatively cleaner as they have different annotation goals. Under these scenarios, it may be desirable to transfer common knowledge and structures from LabelMe tag data to improve tag based search and organization of Flickr images.

The above problems have a common modeling goal that is to model multiple data sources jointly in an unsupervised manner. This can be done using a joint subspace by combining the data from each source. However, learning a joint subspace through such augmentation of data sources risk losing the distinguishing features of each data source. Therefore, to retain the individual features of each source, it is imperative to model individual subspace of each data set separately. This requires a joint framework wherein there is a provision of modeling both the common and the individual structures of multiple data sources.

1.1 Aims and Approach

This thesis aims to develop formal frameworks and models for learning from multiple data sources. Our main objectives are :

- To develop *unsupervised* frameworks for joint modeling of data from multiple sources with the following aims
    - to improve task performance for a target source, leveraging the shared statistical strengths across auxiliary sources.
    - to relate the data from multiple sources, learning common and individual aspects across disparate data sources.
- To apply these joint modeling frameworks to problems in social media and text mining.

To realize our *first* objective, our approach is to extend existing subspace learning techniques to allow joint modeling of multiple data sources. In particular, we propose a shared subspace learning framework, which allows modeling of common aspects of data sources through a shared subspace and the individual aspects through individual subspaces. Our framework, being based on a subspace approach, does not need one-to-one correspondence across different data sources and is therefore applicable to a wider class of problems in general settings. Under this framework, we develop several models meeting the above-mentioned goals formulating them as *matrix factorization* problems. To learn the factorization matrices, we mainly take two approaches : least-squares (LS) error minimization and probabilistic approaches. Many of the models developed under the least-squares approach have non-convex objective functions leading to iterative solutions. While developing these models, we ensure the convergence of the iterative algorithms mathematically. Subspace learning models such as factor analysis, nonnegative matrix factorization have been shown to work well for single data sources. We develop extensions of these models to enable joint modeling of multiple data sources. While developing these extensions, we ensure that computational complexity of the joint modeling does not increase significantly from their single source counterpart.

Learning the extent of sharing needs model selection – a tedious and computationally challenging task. To address this problem, we first specify the extent of sharing through a set of parameters assuming that these parameters are known *a priori*. As a next step, we address the learning of these parameters using Bayesian nonparametric theory, avoiding the need to do model selection and/or *a priori* specification for these parameters.

To realize our *second* objective, we apply the joint modeling frameworks for improving various tasks in social media and text mining. For social media, we utilize user contributed tags associated to the social media items for retrieval. Due to the incomplete, noisy and subjective nature of social media tags, performance of social media retrieval techniques is usually poor. To improve the retrieval performance, we leverage other useful tagging sources that have underlying structures *similar* to the target source but are often *less* noisy. In another application, we use the joint modeling framework for simultaneously retrieving items from multiple disparate media, e.g. using a query term 'Olympics', we retrieve Olympics related weblogs, images and videos. In addition, we relate these items by finding semantically similar items in disparate media using the shared subspace. We refer to this application as 'cross-social media retrieval and correspondence'. In application to text mining, we apply the joint modeling framework for document clustering and show significant improvements in clustering performance for a variety of data sets.

1.2 Contributions and Significance

The overarching theme of this thesis is the development of shared subspace learning for different situations. The key contributions of this thesis are detailed below :

1.2.1 Nonnegative Shared Subspace Learning for *Two* Data Sources

The *first* contribution of this thesis is a joint modeling of two data sources through a novel *shared subspace learning* formulation based on nonnegative matrix factorization. We derive an efficient iterative algorithm for learning the factorization, analyze its complexity, and provide proof of convergence. We further provide efficient algorithms for carrying out social media (image and video) retrieval and validate our framework by improving retrieval tasks on Flickr and YouTube data sets using LabelMe data set as an auxiliary source. The novelty of our approach lies in the flexibility that permits the amount of subspace sharing to be varied from none to full sharing. Whilst these extremes have been considered before [Lee and Seung, 2001, Lin et al., 2009], our results show that the best performance is achieved when both individual *and* joint subspaces are present together. The significance of the proposed shared subspace learning framework is, firstly, increased efficacy of image retrieval in Flickr and video retrieval in YouTube; and secondly, the framework has broader application to unsupervised learning with knowledge transfer from one domain to another of different quality. Unlike multi-view learning approaches, our framework does not require one-to-one correspondences across the two sources and hence has potential for wider applicability in general settings.

1.2.2 Nonnegative Shared Subspace Learning for *Multiple* Data Sources

The *second* contribution of this thesis is a joint matrix factorization framework along with an efficient algorithm for extraction of shared and individual subspaces for *multiple* data sources with *arbitrary* sharing configurations. We provide complexity analysis of the learning algorithm and show that its convergence is provably guaranteed. We further develop algorithms for social media retrieval in a multi-task learning setting and a novel cross-social media retrieval. We provide two real world demonstrations of the proposed framework using three representative social media sources–blogs (Blogspot.com), photos (Flickr) and videos (YouTube). By permitting differential amounts of sharing in the subspaces, our framework can transfer knowledge across multiple data sources and thus, can be applied to a much wider context – it is appropriate wherever one needs to exploit the knowledge across multiple related data sources. It provides efficient means to mine multimedia data, and partly transcends the semantic gap by exploiting the diversity of rich tag metadata from several media domains.

1.2.3 *Regularized* Nonnegative Shared Subspace Learning

The *third* contribution of this thesis is the further extension of the shared subspace learning using regularization to *segregate* the shared and individual subspaces, ensuring that the shared subspace does not capture individual aspects, and the individual subspaces do not capture shared aspects. To achieve the segregation, we model the constituent subspaces to be mutually orthogonal, minimizing the possibility of negative transfer learning [Rosenstein et al., 2005] that may occur if shared subspaces capture domain specific aspects. To learn the joint factorization, we provide an optimization scheme based on multiplicative updates and show that its convergence is theoretically guaranteed. We analyze the computational complexity of the proposed algorithm and show that it remains similar in complexity to standard nonnegative matrix factorization. To demonstrate the usefulness of our framework, we apply it to two applications : tag-based social media retrieval and clustering. Using two real world data sets (Blogspot-Flickr-YouTube data set and BBC-CNN news data set), we show that our approach improves the performance over many state-of-the-art single and multi-task clustering techniques for both applications. In addition, we also provide a comparison of our method with these state-of-the-art techniques using 20 Newsgroup benchmark data set. The significance of our work is its wide applicability to many problems in which our framework can leverage the common knowledge present in auxiliary domains and transfer it to a target domain. Crucially, by segregating the shared subspace from the individual subspaces, our approach is robust against negative transfer learning. In application to clustering, our method differs from other multi-task clustering methods [Gu and Zhou, 2009a, Zhang and Zhang, 2011], which combine both subspace learning and clustering tasks. Our method does this in two steps. In the first step, it learns the regularized common and individual subspaces whilst in the second step, it uses these subspaces for clustering. This provides crucial flexibility to use any state-of-the-art clustering method in conjunction with our shared subspace framework. In application to retrieval, our framework is capable of transferring knowledge from a cleaner data source to a related noisy data source and thus, boosting performance.

1.2.4 *Bayesian* Shared Subspace Learning

The *fourth* contribution of this thesis is the construction of a novel Bayesian shared subspace learning framework for extraction of shared and individual subspaces from multiple data sources with arbitrary sharing configurations using a joint matrix factorization. This work departs from the previous contributions in that it uses a linear-Gaussian framework and can model *mixed-sign* data from multiple sources, not merely the nonnegative data. Efficient inference is developed based on *Rao-Blackwellized* Gibbs sampler for learning the joint factorization. Algorithms are developed for retrieval *within* a target social medium leveraging other auxiliary social media and for joint retrieval *across* multiple social media, using the proposed Bayesian shared subspace learning. Using three popular social media sources (Blogspot, Flickr and YouTube), we demonstrate the utility of the model for two real-world applications (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 probabilistic framework and algorithms for learning *across* diverse data sources. In addition to exploiting the mutual strengths of multiple data sources, our shared and individual subspaces can handle the uncertainties effectively using probability distributions over the subspaces. The significance of our work lies in the fact that theoretical extensions made to Bayesian probabilistic matrix factorization [Salakhutdinov and Mnih, 2008] for multiple data sources allow for the flexible transfer of knowledge across disparate media using tags as a common feature. In addition, *Rao-Blackwellized* Gibbs sampling used for our model has been shown to achieve better Markov chain mixing [Liu, 1994].

1.2.5 *Nonparametric Bayesian* Shared Subspace Learning

The *fifth* contribution of this thesis is the extension of beta process factor analysis [Paisley and Carin, 2009] to hierarchical beta process (HBP) factor analysis, allowing for joint modeling of data from multiple sources. This extension leads to a model that can discover both shared factors and individual factors specific to each source in a nonparametric setting, avoiding the need to specify the dimensionality of the latent subspaces *a priori*. In addition, we propose an efficient Gibbs sampling scheme that provides an alternative to adaptive rejection sampling (ARS) used in Thibaux and Jordan [2007], which advocated for developing better inference algorithms to fully exploit beta process models. We provide details on sampling hyperparameters of the HBP, which was not systematically addressed in Thibaux and Jordan [2007]. Lastly, our work provides an alternative paradigm for statistical sharing across multiple data groups, in a similar realm to the hierarchical Dirichlet process (HDP) [Teh et al., 2006] but we operate at the factor model level instead of mixtures, and therefore it is more appropriate for applications such as retrieval, joint dimensionality reduction and collaborative filtering.

1.3 Outline of the Thesis

- Chapter 2 provides the necessary background for this thesis and presents a brief review of related works in the area of learning from multiple data sources, e.g. transfer learning, multi-task learning, multi-view learning etc.

- Chapter 3 begins with a shared nonnegative matrix factorization (S-NMF) framework for joint modeling of two related data sources. It presents an efficient algorithm for learning the joint factorization, analyzes the algorithm complexity, and discusses convergence. To apply the proposed framework for social image/video retrieval, efficient algorithms using S-NMF are developed. The second half of this chapter focuses on validating the proposed framework on image and video retrieval tasks wherein tags from the LabelMe data set are used to improve image retrieval performance from a Flickr data set and video retrieval performance from a YouTube data set.

- Chapter 4 presents an extension of S-NMF to allow modeling of more than two data sources with any arbitrary sharing configuration (termed as MS-NMF). The effectiveness of MS-NMF is demonstrated using social media retrieval and another novel application termed as ‘cross-social media retrieval’. Efficient algorithms using MS-NMF are developed to carry out these tasks. To perform the experiments, a cross-media data set constructed from three social media sources (Blogspot.com, Flickr and YouTube) is used. This chapter concludes with a comparison of MS-NMF with other baseline methods using entropy and impurity measures.

- Chapter 5 extends the S-NMF framework and describes a regularized shared subspace learning framework that imposes a mutual orthogonality constraint among the constituent subspaces and ensures that shared subspace captures only shared aspects and individual subspaces capture only individual aspects. After developing the regularized model, its applications to social media retrieval and clustering are presented. Through these applications, we clearly demonstrate the benefits of regularization scheme. Before concluding, a model selection procedure is described for learning the parameters used during the modeling.

- Chapter 6 presents a Bayesian framework (termed as BSSL) to learn the shared and individual subspaces from multiple data sources. After setting out the linear-Gaussian modeling assumptions, an efficient *Rao-Blackwellized* Gibbs sampler is developed for learning the subspaces. Similar to the chapter 4, two social media applications of the proposed model are described. Finally, experiments are conducted for the two applications showing the superiority of the proposed model over other non-sharing baseline models.

- Chapter 7 presents a nonparametric shared subspace learning framework for modeling multiple related data sources. Employing nonparametric Bayesian theory of beta processes, this chapter addresses an important problem of learning subspace dimensionalities automatically from the data by embedding it within the framework of model inference and avoids the need for model selection. The proposed model utilizes the hierarchical beta process (HBP) prior for both defining the shared and individual subspaces and inferring their dimensionalities. After presenting the problem formulation, a Gibbs sampling scheme using auxiliary variables is developed for posterior inference and is amenable for sampling both the main variables and the hyperparameters of the model. Next, experiments using a synthetic sharing data set are presented to illustrate the model behavior. Finally, experiments using real data sets are conducted to validate the efficacy of the proposed model for two real-world applications : transfer learning in text and image retrieval.

- Finally, chapter 8 provides a summary of the work presented in this thesis and suggests some possible directions for future works.

Chapter 2

Related Background

In this chapter, we lay out a foundation over which this thesis would develop. We provide background material on subspace learning, its probabilistic formulations and various inference techniques such as Markov chain Monte Carlo (MCMC) and Gibbs sampling. After this, we discuss the use of nonparametric Bayesian priors in mixture and factor models, which is becoming increasingly popular to circumvent the model selection problems arising in many unsupervised applications. We review the works on learning from multiple data sources, e.g. multi-view learning, transfer learning and multi-task learning. Finally, we conclude this chapter with a discussion on some of the applications in social media domain.

2.1 Subspace Learning

Subspace learning is an important research area in data mining with diverse applications - information retrieval, face recognition, and data visualization among many others. Typical subspace learning algorithms are guided by a modeling goal to transform the data to a smaller dimension space, e.g. optimum classification in Fisher linear discriminant analysis (LDA) [Duda et al., 2001], optimum global representation in principal component analysis (PCA) [Jolliffe, 2002] and factor analysis (FA) [Harman, 1976] or local part-based representation in nonnegative matrix factorization (NMF) [Lee and Seung, 2001]. The common theme of all these techniques is the transformation of high-dimensional data to a meaningful reduced dimensional representation. Such reduced dimensional representations usually facilitate many data mining and pattern recognition tasks such as clustering, classification and information retrieval by mitigating noise and the curse of dimensionality [Jimenez and Landgrebe, 1998]. Many subspace learning tasks can be formulated as a matrix factorization problem. In the following, we briefly describe two such techniques to illustrate the main idea.

2.1.1 Factor Analysis

Factor analysis (FA) is one of the basic dimensionality reduction methods used to describe the statistical variations amongst observed variables through potentially a lower number of unobserved variables referred to as “factors”. Given multi-dimensional correlated data vectors, factor analysis discovers an uncorrelated set of factors that can be thought of hidden variables used to explain the data. The observed variables are modeled as linear combinations of the potential factors, plus "error" terms. The information gained about the inter-dependencies between observed variables can be used later to reduce the set of variables in a data set. Factor analysis originated in psychometrics, and since then has been used in behavioral sciences, social sciences, marketing, product management, operations research, and other applied sciences dealing with large quantities of data.

Factor analysis is closely related to principal component analysis (PCA) but there is an important difference. By definition, PCA maximizes the covariance of the data by learning an optimal subspace and provides the “best” approximation of the data through the least number of basis vectors in L_2-norm sense. Factor analysis models the variations of the data in a similar fashion except that it uses a non-isotropic modeling error term. In other words, it has a more general noise model allowing different “specific variances” for different variables. Formally, factor analysis attempts to model any data matrix $\mathbf{X} = [x_1, ..., x_N]$ containing N data vectors (where $x_n \in \mathbb{R}^{M\times 1}$) as a product of two matrices $\Phi \in \mathbb{R}^{M\times K}$ and $\mathbf{H} \in \mathbb{R}^{K\times N}$ plus an error matrix $\mathbf{E} \in \mathbb{R}^{M\times N}$. The matrix $\Phi = [\phi_1, ..., \phi_K]$ contains K factors, which can also be viewed as basis vectors that spans a subspace and the matrix $\mathbf{H} = [h_1, ..., h_N]$ contains the factor loadings or the representation of the data matrix in the subspace. Formally,

$$\mathbf{X} = \Phi\mathbf{H} + \mathbf{E} \tag{2.1}$$

where the factors ϕ_k in the matrix Φ are assumed to have zero mean and isotropic covariance. Usually, the matrix $\mathbf{E}$ contains the residual modeling error and is statistically independent of both Φ and $\mathbf{H}$. If the data matrix $\mathbf{X}$ has been centered to zero by subtracting its mean, then the error matrix can be allowed to have zero mean. Factor analysis solutions are unique only up to an orthogonal transformation. To see this, consider a solution given by the factorization matrices Φ and $\mathbf{H}$, then $\Phi\boldsymbol{Q}$ and $\boldsymbol{Q}^{\mathsf{T}}\mathbf{H}$ are also a solution for any orthogonal matrix $\boldsymbol{Q}$.

2.1.2 Nonnegative Matrix Factorization

Part-based representation has received much interest in machine learning, computer vision and pattern recognition [Li et al., 2001]. To capture these representations, Lee and Seung proposed nonnegative matrix factorization (NMF) and demonstrated its successful use for analyzing face images [Lee and Seung, 1999, 2001]. Due to the nonnegativity constraints, NMF factors tend to capture part-based representation. Mathematically, NMF [Lee and Seung, 2001] aims to factorize a nonnegative data matrix into two other nonnegative matrices such that

$$\mathbf{X} \approx \mathbf{FH}, \, \mathbf{F} \geq 0, \, \mathbf{H} \geq 0 \tag{2.2}$$

where assuming that a $M \times N$ data matrix $\mathbf{X}$ contains N data points in an M-dimensional space, the first $M \times R$ factor matrix $\mathbf{F}$ can be thought of representing a reduced dimensional nonnegative subspace $(R < N)$ whose bases vectors are given by the columns of matrix $\mathbf{F}$ and the second $R \times N$ factor matrix $\mathbf{H}$ contains the representation of each data point in the subspace. Since the above factorization is not unique, the solution depends on the algorithm used and its initialization. To see an instance of non-uniqueness, consider two matrices $\mathbf{F}$ and $\mathbf{H}$ that are a solution to $\mathbf{X} = \mathbf{FH}$, then $\mathbf{FS}$ and $\mathbf{S}^{-1}\mathbf{H}$ are also a solution for any positive diagonal matrix $\mathbf{S}$. Normalization of $\mathbf{F}$ so that each column has a norm (e.g. L_1 or L_2 norm) equal to one eliminates the degree of freedom caused by such diagonal matrices. Therefore, it is a common practice to normalize each column of matrix $\mathbf{F}$ to norm one [Xu et al., 2003, Cai et al., 2011]. To keep the product unchanged, the matrix $\mathbf{H}$ is accordingly adjusted.

The local (part-based) nature of NMF comes from the fact that nonnegative basis vectors (parts) from matrix $\mathbf{F}$ combine in a nonnegative fashion using the coefficients of matrix $\mathbf{H}$ to construct a data point (create "whole") in matrix $\mathbf{X}$. Lee and Seung [1999] have shown that NMF can achieve part-based decomposition of images. Comparing NMF with vector quantization and PCA, they show that the three techniques produce different results due to the imposition of different constraints. There have been proposed many algorithms for learning such factorization [see the survey in Berry et al., 2007] and probably, the most popular one is the approach taken by Lee and Seung [2001] who propose an iterative multiplicative update based solution to the following constrained optimization problem

$$\text{minimize } \|\mathbf{X} - \mathbf{FH}\|_F^2 \text{ , subject to } \mathbf{F}, \mathbf{H} \geq 0$$

Another cost function to obtain NMF solution emerges by an extension of the Kullback-Leibler divergence to positive matrices. Formally,

$$\text{minimize GenKL} \left(\mathbf{X} || (\mathbf{FH})\right) \text{ , subject to } \mathbf{F}, \mathbf{H} \geq 0$$

where $\text{GenKL}\left(A||B\right) \triangleq \sum_{i,j}\left(A_{ij}\text{log}\frac{A_{ij}}{B_{ij}} + A_{ij} - B_{ij}\right)$. Often, the factorization matrices $\mathbf{F}$ and $\mathbf{H}$ may have special interpretation for real world data. For example, consider a text mining application where the data matrix $\mathbf{X}$ is the usual term-document matrix constructed using the term-frequencies. In this case, each column of the matrix $\mathbf{F}$ when normalized to L_1-norm of one, represents a probability distribution on the vocabulary and can be interpreted as a topic [Blei et al., 2003]. Thus, the subspace spanned by columns of matrix $\mathbf{F}$ can be interpreted as a topic space. Given such interpretation, each column of matrix $\mathbf{H}$ provides the mixing proportions of these topics and represents the corresponding document in the topic space.

Optimization and Initialization As discussed above, the most popular optimization technique to obtain NMF solutions is given by Lee and Seung [2001]. The optimization is based on iterative multiplicative updates wherein at each step, the cost function can be shown to be non-increasing. There are alternative ways of optimizing the NMF objective function, e.g. alternating least squares and the active set method [Kim and Park, 2008] or the projected gradients approach [Lin, 2007], which often have better convergence behavior. A good initialization of matrices $\mathbf{F}$ and $\mathbf{H}$ generally leads to quicker convergence of the NMF algorithms. To this end, several methods have been proposed in literature [Wild et al., 2004, Langville et al., 2006, Boutsidis and Gallopoulos, 2008].

Regularization Theoretical analysis of NMF is carried out in Donoho and Stodden [2004] where the authors provide the conditions under which NMF is capable of capturing part-based representations. To enhance the sparsity of the solutions, Hoyer [2004] suggests NMF with sparsity constraints. The regularization based approaches have been very useful in achieving the desired solutions. To predict the labels for unlabeled data, Zhou et al. [2004] propose a semi-supervised learning technique by applying regularization to ensure both local and global consistency. The idea of local consistency is useful because nearby points are likely to have the same labels. Similarly, the global consistency exploits the fact that the points on the same structure (a cluster or a manifold) tend to have the same labels. Bringing the local consistency idea to NMF based algorithms, Gu and Zhou [2009b] optimize a regularized cost function to obtain a nonnegative clustering algorithm with an assumption that the cluster label of each point can be predicted by the points in its neighborhood. Along similar lines, Cai et al. [2011] propose a graph regularized NMF utilizing a graph Laplacian matrix to preserve the geometric structure of the data. Regularization techniques have also been used in conjunction with NMF to obtain orthogonal nonnegative factors. Choi [2008] proposes various schemes where regularization is used to impose orthogonality on either the basis matrix or the feature matrix. To obtain the desired solution, gradient-descent is computed directly in the Stiefel manifold which reduces to multiplicative updates. Further generalizing this concept, Ding et al. [2006] propose bi-orthogonal tri-factor NMF and use it to simultaneously cluster rows and columns of a given data matrix.

Applications Due to its part-based representation property, NMF has been widely used for many machine learning and pattern recognition tasks. Text mining applications have been considered in Shahnaz et al. [2006], Berry and Browne [2005], Abdulla et al. [2009], Park and Ramamohanarao [2011]. NMF or its variants have also been used for image classification [Buchsbaum and Bloch, 2002, Guillamet et al., 2001, 2003], face-expression recognition [Buciu and Pitas, 2004], face detection [Chen et al., 2001] and face/object recognition [Li et al., 2001, Liu and Zheng, 2004a,b, Zhang et al., 2005].

2.1.3 Nonlinear Subspace Learning

Nonlinear subspace learning translates to the problem of nonlinear dimensionality reduction. In recent years, there has been an increasing interest in nonlinear dimensionality reduction techniques. Most of these techniques are developed by preserving *some* desired structure of the data in high dimensional spaces. Some of the popular nonlinear dimensionality reduction techniques are – Isomap [Tenenbaum et al., 2000], Kernel PCA [Schölkopf et al., 1998], locally linear embedding (LLE) [Roweis and Saul, 2000], locality preserving projections (LPP) [He and Niyogi, 2004], Laplacian eigenmaps [Belkin and Niyogi, 2003], multidimensional scaling (MDS) [Cox and Cox, 2001], local tangent space analysis (LTSA) [Zhang and Zha, 2004] etc. Relatively few attempts have been made for developing nonlinear nonnegative counterparts. One such recent work, termed nonlinear nonnegative component analysis, is proposed in Zafeiriou and Petrou [2009].

2.2 Probabilistic Subspace Learning

One of the early attempts towards probabilistic subspace learning was made by the development of probabilistic PCA [Tipping and Bishop, 1999, Roweis, 1998]. Probabilistic PCA (p-PCA) was developed by formulating the factor analysis problem in a maximum likelihood estimation setting. Under the p-PCA model, the columns of matrix **E** in Eq (2.1) are assumed to be isotropic Gaussian distributed random variables with mean zero and covariance of the form $\sigma^2\mathbf{I}$. The n-th column h_n of coefficient matrix **H** is treated as latent or hidden random vector and modeled as a multivariate Gaussian with mean zero and covariance **I**. Formally, the model can be summarized as

$$p\left(h_n\right) = \mathcal{N}\left(h_n \mid \mathbf{0}, \mathbf{I}\right) \text{ and } p\left(x_n \mid h_n\right) = \mathcal{N}\left(x_n \mid \Phi h_n, \sigma^2 \mathbf{I}\right) \tag{2.3}$$

The factor matrix Φ and the variance σ^2 are treated as parameters and are estimated in maximum-likelihood setting. The marginal distribution of the data is given as

$$p\left(x_n \mid \Phi, \sigma^2\right) = \int p\left(x_n \mid h_n\right) p\left(h_n\right) dh_n \tag{2.4}$$

The above marginal distribution reduces to a multivariate Gaussian distribution with mean zero and covariance matrix $\mathbf{C}$ where $\mathbf{C} = \Phi\Phi^{\mathsf{T}} + \sigma^2\mathbf{I}$. Using this, the log-likelihood function which needs to be maximized for estimating the subspace matrix Φ is given as

$$\begin{aligned} \log p\left(\mathbf{X} \mid \Phi, \sigma^2\right) &= \sum_{n=1}^{N} \log p\left(x_n \mid \Phi, \sigma^2\right) \\ &= -\frac{N}{2}\left(M \log 2\pi + \log|\mathbf{C}| + \mathrm{Tr}\left(\mathbf{C}^{-1}\mathbf{S}_N\right)\right) \end{aligned} \tag{2.5}$$

where $\mathbf{S}_N$ is the data covariance matrix computed using N data points. Maximization of the above likelihood with respect to Φ and σ^2 can be done in closed form and has the following solutions [Tipping and Bishop, 1999]

$$\Phi_{ML} = \boldsymbol{U}_K\left(\mathbf{L}_K - \sigma^2\mathbf{I}\right)^{\frac{1}{2}}\boldsymbol{R} \text{ and } \sigma^2_{ML} = \frac{1}{M-K}\sum_{k=K+1}^{M}\lambda_k \tag{2.6}$$

where $\boldsymbol{U}_K$ contains the eigen-vectors of $\mathbf{S}_N$ corresponding to the K largest eigen-values and λ_k denotes the k-th eigen-value. The matrix $\mathbf{L}_K$ is a diagonal matrix containing the K largest eigen-values of $\mathbf{S}_N$ on its diagonal and $\boldsymbol{R}$ is an arbitrary orthogonal matrix of size $K \times K$.

The maximum likelihood model proposed above is prone to over-fitting while learning the parameters [Bishop, 2006]. In addition, it needs cross validation to estimate the appropriate value of K. Addressing these problems, Bishop [1999] derived a Bayesian extension of probabilistic PCA which provides a posterior distribution over the parameters. This is useful in getting rough estimates of subspace dimensionality (K) using automatic relevance determination (ARD) [Bishop, 2006]. Other recent probabilistic subspace learning techniques further generalize the above formulation. A Bayesian probabilistic matrix factorization [Salakhutdinov and Mnih, 2008] has been proposed for collaborative applications where the columns of both the factor matrix Φ and the coefficient matrix $\mathbf{H}$ are modeled as multivariate correlated Gaussian distributed random vectors. In addition, this model being a hierarchical Bayesian model, uses a prior on hyperparameters as well. This avoids the need to tune the model hyperparameters for optimal performance.

2.3 Inference Techniques

A key step in the realization of Bayesian probabilistic models is the computation of posterior distribution and expectations with respect to the posterior distributions. Sometimes, it is possible to derive a closed form expression for these quantities. However, in most cases, it is not possible to do so and as a result, the computations of the posterior distribution and the expectations are not tractable. Even when these closed form expressions exist, the resulting distribution may not take a standard form of known distributions. In such cases, exact inference is intractable and therefore, we have to resort to some form of approximation. In this thesis, we shall encounter such problems and would be using approximate inference techniques. In this section, we provide some background on some of the most popular approximate inference techniques.

2.3.1 Markov Chain Monte Carlo Sampling

A large class of sampling algorithms known as Markov chain Monte Carlo (MCMC) has played a significant role in statistics, machine learning and signal processing recently. MCMC sampling algorithms allow sampling from a large family of distributions and scale well with the dimensionality of the sample space. Due to the advancement in computing technologies, it has become practical to solve many problems by applying MCMC algorithms.

The idea of Monte Carlo simulation is to draw an independent and identically distributed (i.i.d.) set of samples from a desired distribution $p(x)$. We usually need to turn to MCMC algorithms if either we do not have a closed form for the distribution $p(x)$ or it is not easy to sample from $p(x)$. MCMC algorithms approach this problem by constructing a Markov chain that has the desired distribution as its equilibrium distribution. The state of the chain after it reaches equilibrium (in practice, after satisfactory convergence has occurred) is then used as a sample from the desired distribution. If the quality of the sample improves quickly as a function of the number of steps, the Markov chain is said to be "mixing well". Since it is not easy (or possible) to sample from the distribution $p(x)$, MCMC algorithms use a proposal distribution $q(x)$ from which it is easy to draw a sample. We use a proposal distribution which depends on the current state $x^{(t)}$, i.e. $q(x \mid x^{(t)})$. A sample is drawn from the proposal distribution $q(x \mid x^{(t)})$ and accepted/rejected according to some criterion. The sequence of such samples $x^{(1)}, x^{(2)}, \ldots$ form a Markov chain. *Ergodic* theorem guarantees a *unique* invariant distribution for a Markov chain that satisfies *three* properties – irreducibility, aperiodicity and positive recurrence. To ensure that the invariant distribution of the constructed Markov chain is our desired distribution $p(x)$, a sufficient (but not necessary) condition is the following reversibility which is widely known as "detailed balance" condition

$$p(x^{(t)})\mathbf{T}(x^{(t-1)} \mid x^{(t)}) = p(x^{(t-1)})\mathbf{T}(x^{(t)} \mid x^{(t-1)}) \tag{2.7}$$

where $\mathbf{T}$ is a stochastic transition matrix and assumed to satisfy the *ergodic* properties of Markov chains. Summing over $x^{(t-1)}$ on both sides of Eq (2.7), we obtain

$$p(x^{(t)}) = \sum_{x^{(t-1)}} p(x^{(t-1)})\mathbf{T}(x^{(t)} \mid x^{(t-1)}) \tag{2.8}$$

In continuous state space, the above expression translates to the following [Andrieu et al., 2003]

$$p(x^{(t)}) = \int p(x^{(t-1)})\mathbf{T}(x^{(t)} \mid x^{(t-1)})dx^{(t-1)} \tag{2.9}$$

where $\mathbf{T}$ is the conditional density of $x^{(t)}$ given $x^{(t-1)}$.

One of the earliest and most popular MCMC algorithms is the *Metropolis-Hasting* (MH) algorithm, which was first proposed by Metropolis et al. [1953] and later modified by Hastings [1970]. At step t of the algorithm, a new sample $x^{(t+1)}$ is drawn from the proposal distribution $q(x^{(t+1)} \mid x^{(t)})$ and is accepted with probability $\mathbf{A}(x^{(t)}, x^{(t+1)})$ where

$$\mathbf{A}(x^{(t)}, x^{(t+1)}) = \min\left(1, \frac{p(x^{(t+1)})q(x^{(t)} \mid x^{(t+1)})}{p(x^{(t)})q(x^{(t+1)} \mid x^{(t)})}\right) \tag{2.10}$$

The transition kernel $\mathbf{T}(x^{(t+1)} \mid x^{(t)})$ can be written as a function of $\mathbf{A}(x^{(t)}, x^{(t+1)})$ and the proposal distribution $q(x \mid x^{(t)})$, and it can be shown that the desired distribution $p(x)$ is the invariant distribution of the Markov chain constructed by the MH algorithm as it satisfies the required detailed balance condition [Andrieu et al., 2003]. However, at this point, we note that although satisfying detailed balance condition provides us confidence that MH algorithm will converge with our chosen proposal distribution $q(x)$, our choice of $q(x)$ will have enormous effect on the convergence speed. Similarly, a poor initialization of the underlying Markov chain can significantly influence the convergence time. Often, a good initialization can be a value in the center of the distribution, e.g. a solution obtained using maximum likelihood (ML) estimates.

2.3.2 Gibbs Sampling

Gibbs sampling [Geman and Geman, 1984] is one of the MCMC algorithms which allow application of Bayesian analysis to a wide class of problems. Gibbs sampling can be seen as a special case of the MH algorithm where the draw from the proposal distribution is always accepted (i.e. with probability 1). In the case of the MH algorithm, when dealing with high dimensional state spaces, one has to sample from the joint distribution at each step. The key idea in Gibbs sampling is not to sample directly from this high dimensional joint distribution but rather to sample from the *univariate conditional distribution* of one variable in the state space given the remaining variables. Often, it is fairly easy to draw samples from such univariate conditional distributions than from complex joint distributions. Essentially, we draw samples in M-dimensional state space sequentially using M

univariate conditionals rather than drawing directly from the M-dimensional multivariate joint distribution.

Let us consider a M-dimensional state space $(x_1, \ldots, x_M)$ with the univariate conditionals of the form $p\left(x_j \mid x_{-j}\right)$ where $x_{-j} \triangleq (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_M)$, then Gibbs sampling uses the following proposal distributions

$$q\left(x \mid x^{(t)}\right) = \begin{cases} p\left(x_j \mid x_{-j}^{(t)}\right) & \text{If } x_{-j} = x_{-j}^{(t)} \\ 0 & \text{Otherwise} \end{cases} \tag{2.11}$$

Plugging the above proposal distribution $q\left(x \mid x^{(t)}\right)$ in Eq (2.10), it can be easily seen that the draws from $q\left(x \mid x^{(t)}\right)$ are accepted with probability 1. Since Gibbs sampler is a special case of MH algorithm, one can embed the MH acceptance step to draw from nonstandard proposal distributions. Gibbs sampling is a popular technique when doing Bayesian analysis using graphical models. In fact, for such networks, the conditional distribution of a particular node only depends on a set of variables called the *Markov blanket*. For an undirected graph, the Markov blanket of a node comprises the set of neighbouring nodes, while for a directed graph it comprises the parents, the children, and the co-parents [Bishop, 2006]. A popular Gibbs sampling software package called Bayesian Updating with Gibbs sampling (BUGS) has been developed exploiting this dependency [Gilks et al., 1994].

2.3.3 Nuisance Variables and Collapsed Gibbs Sampling

In probabilistic models, a nuisance variable is a random variable that is a key variable for model description, but is of no particular interest in itself. We encounter many such variables that are latent to the model and have a key role in representation of generative models. However, given a particular problem at hand (e.g. prediction, clustering etc), we may often not be interested in inferring the distributions of these nuisance variables. In such cases, it is a standard practice to integrate out such variables by computing the marginals when conducting Bayesian inference. The process of integrating out the nuisance variables is also called “collapsing” the state space, which as a result, is left with a reduced number of variables.

For illustration, consider three dimensional state space (x_1, x_2, x_3) and a desired joint distribution $p\left(x_1, x_2, x_3\right)$. The standard Gibbs sampling constructs a Markov chain by generating samples sequentially, e.g. sample x_1 from $p\left(x_1 \mid x_2, x_3\right)$, then sample x_2 from $p\left(x_2 \mid x_1, x_3\right)$ and finally sample x_3 from $p\left(x_3 \mid x_1, x_2\right)$. This process is repeated until convergence of the sampler. Now, let us assume that the variable x_2 is the nuisance variable, which we are either not interested in sampling at every step or not interested in sampling at all. In such cases, integrating out x_2 is useful as the state-space reduces from three to two dimensions. The dimensionality reduction obtained by collapsing nuisance variables increases the sampling efficiency since first, there is no time involved in sampling these variables and second, this process usually reduces auto-covariance of the samples. The collapsed Gibbs sampler proceeds as follows : sample x_1 from $p(x_1 \mid x_3)$ and then sample x_3 from $p(x_3 \mid x_1)$. The distributions $p(x_1 \mid x_3)$ and $p(x_3 \mid x_1)$ are obtained by marginalizing x_2. Collapsed Gibbs sampling is also known as *Rao-Blackwellized* Gibbs sampling [Casella and Robert, 1996, Sudderth, 2006].

2.3.4 Auxiliary Variable Sampling

In most cases, addition of extra (or auxiliary) variables slows convergence and increases sample auto-covariances [Liu, 1994]. However, in some cases, the careful use of auxiliary variables can convert intractable problems into the tractable ones (e.g. [Damien et al., 1999, Teh et al., 2006]). At times, sampling efficiency may also increase if the augmentation leads to simpler conditional distributions.

The idea of introducing auxiliary variables in probabilistic models is somewhat opposite to the idea of collapsing. There are many problems in which sampling from a set of variables may not be easy because the univariate conditionals may not take standard forms. In such cases, by carefully choosing a set of auxiliary variables and forming a joint distribution with the original set of variables, the conditionals can change into a standard, easy to sample form. The intuition behind this lies in that an apparently complicated problem becomes more tractable when embedded into a higher dimensional framework. The use of auxiliary variables in MCMC algorithms dates back to the work in statistical physics [Swendsen and Wang, 1987]. It was generalized by Edwards and Sokal [1988] and later brought into mainstream statistics by Besag and Green [1993], Mira and Tierney [1997].

Auxiliary variable sampling methods augment the original state space specified by $x \in \mathbb{R}^M$ by one or more additional variables $u \in \mathbb{U}$. Except for some special cases, these auxiliary variables may not have any physical meaning and are used only for the sake of mathematical convenience. The joint distribution of x and u can be defined by multiplying the marginal distribution for x with the conditional distribution $p(u \mid x)$. To assist MCMC based sampling, a Markov chain can be constructed on the product space $\mathbb{R}^M \times \mathbb{U}$ iterating between two different transitions: assuming some value for x at step t as $x^{(t)}$, sampling $u^{(t)}$ with conditional $p\left(u^{(t)} \mid x^{(t)}\right)$ and then sampling a new value for x using a proposal distribution $q\left(x \mid x^{(t)}, u^{(t)}\right)$ which satisfies the detailed balance condition for the desired distribution $p(x \mid u)$. Two well known examples of auxiliary variable methods are hybrid Monte Carlo (HMC) [Duane et al., 1987] and slice sampling [Neal, 2003, Damien et al., 1999]. Other successful demonstrations can be found in Teh et al. [2006].

2.4 Nonparametric Bayesian Priors

The use of nonparametric Bayesian priors has become popular to circumvent the model selection problem as models based on these priors do not assume any *fixed* parametric form and allow the model complexity to grow with the data. The term ‘nonparametric’ is a misnomer as it does not mean that these models are free of parameters. Instead, these models can have a large set of parameters but the number of such parameters automatically gets adjusted with the data.

In contrast to the parametric models which assume a fixed set of parameters, the nonparametric models start with defining a prior distribution over an infinite dimensional parameter space. Using a prior distribution in this way allows the possibility of any arbitrary parameters from this large space. Given observed data, we can compute the posterior distribution over this parameter space weighting the probability of each parameter according to their data likelihood and effectively determining the “active” set of parameters. To define the priors on such infinite-dimensional spaces, one needs infinite dimensional distributions which can be realized using stochastic processes. The stochastic processes (e.g. Gaussian processes, Dirichlet processes etc) can often be characterized by their marginal distributions over finite dimensional subsets of the parameter space. For example, in the case of Gaussian process (GP), the marginal distribution of any finite subset of the parameter space is normally distributed. Similarly, the marginal distribution of Dirichlet process defined over any finite partition of the parameter space follows a Dirichlet distribution. Using this property, Dirichlet process (DP) is used to define a distribution on the probability measures. In other words, every sample of Dirichlet process is a probability measure that can be used to define a discrete (or categorical) distribution over the infinite dimensional parameter space.

2.4.1 Dirichlet Processes

Dirichlet process [Ferguson, 1973, Antoniak, 1974] is one of the widely researched and well understood nonparametric prior in the field of statistics and machine learning. Generalizing the mixture models [Escobar and West, 1995], it has been used as a prior on the latent class to determine the number of clusters. Let $(\Omega, \mathcal{F})$ be a measurable space[1], H be a fixed measure over Ω and α be a scalar (known as concentration parameter), then, a Dirichlet process G is a random probability measure over Ω, denoted by $G \sim \mathrm{DP}(\alpha, H)$. Formally, for every finite partition $(A_1, \ldots, A_K)$ of Ω, G is characterized by

$$\left(G\left(A_1\right), \ldots, G\left(A_K\right)\right) \sim \mathrm{Dir}\left(\alpha H\left(A_1\right), \ldots, \alpha H\left(A_K\right)\right) \tag{2.12}$$

[1] The set Ω is the sample space and $\mathcal{F}$ is a sigma algebra constructed from the collection of subsets of Ω

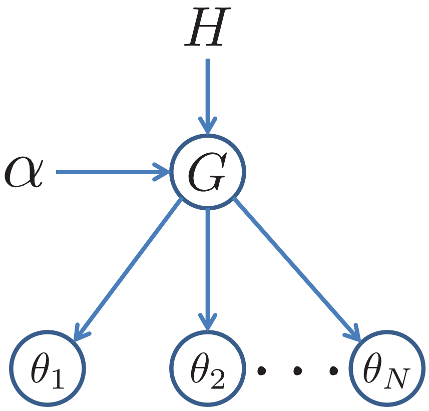


Figure 2.1: Directed graphical representation of Dirichlet process prior.

The consistency of above characterization can be shown using the *aggregation* property of Dirichlet distribution. Since the Dirichlet process G defines a prior distribution over Ω, we can draw samples from G. Let $\theta_1, \ldots, \theta_N$ be i.i.d. samples from G, then given these samples, the posterior distribution of G can be easily computed using Bayes rule. A directed graphical representation of this generative process is shown in Figure 2.1. In particular, let $(A_1, \ldots, A_K)$ be a finite partition of Ω and let n_k be the number of θ_i's belonging to partition A_k, i.e., $n_k = \#\{i \mid \theta_i \in A_k\}$, then using the conjugacy between Dirichlet and multinomial distribution, we can write the posterior as

$$(G(A_1), \ldots, G(A_K)) \mid \{\theta_1, \ldots, \theta_N\} \sim \mathrm{Dir}(\alpha H(A_1) + n_1, \ldots, \alpha H(A_K) + n_K) \qquad (2.13)$$

Since the above characterization is true for any finite partition $(A_1, \ldots, A_K)$ with arbitrary K, Eq (2.13) implies that the posterior is another Dirichlet process with modified concentration parameter and the base measure. We note that there are repetitions in the collection $\{\theta_1, \ldots, \theta_N\}$ as the draws from a DP are discrete with probability one [Sethuraman, 1994]. Assuming that the set $\{\phi_1, \ldots, \phi_K\}$ denotes the unique values in $\{\theta_1, \ldots, \theta_N\}$, the posterior measure of Eq (2.13) can be written as the following

$$G \mid \{\theta_1, \ldots, \theta_N\} \sim \mathrm{DP}\left(\alpha + N, \frac{1}{\alpha + N}\left[\alpha H + \sum_{k=1}^{K} n_k \delta_{\phi_k}\right]\right) \qquad (2.14)$$

Both, the concentration parameter and the base measure of the posterior DP are set to a value by making a compromise between the prior DP and the observed data. To relate $\{\theta_1, \ldots, \theta_N\}$ with $\{\phi_1, \ldots, \phi_K\}$, we can use a set of indicator variables z_i for each θ_i such that $z_i = k$ if $\theta_i = \phi_k$. To avoid the difficulty of dealing with an infinite dimensional prior, it is customary to integrate out G and deal directly with predictive distribution $p\left(\theta_{N+1} \mid \theta_1, \ldots, \theta_N\right)$ if the posterior of G is not needed explicitly. For this, consider

$$\begin{aligned} p\left(\theta_{N+1} \mid \theta_1, \ldots, \theta_N\right) &= \int_G p\left(\theta_{N+1}, G \mid \theta_1, \ldots, \theta_N\right) dG \\ &= \int_G p\left(\theta_{N+1} \mid G\right) p\left(G \mid \theta_1, \ldots, \theta_N\right) dG \\ &= \mathbb{E}_{p(G \mid \theta_1, \ldots, \theta_N)}\left[p\left(\theta_{N+1} \mid G\right)\right] \end{aligned} \tag{2.15}$$

Using Eqs (2.14) and (2.15), the predictive distribution can be shown to take the following form

$$\theta_{N+1} \mid \{\theta_1, \ldots, \theta_N\} \sim \frac{1}{\alpha + N}\left[\alpha H + \sum_{k=1}^{K} n_k \delta_{\phi_k}\right] \tag{2.16}$$

The above predictive distribution can be simulated incrementally using a simple scheme known as the *Chinese Restaurant Process* (CRP) [Pitman, 2006]. The Chinese restaurant process constructs a distribution over the sequence $\theta_1, \theta_2, \ldots$ by incrementally drawing θ_i given $\theta_1, \ldots, \theta_{i-1}$. Importantly, the Chinese restaurant process induces an exchangeable distribution on partitions meaning that the joint distribution of $\{\theta_1, \ldots, \theta_N\}$ is invariant to the order of construction, i.e. $p\left(\theta_1, \ldots, \theta_N\right) = p\left(\theta_{\sigma(1)}, \ldots, \theta_{\sigma(N)}\right)$ for any permutation σ. Given this exchangeability result, another way to show the existence of Dirichlet process is through appealing to de Finetti's theorem [De Finetti, 1990] – a classic theorem which establishes that any collection of exchangeable random variables has a representation as an infinite mixture distribution (see the infinite mixture in Figure 2.1).

2.4.2 Dirichlet Process Mixture Models

The Dirichlet process has been successfully applied for clustering data using mixture models. The use of Dirichlet process as a nonparametric prior enables the mixture model to accommodate infinitely many components. In the mixture model setting, assume that we have a set of observations as $\{x_1, \ldots, x_N\}$ with the corresponding component parameters $\{\theta_1, \ldots, \theta_N\}$ where θ_i's are drawn from a Dirichlet process G with concentration parameter α and base measure H, as shown in Figure 2.1. We also have a observation model which follows a parametric distribution denoted by $F(\theta)$. The whole model can be summarized as below

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ \theta_i &\sim G \\ x_i \mid \theta_i &\sim F(\theta_i) \end{aligned} \tag{2.17}$$

Sethuraman [1994] provided a stick-breaking construction of Dirichlet process which makes it immediately clear that the draws from Dirichlet process are discrete with probability one. Recall that, to relate $\{\theta_1, \ldots, \theta_N\}$ with $\{\phi_1, \ldots, \phi_K\}$, we used a set of indicator variables z_i for each θ_i such that $z_i = k$ if $\theta_i = \phi_k$. In applications of Dirichlet process mixture model (DPMM) to clustering, z_i is used to denote cluster indicator for data x_i. Using $\{\phi_1, \ldots, \phi_K\}$ and the stick-breaking construction, we can write G as the following

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}(\theta) \tag{2.18}$$

where the stick breaking weights π_k are drawn in the following manner

$$\begin{aligned} v_k \mid \alpha &\sim \text{beta}(1, \alpha) \quad \text{for } k = 1, 2, \ldots & (2.19)\\ \pi_k &= v_k \Pi_{r=1}^{k-1} (1 - v_r) & (2.20) \end{aligned}$$

The above construction is also known as GEM distribution. Using the above representation, we can describe DPMM in an alternative manner as the following

$$\begin{aligned} \pi &\sim \text{GEM}(\alpha) \\ z_i &\sim \text{Discrete}(\pi) \\ \phi_k &\sim H \\ x_i \mid z_i, \{\phi_1, \ldots, \phi_K\} &\sim F(\phi_{z_i}) \end{aligned} \tag{2.21}$$

where $\pi \triangleq [\pi_1, \pi_2, \ldots, \pi_K]$. Both the representations of DPMM are shown in Figure 2.2. It can be seen that Dirichlet process mixtures can control complexity of the model using stick-breaking prior instead of choosing a finite number of components K. An important property of stick breaking weights π_k is that these weights decrease exponentially (although appearing in size-biased order). This ensures that only a finite number of clusters are used to model the data. If there is evidence for a new cluster through observed data, the number of clusters grow automatically.

2.4.3 Infinite Factor Analysis Models

Dirichlet process mixture models avoid the need of model selection for determining the number of clusters. While mixture models have found many natural applications, there are other applications, which are difficult to be modeled that way. A different assumption in many problem settings is that objects can be described in terms of a *set* of features or attributes. As an example, consider a collaborative filtering application where users

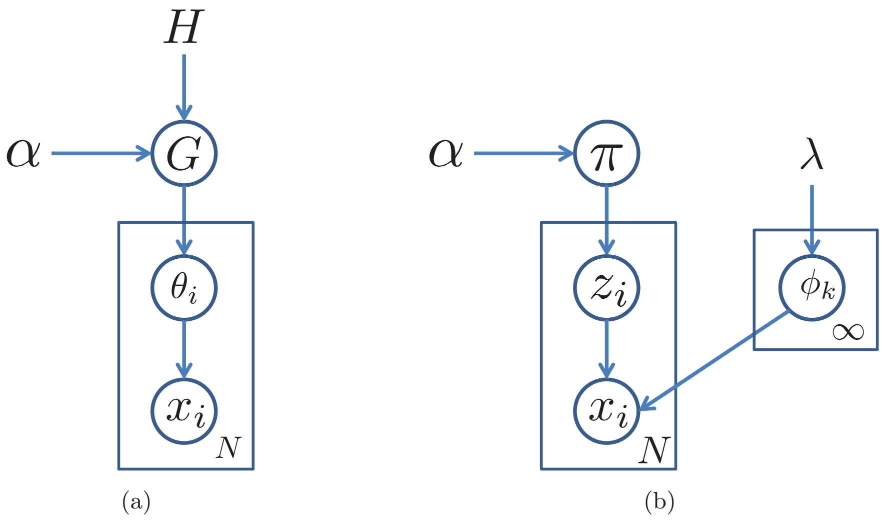


Figure 2.2: Directed graphical representations of Dirichlet process mixture model (DPMM) (a) stochastic process representation (b) stick-breaking representation; akin to finite mixture model except that number of components (K) can grow up to infinity.

assign ratings to items. When a user rates an item, there are many latent factors that determine the level of rating, e.g. while rating 'cars', the user might think about its 'price', 'sporty-look', 'efficiency' along with many other features or factors which an item may or may not be good at. Mixture models consider each attribute as one mixture component and associate an item with only 1 out of K such components. This causes a difficulty in modeling the problems that require more than 1 feature out of K features. As an alternative to the "1-out-of-K" representation underlying the mixture models, factor models can associate a *set* of latent components to each data point.

Similar to the mixture model setting, using nonparametric Bayesian priors in combination with the factor models avoids the need for model selection in determining the number of latent components. The first significant attempt towards this goal was made by Griffiths and Ghahramani [2006] who proposed a nonparametric predictive prior that can be used as a generative model for infinite binary matrices. This predictive prior is known as *Indian Buffet Process* (IBP). The predictive distribution of IBP is analogous to the Chinese restaurant process used for mixture models and can be utilized for factor models to associate the data to a set of factors.

Before discussing IBP further, it is useful to recall the factor analysis model of Eq (2.1). IBP can be used to infer the number of factors K automatically as can be seen from the following formulation. Consider $\mathbf{X} = \Phi\mathbf{H} + \mathbf{E}$, and rewrite it as $\mathbf{X} = \Phi\left(\mathbf{Z} \odot \mathbf{W}\right) + \mathbf{E}$ where $\mathbf{H} \triangleq \mathbf{Z} \odot \mathbf{W}$ and the matrix $\mathbf{Z}$ is a binary matrix of dimension $K \times N$ and K can grow to an arbitrary large number as may be required to explain the data. Since the matrix $\mathbf{Z}$

only indicates the presence or absence of the factors, the value of the feature matrix $\mathbf{H}$ is now represented by the matrix $\mathbf{W}$. The IBP prior is used on the matrix $\mathbf{Z}$ through which an object is allowed to have a set of factors in contrast to the Chinese restaurant process that allows an object to have only one component.

Several latent factor analysis models have been proposed using IBP. Knowles and Ghahramani [2007] propose a nonparametric Bayesian extension of Independent Components Analysis (ICA) where the observed data matrix is modeled as a linear superposition of a potentially infinite number of hidden sources. A variational inference procedure for IBP based models is derived in Doshi-Velez et al. [2009b]. The works in [Doshi-Velez and Ghahramani, 2009a, Doshi-Velez et al., 2009a] have successfully shown the use of IBP based models for large scale applications. Extending the use of IBP to allow modeling non-conjugate distributions, a slice sampler is derived in Teh et al. [2007]. Another extension to model the correlated binary matrices (e.g. correlation due to the co-occurring words in text data or co-occurring objects in images) is described in Doshi-Velez and Ghahramani [2009b].

Similar to the Chinese restaurant process, IBP also satisfies the exchangeable property, i.e. if $\mathbf{Z} = [z_1, \ldots, z_N]$ where each $z_i \in \{0,1\}^{K\times 1}$ then $p\left(z_1, \ldots, z_N\right) = p\left(z_{\sigma(1)}, \ldots, z_{\sigma(N)}\right)$. Given the exchangeability property of IBP, it is desirable to inquire the de Finetti mixing stochastic process underlying IBP. Thibaux and Jordan [2007] investigated this question and found that the de Finetti stochastic process underlying IBP is beta process. We note that the similar process underlying Chinese restaurant process is Dirichlet process. As pointed out by Thibaux and Jordan [2007], establishing this connection allows us to develop effective machinery for the beta process analogous to those developed for Dirichlet processes.

2.4.4 Beta Processes

Let $(\Omega, \mathcal{F})$ be a measurable space, B_0 be a fixed measure over Ω and γ_0 be a positive function over Ω, then, a beta process B is a positive random measure over Ω, denoted by $B \sim \mathrm{BP}\left(\gamma_0, B_0\right)$ [Hjort, 1990]. If $S_1, \ldots, S_r$ are disjoint subsets of Ω, the measures $B\left(S_1\right), \ldots, B\left(S_r\right)$ are independent. This implies that beta process is a positive Lévy process and can be uniquely characterized using a Lévy measure (for details about the Lévy measure of beta process , see Thibaux and Jordan [2007]).

If B_0 is discrete and given as $B_0 = \sum_k \lambda_k \delta_{\phi_k}$ (where ϕ_k is an atom such that $\phi_k \in \Omega$), then the draws from the beta process B are also discrete and can be written in the following form [Thibaux and Jordan, 2007]

$$B = \sum_k \beta_k \delta_{\phi_k} \tag{2.22}$$

where δ_{ϕ_k} is Dirac measure and equals to one if ϕ_k belongs to its argument set. The variable β_k is a random measure associated to an atom ϕ_k such that

$$\beta_k \sim \text{beta}\left(\gamma_0 \lambda_k, \gamma_0\left(1-\lambda_k\right)\right), \, k = 1, 2, \ldots \tag{2.23}$$

In the case of continuous B_0, the measure associated to a single atom (i.e. any ϕ_k) is zero. Therefore, instead of using a single atom ϕ_k, we consider infinitesimally small region $d\phi_k$ in Ω around ϕ_k. [Hjort, 1990] has shown that increments of such infinitesimal form for the beta process can be shown to be independent and follow beta distribution. Thus, for continuous B_0, we can partition Ω into L equal parts such that $\cup_{k=1}^{L} d\phi_k = \Omega$ [Paisley and Carin, 2009]. Using this partition, we can write : $B = \sum_{k=1}^{L} \beta_k \delta_{d\phi_k}$. Following Theorem 3.1 in [Hjort, 1990], for large L, β_k can be approximately drawn from a beta distribution as below

$$\beta_k \sim \text{beta}\left(\gamma_0 b_k, \gamma_0\left(1-b_k\right)\right), \, k = 1, 2, \ldots L \tag{2.24}$$

where $b_k \triangleq B_0\left(d\phi_k\right)$. If z_i is a draw from a Bernoulli process parametrized by B, i.e. $z_i \mid B \sim \text{BeP}\left(B\right)$ then the posterior $B\left(S\right) \mid z_1, \ldots, z_N$ for $S \in \mathcal{F}$, using the "conjugacy" property of beta process [Hjort, 1990, Kim, 1999], is given as

$$B\left(S\right) \mid z_1, \ldots, z_N \sim \text{BP}\left(\gamma_0 + N, \frac{\gamma_0}{\gamma_0 + N} B_0 + \frac{1}{\gamma_0 + N} \sum_{\{k \mid \phi_k \in S\}} \sum_{i=1}^{N} z_i^{(k)}\right) \tag{2.25}$$

Given the above posterior, predictive distribution of $z_{N+1} \mid z_1, \ldots, z_N$, computed under the expectation $\mathbb{E}_{B \mid z_1, \ldots, z_N}\left[p\left(z_{N+1} \mid B\right)\right]$, can be written as

$$z_{N+1}^{(k)} \mid z_1, \ldots, z_N \sim \text{BeP}\left(\frac{\gamma_0}{\gamma_0 + N} B_0 + \frac{1}{\gamma_0 + N} \sum_{i=1}^{N} z_i^{(k)}\right) \tag{2.26}$$

The above predictive process can be understood as the following : z_{N+1} has a value of 1 for an existing factor in proportion to their evidence shown by data, i.e. $\text{BeP}\left(\frac{1}{\gamma_0+N} \sum_{i=1}^{N} z_i\right)$ and a value of 1 for a new factor $\text{BeP}\left(\frac{\gamma_0}{\gamma_0+N} B_0\right)$ which can be shown to be $\text{Poi}\left(\frac{\gamma_0}{\gamma_0+N}\right)$ where $\text{Poi}\left(.\right)$ denotes the Poisson distribution. Thibaux and Jordan [2007] show that the predictive process in Eq (2.26) is the two-parameter generalization of the IBP [Griffiths and Ghahramani, 2006], observing the underlying de Finetti mixing beta process. A nonparametric Bayesian factor analysis model employing variational inference and explicit sampling from beta process has been developed in Paisley and Carin [2009].

There have been proposed many algorithms to generate the samples from a beta process. One of the early attempts towards this was made by Inverse Lévy Measure (ILM) algorithm [Wolpert and Ickstadt, 1998]. This algorithm is efficient in that it generates the samples in a decreasing order. However, a key drawback is that at each step, it requires a computationally expensive inversion of incomplete beta functions. Improving upon this algorithm, Lee and Kim [2004] proposed an approximation of the underlying beta process by a sequence of compound Poisson processes. Due to the approximation used in their algorithm, their algorithm converges only in distribution. Thibaux and Jordan [2007] proposed a simple iterative algorithm to generate the samples from the beta process but it generates the samples in size-biased order. Similar to the stick-breaking construction of Dirichlet processes, an equivalent construction of the beta process is derived in Paisley et al. [2010]. However, this stick-breaking construction is more involved than that of Dirichlet process.

2.4.5 Hierarchical Nonparametric Priors

One of the strengths of probabilistic modeling is to be able to express the dependencies through hierarchies. Taking this into consideration, MacEachern [1999] proposed a *dependent Dirichlet process* (DDP) framework extending nonparametric Bayesian priors to allow hierarchical modeling. DDP suggests to use a covariate space $\mathcal{X}$ for capturing the evolution of random distributions $G_x,\ x \in \mathcal{X}$ and defines a prior over this dependent family. One of the most popular nonparametric hierarchical prior in this category is the hierarchical Dirichlet process [Teh et al., 2006]. Hierarchical Dirichlet process (HDP) allows us to model a generative process involving multiple groups of data sharing a common set of parameters. In particular, it provides a framework to combine J different Dirichlet process models (indexed as $j = 1, \ldots, J$) with random distributions G_j as

$$G_j \mid G_0, \alpha_j \sim \mathrm{DP}\left(\alpha_j, G_0\right) \tag{2.27}$$

through a common base measure G_0, which itself is a draw from another higher level Dirichlet process with base measure H where

$$G_0 \mid H, \gamma \sim \mathrm{DP}\left(\gamma, H\right) \tag{2.28}$$

Intuitively, it seems to be enough to have a common base measure G_0 to generate a family of random distributions G_j which are conditionally independent given G_0. However, a closer look makes it immediately clear that unless G_0 is discrete, the random distributions G_j's do not share any atoms. For continuous G_0, the random distributions G_j, for each j, have atoms at different locations with probability one.

Using HDP prior for mixture model, Teh et al. [2006] propose a model that can share a set of clusters across multiple data groups. This model can also be understood as a nonparametric extension of the popular topic model known as *latent Dirichlet allocation* (LDA) [Blei et al., 2003]. In the context of language model, HDP provides a generative model according to which every document can be associated with multiple topics and the

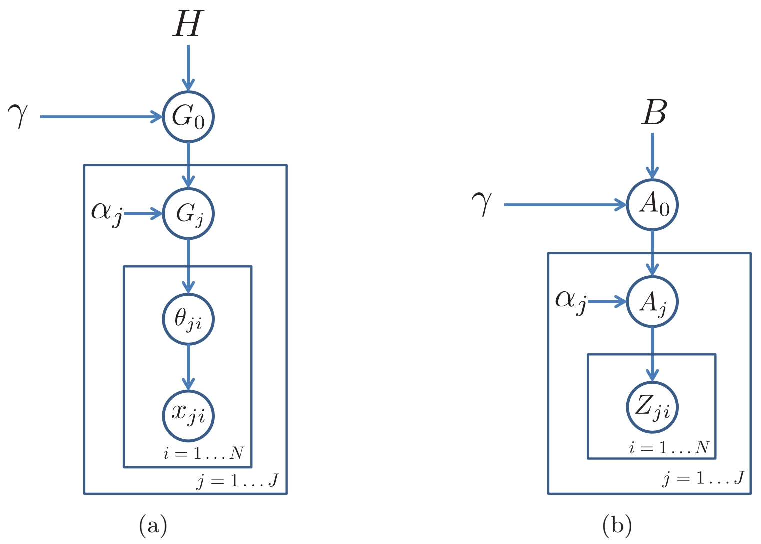


Figure 2.3: Directed graphical representation (a) hierarchical Dirichlet process mixture model (HDP-MM) (b) generative model of infinite binary matrices using hierarchical beta process (HBP).

number of topics used across the corpus can be learnt automatically and grow with the data. A directed graphical representation of HDP is shown in Figure 2.3a.

In an analogous manner to the hierarchical Dirichlet process (HDP), Thibaux and Jordan [2007] propose a prior based on a hierarchy of beta processes – termed as hierarchical beta processes (HBP) – that allows sharing of atoms drawn from different beta processes. If $A_0 \sim \text{BP}(\gamma_0, B)$ be a draw from beta process and there exist J sources, then the hierarchical beta process imposes the following hierarchy for each $j = 1, \ldots, J$:

$$A_j \sim \text{BP}(\alpha_0, A_0) \quad \text{and} \quad \mathbf{Z}_{ji} \mid A_j \sim \text{BeP}(A_j) \tag{2.29}$$

Irrespective of whether B is continuous or discrete, A_0 has the support which is a subset of the support of B. Since support of A_j's, for each j, are subsets of the support of A, A_j's share some of their atoms.

The above property of HBP can be *exploited* for sharing the factors across multiple sources. In fact, similar to the use of HDP in mixture models to share clusters (or the topics in the language modeling context), we can use HBP in combination of factor analysis model to share factors (see Figure 2.3). This is useful in applications where each object in a group can be associated to a subset of a set of factors shared across groups. HBP based factor analysis can potentially discover shared factors and individual factors specific to each source in a nonparametric setting, avoiding the need to specify the dimensionality of the latent subspaces *a priori*. The significance of this extension lies in an alternative paradigm for statistical sharing across multiple data groups, in a similar realm to HDP, but the HBP based model operates at the factor level instead of mixtures, and therefore it is more appropriate for applications such as retrieval, joint dimensionality reduction, collaborative filtering and so on.

The adaptive rejection sampling (ARS) based inference scheme proposed for HBP is elementary (as noted by Thibaux and Jordan [2007]) and further research in developing more advanced inference algorithms is *desirable* for fully exploiting beta process models. In addition to sampling the main variables of HBP based models, it is also required to sample the hyperparameters as these parameters play a direct role in determining the *extent of sharing* across different data groups. Thibaux and Jordan [2007] address this problem in a limited way by estimating these parameters either separately from the data without embedding this procedure in the sampling scheme or through model selection. It is desirable to avoid the model selection as Bayesian priors over these hyperparameters can be used to infer them *within* the model framework.

2.5 Learning from Multiple Data Sources

The proliferation of sensor networks and the Internet has created a plethora of data sources. Often, these data sources are related and can strengthen one another to improve performance of many machine learning and data mining tasks. However, improving the performance by combining them is not straightforward as these sources, usually, also have different characteristics. Traditional subspace learning methods, when applied to multiple sources separately, do not exploit their shared statistical strengths. Moreover, combining multiple data sources as a single data set and applying subspace learning does not deliver satisfactory results either, as it is unable to capture individual variations. Thus, there is a *need to develop* models which, not only can exploit the shared strength of multiple sources, but also can retain the individual variations of each source.

There are open questions such as how to analyze data with more than one set of observations/views representing the same phenomenon *or* in a more general setting, how can we extract the “useful” information that is shared across the multiple data sources? Given multiple related data sources, how can we build models for improved prediction in one data source given others? Closely related questions are how to improve the performance of unsupervised machine learning tasks (e.g. information retrieval and clustering) for one data source using the knowledge from other related data sources. Taking a broad perspective, the previous works answering the above questions can be divided into three main categories : multi-view learning, transfer learning and multi-task learning. In the following subsections, we discuss related works in these areas.

2.5.1 Multi-view Learning

In many real world applications, such as image annotation [Blei and Jordan, 2003] and web page classification [Blum and Mitchell, 1998], data is observed from multiple aspects or views. For example, a collection of images with captions can be represented with two views, one describing the content of the image and the other describing its meaning. Dependencies between these representations reveal more information on the intended semantics than either view alone. Traditional supervised and unsupervised models are built without taking different views into consideration and thus, they are not only incapable of performing joint view-level analysis (e.g. predicting the view correspondences or discovering the related clusters across different views) but also suffer in terms of task performance as they do not exploit the shared knowledge across multiple views. Addressing these problems, there has been significant amount of work exploring various multi-view formulations under different settings.

Multi-view Clustering Kailing et al. [2004] propose a density-based approach to cluster data points with multiple representations using all available views of the data. The authors extend DBSCAN algorithm to enable its application on data having multiple views or representations. Bickel and Scheffer [2004] propose a partitioning and agglomerative, multi-view clustering algorithm by utilizing different views of the data. The authors empirically show that the multi-view versions of clustering algorithms, especially K-means and EM, greatly improve their single-view counterparts. These techniques assume that all the representations of the data have the same clustering structure and attempt to find a *consensus* among different views. Real world data, on the other hand, may not fully share the common structures across different views and under these scenarios, the above assumption is only partially true. Recently, Wiswedel et al. [2010] propose a technique to cluster data with multiple views. Instead of assuming a common structure over each view, their work combines local models from individual views into a global model learnt across all views. In another related work, Salzmann et al. [2010] propose a robust approach to factorizing the latent space into shared and private spaces by introducing orthogonality constraints. A more general approach is taken in Jia et al. [2010] that allows sharing latent dimensions across any subset of the views instead of a common sharing across all views.

CCA based Approaches There has been a parallel approach to modeling the correlated data using canonical correlation analysis (CCA) by learning two orthogonal transformations such that transformed data pairs are maximally correlated. With this intuition, CCA is formulated as a multi-view learning approach in Hardoon et al. [2004] who demonstrate its use in learning a semantic representation between the web images and the associated text. A probabilistic formulation of CCA is proposed in Bach and Jordan [2005] making it possible to deal with missing data and also enabling its use as a building block in other complex probabilistic models. Bach and Jordan [2005] draw the connection between the probabilistic model and traditional CCA, showing that the maximum-likelihood solution of the probabilistic model is the same as that of the traditional CCA up to an arbitrary rotation in the latent space and projection matrices. A Bayesian extension of the maximum likelihood CCA, detecting dependency between two data sets is proposed in Klami and Kaski [2008]. A more efficient Bayesian model using variational inference was developed in Viinikanoja et al. [2010], Virtanen et al. [2011]. In addition to developing a robust method for Bayesian CCA, Viinikanoja et al. [2010] also extended it to a robust Bayesian mixture of CCA. A critical limitation of CCA is that the linear correlation between the views is computed *globally*, which captures an overall dependency across the views. Fern et al. [2005] address this limitation by constructing a mixture of *locally linear* CCA models and refer it as *correlation clustering*. Correlation clustering constructs locally linear correlation models and thus can simultaneously cluster any two data sets according to the their *local* correlation structure.

A theoretical work [Sridharan and Kakade, 2008] explicitly formalizes an information theoretic, multi-view assumption and studies the multi-view paradigm in the PAC style semi-supervised framework of Balcan and Blum [2005]. It also provides a normative justification for CCA as a dimensionality reduction technique, especially, for strictly convex loss functions. Exploiting CCA for different applications, there have been some works [Foster et al., 2008, Kakade and Foster, 2007] on multi-view dimensionality reduction framework providing sufficient conditions under which the dimensionality of data from the two views can be reduced via a projection without sacrificing their predictive power. Since clustering data in high dimensions is believed to be a hard problem in general, a common practice is to project the data into a lower dimensional subspace, e.g. via Principal Components Analysis (PCA) or random projections before clustering. Constructing such projections for data having multiple views, a multi-view clustering method using CCA is proposed in Chaudhuri et al. [2009].

Non-linear Approaches There have been relatively few attempts on nonlinear modeling of multi-view data. Globerson et al. [2005] present an Euclidean embedding of co-occurrence data such as images and text, into a single common Euclidean space, based on their co-occurrence statistics. The local structure of the embedding corresponds to the statistical correlations via random walks in the Euclidean space. In other works [Kuss and Graepel, 2003, Hardoon et al., 2004], a kernel approach is taken where the CCA subspaces are learnt after mapping the data from different views into kernel feature spaces. Taking a shared and private subspace approach, a Gaussian process latent variable model (GPLVM)

based multi-view scheme is proposed in Leen and Fyfe [2008b].
When supervised information of multiple views is available, it can be used to improve the predictive performance of the methods dealing with multi-view data. To discover the shared subspace representation of multi-view data, unsupervised methods such as CCA and other extensions, in their standard form, are not capable of incorporating label information. The discriminant approaches such as Diethe et al. [2008] do incorporate the supervised information but are unable to provide view-level predictions. Addressing this problem, a predictive subspace learning model for multi-view data is proposed in Chen et al. [2010].

Bayesian Nonparametric Approaches An important problem in multi-view learning based on CCA is to determine the number of correlation components. To this end, Rai and Daumé III [2009] propose a nonparametric, fully Bayesian framework that can automatically select the number of correlation components, and effectively capture the sparsity underlying the projections. Another nonparametric Bayesian method [Leen and Fyfe, 2008a] is proposed using a Dirichlet process mixture model of probabilistic canonical correlation analyzers for the problem of learning from two related data sets. The model can learn the number of mixtures automatically from the data, avoiding the need for separate model selection. A hierarchical nonparametric Bayesian mixture model for multi-view learning is proposed in Rogers et al. [2010]. This has the flexibility of having non-parametric models for each of the views, coupled by a flexible model for the interactions.

Distinction from Alternate Clustering We would like to distinguish between learning from multiple data sources and the *alternate clustering* paradigm [Bae and Bailey, 2006, Cui et al., 2007, Qi and Davidson, 2009, Niu et al., 2010]. Often, high dimensional data can be visualized in different ways and thus leads to multiple ways of clustering. In these scenarios, instead of clustering the data in a single specific way, it is useful to consider various non-redundant alternate clusters. Bae and Bailey [2006] present a technique to find an alternate clustering partition given an existing clustering partition. The idea is to get another clustering partition which can discover patterns distinctively different from the patterns provided by the existing clustering. Cui et al. [2007] propose a clustering scheme to discover multiple non-redundant clustering partitions by using orthogonality either in the cluster or the feature space. Non-redundant clusters are obtained by first clustering the data in a discriminating subspace and then, alternate clusters are obtained by projecting the data to subspaces orthogonal to the former subspace. Qi and Davidson [2009] propose an alternate clustering technique which can discover an alternate clustering partition given an existing clustering partition while specifying properties of the existing clustering partition that should be retained or omitted. Niu et al. [2010] propose a technique to discover multiple clustering partitions by learning non-redundant subspaces that can provide multiple views of the data and find a clustering partition using each subspace. In contrast to the above approaches, this approach optimizes a spectral clustering objective in each subspace to obtain multiple clusters while also penalizing for the redundancy among the different clusters. We note that the above set of techniques aim to achieve alternate clusters providing complementary views and are different from learning from multiple data sources as well as multi-view learning.

We note that all above approaches confine themselves to a *single* data source and focus on modeling data points that have multiple views or correspondences. These correspondences may not be available in the general setting, e.g. the textual tags of Flickr images and YouTube videos, generated in response to some real world events, may be highly correlated, however, direct correspondences between the two sources may not be available as these items (images and videos) reside on different websites. Due to this, although being highly correlated, the data from these sources should not be treated as two different views of the same event or phenomenon. In such cases, the whole class of multi-view learning algorithms become inapplicable and there is a *need to develop* frameworks which can model the data from multiple related sources without needing multiple views or correspondences. In fact, these methods are expected to relate the objects from different sources and predict the latent correspondences [Tripathi et al., 2011].

2.5.2 Transfer Learning/Multi-task Learning

Another set of techniques that deal with learning from more than one data source come under the umbrella of multi-task learning or transfer learning. Transfer learning involves related learning problems with the goal of improving performance on a particular task with the help of knowledge available across multiple related tasks. The idea is that there are some basic patterns that can be learnt from the related tasks especially when there is less training data available in the domain of interest. Closely related is the problem of multi-task learning. The main difference between transfer learning and multi-task learning is that the former focuses on improving performance of only the target task whereas the later aims at improving performance of each task with mutual help from other tasks. There are various forms of transfer learning (see the survey in Pan and Yang [2008]), e.g. inductive transfer learning [Wu and Dietterich, 2004, Liao et al., 2005, Dai et al., 2007b, Raina et al., 2007], transductive transfer learning [Dai et al., 2007a, Ling et al., 2008, Pan et al., 2008, Gao et al., 2008] or unsupervised transfer learning [Si et al., 2009, Gu and Zhou, 2009a, Dai et al., 2008].

Supervised Learning using Auxiliary Sources Increasing availability of data from related sources has given rise to their joint analysis. Previous works [Thrun, 1996, Caruana, 1997, Baxter, 2000, Ben-David and Schuller, 2003] provide a foundation for such analysis in multi-task learning. As a result, there has been lot of work to improve performance of both supervised and unsupervised tasks using data from related auxiliary sources. Ando and Zhang [2005] discover predictive structures shared among various classification tasks by learning a subspace of parameters and use it to improve target data classification. Ji et al. [2008] provide a framework for extracting shared structures in multi-label classification by learning a shared subspace which is assumed to be shared among multiple labels. Since multiple labels share the same input space, and semantics conveyed by different labels are often correlated, the authors exploit the correlation information contained in different labels. Yan et al. [2007] use a shared subspace boosting algorithm for multi-label classification combining a number of base models across multiple labels to reduce the information overlap. The base models are learnt using random subspace method and bootstrap data samples [Friedman et al., 2009]. In early attempts to multi-task learning, [Caruana, 1997] trains neural networks for jointly modeling multiple tasks. The idea of joint training was to have improved learning of some structures available across different related tasks. Evgeniou and Pontil [2004] propose a regularized multi-task learning approach that can naturally extend existing single-task *kernel* based learning methods, e.g. such as Support Vector Machines (SVM) to multi-task learning. Micchelli and Pontil [2005] suggest a way of using kernels for multi-task learning by modeling relations between the tasks and present multi-task learning algorithms. Argyriou et al. [2007] present a method for learning a low-dimensional representation shared across a set of multiple related tasks. The underlying assumption is that the classification/regression functions are related so that they all share a set of features. Building upon L_1-norm regularization problem, a new regularizer is used that controls the number of learnt features common for all tasks. Similarly, Chen et al. [2009] present a formulation for multi-task learning based on the non-convex alternating structure optimization (ASO), wherein all tasks are related by a shared feature representation.

There have also been many works that combine unlabeled data from auxiliary sources to improve the classification/recognition performance on target sources, and *vice versa*. Yang et al. [2010] propose a classification model for a target class in the absence of any labeled training example from that class while assuming that a collection of labeled examples belonging to other auxiliary classes are available. In addition, they also assume the knowledge of the prior distribution of the target class and the correlation between the target class and the auxiliary classes. Opposite to this, Zheng et al. [2011] propose a transfer learning framework that utilizes unlabeled auxiliary data to select the relevant transferable knowledge for recognizing a target object class from the background given a limited set of target training samples. In a similar approach, Raina et al. [2007] use a large number of unlabeled images (or audio samples or text documents) randomly downloaded from the

Internet to improve performance on a given image (or audio, or text) classification task without assuming a common generative model between target and auxiliary sources.

Unsupervised Learning using Auxiliary Sources The above methods focus on improving classification/regression tasks and learn joint subspaces to model either labels or classification parameters in supervised settings. There have been relatively few attempts [Si et al., 2009, Gu and Zhou, 2009a, Dai et al., 2008] towards unsupervised transfer learning and even fewer attempts towards unsupervised multi-task learning (see survey in [Pan and Yang, 2008]). Multi-task learning has been addressed mostly in supervised settings. Similar models for unsupervised case are *desired*. Dai et al. [2008] investigate an unsupervised transfer learning problem called self-taught clustering, and develop a co-clustering algorithm by using unlabeled auxiliary data to help improve the target clustering results. Si et al. [2009] propose to learn a shared subspace by minimizing the *Bregman* divergence between the distributions of the related data sources. This approach is generic to learn subspaces with different modeling goals. Although the shared subspace learnt using this approach is appropriate to model the commonalities between two related sources, it has no explicit provision for modeling individual variations of each data source. Modeling both shared and individual variations explicitly ensures that knowledge from auxiliary sources is exploited without sacrificing the knowledge available locally at each data source. Recently, a framework, which uses both shared and individual subspaces for jointly modeling data from related sources is proposed in Gu and Zhou [2009a] for a multi-task clustering application. Authors exploit the relationship among similar tasks to enhance the clustering performance of each task and present a transductive transfer classification method under the proposed framework. However, this framework needs to maintain the *same* number of clusters for each data source in their individual subspaces and forces *identical* cluster centroids in the shared subspace. This *limitation* renders the framework too restrictive for modeling real world data sources.

There has been a growing interest toward the use of prior knowledge in clustering, e.g semi-supervised clustering, constrained clustering and co-clustering. Although this does not deal with multiple data sources, it can be put under the category of transfer learning as it utilizes prior information through constraints or similarities. Wagstaff et al. [2001] develop a K-means variant that can incorporate background knowledge in the form of instance-level constraints (e.g. must link or can not link). Basu et al. [2004] propose a probabilistic framework based on Hidden Markov Random Fields (HMRFs) for semi-supervised clustering that combines the constraint-based and distance-based approaches in a unified model. Generalizing the constrained clustering idea for graphs, Kulis et al. [2005] propose a kernel based semi-supervised graph clustering and show that the method presented in Basu et al. [2004] is a special case of their framework. Li et al. [2007] combine semi-supervised clustering with consensus clustering within the framework of nonnegative matrix factorization (NMF). Utilizing NMF, Wang et al. [2008b] extend semi-supervised constrained clustering to constrained co-clustering for simultaneously clustering the dyadic data [Dhillon, 2001].

Bayesian Nonparametric Approaches A key problem to be addressed in both multi-task learning and transfer learning is how to determine the *extent* of knowledge which should be shared across multiple tasks. For example, when learning a shared subspace for the task parameters, it is important to know the optimal dimensionality of the subspace. In a supervised setting, multi-task learning has been addressed using Bayesian nonparametrics theory [Rai and Daumé III, 2010]. However, similar models for unsupervised case are *yet to be explored*. Inaccurate estimation of the sharing extent results in negative transfer learning as discussed in Pan and Yang [2008], Rosenstein et al. [2005]. The related concepts of task relatedness has also been analyzed theoretically [Ben-David and Schuller, 2003].
Recently, a nonparametric factor analysis model using beta process prior is proposed in Paisley and Carin [2009] wherein the data matrix is decomposed as a product of two matrices containing factors and features. The feature matrix is further decomposed into a product of a binary matrix (say $\mathbf{Z}$) (indicating absence or presence of a feature) and a weight matrix (feature values). The binary matrix $\mathbf{Z}$ is modeled using a Bernoulli process parametrized by a beta process. A more general covariate model, without an exchangeability assumption, is proposed in Zhou et al. [2011]. Although these nonparametric methods allow the number of factors to grow with the data (i.e. learn the subspace dimensionality automatically from the data), they are restricted to modeling a *single* data source. Extension is thus *required* for multiple data sources.
Other nonparametric models addressing multiple data sources include the work of Fox et al. [2009], who use the Indian Buffet Process (IBP) representation of the beta process, to share features amongst dynamical systems. Bernoulli processes for different dynamical objects are parametrized by a common beta process in a non-hierarchical manner. The potential benefits of a hierarchical model are yet to be explored. In another work, Saria et al. [2010] employ the hierarchical Dirichlet process (HDP) as the underlying stochastic process to model common aspects of multiple time series. However, this model is limited in its applications due to using HDP as its underlying stochastic process, and focuses on topical modeling. A multi-scale convolutional factor model using the hierarchical beta process (HBP) is presented in Chen et al. [2011]. Again, its focus is towards deep learning instead of modeling multiple data sources.

2.6 Applications to Social Media

Social tagging by users is a defining characteristic of Web 2.0 and has had a huge impact on the way we use the Web in a relatively short time. Social tagging systems exist that allow users to annotate and retrieve *any* Web-accessible item of interest, including web-pages, images, sounds, videos, blog posts, tweets, URLs, locations, and even people. Similar to keywords in information retrieval (IR), short textual descriptors, termed tags, provide concise summarization of resources, often at topical or conceptual levels, which are difficult to infer using automatic, content-based IR methods. The resulting aggregation of tags forms a *folksonomy*, which acts as a proxy for a controlled taxonomy created by information science experts, and can be used to facilitate retrieval from the resources covered by the folksonomy [Vanderwal, 2005, Peters, 2009].

Folksonomy-enabled search has been instrumental in the rising popularity of social image and video sharing platforms, such as Del.icio.us, Flickr, Picassa and YouTube. However, the use of tags poses serious challenges; the lack of constraints when creating free-text tags are part of their appeal, but as a result they tend to be noisy, ambiguous and incomplete [Marlow et al., 2006, Golder and Huberman, 2006], and can seriously degrade retrieval performance. Research has attempted to improve the accuracy of tags, but a common characteristic of the proposed solutions is a focus solely *within* the internal structure of a given tagging system. However, so long as only internal data is sought, these methods are less likely to be able to break the retrieval barriers caused by the uncertainty and noise inherent within the tags.

From a broad perspective, previous works on tagging systems have been aimed at finding tag relevance or refinement, often improving tags through modification or recommending additional tags. Marlow et al. [2006] present a taxonomy of social tagging systems, and highlight the effect of different design parameters and user communities on the make-up of the resulting tags. Folksonomies acquire differing characteristics by virtue of the myriad contextual factors and the design decisions that lead to their creation, e.g., tagging systems that allow users to see the tags of others allow for rapid vocabulary convergence as opposed to “blind” systems. Users who tag solely for the purposes of retrieving their own resources are more likely to tag with idiosyncratic terms. Some folksonomies may be rich in conceptual labels or subsidiary descriptions, whilst others may consist predominantly of precise labels for visible objects. They also present information on potential evaluative frameworks by providing a simple taxonomy of incentives and contribution models. Sigurbjörnsson and Van Zwol [2008] investigate how users tag photos and the information contained in Flickr tags. Analyzing the characteristics of tags in social web sites such as Flickr, they provide a method to recommend a set of relevant tags at different levels of exhaustiveness from the original tags. Recent works of Li et al. [2009a,b] present a method to learn social tag relevance by finding visual neighbors and voting based on tag frequency. The authors list all images for a given tag and then count the number of images which are visually similar to calculate tag relevance. Wang et al. [2008a] propose a search-based method which uses content-based retrieval method to get visually similar images from an image collection and fuse it with text-based retrieval results. These methods perform poorly in practice as visual similarity techniques have not yet matured. In addition, due to the need to search for visual neighbors, these methods are computationally expensive. In another work, Wu et al. [2009] note problems caused by irrelevant tags and describe the semantic loss caused when users do not tag a complete image but only one or two objects in the image. They propose a multi-modality tag recommendation method based on both tag and visual correlation by generating a ranking feature that uses each modality, and therefore, again suffers from the problems faced by content-based methods.

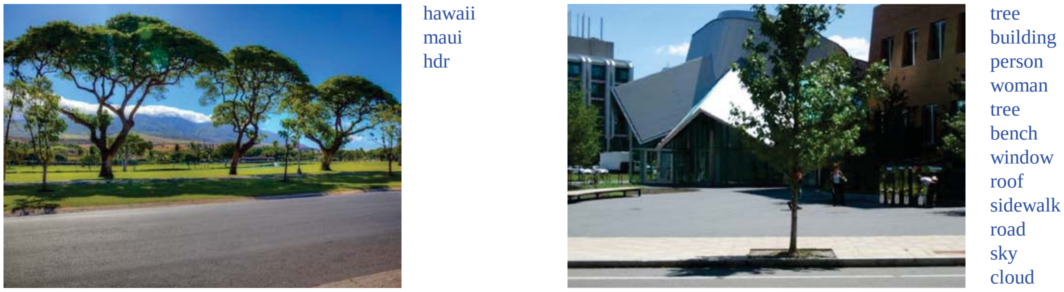


Figure 2.4: Example of images with their associated tags from Flickr (left) and LabelMe (right) data sets. As noted, Flickr tags are often noisy and incomplete (e.g., tags are absent for many objects such as ‘road’, ‘tree’, ‘sky’ or ‘cloud’, which are plainly visible in the image), whereas LabelMe tags are mostly consistent with the image existents.

Most of the social media sharing web sites have tags appearing in random order and often without any relevance information. This limits their effectiveness in social media search and mining tasks. Addressing this problem, Liu et al. [2009] propose a tag ranking scheme for Flickr images, aiming to rank the tags according to their relevance to the image content. Begelman et al. [2006] discuss the use of clustering techniques to improve the search and exploration in the tagging space and enhance the user experience on collaborative tagging systems. Many social media sharing websites encourage their users to create virtual social networks, which acts as a medium for the exchange of data and further boosts media sharing. Exploiting the social graph, Konstas et al. [2009] develop a collaborative track recommendation system that adapts to the personal information needs of each user by taking into account both the social annotation and links in the social graph established among users, items and tags. Golder and Huberman [2006] analyze popularly annotated links in Del.icio.us for understanding usage patterns of a social annotation system and find that usually after nearly the first one hundred bookmarks, the frequency of a tag saturates to a fixed proportion of the total number of tags. Halpin et al. [2007] investigate the stability of tags and show that it follows a stable power law distribution. Hsu and Chen [2008] propose a tag normalization algorithm to unify the users' annotations and predict the stabilized tag set when only a small amount of early user annotations are available.

Note that all the aforementioned approaches are confined to the noisy tags within the primary data set of interest. Often, the quality of user-contributed tags is so poor that none of these methods work well and an alternative method must be sought to improve performance of retrieval and mining tasks. One such alternative is to use auxiliary sources of information, resonating with the call in Kankanhalli and Rui [2008], which emphasizes the importance of the use of auxiliary source in the form of web-data and propose that the judicious use of such easily available data can substantially improve precision and recall. They also emphasize the need to formalize the notion of auxiliary sources of information in a rigorous framework. Theoretical justification for the use of side information is discussed in Chechik and Tishby [2003]. A successful example was shown in Raina et al. [2007] but the unlabeled data from auxiliary sources was used for a supervised task instead of unsupervised task. Another motivation is that when the target data set is noisy, the use of auxiliary data which is less noisy may help [Crammer et al., 2006]. There exist many tagged data sets which are suitable for the purpose of enhancing performance of social media retrieval from a primary data set, and are readily available. For example, LabelMe [Russell et al., 2008] and Caltech-101 [Fei-Fei et al., 2006] are often used for evaluating and benchmarking object detection and classification techniques. In these data sets, users follow certain guidelines and/or a controlled vocabulary to construct ground-truth labels or tags, making them ideal auxiliary sources of information[2].

The use of auxiliary information can be perceived as an application of unsupervised transfer learning. In particular, the knowledge from other related auxiliary sources can be transferred to improve performance. However, this should be carried out under general settings without needing the correspondences among the target and the auxiliary sources. Developing such frameworks would help transfer knowledge from one domain to another of different quality. For example, the specific task we discuss above can be perceived as transferring the view of LabelMe images provided by its folksonomy–static, visual objects, typified by nouns–to the Flickr images; But it might be just as desirable to transfer learning from an event or action-oriented folksonomy, typified by verbs, onto the LabelMe data

[2] For example, a comparative analysis of the LabelMe and Flickr data sets reveals the Flickr tags to contain five times more function words (which exist only for grammatical purposes, and do not correspond directly to visual existents within an image) than LabelMe, and also contains less nouns.

set. The Web continues to spawn new and specialized folksonomies, and the ability to cross-leverage these otherwise fragmented resources is *desirable*.

2.6.1 Vector Space Model and Social Media Retrieval

Vector space model [Raghavan and Wong, 1986, Manning and Schütze, 1999] is one of the earliest approaches taken toward information retrieval. In this model, a document is represented by a vector with the dimensionality as the size of the vocabulary. Given the vector space representation, the similarity between two documents can be computed using any standard metric in Euclidean space. However, a similarity which is the most widely used is cosine similarity.

Given a document $\mathcal{D}$, its vector space representation d is constructed by setting the i-th element, i.e. d_i to the frequency of i-th vocabulary word in the document $\mathcal{D}$. Similarly, given a set of query keywords S_q, a query vector q is constructed in vector space by setting its i-th element to the frequency of i-th vocabulary word in the document S_q. Relevance of the document $\mathcal{D}$ with respect to the set of query keywords S_q is determined using the similarity function such as

$$\text{Relevance}\left(S_q, \mathcal{D}\right) = \frac{q^{\mathsf{T}} d}{\|q\|_2 \, \|d\|_2} \tag{2.30}$$

Note that cosine similarity is often used when the magnitude of vectors is not an important factor while considering similarity. It is appropriate for text based retrieval because two documents may be very similar due to the use of common words, irrespective of their document lengths. Generally, the set of query keywords has few words while the documents to be retrieved are much longer in length.

The above model can be directly used for tag based social media retrieval by treating the list of tags as a document. However, because of the above-discussed noisy, subjective and incomplete nature of the tags, performance of vector space model is poor and unsatisfactory. Use of subspace learning can deal with these issues up to some extent. By capturing co-occurrences through latent basis vectors, the problem of subjectivity can be addressed similar to the 'polysemy' problem in text mining. The incomplete nature of the tags can be dealt with using the framework of probabilistic subspace learning that handle the uncertainties of data effectively. Additional improvement in the performance of social media mining tasks can be obtained by the use of auxiliary sources using the framework of unsupervised transfer learning.

2.6.2 Cross-Media Retrieval

Often, social media users are interested in searching and organizing different media genres simultaneously. When asked to aggregate information related to some event, one usually wants to look on more than a single source to get relatively better and more complete information. Despite a great deal of research work dedicated to the exploration of content-based information retrieval, the performance is not satisfactory due to the semantic gap. The problem is aggravated further when the items across *multiple* heterogeneous media need to be searched. Previous approaches to content-based cross-media retrieval [Yang et al., 2009, Yi et al., 2008, Zhuang et al., 2008] use the concept of a Multimedia Document (MMD), which is a set of *co-occurring* multimedia objects that are of different modalities but carry the same semantics. The two multimedia objects can be regarded as context for each other if they are in the same MMD, and thus the combination of content and context is likely to enable multimedia retrieval methods to overcome the semantic gap, as the system acquires richer representations from co-occurring multi-modal items. However, this line of research crucially depends on a set of *co-occurring* multimedia objects to define the MMDs, which often may not be available. For example, many YouTube videos occur alone on a web page without the context of photos or weblogs and therefore contextual relations are not available for MMD-based methods.

The above problem necessitates the development of formal frameworks for modeling multiple heterogeneous sources simultaneously. However, modeling data across multiple heterogeneous and disparate sources is challenging. For example, how do we model text, image and video together? It is clearly understood that at the semantic level, they provide much richer information together. The question is, can we exploit these strengths? 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 [Datta et al., 2008].

Although tags provide a unified way to search media items across different media, when used in conjunction with content-based methods, retrieval is limited to the medium for which content is processed. For example, given Oscars related photos, videos and blogs, it is obvious that these media items are semantically related but correlating them using content is not possible due to the semantic gap. Therefore, we need a framework modeling multiple folksonomies (providing rich sources of information at semantic level) simultaneously to improve the retrieval within one domain as well as enabling cross-media retrieval across an arbitrary number of media where tags are available. A naïve way of doing cross-media retrieval using these textual tags is to use some similarity function akin to text retrieval such as *Jaccard* coefficient [Tan et al., 2006], cosine similarity or Okapi BM25 [Maron and Kuhns, 1960] for relating items from different media. However, performance of this method is generally poor due to the noisy tags. Subspace learning techniques can help to deal with these noisy tags effectively but these techniques, in their standard form, can not be directly used as they have been developed for modeling data from a single source. This gives rise to a *need for developing* subspace learning frameworks, to model multiple media jointly on a common ground provided by textual tags.

2.7 Conclusion

In this chapter, we have discussed the necessary background over which the material of the following chapters would develop. We described the subspace learning techniques such as nonnegative matrix factorization (NMF), factor analysis and briefly touch upon probabilistic extensions. For probabilistic techniques, we discussed many inference techniques such as Metropolis-Hasting sampling, Gibbs sampling along with their collapsed variants. We also discussed the use of auxiliary variables in the sampling techniques. Addressing the problem of model selection, probabilistic techniques use the theory of Bayesian nonparametrics. We briefly covered some of the background on Bayesian nonparametric priors which would be useful for the work done in this thesis. In addition to these background materials, we extensively reviewed the literature in the area of learning from multiple data sources. In this endeavor, we mainly focused on multi-view learning, transfer learning and multi-task learning. We discussed some open problems in these areas and motivated the requirement to develop formal frameworks for unsupervised modeling of multiple data sources. Finally, we discussed some unsupervised applications in social media and motivated the need of joint modeling. In the following chapters, we shall develop a generic framework of shared subspace learning to model multiple related data sources by extending the traditional single source subspace learning techniques.

Chapter 3

S-NMF : Nonnegative Shared Subspace Learning for Two Data Sources

In the previous chapter, we reviewed the works in transfer learning and other related areas and identified open problems. Unsupervised joint modeling of related data sources is one such problem that has been addressed in a limited way, except in the context of multi-view learning and remains open in general settings. Addressing this gap, we present a shared subspace learning framework for *jointly* modeling two data sources. This framework is useful for transfer learning or multi-task learning applications where the shared subspace is used to transfer knowledge from auxiliary to target domains. A key advantage of the proposed framework is the modeling flexibility to *explicitly* vary the level of sharing between data sources. Additionaly, the proposed framework learns the individual subspaces - crucial for retaining domain specific knowledge.

To learn the shared and individual subspaces, we propose a novel joint matrix factorization extending nonnegative matrix factorization (NMF) [Lee and Seung, 2001]. Nonnegative constraints imposed on the factorization matrices yield the subspace basis vectors that are part-based and provide locally meaningful representation of the data. We provide an efficient algorithm for learning the factorization, analyze the algorithmic complexity, and provide a proof of convergence. The proposed framework is used to improve the performance of tag based social media retrieval on a target data set by exploiting the knowledge present in the tags of another *related* and *less noisy* auxiliary data set. The effectiveness of the proposed approach is validated on image and video retrieval tasks in which tags from the LabelMe data set are used to improve image retrieval performance for a Flickr data set and video retrieval performance for a YouTube data set.

This chapter is organized as follows. Section 3.1 presents the Shared NMF (S-NMF) framework and describes the shared subspace learning. Section 3.2 describes the tag based social image/video retrieval using S-NMF. Section 3.3 presents the experimental results and conclusions are drawn in Section 3.4.

3.1 Nonnegative Shared Subspace Learning

3.1.1 Problem Formulation

We present a framework that aims to capture both shared and individual basis vectors for two data sources. This leads to a partitioning of the data subspace into two parts. The first one is common to the data sources while the second is specific to a data source. Let us represent the two data sources by matrices $\mathbf{X}$ and $\boldsymbol{Y}$ with dimensions $M \times N_1$ and $M \times N_2$ respectively (where $\mathbf{X} = [x_1, \ldots, x_{N_1}]$, $\boldsymbol{Y} = [y_1, \ldots, y_{N_2}]$ and $x_i, y_i \in \mathbb{R}^{M \times 1}$) and write the decomposition and partition of the matrices as

$$\mathbf{X} \approx \underbrace{[\mathbf{W} \mid \boldsymbol{U}]}_{\mathbf{F}} \mathbf{H} = \mathbf{FH} \tag{3.1}$$

$$\boldsymbol{Y} \approx \underbrace{[\mathbf{W} \mid \boldsymbol{V}]}_{\mathbf{G}} \mathbf{L} = \mathbf{GL} \tag{3.2}$$

where $\mathbf{W}$ is a $M \times K$ matrix whose columns span the common or shared subspace; $\boldsymbol{U}$ and $\boldsymbol{V}$ span the remaining individual subspaces and have dimensions $M \times (R_1 - K)$ and $M \times (R_2 - K)$ respectively. R_1 and R_2 are the dimensionality of low-rank underlying subspaces for $\mathbf{X}$ and $\boldsymbol{Y}$; K denotes the number of shared basis vectors. $\mathbf{H}$ and $\mathbf{L}$ are the encoding matrices and have dimensions $R_1 \times N_1$ and $R_2 \times N_2$ respectively. We define $\mathbf{F}$ and $\mathbf{G}$ as $\mathbf{F} \triangleq [\mathbf{W} \mid \boldsymbol{U}]$ and $\mathbf{G} \triangleq [\mathbf{W} \mid \boldsymbol{V}]$. We note that although $\mathbf{X}$ and $\boldsymbol{Y}$ have different vocabularies in general, they can be merged to construct a common vocabulary with M words.

We further impose a constraint of nonnegativity to achieve part-based representation. By nonnegativity, we mean that the elements of matrices $\mathbf{W}, \boldsymbol{U}, \boldsymbol{V}, \mathbf{H}$, and $\mathbf{L}$ are restricted to take only nonnegative values.

3.1.2 Optimization and Algorithm

To learn the required subspaces, we minimize the Frobenius norm of the joint decomposition error in the following manner

$$\min_{\mathbf{W}, \boldsymbol{U}, \boldsymbol{V}, \mathbf{H}, \mathbf{L} \geq 0} \left\{ \frac{\|\mathbf{X} - [\mathbf{W} \mid \boldsymbol{U}]\mathbf{H}\|_F^2}{\|\mathbf{X}\|_F^2} + \frac{\|\boldsymbol{Y} - [\mathbf{W} \mid \boldsymbol{V}]\mathbf{L}\|_F^2}{\|\boldsymbol{Y}\|_F^2} \right\} \tag{3.3}$$

which translates to a constrained optimization problem with the following objective function

$$\begin{aligned}&\min D\left(\mathbf{W},\boldsymbol{U},\boldsymbol{V},\mathbf{H},\mathbf{L}\right)\triangleq\frac{1}{2}\left\{\|\mathbf{X}-[\mathbf{W}\mid\boldsymbol{U}]\mathbf{H}\|_F^2+\lambda\left\|\boldsymbol{Y}-[\mathbf{W}\mid\boldsymbol{V}]\mathbf{L}\right\|_F^2\right\}\\&\text{subject to }\mathbf{W},\boldsymbol{U},\boldsymbol{V},\mathbf{H},\mathbf{L}\geq 0\end{aligned}\tag{3.4}$$

where $\|.\|_F$ denotes the Frobenius norm and $\lambda=\|\mathbf{X}\|_F^2\,/\,\|\boldsymbol{Y}\|_F^2$ is defined as the relative ratio between Frobenius norms of the two data matrices.

Expressing $D\left(\mathbf{W},\boldsymbol{U},\boldsymbol{V},\mathbf{H},\mathbf{L}\right)$ elementwise, this optimization can be efficiently solved in similar fashion to the original formulation of NMF [Lee and Seung, 2001], yielding the following multiplicative update equations for $\mathbf{W}$

$$(\mathbf{W})_{ab}\leftarrow(\mathbf{W})_{ab}\frac{\left(\mathbf{X}\mathbf{H}_w^{\mathsf{T}}+\lambda\boldsymbol{Y}\mathbf{L}_w^{\mathsf{T}}\right)_{ab}}{(\mathbf{W}\mathbf{H}_w\mathbf{H}_w^{\mathsf{T}}+\boldsymbol{U}\mathbf{H}_u\mathbf{H}_w^{\mathsf{T}})_{ab}+\lambda\left(\mathbf{W}\mathbf{L}_w\mathbf{L}_w^{\mathsf{T}}+\boldsymbol{V}\mathbf{L}_v\mathbf{L}_w^{\mathsf{T}}\right)_{ab}}\tag{3.5}$$

where $\mathbf{H}\triangleq\left[\mathbf{H}_w^{\mathsf{T}}\mid\mathbf{H}_u^{\mathsf{T}}\right]^{\mathsf{T}}$and $\mathbf{L}\triangleq\left[\mathbf{L}_w^{\mathsf{T}}\mid\mathbf{L}_v^{\mathsf{T}}\right]^{\mathsf{T}}$. Similar multiplicative update equations are obtained for $\boldsymbol{U},\boldsymbol{V},\mathbf{H}$ and $\mathbf{L}$ and given by

$$(\mathbf{H})_{ab}\leftarrow(\mathbf{H})_{ab}\frac{\left(\mathbf{F}^{\mathsf{T}}\mathbf{X}\right)_{ab}}{\left(\mathbf{F}^{\mathsf{T}}\mathbf{F}\mathbf{H}\right)_{ab}}\tag{3.6}$$

$$(\mathbf{L})_{ab}\leftarrow(\mathbf{L})_{ab}\frac{\left(\mathbf{G}^{\mathsf{T}}\boldsymbol{Y}\right)_{ab}}{\left(\mathbf{G}^{\mathsf{T}}\mathbf{G}\mathbf{L}\right)_{ab}}\tag{3.7}$$

$$(\boldsymbol{U})_{ab}\leftarrow(\boldsymbol{U})_{ab}\frac{\left(\mathbf{X}\mathbf{H}_u^{\mathsf{T}}\right)_{ab}}{\left(\mathbf{W}\mathbf{H}_w\mathbf{H}_u^{\mathsf{T}}+\boldsymbol{U}\mathbf{H}_u\mathbf{H}_u^{\mathsf{T}}\right)_{ab}}\tag{3.8}$$

$$(\boldsymbol{V})_{ab}\leftarrow(\boldsymbol{V})_{ab}\frac{\left(\boldsymbol{Y}\mathbf{L}_v^{\mathsf{T}}\right)_{ab}}{\left(\mathbf{W}\mathbf{L}_w\mathbf{L}_v^{\mathsf{T}}+\boldsymbol{V}\mathbf{L}_v\mathbf{L}_v^{\mathsf{T}}\right)_{ab}}\tag{3.9}$$

The above multiplicative update equations[1] of the proposed joint subspace learning carry an intuition similar to that in NMF: if the perfect factorization is achieved, the multiplicative factors in the update equations reduce to unity. In other words, it can be verified by inspection from the update equations (3.5-3.9) that when the factorization for the two data sets $\mathbf{X}$ and $\boldsymbol{Y}$ in (3.1) and (3.2) are exact, then the multiplicatives on the RHS become unity, as expected at convergence. A pseudo code for the proposed nonnegative joint subspace factorization is shown in Algorithm 3.1.

We note two special cases from our joint factorization framework. When there is no sharing (i.e., $K=0$ or $\mathbf{W}$ does not exist, and hence no common basis vectors between the two subspaces), the update equations for $\boldsymbol{U}$, $\boldsymbol{V}$, $\mathbf{H}$ and $\mathbf{L}$ reduce to individual NMF for $\mathbf{X}$ and

[1] Note that in the implementation, a common practice is to add a small number δ (we used $\delta=10^{-9}$) in the denominator to avoid division by zero.

Algorithm 3.1 Shared Nonnegative Matrix Factorization (S-NMF).

```
**Input**: Data matrices $\mathbf{X}$, $\boldsymbol{Y}$, Parameters $R_1$, $R_2$, $K$ and a threshold $\epsilon$.
let $\lambda = \|\mathbf{X}\|_F^2 \,/\, \|\boldsymbol{Y}\|_F^2$.
initialize $\mathbf{W}_0, \boldsymbol{U}_0, \boldsymbol{V}_0, \mathbf{H}_0, \mathbf{L}_0$ randomly.
set $r = 1$.
**while** ($r <$MaxNumIters) or ($C < \epsilon$) **do**
    update $\mathbf{W}_r, \boldsymbol{U}_r, \boldsymbol{V}_r, \mathbf{H}_r, \mathbf{L}_r$ according to Eqs (3.5)– (3.9)
    normalize each column of $\mathbf{W}_r$, $\boldsymbol{U}_r$ and $\boldsymbol{V}_r$ to 1.
    let $\mathbf{X}_r = [\mathbf{W}_r \mid \boldsymbol{U}_r]\mathbf{H}_r$ and $\boldsymbol{Y}_r = [\mathbf{W}_r \mid \boldsymbol{V}_r]\mathbf{L}_r$
    compute error $C = \|\mathbf{X} - \mathbf{X}_r\|_F^2 + \lambda \|\boldsymbol{Y} - \boldsymbol{Y}_r\|_F^2$
    $r = r + 1$
**end while**
**Output**: return $\mathbf{W}, \boldsymbol{U}, \boldsymbol{V}, \mathbf{H}, \mathbf{L}$.
```

$\boldsymbol{Y}$ described in Lee and Seung [2001]. Similarly, when we force a single shared subspace for the two datasets, i.e., $\dim(\boldsymbol{U}) = 0$ and $\dim(\boldsymbol{V}) = 0$, the update equations reduce to the “fully shared” formulation recently studied in Wu et al. [2008].

3.1.3 Convergence Analysis and Multiplicative Updates

To prove the convergence of above multiplicative updates, we need to extend the proof of convergence of NMF to our case. The proof of convergence makes use of an auxiliary upper bound function akin to the auxiliary lower bound used in the EM-algorithm [Dempster et al., 1977]. The main objective function is indirectly minimized by iteratively finding a solution that minimizes the auxiliary function. Let x_i, y_i, w_i, u_i and v_i denote the i-th row of matrices $\mathbf{X}$, $\boldsymbol{Y}$, $\mathbf{W}$, $\boldsymbol{U}$ and $\boldsymbol{V}$ respectively.

Definition. [Lee and Seung, 2001] Let $D(w)$ denote the cost function defined in Eq (3.4) for w (a row of $\mathbf{W}$) when all other rows of matrix $\mathbf{W}$ and matrices $\boldsymbol{U}$, $\boldsymbol{V}$, $\mathbf{H}$ and $\mathbf{L}$ are held fixed. $G(w, w')$ is defined as an upper bound function for $D(w)$ if $G(w, w') \geqslant D(w)$ and equality is satisfied iff $w = w'$.

Note that minimizing the upper bound function G at every update of w leads to non-increasing function D on every update. In other words, if $w^{t+1} = \underset{w}{\operatorname{argmin}}\, G(w, w^t)$, then

$$D\left(w^{t+1}\right) \leq G\left(w^{t+1}, w^t\right) \leq G\left(w^t, w^t\right) = D\left(w^t\right)$$

Also note that when $D\left(w^{t+1}\right) = D\left(w^t\right)$, it implies that w^t is a local minimum of $G\left(w, w^t\right)$ and if derivatives of D exist and are continuous in a small neighborhood $\left\|w^t - \delta w\right\| < \epsilon_0$, this also implies that $\nabla_w D\left(w^t\right) = 0$.

Lemma 3.1. *If $K(w_i)$ is the diagonal matrix with its (a,b)-th element given by*

$$K_{ab}(w_i) = \mathbf{1}_{a,b}\frac{\left(\mathbf{H}_w\mathbf{H}_w^\mathsf{T} w_i + \mathbf{H}_w\mathbf{H}_u^\mathsf{T} u_i\right)_a + \lambda\left(\mathbf{L}_w\mathbf{L}_w^\mathsf{T} w_i + \mathbf{L}_w\mathbf{L}_v^\mathsf{T} v_i\right)_a}{\left(w_i\right)_a}$$

then $G\left(w_i, w_i^t\right)$ is an auxiliary function for $D\left(w_i\right)$ where

$$G\left(w_i, w_i^t\right) = D\left(w_i^t\right) + \left(w_i - w_i^t\right)^\mathsf{T}\nabla_{w_i}D\left(w_i^t\right) + \frac{1}{2}\left(w_i - w_i^t\right)^\mathsf{T}K\left(w_i^t\right)\left(w_i - w_i^t\right)$$

and $\mathbf{H} \triangleq \left[\mathbf{H}_w^\mathsf{T} \mid \mathbf{H}_u^\mathsf{T}\right]^\mathsf{T}$, $\mathbf{L} \triangleq \left[\mathbf{L}_w^\mathsf{T} \mid \mathbf{L}_v^\mathsf{T}\right]^\mathsf{T}$. $\mathbf{1}_{a,b}$ denotes the identity function, i.e., equal 1 if $a = b$ and 0 otherwise.

Proof. For $G\left(w_i, w_i^t\right)$ to be an auxiliary function for $D\left(w_i\right)$, we need to show that $G\left(w_i, w_i^t\right) \geqslant D(w_i)$ for all $w_i^t \neq w_i$. The Taylor series expansion of $D\left(w_i\right)$ around w_i^t can be written as the following

$$D\left(w_i\right) = D\left(w_i^t\right) + \left(w_i - w_i^t\right)^\mathsf{T}\nabla_{w_i}D\left(w_i\right) + \frac{1}{2}\left(w_i - w_i^t\right)^\mathsf{T}\nabla_{w_i}^2 D\left(w_i\right)\left(w_i - w_i^t\right) \quad (3.10)$$

The first and second derivatives of $D\left(w_i\right)$ can be written as

$$\nabla_{w_i}D\left(w_i\right) = -\mathbf{H}_w x_i + \mathbf{H}_w\mathbf{H}_w^\mathsf{T} w_i + \mathbf{H}_w\mathbf{H}_u^\mathsf{T} u_i + \lambda\left(-\mathbf{L}_w y_i + \mathbf{L}_w\mathbf{L}_w^\mathsf{T} w_i + \mathbf{L}_w\mathbf{L}_v^\mathsf{T} v_i\right) \quad (3.11)$$

$$\nabla_{w_i}^2 D\left(w_i\right) = \mathbf{H}_w\mathbf{H}_w^\mathsf{T} + \lambda\mathbf{L}_w\mathbf{L}_w^\mathsf{T} \quad (3.12)$$

Comparing $D\left(w_i\right)$ with $G\left(w_i, w_i^t\right)$, it can be seen that we need to prove the following inequality

$$\left(w_i - w_i^t\right)^\mathsf{T}\left(K\left(w\right) - \mathbf{H}_w\mathbf{H}_w^\mathsf{T} - \lambda\mathbf{L}_w\mathbf{L}_w^\mathsf{T}\right)\left(w_i - w_i^t\right) \geq 0 \quad (3.13)$$

which turns out to be the positive semi-definite property of the matrix $K\left(w\right) - \mathbf{H}_w\mathbf{H}_w^\mathsf{T} - \lambda\mathbf{L}_w\mathbf{L}_w^\mathsf{T}$. To prove this, consider the matrix with elements

$$M_{ab}\left(w^t\right) = \left(w_i^t\right)_a\left(K\left(w^t\right) - \mathbf{H}_w\mathbf{H}_w^\mathsf{T} - \lambda\mathbf{L}_w\mathbf{L}_w^\mathsf{T}\right)_{ab}\left(w_i^t\right)_b$$

which is essentially a rescaling of the elements of matrix $K\left(w\right) - \mathbf{H}_w\mathbf{H}_w^\mathsf{T} - \lambda\mathbf{L}_w\mathbf{L}_w^\mathsf{T}$, and thus does not affect the positive definiteness. Equivalently, we need to prove $\nu^\mathsf{T}M\nu \geq 0, \forall\nu$. By explicitly expressing $\nu^\mathsf{T}M\nu$, we can show that it is a sum of nonnegative terms and thus greater than zero. Formally,

$$\begin{aligned}
\nu^{\mathsf{T}} M \nu &= \sum_{a,b} \nu_a \left(w_i^t\right)_a \left(K\left(w^t\right) - \mathbf{H}_w \mathbf{H}_w^{\mathsf{T}} - \lambda \mathbf{L}_w \mathbf{L}_w^{\mathsf{T}}\right)_{ab} \left(w_i^t\right)_b \nu_b \\
&= \sum_{a,b} \left(w_i^t\right)_a \left(\mathbf{H}_w \mathbf{H}_w^{\mathsf{T}}\right)_{ab} \left(w_i^t\right)_b \frac{\left(\nu_a - \nu_b\right)^2}{2} \\
&+ \lambda \sum_{a,b} \left(w_i^t\right)_a \left(\mathbf{L}_w \mathbf{L}_w^{\mathsf{T}}\right)_{ab} \left(w_i^t\right)_b \frac{\left(\nu_a - \nu_b\right)^2}{2} \\
&+ \sum_{a,b} \nu_a^2 \left(\mathbf{H}_w \mathbf{H}_u^{\mathsf{T}}\right)_{ab} \left(w_i^t\right)_a \left(w_i^t\right)_b \\
&+ \lambda \sum_{a,b} \nu_a^2 \left(\mathbf{L}_w \mathbf{L}_v^{\mathsf{T}}\right)_{ab} \left(w_i^t\right)_a \left(w_i^t\right)_b \\
&\geq 0
\end{aligned}$$

□

Theorem 3.1. *The optimization function $D\left(\mathbf{W}, \boldsymbol{U}, \boldsymbol{V}, \mathbf{H}, \mathbf{L}\right)$ is non-increasing under the multiplicative updates of Eq (3.5).*

Proof. Keeping all other variables fixed, the auxiliary function $G\left(w_i, w_i^t\right)$ indirectly minimizes the cost function $D\left(w_i\right)$. The gradient descent update of w_i can be written as

$$w_i^{t+1} = w_i^t - \eta\left(w_i^t\right) \nabla_{w_i} D\left(w_i^t\right)$$

where $\eta\left(w_i^t\right)$ is the gradient descent step size. Using the first derivative of $G\left(w_i, w_i^t\right)$ in the gradient descent update rule, we obtain $\eta_{w_i^t} = K^{-1}\left(w_i^t\right)$ where $K\left(w_i^t\right)$ is defined in Lemma 3.1. Using this step size, the above update equation can be written as

$$w_i \leftarrow w_i - K^{-1}\left(w_i\right) \nabla_{w_i} D\left(w_i\right)$$

where we have dropped the iteration superscript for notational clarity. After substituting for $\nabla_{w_i} D\left(w_i\right)$ from Eq. (3.11), we get the following element-wise update for w_i

$$\begin{aligned}
\left(w_i\right)_a \quad \leftarrow \quad & \left(w_i\right)_a - \frac{\left(w_i\right)_a}{\left(\mathbf{H}_w \mathbf{H}_w^{\mathsf{T}} w_i + \mathbf{H}_w \mathbf{H}_u^{\mathsf{T}} u_i\right)_a + \lambda\left(\mathbf{L}_w \mathbf{L}_w^{\mathsf{T}} w_i + \mathbf{L}_w \mathbf{L}_v^{\mathsf{T}} v_i\right)_a} \times \\
& \left[\left(-x_i^{\mathsf{T}} \mathbf{H}_w^{\mathsf{T}} + w_i^{\mathsf{T}} \mathbf{H}_w \mathbf{H}_w^{\mathsf{T}} + u_i^{\mathsf{T}} \mathbf{H}_u \mathbf{H}_w^{\mathsf{T}}\right)_a + \lambda\left(-y_i^{\mathsf{T}} \mathbf{L}_w^{\mathsf{T}} + w_i^{\mathsf{T}} \mathbf{L}_w \mathbf{L}_w^{\mathsf{T}} + v_i^{\mathsf{T}} \mathbf{L}_v \mathbf{L}_w^{\mathsf{T}}\right)_a\right]
\end{aligned}$$

Using matrix notation, the above expression can be written as the following

$$\left(\mathbf{W}\right)_{ab} \leftarrow \left(\mathbf{W}\right)_{ab} \frac{\left(\mathbf{X}\mathbf{H}_w^{\mathsf{T}} + \lambda \boldsymbol{Y} \mathbf{L}_w^{\mathsf{T}}\right)_{ab}}{\left(\mathbf{W}\mathbf{H}_w \mathbf{H}_w^{\mathsf{T}} + \boldsymbol{U}\mathbf{H}_u \mathbf{H}_w^{\mathsf{T}}\right)_{ab} + \lambda\left(\mathbf{W}\mathbf{L}_w \mathbf{L}_w^{\mathsf{T}} + \boldsymbol{V}\mathbf{L}_v \mathbf{L}_w^{\mathsf{T}}\right)_{ab}}$$

which is the multiplicative update of Eq (3.5). □

Lemma 3.2. *If $K(u_i)$ is the diagonal matrix with its (a,b)-th element given by*

$$K_{ab}\left(u_i\right) = \mathbf{1}_{a,b}\frac{\left(\mathbf{H}_u\mathbf{H}_w^{\mathsf{T}}w_i + \mathbf{H}_u\mathbf{H}_u^{\mathsf{T}}u_i\right)_a}{\left(u_i\right)_a}$$

then $G\left(u_i, u_i^t\right)$ is an auxiliary function for $D\left(u_i\right)$ where

$$G\left(u_i, u_i^t\right) \quad = \quad D\left(u_i^t\right) + \left(u_i - u_i^t\right)^{\mathsf{T}}\nabla_{u_i}D\left(u_i^t\right) + \frac{1}{2}\left(u_i - u_i^t\right)^{\mathsf{T}}K\left(u_i^t\right)\left(u_i - u_i^t\right)$$

Proof. Proof is similar to that of Lemma 3.1. □

Theorem 3.2. *The optimization function $D\left(\mathbf{W}, \boldsymbol{U}, \boldsymbol{V}, \mathbf{H}, \mathbf{L}\right)$ is non-increasing under the multiplicative updates of Eq (3.8).*

Proof. The first and second derivatives of $D\left(u_i\right)$ are given by

$$\nabla_{u_i}D\left(u_i\right) = -\mathbf{H}_u x_i + \mathbf{H}_u\mathbf{H}_w^{\mathsf{T}}w_i + \mathbf{H}_u\mathbf{H}_u^{\mathsf{T}}u_i$$
$$\nabla_{u_i}^2 D\left(u\right) = \mathbf{H}_u\mathbf{H}_u^{\mathsf{T}}$$

Taking $K\left(u_i^t\right)$ as defined in Lemma 3.2 and using the gradient descent step size $\eta_{u_i^t} = K^{-1}\left(u_i^t\right)$ (as obtained from the first derivative of $G\left(u_i, u_i^t\right)$), we get the following element-wise update for u_i

$$\left(u_i\right)_a \quad \leftarrow \quad \left(u_i\right)_a - \left(-\mathbf{H}_u x_i + \mathbf{H}_u\mathbf{H}_w^{\mathsf{T}}w_i + \mathbf{H}_u\mathbf{H}_u^{\mathsf{T}}u_i\right)_a \times \frac{\left(u_i^t\right)_a}{\left(\mathbf{H}_u\mathbf{H}_w^{\mathsf{T}}w_i + \mathbf{H}_u\mathbf{H}_u^{\mathsf{T}}u_i^t\right)_a}$$

Using matrix notation, the above expression can be written as

$$\left(\boldsymbol{U}\right)_{ab} \leftarrow \left(\boldsymbol{U}\right)_{ab}\frac{\left(\mathbf{X}\mathbf{H}_u^{\mathsf{T}}\right)_{ab}}{\left(\mathbf{W}\mathbf{H}_w\mathbf{H}_u^{\mathsf{T}} + \boldsymbol{U}\mathbf{H}_u\mathbf{H}_u^{\mathsf{T}}\right)_{ab}}$$

which is the multiplicative update of Eq (3.8). □

The multiplicative update expression for $\boldsymbol{V}$ and the convergence result in this case can be derived by symmetry between $\boldsymbol{U}$ and $\boldsymbol{V}$. The updates and the convergence analysis for $\mathbf{H}$ and $\mathbf{L}$ are identical to the basic NMF updates since the modeling changes in S-NMF do not affect these matrices.

3.1.4 Complexity Analysis and Subspace Dimensionality

For complexity analysis, we compare the computational cost of the basic NMF algorithm in Lee and Seung [2001] and our proposed S-NMF. For an $M \times N_1$ matrix $\mathbf{X}$ and an $M \times N_2$ matrix $\boldsymbol{Y}$, assuming that K basis vectors are shared and that the latent space dimensionalities for decomposition of $\mathbf{X}$ and $\boldsymbol{Y}$ are R_1 and R_2 respectively, the computational complexity for the S-NMF per iteration is $\mathcal{O}\left(\max\left\{MN_1R_1, MN_2R_2\right\}\right)$. The basic NMF algorithm applied for $\mathbf{X}$ and $\boldsymbol{Y}$ separately will have complexity of $\mathcal{O}\left(MN_1R_1\right)$ and $\mathcal{O}\left(MN_2R_2\right)$ respectively. This shows that computational complexity of S-NMF remains the same as that of the basic NMF.

As far as the shared subspace dimensionality (K) is concerned, there is no method to determine its exact value. However, empirical 'rule of thumb' methods have been used with success. In our case, the value of K is bounded between 0 and $\min\left(R_1, R_2\right)$, i.e. ranges from no sharing to full sharing. Its optimal value depends on the nature of the target and auxiliary data. Intuitively, K increases with the level of sharing between the two data sources. For a rough estimate on K, we find the number of the common features (tags in our case) between the two data sets, say M_{xy}, then the rule of thumb is to use $K = \sqrt{M_{xy}/2}$ as suggested in Mardia et al. [1979].

3.2 Improving Social Media Retrieval with Auxiliary Sources

For social media retrieval, our key intuition is that tag ambiguities in a target domain are partially resolved by jointly modeling the tags of both the target and the auxiliary data. This is because any two data sources that share underlying structures when analyzed together, often provide richer information and thus have the potential to disambiguate the context for each other. S-NMF exploits this aspect by joint modeling and is able to learn the shared structures of the two data sources. Continuing on from the S-NMF framework described in the previous section, we use matrix $\mathbf{X}$ to denote *tf-idf* weighted [Salton and Buckley, 1988] tag-item matrix (akin to the term-document matrices where the tag list of each media item is considered as a document) from the target data source and matrix $\boldsymbol{Y}$ to denote the similar counterpart from the auxiliary source. Using S-NMF framework, we learn common subspace (spanned by matrix $\mathbf{W}$), individual subspaces (spanned by matrices $\boldsymbol{U}$ and $\boldsymbol{V}$) and the matrices $\mathbf{H}$ and $\mathbf{L}$ which contain the representations of the data $\mathbf{X}$ and $\boldsymbol{Y}$ in the learnt subspaces as expressed in Eqs (3.1-3.2).

Given a set of query keywords S_Q, we construct a query vector q_x by setting its elements to the *tf-idf* values at each index if the vocabulary contains a word from the keywords set S_Q, otherwise setting them to zero. Next, we project the vector q_x onto the subspace (refer steps 4-6 in Algorithm 3.2) spanned by matrix $[\mathbf{W} \mid \boldsymbol{U}]$ to find its subspace representation,

Algorithm 3.2 Image/Video Retrieval using S-NMF.

```
Input: Given $\mathbf{W}$, $\boldsymbol{U}$, $\mathbf{H}$ (learnt using Algorithm 3.1), query sentence $S_Q$, number of images (or videos) to be retrieved $N$ and image (or video) data set $\mathcal{I} = \{I_1, I_2, ......., I_{N_1}\}$.
prepare $q_x$ using tf-idf method from $S_Q$.
set $\delta = 10^{-9}$, $\epsilon = 10^{-2}$, $\mathbf{F} \triangleq [\mathbf{W} \mid \boldsymbol{U}]$ and project $q_x$ onto $\mathbf{F}$ to get $q_h$ by an initialization then looping as below
while ($\|\mathbf{F}q_h - q_x\|_2 \geq \epsilon$) do
    $q_h^{r+1} \leftarrow q_h^r \odot \left(\mathbf{F}^\top q_x\right) \oslash \left(\mathbf{F}^\top \mathbf{F} q_h^r + \delta\right)$
end while
compute cosine similarities $\mathrm{sim}\left(q_h, h_i\right)$ between query sentence and each tag document (image or video tags) according to the following
    $$\mathrm{sim}\left(q_h, h_i\right) = \frac{q_h^\top h_i}{\|q_h\|_2 \|h_i\|_2}$$
sort the cosine similarities $\mathrm{sim}\left(q_h, h_i\right)$ in descending order and get the ranking indices set $\{s_1, s_2, \ldots, s_{N_1}\}$.
Output: return the top $N$ retrieved images (or videos) as $\{I_{s_k} \mid k = 1, ..., N\}$.
```

denoted by q_h. Now, for retrieval, cosine similarities are computed between the vector q_h and each column of the matrix $\mathbf{H}$ to find similarly tagged items from the target medium. The retrieval results are obtained by ranking these similarity scores.

More formally, let the image (or video) data set on which retrieval is to be performed, be represented as $\mathcal{I} = \{I_1, I_2, ..., I_{N_1}\}$ and the collection of associated tags as $\mathcal{T}$. We prepare tag-item matrix $\mathbf{X} = [x_1, \ldots, x_{N_1}]$ where x_k is constructed from the tags of image (or video) I_k. We decompose matrix $\mathbf{X}$ into the matrices $\mathbf{W}$, $\boldsymbol{U}$ and $\mathbf{H}$ using the S-NMF framework described in section 3.1. Algorithm 3.2 provides pseudo-code for the social image/video retrieval using the proposed shared subspace learning framework where $\odot$ and $\oslash$ denote element-wise matrix multiplication and division respectively.

3.3 Experiments

3.3.1 Experimental Setup

Our experimental setup utilizes tags from LabelMe (auxiliary) to improve two social web retrieval tasks (target) in two media domains : image (Flickr) and video (YouTube). We denote the target data set from which retrieval is to be performed by $\mathbf{X}$ and the auxiliary data source by $\boldsymbol{Y}$. Following Algorithm 3.1, we learn a shared subspace $\mathbf{W}$ and the two discriminant subspaces $\boldsymbol{U}$ and $\boldsymbol{V}$ using the two data sets.

For comparison, we consider two baseline performances corresponding to two existing methods. For the first baseline, Flickr (or YouTube) data alone is used to learn the latent subspace $\boldsymbol{U}$ using the NMF algorithm described in Lee and Seung [2001]. For the second baseline, we use a recently proposed NMF based social semantic analysis method [Lin et al., 2009]. This method forces the whole latent subspace to be shared between the two data sets, i.e., $\boldsymbol{U}$ and $\boldsymbol{V}$ vanishes. We call these two baselines $BaselineI$ and $BaselineII$ respectively. Both these baselines are extreme cases – the former baseline has no provision of sharing with auxiliary data and thus unable to exploit external knowledge, while the later does not allow to maintain individual aspects, crucial for data faithfulness. This is where the attractiveness of our framework lies: it provides a freedom to exploit as much sharing information as required, but at the same time, allows to retain the individual differences of each domain. Our experiments are designed to evaluate this point for retrieval applications.

We evaluate the proposed S-NMF framework against the abovementioned baseline methods in regard to the two aspects (1) the effect of different levels of sharing on the retrieval performance and (2) improvement in performance when the auxiliary source is used under S-NMF framework compared to the use of additional data from within the *same* source.

Data Collection and Groundtruth

Since there is no standard groundtruth data available to evaluate the proposed tasks, we construct a subset of images/videos crawled from Flickr, YouTube and LabelMe and manually evaluate the retrieval results. To obtain the data, we design a set of concepts varying from indoor (e.g., 'chair', 'computer', 'cup', 'door', 'desk', 'microwave') to outdoor (e.g., 'beach', 'boat', 'building', 'plane', 'ship', 'sky', 'tree') and generic ('book', 'car', 'pen', 'person', 'phone', 'picture', 'window').

Flickr Data Set We download 50000 images from Flickr website using its API service[2]. On average, the number of distinct tags per image is 8. We remove the rare tags (appearing less than 5 times in the entire corpus), images with no tags and the images having non-English tags. After cleaning, we obtain around 20,000 tagged images. From this data set, we keep aside 7000 examples, which we use later as an auxiliary data set (termed as "internal" auxiliary data).

[2] Accessed between July 2009 to Dec 2009 from the website http://www.flickr.com/services/api/

Table 3.1: Jensen-Shannon divergence between the tag distributions of different data set pairs. For example, “LabelMe-LabelMe” means two different subsets of LabelMe data.

Data Set Pairs	*Jensen-Shannon Divergence*
LabelMe-Flickr	0.5237
LabelMe-YouTube	0.5603
LabelMe-LabelMe	0.0758
YouTube-YouTube	0.1985
Flickr-Flickr	0.4511

YouTube Data Set We download 18000 videos’ metadata (including tags, URL, category, title, comments etc) using YouTube API service[3]. On average, YouTube folksonomy has 7 distinct tags per video. As above, we remove the rare tags (appearing less than 2 times in the entire corpus), videos with no tags and non-English tags. After the cleaning, we obtain a data set having tag-list of around 12000 videos. Again, we keep aside 7000 examples to be used as an auxiliary data set (termed as “internal” auxiliary data).

LabelMe Data Set Further we add around 7000 images with their tags from LabelMe. This is a relatively less noisy data set with respect to the tag accuracy. On average, there are 32 distinct tags per image. We remove the rare tags appearing less than 2 times in the entire data corpus. This process does not reduce the size of data set.

To get an estimate of how much sharing some of these target-auxiliary data sets have in terms of their tags, Table 3.1 provides the Jensen-Shannon divergence [Lin, 1991] between their tag distributions.

3.3.2 Evaluation Measures

For the purpose of evaluation, we define a query set $\mathbb{Q} = \{$‘cloud’, ‘man’, ‘street’, ‘water’, ‘road’, ‘leg’, ‘table’, ‘plant’, ‘girl’, ‘drawer’, ‘lamp’, ‘bed’, ‘cable’, ‘bus’, ‘pole’, ‘laptop’, ‘plate’, ‘kitchen’, ‘river’, ‘pool’, ‘flower’$\}$. For the purpose of evaluation, we construct the ground truth by manually annotating each example in the two data sets (Flickr and YouTube) with respect to the query set $\mathbb{Q}$. For annotation, we consider a query term and an image (or video) relevant if the concept is clearly visible in the image (or video).

To evaluate the overall retrieval performance, we use the popular *11-point average precision-recall curve*[Baeza-Yates and Ribeiro-Neto, 1999, Manning et al., 2008]. For this evaluation

[3] Accessed between Oct 2009 to Dec 2009. API details can be found at the website http://code.google.com/apis/youtube/overview.html

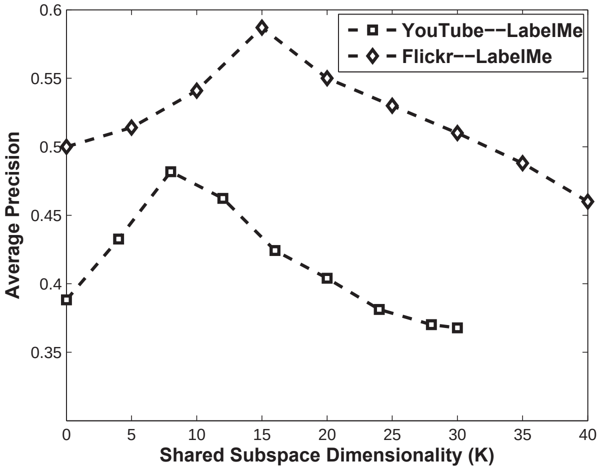


Figure 3.1: Retrieval performance with respect to shared subspace dimensionality for Flickr and YouTube.

measure, recall is fixed at 11 uniformly spaced values between 0 and 1, i.e. recall values 0, 0.1, . . ., 1. For each query, the precision is computed at the 11 recall values giving rise to a 11-point precision-recall curve for each query. The 11-point precision-recall curves are averaged over all queries to obtain the *11-point average precision-recall curve*.

For the web retrieval task, typically web users would like every item (images or videos) to be highly relevant with respect to the query in the first few retrieved results. Therefore, we also report our results in terms of the *precision-scope (P@N)* curve[4] to demonstrate the ranking capability of our proposed method. By looking at the retrieval results at various scope levels, it is easier to appreciate the ranking performance of a method as precision-scope measure is calculated for only top N retrieved items.

3.3.3 Flickr Image Retrieval Results

To compare the retrieval performance at different levels of sharing K, we consider the popular precision-at-fixed-recall metric in information retrieval. We fix the recall at 0.1, which is adequate, since users are mostly interested in only the first few results. For the subspace learning, we set $R_1 = 60$ and $R_2 = 40$ respectively, interpreted as the latent dimensionalities in the data. Figure 3.1 shows the average precision (averaged over the query set $\mathbb{Q}$) with respect to increasing values of the shared subspace dimensionality K.

[4] In this curve N represents the number of top retrieved images with which we compute the precision. For example P@20 is the retrieval precision when considering only the first 20 images retrieved.

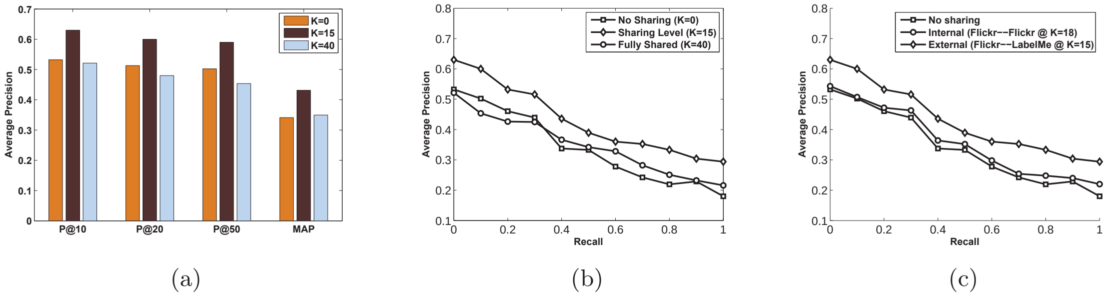
As can be seen, the average precision follows an interesting bell-shape curve when K increases, suggesting that there is an optimal level of sharing that achieves the best level of transfer from one data source to another. It also indicates that the two baselines – with no or total sharing – are highly non-optimal approaches. On average, across the query set $\mathbb{Q}$, $K = 15$ results in 58% precision and is the optimal level of sharing for our data set. This contrasts with 50% ($K = 0, BaselineI$) and 46% ($K = 40, BaselineII$) and thus our method delivers around 10% improvement.
The above results can be explained in the following manner; when K is much smaller than 15, the Flickr subspace shares very few basis vectors with LabelMe data set and therefore does not benefit fully by its accurately tagged nature – this is caused by an *under-representation* of sharing between the two data sets. On the other hand, when K becomes much larger than 15, it forces many basis vectors in the learnt subspace to represent both the data sets which can be difficult and therefore, the approximation in S-NMF factorization becomes poor – and this is caused by an *over-representation* of sharing between the two data sets. Neither extreme is desirable. The optimal sharing $K = 15$ in our framework represents a case in which an appropriate level of sharing for the two data sets is used. To further highlight this effect, the retrieval performance in terms of average precision for different values of shared subspace dimensionality K is plotted in Figure 3.1 where one can observe the bell-shape behaviors. We also note the correlation between the number of common basis vectors with the Jensen-Shannon divergence values of Table 3.1.
To demonstrate the ranking capabilities of the S-NMF based retrieval, we compute the precision at various scope levels. Figure 3.2a depicts the retrieval performance in terms of precision-scope ($P@N$) and MAP metrics and it is evident from the graph that S-NMF with $K = 15$ achieves much better performance than both $BaselineI$ and $BaselineII$.

Precision-Recall Curve

To examine the retrieval performance at different recall levels, we present the standard 11-point average precision-recall curve for $K = 15$. To compare the performance against the two baselines, we also present these curves for $BaselineI$ and $BaselineII$. Figure 3.2b depicts these precision-recall curves and shows that across all recall values, the S-NMF framework consistently outperforms the two baseline methods.

3.3.4 YouTube Video Retrieval Results

To demonstrate the flexibility of our proposed framework, we conduct YouTube experiments in the same way as for the Flickr data set. Again we fix the recall at 0.1 (due to



(a) (b) (c)

Figure 3.2: Flickr image retrieval results (a) precision-scope and MAP plots for *BaselineI* ($K = 0$), optimally shared subspace ($K = 15$) and *BaselineII* ($K = 40$) (b) 11-point interpolated precision-recall curve for *BaselineI* ($K = 0$), optimally shared subspace ($K = 15$) and *BaselineII* ($K = 40$) (c) comparison with case when auxiliary data is chosen from within Flickr domain instead of LabelMe.

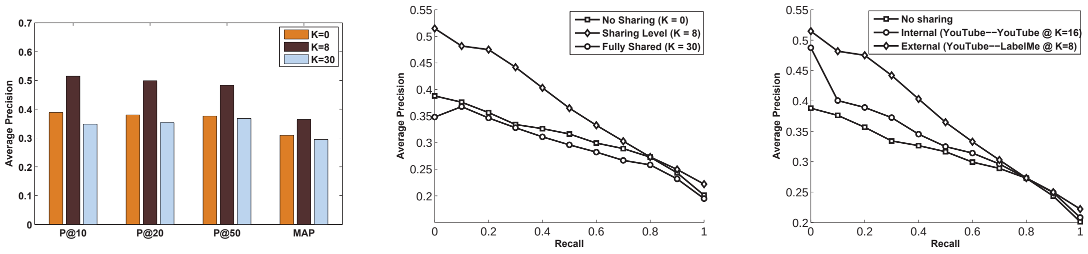
users' interest in only first few results) and generate results at various levels of subspace sharing by varying K. This time the latent dimensionality for YouTube (R_1) was set to be 30 and that of LabelMe (R_2) was set to be 40 as before. Figure 3.1 shows the average precision (averaged over the query set $\mathbb{Q}$) with respect to increasing values of the shared subspace dimensionality (K).

Similar to the Flickr case, the average precision follows a curve indicating an optimal level of sharing corresponding to $K = 8$. Using this optimal level of sharing between YouTube and LabelMe, we achieve improvements in precision performance by around 10% compared to *BaselineI* and by around 12% compared to *BaselineII*. This follows the *under-representation and over-representation* observations made for the Flickr case.

To demonstrate the ranking capability along with the overall performance, we present the Precision-Scope (P@N) and MAP results in Figure 3.3a. Note that the best performance in terms of both metrics has been achieved at the optimum sharing level ($K = 8$) and this result is consistently better than the two baselines.

Precision-Recall Curve

Similar to the Flickr case, we examine the retrieval performance at different recall levels and present the standard 11-point average precision-recall curve for $K = 8$, found above to be the optimal sharing level. To compare the performance against the two baselines, we also present these curves for *BaselineI* and *BaselineII*. Again, it can be seen from Figure 3.3b that across all recall values, S-NMF consistently outperforms the two baseline methods.



(a) (b) (c)

Figure 3.3: YouTube video retrieval results (a) precision-scope and MAP plots for $BaselineI$ ($K = 0$), optimally shared subspace ($K = 8$) and $BaselineII$ ($K = 30$) (b) 11-point interpolated precision-recall curve for $BaselineI$ ($K = 0$), optimally shared subspace ($K = 8$) and $BaselineII$ ($K = 30$) (c) comparison with case when auxiliary data is chosen from within YouTube domain instead of LabelMe.

External vs. Internal Auxiliary Source

We also investigate the benefits of external auxiliary sources (less noisy) vis-à-vis the additional data used from *within* the target source (e.g. Flickr or YouTube). For this, we compare the performance improvement obtained using LabelMe (external auxiliary source) with that obtained using the additional data from the same domain (internal auxiliary source). We denote YouTube data set as $\mathbf{X}$ and the "internal" auxiliary YouTube data set as $\boldsymbol{Y}$. Then S-NMF algorithm is used to learn both $\mathbf{W}$ and $\boldsymbol{U}$. We repeat the retrieval experiment similar to the one conducted for YouTube and LabelMe pair and find that the optimal sharing occurs when $K = 16$. Figure 3.3c clearly shows that improvement due to LabelMe data is much better than that achieved by "internal" auxiliary source. We believe that this improvement is due to the more complete and objective nature of LabelMe tags which helps in discovering the right term co-occurrences.

Similar results are shown in Figure 3.2c for the Flickr data set where the use of "internal" auxiliary Flickr data set is compared with "external" auxiliary LableMe data set. The optimum sharing with the "internal" auxiliary data set is achieved at $K = 18$. As seen from the figure, similar to YouTube experiments, using LableMe as auxiliary source yields much better performance than that using the noisy auxiliary auxiliary data from Flickr.

3.4 Discussion

In this chapter, we have presented a novel shared nonnegative matrix factorization (S-NMF) framework and applied it to improve tag-based image (Flickr) and video (YouTube) retrieval for online social media, leveraging information from an auxiliary source (LabelMe). Apart from inheriting the text analysis benefits of NMF such as the ability to capture tag co-occurrences and part-based decomposition, the key features of our proposed S-NMF is the ability to discover the shared structures between the two data sets and the flexibility in controlling the optimal level of sharing. This flexibility is important in dealing with real-world data sets since forcing the subspaces to be identical or keeping them completely different, as practiced in current existing works, is quite unrealistic. Our experimental results validate our claims, demonstrating an appropriate level of subspace sharing significantly boosts the retrieval performance – an average of over 10% for retrieval precision on the Flickr data set and an average of over 11% for retrieval precision on the YouTube data set when using LabelMe as the auxiliary source. Although applied to image and video retrieval in the social media context, our shared subspace learning framework is generic and can be applied in a wider setting to other machine learning and data mining tasks.

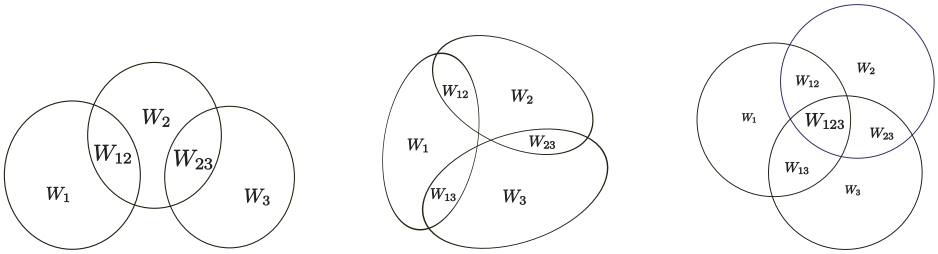
In the current form, S-NMF uses only one auxiliary source. There could be situations when one needs to use multiple auxiliary sources, e.g. when retrieving Flickr photos generated in reponse to some real world event, it may be useful to learn patterns from the related weblogs and YouTube videos. The tag metadata from multiple social media sources has strong underlying correlations and should be exploited to improve social media mining tasks. This necessitates the extension of S-NMF to a model that can incorporate multiple auxiliary sources.

Chapter 4

MS-NMF : Nonnegative Shared Subspace Learning for Multiple Data Sources

In the previous chapter, we presented the shared nonnegative matrix factorization (S-NMF) framework by extending NMF for joint modeling of two data sources. If S-NMF is used for joint modeling of more than two data sources, then pairwise analysis needs to be performed. However, there are sharing configurations even for three data sources that are difficult to be modeled appropriately using S-NMF (see Figure 4.1). We note that S-NMF is unable to handle the sharing configuration in Figure 4.1c. Moreover, consider three sources (denote by D_1, D_2 and D_3); jointly modeling (D_1, D_2, D_3) differs from the pairwise modeling (D_1, D_2) or (D_2, D_3). Therefore, S-NMF is limited and unusable for many real-world applications where one needs to include several auxiliary sources to achieve meaningful improvements in performance.

Addressing this limitation, in this chapter, we extend the S-NMF framework to allow modeling of multiple data sources. For learning shared and individual subspaces, we provide an efficient iterative algorithm that has the same order of computational complexity as NMF and S-NMF. Similar to S-NMF, a proof is provided to guarantee the convergence of the iterative algorithm. In addition to the modeling extensions, we empirically show that the use of multiple related auxiliary sources may improve performance. We demonstrate this point through the application of the proposed framework on tag based social media retrieval task by incorporating the knowledge from *multiple* auxiliary sources. As a second application, using shared subspaces, the proposed model establishes the correspondences amongst data items of multiple sources. This has been used to perform tag based cross-media retrieval. We develop efficient algorithms for both the applications and demonstrate them using the data from three representative social media sources – blogs (Blogspot.com),



(a) chain sharing (b) pairwise sharing (c) general sharing

Figure 4.1: Some possible sharing configurations for three data sources. An extension of S-NMF is required to model the general sharing configuration in (c).

images (Flickr) and videos (YouTube).

This chapter is organized as follows. Section 4.1 presents a multiple shared nonnegative matrix factorization (MS-NMF) framework for extraction of shared and individual subspaces across arbitrary numbers of data sources. To learn the factorization, an efficient algorithm is provided along with its complexity analysis. A proof of convergence is provided in section 4.1.3. Section 4.2 presents algorithms for social media retrieval incorporating multiple auxiliary sources and cross-social media retrieval across multiple sources. Section 4.3 demonstrates the proposed framework for the two applications using three real world social media sources.

4.1 Multiple Shared Subspace Learning

4.1.1 Problem Formulation

In this section, we describe a framework for learning shared and individual subspaces from multiple nonnegative data sources. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ represent the feature matrices constructed from a set of n data sources. For example, in collaborative filtering applications, *user-item rating* matrices are used where each row corresponds to a user, each column corresponds to an item and the features are ratings assigned by users. Similarly, *term-document* matrices are used in tag based social media retrieval applications where each row corresponds to a tag, each column corresponds to an item and features are usual tf-idf weights and so on. Given $\mathbf{X}_1, \ldots, \mathbf{X}_n$, we decompose each data matrix $\mathbf{X}_i$ as a product of two matrices $\mathbf{X}_i = \mathbf{W}_i \cdot \mathbf{H}_i$ such that the subspace spanned by the columns of matrix $\mathbf{W}_i$ explicitly represents arbitrary sharing among n data sources through *shared subspaces* and individual data by preserving their *individual subspaces*. For example, when $n = 2$, we create three subspaces: a shared subspace spanned by matrix W_{12} and two individual subspaces spanned by matrices W_1, W_2. Formally,

$$\mathbf{X}_1 = \underbrace{[W_{12} \mid W_1]}_{\mathbf{W}_1} \underbrace{\begin{bmatrix} H_{1,12} \\ \\ H_{1,1} \end{bmatrix}}_{\mathbf{H}_1} = W_{12} \cdot H_{1,12} + W_1 \cdot H_{1,1} \tag{4.1}$$

$$\mathbf{X}_2 = \underbrace{[W_{12} \mid W_2]}_{\mathbf{W}_2} \underbrace{\begin{bmatrix} H_{2,12} \\ \\ H_{2,2} \end{bmatrix}}_{\mathbf{H}_2} = W_{12} \cdot H_{2,12} + W_2 \cdot H_{2,2} \tag{4.2}$$

Notationwise, we use *bold capital letters* $\mathbf{W}$,$\mathbf{H}$ to denote the decomposition at the data source level and *normal capital letters* W, H to denote the partial subspaces. In the above expressions, the shared basis vectors are contained in W_{12} while individual basis vectors are captured in W_1 and W_2 respectively, giving rise to the full subspace representation $\mathbf{W}_1 = [W_{12} \mid W_1]$ and $\mathbf{W}_2 = [W_{12} \mid W_2]$ for the two data sources. However, note that the encoding coefficients of each data source in the shared subspace corresponding to W_{12} are different, and thus, an extra subscript is used to make it explicit as $H_{1,12}$ and $H_{2,12}$.
To generalize these expressions for n data sets in arbitrary configuration, we continue with this example ($n = 2$) and consider the power set over $\{1, 2\}$ given as

$$S\left(2\right) = \left\{\emptyset, \left\{1\right\}, \left\{2\right\}, \left\{1, 2\right\}\right\}$$

We can use the power set $S(2)$ to create an index set for the subscripts ‘1’, ‘2’ and ‘12’ used in matrices of Eqs (4.1) and (4.2). This helps in writing the factorization conveniently using a summation. We further use $S(2, i)$ to denote the subset of $S(2)$ in which only elements involving i are retained, i.e.

$$S\left(2,1\right) = \{\{1\}, \{1,2\}\} \text{ and } S\left(2,2\right) = \{\{2\}, \{1,2\}\}$$

With a slight abuse of notation, we re-write them as $S(2,1) = \{1, 12\}$ and $S(2,2) = \{2, 12\}$. Now, using these sets, Eqs (4.1) and (4.2) can be re-written as

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

Next, we generalize the above example ($n = 2$) for an arbitrary value of n. For a set of n data sets having an arbitrary sharing configuration, let $S(n)$ denote the power set of $\{1, 2, \ldots n\}$. For each $i = 1, \ldots, n$, let the index set associated with the i-th data source be defined as $S(n, i) = \{v \in S(n) \mid i \in v\}$. Using this notation, our proposed joint matrix factorization for n data sources can then be written as

$$\mathbf{X}_i = \mathbf{W}_i \cdot \mathbf{H}_i = \sum_{v \in S(n,i)} W_v \cdot H_{i,v} \tag{4.3}$$

The above expression is in its most generic form and considers all possible sharing configurations that can occur. In fact, the total number of subspaces equates $2^n - 1$ which is the cardinality of the power set $S(n)$ minus the empty set $\emptyset$. We consider this generic form here. However, our framework is directly applicable where we can customize the index set $S(n,i)$ to tailor any combination of sharing one wishes to model. Figure 4.1 illustrates some possible scenarios when there are three data sources ($n = 3$).

If we explicitly list the elements of $S(n,i)$ as $S(n,i) = \{v_1, v_2, \ldots, v_Z\}$ then $\mathbf{W}_i$ and $\mathbf{H}_i$ are

$$\mathbf{W}_i = \left[W_{v_1} \mid W_{v_2} \mid \ldots \mid W_{v_Z}\right] \ , \mathbf{H}_i = \begin{bmatrix} H_{i,v_1} \\ \vdots \\ H_{i,v_Z} \end{bmatrix} \tag{4.4}$$

4.1.2 Learning and Optimization

Our goal is to achieve sparse part-based representation of the subspaces and therefore, we impose nonnegative constraints on $\{\mathbf{W}_i, \mathbf{H}_i\}_{i=1}^n$. We formulate an optimization problem to minimize the Frobenius norm of joint decomposition error. The objective function accumulating normalized decomposition error across all data matrices can be written as

$$\begin{aligned} J(\mathbf{W}, \mathbf{H}) &= \frac{1}{2}\left\{\sum_{i=1}^{n} \lambda_i \left\|\mathbf{X}_i - \mathbf{W}_i \cdot \mathbf{H}_i\right\|_F^2\right\} \\ &= \frac{1}{2}\left\{\sum_{i=1}^{n} \lambda_i \left\|\mathbf{X}_i - \sum_{v \in S(n,i)} W_v \cdot H_{i,v}\right\|_F^2\right\} \end{aligned} \tag{4.5}$$

where $\|.\|_F$ is the Frobenius norm and $\lambda_i \triangleq \|\mathbf{X}_i\|_F^{-2}$ is the normalizing factor for data $\mathbf{X}_i$. Thus, the final optimization is given as

$$\begin{aligned} &\text{minimize } J(\mathbf{W}, \mathbf{H}) \\ &\text{subject to } W_v, H_{i,v} \geq 0 \text{ for all } 1 \leq i \leq n \text{ and } v \in S(n,i) \end{aligned} \tag{4.6}$$

where $J(\mathbf{W}, \mathbf{H})$ is defined in Eq (4.5). A few directions can be followed to solve the above constrained optimization problem, e.g. gradient-descent based multiplicative updates [Lee and Seung, 2001] or projected gradient [Lin, 2007]. We found that optimization of $J(\mathbf{W}, \mathbf{H})$ using multiplicative updates provides a good trade off between automatically selecting the gradient-descent step size and fast convergence for both synthetic and real data sets, and therefore, will be used here. Expressing the objective function element-wise, we shall show that multiplicative update equations for W_v and $H_{i,v}$ can be formulated efficiently as in the standard NMF [Lee and Seung, 2001]. Since the cost function of Eq (4.5) is non-convex jointly for all W_v and $H_{i,v}$, the multiplicative updates lead to a local minima solution. However, unlike NMF, this problem is less ill-posed as the constraint of partitioning the total subspace into common and individual subspaces eliminates some degree of freedom of factorization. The gradient of the cost function in Eq (4.5) w.r.t. W_v is given by

$$\nabla_{W_v} J(\mathbf{W},\mathbf{H}) = \sum_{i \in v} \lambda_i \left[-\mathbf{X}_i H_{i,v}^{\mathsf{T}} + \mathbf{X}_i^{(t)} H_{i,v}^{\mathsf{T}} \right]$$

where $\mathbf{X}_i^{(t)}$ is defined as

$$\mathbf{X}_i^{(t)} = \sum_{v \in S(n,i)} W_v \cdot H_{i,v} \tag{4.7}$$

Using Gradient-Descent optimization, we update matrix W_v as the following

$$(W_v)_{lk}^{t+1} \leftarrow (W_v)_{lk}^{t} + \eta_{(W_v)_{lk}^t} \left(-\nabla_{(W_v)_{lk}^t} J(\mathbf{W},\mathbf{H}) \right) \tag{4.8}$$

where $\eta_{(W_v)_{lk}^t}$ is the optimization step-size and given by

$$\eta_{(W_v)_{lk}^t} = \frac{(W_v)_{lk}^t}{\sum_{i \in v} \lambda_i \left(\mathbf{X}_i^{(t)} H_{i,v}^{\mathsf{T}} \right)_{lk}^t} \tag{4.9}$$

In the next section, we prove that the updates in Eq (4.8) when combined with step-size of Eq (4.9), converge to a locally optimum solution of the optimization problem given in Eq (4.6). Plugging the value of $\eta_{(W_v)_{lk}^t}$ from Eq (4.9) in Eq (4.8), we obtain the following multiplicative update equation for W_v

$$(W_v)_{lk} \leftarrow (W_v)_{lk} \frac{\left(\sum_{i \in v} \lambda_i \mathbf{X}_i \cdot H_{i,v}^{\mathsf{T}} \right)_{lk}}{\left(\sum_{i \in v} \lambda_i \mathbf{X}_i^{(t)} \cdot H_{i,v}^{\mathsf{T}} \right)_{lk}} \tag{4.10}$$

Multiplicative updates for $H_{i,v}$ can be obtained similarly and given by

$$(H_{i,v})_{km} \leftarrow (H_{i,v})_{km} \frac{\left(W_v^{\mathsf{T}} \cdot \mathbf{X}_i \right)_{km}}{\left(W_v^{\mathsf{T}} \cdot \mathbf{X}_i^{(t)} \right)_{km}} \tag{4.11}$$

As an example, for the case of $n = 2$ data sources mentioned earlier, the update equations for the shared subspace W_{12} (corresponding to $v = \{1, 2\}$) reduce to

$$(W_{12})_{lk} \leftarrow (W_{12})_{lk} \frac{\left(\lambda_1 \mathbf{X}_1 \cdot H_{1,12}^{\mathsf{T}} + \lambda_2 \mathbf{X}_2 \cdot H_{2,12}^{\mathsf{T}} \right)_{lk}}{\left(\lambda_1 \mathbf{X}_1^{(t)} \cdot H_{1,12}^{\mathsf{T}} + \lambda_2 \mathbf{X}_2^{(t)} \cdot H_{2,12}^{\mathsf{T}} \right)_{lk}} \tag{4.12}$$

and the update equations for the individual subspaces W_1 (when $v = \{1\}$) and W_2 (when $v = \{2\}$) become:

$$(W_1)_{lk} \leftarrow (W_1)_{lk} \frac{\left(\mathbf{X}_1 \cdot H_{1,1}^{\mathsf{T}}\right)_{lk}}{\left(\mathbf{X}_1^{(t)} \cdot H_{1,1}^{\mathsf{T}}\right)_{lk}} \tag{4.13}$$

$$(W_2)_{lk} \leftarrow (W_2)_{lk} \frac{\left(\mathbf{X}_2 \cdot H_{1,2}^{\mathsf{T}}\right)_{lk}}{\left(\mathbf{X}_2^{(t)} \cdot H_{1,2}^{\mathsf{T}}\right)_{lk}} \tag{4.14}$$

We note the intuition carried in these update equations. First, it can be verified by inspection that at the ideal convergence point when $\mathbf{X}_i = \mathbf{X}_i^{(t)}$, the multiplicative factors (second term on the RHS) in these equations become unity, thus no more updates are necessary. Secondly, updating a particular shared subspace W_v involves only relevant data sources for that share (sum over its index set $i \in v$, cf. Eq 4.10). For example updating W_{12} in Eq (4.12) involves both $\mathbf{X}_1$ and $\mathbf{X}_2$ but updating W_1 in Eq (4.13) involves only $\mathbf{X}_1$; the next iteration takes into account the joint decomposition effect and regularizes the parameter via Eq (4.7). From this point onwards, we refer to our framework as *Multiple Shared Nonnegative Matrix Factorization* (MS-NMF).

4.1.3 Convergence Analysis

In this subsection, we prove that the updates in Eq (4.8) when combined with step-size of Eq (4.9), converge to a locally optimum solution of the optimization problem given in Eq (4.6). Extending the Lemma 3.1 of chapter 3 for multiple source factorization, we prove the result stated in Theorem 4.1.

Definition 4.1. The auxiliary function $G\left((W_v)_p, (W_v)_p^t\right)$ is defined as an upper bound function for $J\left((W_v)_p\right)$ if $G\left((W_v)_p, (W_v)_p^t\right) \geqslant J\left((W_v)_p\right)$ and equality is satisfied iff $(W_v)_p = (W_v)_p^t$.

Theorem 4.1. *If $(W_v)_p$ is p-th row of matrix W_v, $v \in S(n,i)$ and $C\left((W_v)_p\right)$ is the diagonal matrix with its $(l,k)^{th}$element*

$$C_{lk}\left((W_v)_p\right) = \mathbf{1}_{l,k} \frac{\left(\sum_{i\in v} \lambda_i H_{i,v} \left(\sum_{u\in S(n,i)} H_{i,u}^{\mathsf{T}} (W_u)_p\right)\right)_l}{(W_v)_{pl}}$$

then $G\left((W_v)_p, (W_v)_p^t\right)$ is is an auxiliary function for $J\left((W_v)_p\right)$, the cost function defined for p-th row of the data where

$$\begin{aligned} G\left((W_v)_p, (W_v)_p^t\right) = J\left((W_v)_p^t\right) + \left((W_v)_p - (W_v)_p^t\right)^{\mathsf{T}} \nabla_{(W_v)_p^t} J\left((W_v)_p^t\right) \\ + \frac{1}{2}\left((W_v)_p - (W_v)_p^t\right)^{\mathsf{T}} C\left((W_v)_p^t\right)\left((W_v)_p^t - (W_v)_p^t\right) \end{aligned} \tag{4.15}$$

Proof. The second derivative of $J\left(\left(W_v\right)_p\right)$, i.e. $\nabla^2_{\left(W_v\right)_p} J\left(\left(W_v\right)_p\right) = \sum_{i\in v}\lambda_i H_{i,v}H_{i,v}^{\mathsf{T}}$. Comparing the expression of $G\left(\left(W_v\right)_p, \left(W_v\right)_p^t\right)$ in Eq (4.15) with the Taylor series expansion of $J\left(\left(W_v\right)_p\right)$ at point $\left(W_v\right)_p^t$, it can be seen that all we need to prove is the following

$$\left(\left(W_v\right)_p - \left(W_v\right)_p^t\right)^{\mathsf{T}} T_{W_v}\left(\left(W_v\right)_p^t - \left(W_v\right)_p^t\right) \geq 0$$

where $T_{W_v} \triangleq C\left(\left(W_v\right)_p^t\right) - \sum_{i\in v}\lambda_i H_{i,v}H_{i,v}^{\mathsf{T}}$. As in the proof of standard NMF, instead of showing it directly, we show the positive semi-definiteness of matrix E with (l,k)-th elements

$$E_{lk}\left(\left(W_v\right)_p^t\right) \quad = \quad \left(W_v\right)_{pl}^t \left(T_{W_v}\right)_{lk} \left(W_v\right)_{pk}^t$$

where $\left(W_v\right)_{pl}^t$ is the (p,l)-th element of matrix $\left(W_v\right)^t$. For positive semi-definiteness of matrix E, we have to show that for every nonzero z, the value of $z^{\mathsf{T}}Mz$ is nonnegative.

$$\begin{aligned}
z^{\mathsf{T}}Mz &= \sum_{l,k} z_l \left(W_v\right)_{pl}^t \left(T_{W_v}\right)_{lk} \left(W_v\right)_{pk}^t z_k \\
&= \sum_{l,k} z_l^2 \left(W_v\right)_{pl}^t \left(\sum_{u\in S(n,i),u\neq v} H_{i,u}^{\mathsf{T}} \left(W_u\right)_p\right)_l \\
&+ \lambda \sum_{l,k} \left(W_v\right)_{pl}^t \left(\sum_{i\in v}\lambda_i \left(H_{i,v}H_{i,v}^{\mathsf{T}}\right)_{lk}\right) \left(W_v\right)_{pk}^t \frac{\left(z_l - z_k\right)^2}{2} \geq 0
\end{aligned}$$

□

At the local minimum of $G\left(\left(W_v\right)_p, \left(W_v\right)_p^t\right)$, comparing $\nabla_{\left(W_v\right)_p^t} G\left(\left(W_v\right)_p, \left(W_v\right)_p^t\right)$ with gradient-descent update of Eq (4.8), we get the step size $\eta_{\left(W_v\right)_{lk}^t}$ as in equation (4.9).

4.1.4 Subspace Dimensionality and Complexity Analysis

Let M be the number of rows for each $\mathbf{X}_i$ (although $\mathbf{X}_i$'s usually have different vocabularies but they can be merged together to construct a common vocabulary that has M words) and N_i be the number of columns. Then, the dimensions for $\mathbf{W}_i$ and $\mathbf{H}_i$ are $M \times R_i$ and $R_i \times N_i$ respectively using R_i as reduced dimension. Since each $\mathbf{W}_i$ is an augmentation of individual and shared subspace matrices W_v, we further use K_v to denote the number of columns in W_v. Next, from Eq (6.2), it implies that $\sum_{v\in S(n,i)} K_v = R_i$. The value of K_v depends upon the sharing level among the involved data sources. Similar to S-NMF, we can use a rule of thumb as $K_v \approx \sqrt{M_v/2}$ where M_v is equal to the number of features common in data configuration specified by v. For example, if $v = \{1, 2\}$, M_v is equal to the number of common tags between source-1 and source-2.

Using above notation, the computational complexity of MS-NMF is $\mathcal{O}\left(M \times N_{\max} \times R_{\max}\right)$ per iteration where $N_{\max} = \max_{i\in[1,n]}\{N_i\}$ and $R_{\max} = \max_{i\in[1,n]}\{R_i\}$. The standard NMF algorithm [Lee and Seung, 2001] when applied on each matrix $\mathbf{X}_i$ with parameter R_i will have a complexity of $\mathcal{O}\left(M \times N_i \times R_i\right)$ and total complexity of $\mathcal{O}\left(M \times N_{\max} \times R_{\max}\right)$ per iteration. Therefore, computational complexity of MS-NMF remains equal to that of standard NMF.

4.2 Social Media Applications

Focusing on the social media domain, we show the utility of MS-NMF framework through two applications

1. Improving social media retrieval in target medium with the help of *multiple* related auxiliary social media sources.

2. Retrieving items across multiple social media sources.

The *first* application has been discussed before in chapter 3 but only one auxiliary source was used. Our key intuition here in using multiple related auxiliary sources is that different sources are rich in different patterns and structures. When jointly analyzed, they provide complementary information and help in boosting retrieval performance by learning improved tag co-occurrences through the shared subspace. Intuitively, improvement is expected when auxiliary sources share underlying structures with the target medium. These auxiliary sources can be readily found from the Web. The *second* application is “cross-media retrieval and correspondence”, where the shared subspace among multiple media provides a common representation across each medium and enables us to compute cross-media similarity between items of different media.

Social media users assign tags to their content (blog, images and videos) to retrieve them later and share them with other users. Often these user-generated content are associated with real world events, e.g., travel, sports, wedding receptions etc. In such a scenario, when users search for items from one medium, they are also interested in semantically similar items from other media to obtain more information. For example, one might be interested in retrieving ‘olympics’ related blogs, images and videos at the same time (cross-media retrieval). Similarly, given an ‘oscars’ related video, one may be interested in semantically similar blogs, images or any other form of social media (cross-media correspondence).

A naïve method of cross-media retrieval is to match the query keywords with the tag lists of items of different media. Performance of this method is usually poor due to poor semantic

Algorithm 4.1 Social Media Retrieval using MS-NMF.

```
Input: target $\mathbf{X}_k$, auxiliary $\mathbf{X}_j$ $(\forall j \neq k)$, query $q$, number of items to be retrieved $N$.

learn $\mathbf{X}_k = \mathbf{W}_k\mathbf{H}_k$ using Eqs.(4.10–4.11).
set $\epsilon = 10^{-2}$, project $q$ onto $\mathbf{W}_k$ to get $h$ by an initialization then looping as below
while $(\|\mathbf{W}_k h - q\|_2 \geq \epsilon)$ do
    $(h)_a \leftarrow (h)_a \left(\mathbf{W}_k^{\mathsf{T}} q\right)_a / \left(\mathbf{W}_k^{\mathsf{T}} \mathbf{W}_k h\right)_a$
end while
for each media item (indexed by $r$) in $\mathbf{X}_k$, with representation $h_r$ = $r$-th column of $\mathbf{H}_k$, compute its similarity with query projection $h$ as following
    $$\text{sim}\left(h, h_r\right) = \frac{h^{\mathsf{T}} h_r}{\|h\|_2 \|h_r\|_2}$$
Output: return the top $N$ items in decreasing order of similarities.
```

indexing caused by noisy tags, polysemy and synonymy. Subspace methods such as LSI or NMF, although robust against these problems, do not support cross-media retrieval in their standard form. Interestingly, MS-NMF provides solutions to both the problems. First, being a subspace based method, it is less affected by the problems caused by noisy tags, 'polysemy' and 'synonymy' and second, it is appropriate for cross-media retrieval as it represents items from each medium in a common subspace enabling to define a similarity for cross-media retrieval. In the following, we detail the use of MS-NMF for these applications.

4.2.1 Improving Social Media Retrieval with Auxiliary Sources

Let the target medium for which retrieval is to be performed be $\mathbf{X}_k$. Further, let us assume that we have other auxiliary media sources $\mathbf{X}_j$, $j \neq k$, which share some underlying structures with the target medium. We use these auxiliary sources to improve the retrieval precision from the target medium. 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 vocabulary contains a word from the keywords set or else putting zero. Algorithm 4.1 provides the details of social media retrieval (incorporating knowledge from multiple auxiliary sources) using MS-NMF.

4.2.2 Cross-Media Retrieval and Correspondence

To relate items from medium i and j, we use the common subspace spanned by $\mathbf{W}_{ij}$. As an example, $\mathbf{W}_{12} = [W_{12} \mid W_{123}]$, $\mathbf{W}_{23} = [W_{23} \mid W_{123}]$ and $\mathbf{W}_{13} = [W_{13} \mid W_{123}]$ for three data

Algorithm 4.2 Cross-Social Media Retrieval using MS-NMF.

```
Input: data $\mathbf{X}_1, \ldots, \mathbf{X}_n$, query $q$, number of items to be retrieved from medium $i, j$ as $N^i$ and $N^j$.
learn $\mathbf{X}_i = \mathbf{W}_i \mathbf{H}_i$ for every $i$ using Eqs.(4.10–4.11).
set $\epsilon = 10^{-2}$, project $q$ onto $\mathbf{W}_{ij}$ to get $h$ by an initialization then looping as below
while $\left(\|\mathbf{W}_{ij} h - q\|_2 \geq \epsilon\right)$ do
    $(h)_a \leftarrow (h)_a \left(\mathbf{W}_{ij}^{\mathsf{T}} q\right)_a / \left(\mathbf{W}_{ij}^{\mathsf{T}} \mathbf{W}_{ij} h\right)_a$
end while
for each item (indexed by $r$) in medium $i$ with the representation in shared subspace as $\mathbf{H}_{i,ij}(:,r)$, compute its similarity with query projection $h$ as
    $$\text{sim}\left(h, \mathbf{H}_{i,ij}(:,r)\right) = \frac{h^{\mathsf{T}} \mathbf{H}_{i,ij}(:,r)}{\|h\|_2 \|\mathbf{H}_{i,ij}(:,r)\|_2}$$
for each item (indexed by $r$) in medium $j$, compute $\text{sim}\left(h, \mathbf{H}_{j,ij}(:,r)\right)$ similar to step 7.
Output: return the top $N^i$ and $N^j$ items in decreasing order of similarities from medium $i$ and $j$ respectively.
```

sources, illustrated in Figure 4.1c. More generally, if $S(n,i,j)$ is the set of all subsets in $S(n)$ involving both i and j, i.e. $S(n,i,j) \triangleq \{v \in S(n) \mid i,j \in v\}$, the common subspace between i-th and j-th medium $\mathbf{W}_{ij}$ is then given by horizontally augmenting all W_v such that $v \in S(n,i,j)$. Similarly, representation of $\mathbf{X}_i$ (or $\mathbf{X}_j$) in this common subspace, i.e. $\mathbf{H}_{i,ij}$ (or $\mathbf{H}_{j,ij}$), is given by vertically augmenting all $H_{i,v}$ (or $H_{j,v}$) such that $v \in S(n,i,j)$. For $n = 3$, $\mathbf{H}_{1,12}^{\mathsf{T}} = \left[H_{1,12}^{\mathsf{T}} | H_{1,123}^{\mathsf{T}}\right]$, $\mathbf{H}_{2,12}^{\mathsf{T}} = \left[H_{2,12}^{\mathsf{T}} | H_{2,123}^{\mathsf{T}}\right]$ and so on.

Given the set of query keywords S_Q, we prepare the query vector q as described in subsection 4.2.1. Given query vector q, we wish not only to retrieve relevant items from i-th domain, but also from j-th domain. In the language of MS-NMF, this is performed by projecting q onto the common subspace matrix $\mathbf{W}_{ij}$ to get its representation h in the common subspace. Next, we compute similarity between h and the columns of matrix $\mathbf{H}_{i,ij}$ and $\mathbf{H}_{j,ij}$ (the representation of media items in the common subspace) to find out similar items from medium i and j respectively and the results are ranked based on these similarity scores either individually or jointly (see Algorithm 4.2).

Cross-media correspondence can be perceived as a special case of cross-media retrieval. To find the *correspondences* of an item from medium i to items in the medium j, we use the tags of the item (from medium i) as the set of query keywords S_Q in cross-media retrieval Algorithm 4.2 to get a ranked list of items from medium j.

Table 4.1: Description of Blogspot, Flickr and YouTube data sets.

Data Source	*Data Set Size*	*Concepts Used for Creating Data Set*	*Avg-Tags/Item* (rounded)
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

4.3 Experiments

4.3.1 Data Set

We conduct our experiments on a cross-social media data set consisting of the textual tags of three disparate media genres : *text*, *image* and *video*. To create the data set, three popular social media websites namely, Blogspot[1], Flickr[2] and YouTube[3], were used. To obtain the data, we queried all three websites using common concepts - 'Academy Awards', 'Australian Open', 'Olympic Games', 'US Election'. To have pairwise sharing in the data, we additionally queried Blogspot and Flickr with concept 'Christmas', YouTube and Flickr with concept 'Terror Attacks' and Blogspot and YouTube with concept 'Earthquake'. Lastly, to have some individual data of each medium, we queried Blogspot, Flickr and YouTube with concepts 'Cricket World Cup', 'Holi' and 'Global Warming' respectively. Total number of unique tags (M) combined from the three data sets were 3740. Further details of the three data sets are provided in Table 4.1.

4.3.2 Parameter Setting

We denote YouTube, Flickr and Blogspot *tf-idf* weighted tag-item matrices[4] by $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ respectively. For learning MS-NMF factorization, recall the notation K_v which is the dimensionality of the subspace spanned by W_v; following this notation, we use the individual subspace dimensions as $K_1 = 6, K_2 = 8, K_3 = 8$, pair-wise shared subspace

[1] http://www.blogger.com/

[2] http://www.flickr.com/services/api/

[3] http://code.google.com/apis/youtube/overview.html

[4] similar to widely known term-document matrices but generated from the tag-lists

dimension as $K_{12} = 15, K_{23} = 18, K_{13} = 12$ and all sharing subspace dimension as $K_{123} = 25$. To learn these parameters, we first initialize them using the heuristic described in subsection 4.1.4 based on the number of common and individual tags and then do cross-validation based on retrieval precision performance.

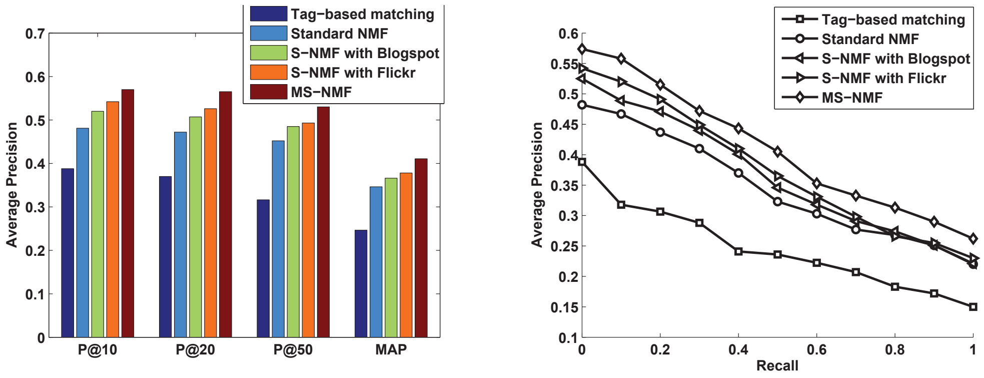
4.3.3 Experiment I : Improving Social Media Retrieval using Auxiliary Sources

To demonstrate the usefulness of MS-NMF for social media retrieval application, we carry out our experiments in a transfer learning setting. Focusing on YouTube video retrieval task, we choose YouTube as target data set while Blogspot and Flickr are used as auxiliary data sets. To perform retrieval using MS-NMF, we follow Algorithm 4.1.

Baseline Methods and Evaluation Measures

- The first baseline performs retrieval by matching the query with the tag-lists of videos (using vector-space model) without learning any subspace.
- The second baseline is the retrieval based on standard NMF. The retrieval algorithm using NMF remains similar to the retrieval using MS-NMF as it becomes a special case of MS-NMF when there is no sharing, i.e. $\mathbf{W}_1 = W_1$, $\mathbf{H}_1 = H_{1,1}$ and $R_1 = 56$.
- The third baseline is the S-NMF (described in the previous section) which learns shared and individual subspaces but allows only one auxiliary source at a time. Therefore, we use two instances of S-NMF (1) with Blogspot as auxiliary source (2) with Flickr as auxiliary source. Following model selection procedure described for S-NMF, we obtained its best performance with parameters setting : $R_Y = 56, R_F = 65, R_B = 62$ and $K_{YB} = 37, K_{YF} = 40, K_{BF} = 43$ where R_Y, R_F, R_B are total subspace dimensionalities of YouTube, Flickr and Blogspot respectively and K_{YB}, K_{YF}, K_{BF} are the shared subspace dimensionalities.

To compare the above baselines with the proposed MS-NMF, we use *precision-scope* (P@N), *mean average precision* (MAP) and *11-point interpolated precision-recall* [Baeza-Yates and Ribeiro-Neto, 1999]. The performance of MS-NMF is compared with the baselines by averaging the retrieval results over a query set of 20 concepts given by $\mathbb{Q} = \{$‘beach’, ‘america’, ‘bomb’, ‘animal’, ‘bank’, ‘movie’, ‘river’, ‘cable’, ‘climate’, ‘federer’, ‘disaster’, ‘elephant’, ‘europe’, ‘fire’, ‘festival’, ‘ice’, ‘obama’, ‘phone’, ‘santa’, ‘tsunami’$\}$.



(a) (b)

Figure 4.2: YouTube retrieval results with Flickr and Blogspot as auxiliary sources (a) precision-scope and MAP (b) 11-point interpolated precision-recall; for tag-based matching (baseline 1), standard NMF (baseline 2), S-NMF with Blogspot (baseline 3a), S-NMF with Flickr (baseline 3b) and the proposed MS-NMF.

Experimental Results

Figure 4.2 compares the retrieval performance of MS-NMF with the three baselines. It can be seen that MS-NMF outperforms the baselines in terms of all three evaluation criteria. Since the tag based matching method does not learn any subspaces, its performance suffers from the 'polysemy' and 'synonymy' problems prevalent in tag space. NMF, being a subspace learning method, performs better than tag based method but does not perform better than shared subspace methods (S-NMF and MS-NMF) as it is unable to exploit the knowledge from auxiliary sources. When comparing S-NMF with MS-NMF, we see that MS-NMF clearly outperforms both settings of S-NMF. This is due to the fact that S-NMF can use only one auxiliary source and thus does not exploit the knowledge available across multiple data sources. Although, S-NMF uses one auxiliary source at a time and improves the performance over NMF but the real strength of the three media sources is exploited by MS-NMF which performs the best. Better performance of MS-NMF can be attributed to the shared subspace model finding better term co-occurrences and reducing tag subjectivity by exploiting knowledge across three data sources. Further insight into the improvement is provided through entropy and impurity results given in subsection 4.3.5.

4.3.4 Experiment II : Cross-Media Retrieval and Correspondence

For cross-media retrieval experiments, we use the same data set as used in our first experiment but choose more appropriate baselines and evaluation measures. Subspace learning using MS-NMF remains the same, as the factorization is carried out on the same data set using the same parameter setting. We follow Algorithm 4.2 which utilizes MS-NMF framework to return the ranked list of cross-media items.

Baseline Methods and Evaluation Measures

To see the effectiveness of MS-NMF for cross-media retrieval, the *first* baseline is tag-based matching performed in a typical vector-space model setting. The *second* baseline is the framework of Lin et al. [2009], which can be adapted to learn a subspace that is fully shared among three media but does not retain any individual subspace. We shall denote this baseline as LIN_ETAL09. We present cross-media results for both pair-wise and across all three media. When presenting pair-wise results, we choose S-NMF (subspace learning remains same as in the first experiment) as a *third* baseline by applying it on the media pairs.

To evaluate our cross-media algorithm, we use *precision-scope* (P@N*),* MAP and *11-point interpolated precision-recall* measures. To explicitly state these measures for cross-media retrieval, we define precision and recall in a cross-media scenario. Consider a query term $q \in \mathbb{Q}$, let its ground truth set be G_i for i-th medium. If a retrieval method used with query q results in an answer set A_i from i-th medium, the precision and recall measures across n media are defined as

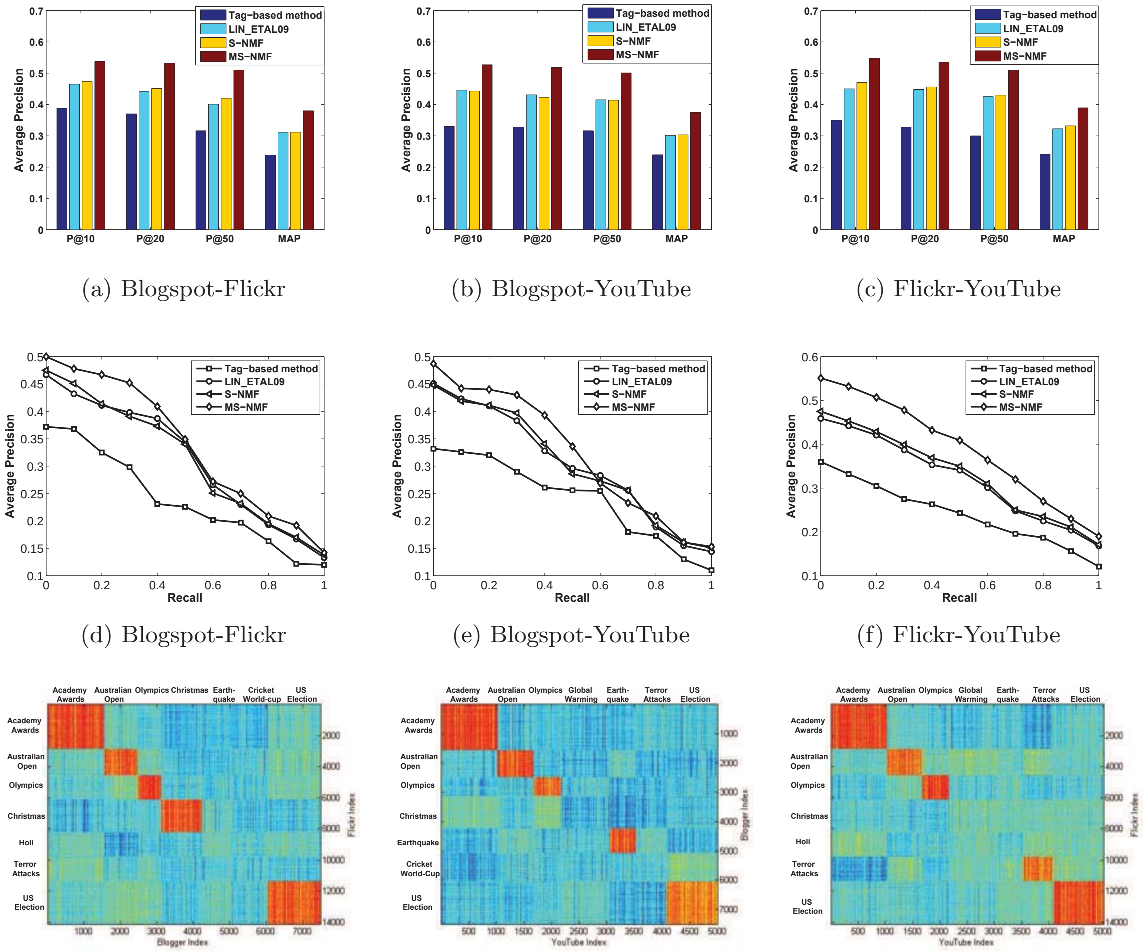
$$\text{Precision} = \frac{\sum_{i=1}^{n} |A_i \cap G_i|}{\sum_{i=1}^{n} |A_i|}, \text{ Recall} = \frac{\sum_{i=1}^{n} |A_i \cap G_i|}{\sum_{i=1}^{n} |G_i|} \tag{4.16}$$

Experimental Results

Cross-media retrieval results across media pairs are shown in Figure 4.3 whereas those from across all three media (Blogspot, Flickr and YouTube) are shown in Figure 4.4. To generate the graphs, we average the retrieval results over the *same* query set $\mathbb{Q}$ as defined for YouTube retrieval task in subsection 4.2.1. It can be seen from Figure 4.3 that MS-NMF significantly outperforms all baselines including S-NMF on cross-media retrieval task for each media-pair. This performance improvement is consistent in terms of all three evaluation measures. Note that, to learn the subspaces, MS-NMF uses all three media whereas S-NMF uses the data only from the media pair being considered. The ability to exploit knowledge from multiple media helps MS-NMF to achieve better performance. When retrieval precision and recall are calculated across all three media domains simultaneously (see Figure 4.4), MS-NMF still performs better than the tag-based matching and LIN_ETAL09. Note that S-NMF is not applicable in this case.

To show the efficacy of the projection of the multiple media tag data sets onto the shared subspace, we compute the similarity matrix across each pair of domains (the third row in Figure 4.3). The three plots show the relation between Blogger vs. Flickr, Blogger vs.

YouTube and Flickr vs. YouTube respectively. For each medium, items are ordered according to the concepts, i.e. the items related to a particular concept are indexed consecutively. It can be seen that relating the items from any two media (cross-media correspondence) through their shared subspace yields good results as inferred by block-diagonal nature of similarity matrices. To generate the similarity plots, we use cosine similarity. The missing diagonal blocks indicate that these concepts are different in the two media sources.



(a) Blogspot-Flickr (b) Blogspot-YouTube (c) Flickr-YouTube

(d) Blogspot-Flickr (e) Blogspot-YouTube (f) Flickr-YouTube

(g) Blogspot-Flickr (h) Blogspot-YouTube (i) Flickr-YouTube

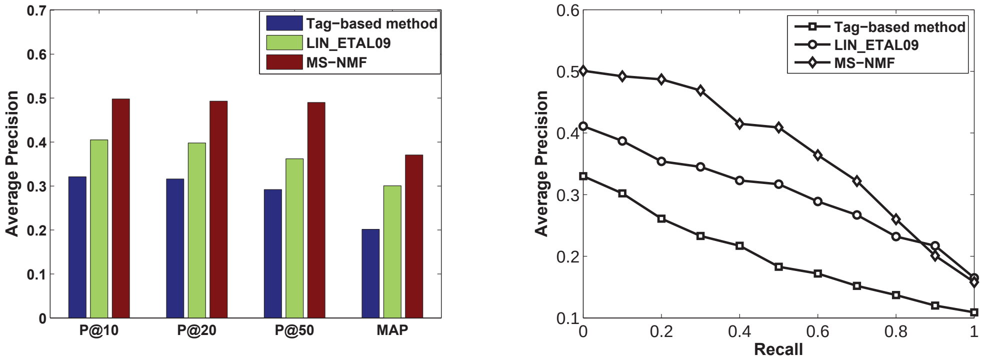
Figure 4.3: *Pairwise* cross-media retrieval and correspondence results : precision-scope, MAP, 11-point interpolated precision-recall and similarity plots of Blogspot-Flickr (first column), Blogspot-YouTube (second column) and Flickr-YouTube (third column). The precision results are shown for tag-based matching (baseline 1), LIN_ETAL09 (baseline 2), S-NMF (baseline 3) and MS-NMF. For similarity plots, ‘red’ color depicts high similarity while ‘blue’ depicts low similarity.

4.3.5 Topical Analysis

To provide further insights into the benefits achieved by MS-NMF, we examine the results at topical level. Every basis vector of the subspace, when normalized to sum to one, defines a distribution over the vocabulary and can be interpreted as a topic. We define a metric for measuring the *impurity* of a topic as

$$P\left(T\right)=\frac{1}{L\left(L-1\right)}\sum_{\substack{x,y\\x\neq y}}\mathrm{NGD}\left(t_x,t_y\right) \tag{4.17}$$

where L denotes the number of tags in a topic T for which corresponding basis vector element greater than a threshold[5] and $\mathrm{NGD}\left(t_x,t_y\right)$ is Normalized Google Distance between tags t_x and t_y [Calibrasi et al., 2007].



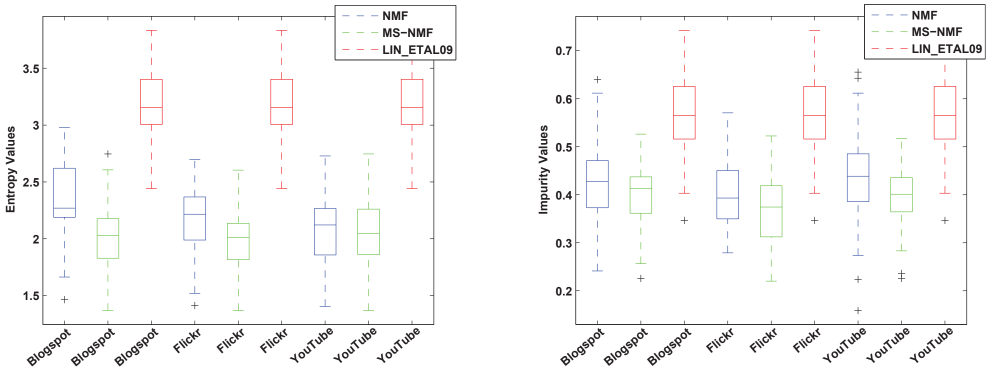
(a) precision-scope/MAP

(b) 11-point precision-recall curve

Figure 4.4: Cross-media retrieval results plotted across *all three data sources* (Blogspot, Flickr and YouTube) for tag-based matching (baseline 1), LIN_ETAL09 (baseline 2) and proposed MS-NMF.

We compute the entropy and impurity for each topic (normalized subspace basis) and plot their distributions in Figure 4.5 using the box-plots. It can be seen that topics learnt by MS-NMF have, on average, lesser entropy and impurity than their NMF and LIN_ETAL09 counterparts for all three social media sources. Although, LIN_ETAL09 can model multiple data sources, it uses a single subspace to model each source without retaining differences. As a result, the variabilities of the three sources get averaged out, and thereby increase the entropy and impurity of the resulting topics. In contrast, MS-NMF allowing the flexibility of partial sharing, averages the commonalities of the three data sources only up to their true sharing extent and results in purer and compact (less entropy) topics.

[5] fixed at 0.05 for selecting the tags with more than 5% weight in a topic.



(a) entropy distribution

(b) impurity distribution

Figure 4.5: A comparison of MS-NMF with NMF and LIN_ETAL09 in terms of entropy and impurity distributions.

Table 4.2 lists some of the topics discovered by both MS-NMF and NMF and illustrates the improvement due to sharing using the entropy and impurity measures. This accords with our expectation that joint modeling of multiple data sources should reduce the number of bits required for encoding the data compared with independently done. Though not proven theoretically, we expect that our modeling framework shares a close theoretical underpinning with joint coding in information theory, such as the information bottleneck [Tishby et al., 2000].

4.4 Discussion

We have extended the idea of shared subspace learning to a multiple shared nonnegative matrix factorization (MS-NMF) framework, which can model multiple data sources with arbitrary sharing configurations. We provided an efficient algorithm to learn the joint factorization and proved its convergence. We have shown the application of MS-NMF to two social media problems – tag based social media retrieval (by incorporating the knowledge from multiple auxiliary sources) and cross-media retrieval/correspondence. Our first application demonstrates that MS-NMF can help improving retrieval in YouTube by transferring knowledge from the tags of Flickr and Blogspot. Improving further over S-NMF, it justifies the need for a framework that can simultaneously model multiple data sources with arbitrary sharing. The second application shows the utility of MS-NMF for cross-media retrieval by demonstrating its superiority over existing methods using Blogspot, Flickr and YouTube data sets. Although we applied MS-NMF to the social media retrieval problem, the proposed MS-NMF model is generic and can also be applied to other data mining tasks requiring transfer learning and/or multi-task learning.

Table 4.2: Entropy/Impurity based comparison between topics discovered by MS-NMF and NMF.

	MS-NMF		*NMF*	
Concepts	Topics	Entropy/ Impurity	Topics	Entropy/ Impurity
Global Warming	globalwarming, carbon, climatechange, inconvenient, co2, green, energy, weather, science, monckton	2.13/0.394	globalwarming, climatechange, debate, climategate, e-mails, hackers, fraud, leaked, shock, skeptics, russian, copenhagen	2.47/0.457
US elections	election, us, president, politics, vote, obama, 2008, compaign, elections, america, usa, democrat, elect, political, candidate, history, senator, voting, democracy, republican	1.96/0.215	obama, president, speech, mccain, biography, bush, bloody, oil, michelle, 2008, victory, clinton, democrat, inauguration administration, iran	2.25/0.323
Olympics	olympic, 2008, beijing, games, olympics, summer, gymnastic, swimming, team, opening, commonwealth, chinese, phelps, vancouver, winter	2.17/0.372	olympics, beijing, vancouver, winter, 2008, atheletes, canada, summer, 2012, lago, cctv, nbc, earning, sailing, nest, white, london, top, media	2.20/0.411

In the current form of S-NMF/MS-NMF, there are no explicit constraints on shared and individual subspaces which ensure that the shared subspaces capture no more than shared aspects and the individual subspaces capture no more than individual aspects. In the absence of such constraints, shared subspace may capture some individual aspects, especially if there are some strong individual topics. As a result, when using this "noisy" shared subspace for knowledge transfer from auxiliary to source domain, negative knowledge transfer may occur [Rosenstein et al., 2005]. However, the explicit use of constraints, ensuring the segregation of shared subspace from individual subspace, can avoid this problem to some extent. In the next chapter, we extend S-NMF by incorporating these constraints and show that the resulting model outperforms S-NMF on retrieval and clustering tasks using a number of data sets. Additionally, it is shown to be *less* affected by negative transfer learning.

Chapter 5

RS-NMF : Regularized Nonnegative Shared Subspace Learning

The joint subspace learning approach considered in chapter 3 (or 4) uses shared and individual subspaces for joint modeling of two (or multiple) data sources. However, in this model, there are no explicit constraints on the shared and individual subspaces, which can ensure that shared subspace captures only shared aspects and individual subspaces capture only individual aspects. For real world data sets, without these constraints, the shared subspace may capture some individual aspects of a particular domain. As a result, when using this "noisy" shared subspace to transfer knowledge from auxiliary to target data source, negative transfer knowledge may occur [Rosenstein et al., 2005]. In a similar fashion, individual subspace may capture some of the common aspects and thus loses its discrimination capability.

To overcome these problems, in this chapter, we propose a regularized shared subspace learning framework that imposes a mutual orthogonality constraint on the constituent subspaces, ensuring that the shared subspace captures only shared aspects and the individual subspaces capture only individual aspects. We provide an efficient iterative solution to the constrained optimization problem along with a proof of convergence. In addition, we present a systematic way to learn the parameters of the model using nonnegative matrix factorization and rank criteria. We validate the proposed regularized model on social media retrieval and clustering tasks using a variety of datasets and demonstrate the benefits of the regularization scheme. In particular, we show that the proposed model compared to its unregularized counterpart is less affected by the problem of negative knowledge transfer. For clustering applications, the proposed model outperforms many state-of-the-art techniques.

This chapter is organized as follows. Section 5.1 extends the S-NMF framework (described in the previous chapter) to a new model that can learn regularized subspaces. This model is referred to as RS-NMF. In addition to formulating joint factorization with additional constraints, this section also presents an iterative solution to the constrained optimization problem. Section 5.2 discusses two real world applications of the proposed RS-NMF along with their algorithms. Section 5.3 describes the experimental results conducted for the two applications and compares them with the state-of-the-art methods along with further analysis and discussion. Section 5.4 provides the details of learning model parameters along with the empirical study of convergence. Section 5.5 extends the RS-NMF framework from two data sources to multiple data sources with arbitrary sharing configuration akin to MS-NMF. Concluding remarks are given in Section 5.6.

5.1 Regularized Nonnegative Shared Subspace Learning

In this section, we continue from the joint factorization (considered in Eqs (3.1-3.2)) and the notations used in chapter 3. To address the segregation problems discussed above, we introduce a regularization scheme on the shared and individual subspaces. It ensures that the two individual subspaces (spanned by $\boldsymbol{U}$ and $\boldsymbol{V}$) and the shared subspace (spanned by $\mathbf{W}$) capture only their respective basis vectors. With this regularization, the optimization cost function of Eq (3.3) takes the following form

$$\min_{\mathbf{W},\boldsymbol{U},\boldsymbol{V},\mathbf{H},\mathbf{L}\geq 0}\left\{\lambda_{\mathbf{X}}\left\|\mathbf{X}-[\mathbf{W}\mid\boldsymbol{U}]\mathbf{H}\right\|_F^2+\lambda_{\boldsymbol{Y}}\left\|\boldsymbol{Y}-[\mathbf{W}\mid\boldsymbol{V}]\mathbf{L}\right\|_F^2+R\left(\mathbf{W},\boldsymbol{U},\boldsymbol{V}\right)\right\} \tag{5.1}$$

where $\lambda_{\mathbf{X}}\triangleq\|\mathbf{X}\|_F^{-2}$, $\lambda_{\boldsymbol{Y}}\triangleq\|\boldsymbol{Y}\|_F^{-2}$ and $R\left(\mathbf{W},\boldsymbol{U},\boldsymbol{V}\right)$ is a regularization term used to penalize the "similarity" between subspaces spanned by matrices $\mathbf{W}$, $\boldsymbol{U}$ and $\boldsymbol{V}$. In this special case, when $R\left(\mathbf{W},\boldsymbol{U},\boldsymbol{V}\right)=0$, the model reduces to the S-NMF, discussed in chapter 3.

When seeking shared and individual subspaces that do not capture similar basis vectors (e.g. similar topics in case of text data) and that are complementary to each other, one solution is to constrain them to be mutually orthogonal. Note for example that if the subspaces spanned by matrices $\mathbf{W},\boldsymbol{U}$, are mutually orthogonal[1], we have $\mathbf{W}^\mathsf{T}\boldsymbol{U}=\mathbf{0}$. To impose this constraint, we choose to minimize the sum-of-squares of entries of the matrix $\mathbf{W}^\mathsf{T}\boldsymbol{U}$, i.e. $\left\|\mathbf{W}^\mathsf{T}\boldsymbol{U}\right\|_F^2$ which uniformly optimizes each entry of $\mathbf{W}^\mathsf{T}\boldsymbol{U}$. With this choice, the regularization term of Eq (5.1) is given by

$$R\left(\mathbf{W},\boldsymbol{U},\boldsymbol{V}\right)=\alpha\left\|\mathbf{W}^\mathsf{T}\boldsymbol{U}\right\|_F^2+\beta\left\|\mathbf{W}^\mathsf{T}\boldsymbol{V}\right\|_F^2+\gamma\left\|\boldsymbol{U}^\mathsf{T}\boldsymbol{V}\right\|_F^2 \tag{5.2}$$

[1] To see this, consider any two vectors p_i and q_j from the subspaces spanned by $\mathbf{W}$ and $\boldsymbol{U}$ respectively and note that $p_i=\mathbf{W}r_i$ and $q_j=\boldsymbol{U}s_j$. For the two subspaces to be mutually orthogonal, we have $p_i^\mathsf{T}q_j=0,\forall p_i,q_j$ which leads to $r_i^\mathsf{T}\mathbf{W}^\mathsf{T}\boldsymbol{U}s_j=0$. Since this relation has to hold for every r_i and s_j, we have $\mathbf{W}^\mathsf{T}\boldsymbol{U}=\mathbf{0}$.

where α, β and γ are the regularization parameters.

We note that an alternative formulation could have been based on minimizing the sum of entries[2] in matrix $\mathbf{W}^{\mathsf{T}}\boldsymbol{U}$, i.e. $\mathbf{1}_{\mathbf{w}}^{\mathsf{T}}\mathbf{W}^{\mathsf{T}}\boldsymbol{U}\mathbf{1}_{u}$ where $\mathbf{1}_{w}$ and $\mathbf{1}_{u}$ are the vector of ones of appropriate lengths. This will tend to give solutions such that entries of $\mathbf{W}^{\mathsf{T}}\boldsymbol{U}$ are sparse, meaning that most of the entries in $\mathbf{W}^{\mathsf{T}}\boldsymbol{U}$ would be either small or zero but there could be few large entries in $\mathbf{W}^{\mathsf{T}}\boldsymbol{U}$. However, since our goal here is to get a solution such that shared subspace spanned by matrix $\mathbf{W}$ does not capture any basis vector similar to the basis vectors of individual subspaces (spanned by $\boldsymbol{U}$ or $\boldsymbol{V}$), none of the entries in $\mathbf{W}^{\mathsf{T}}\boldsymbol{U}$ or $\mathbf{W}^{\mathsf{T}}\boldsymbol{V}$ are desired to be large[3].

5.1.1 Unified Formulation

Choosing the regularization term $R(\mathbf{W},\boldsymbol{U},\boldsymbol{V})$ as in Eq (5.2) and substituting into Eq (5.1) , our final optimization formulation is given as

$$\Pi \quad : \quad \begin{cases} \text{minimize} & \left\{\lambda_{\mathbf{X}}\left\|\mathbf{X}-[\mathbf{W}\mid\boldsymbol{U}]\mathbf{H}\right\|_F^2+\lambda_{\boldsymbol{Y}}\left\|\boldsymbol{Y}-[\mathbf{W}\mid\boldsymbol{V}]\mathbf{L}\right\|_F^2\right\} \\ & +\left\{\alpha\left\|\mathbf{W}^{\mathsf{T}}\boldsymbol{U}\right\|_F^2+\beta\left\|\mathbf{W}^{\mathsf{T}}\boldsymbol{V}\right\|_F^2+\gamma\left\|\boldsymbol{U}^{\mathsf{T}}\boldsymbol{V}\right\|_F^2\right\} \\ \text{subject to} & \mathbf{W},\boldsymbol{U},\boldsymbol{V},\mathbf{H},\mathbf{L}\geq 0 \end{cases}$$

In the above formulation, there is a trade-off between data fitting and obtaining subspaces that are mutually orthogonal. This trade-off is controlled by the regularization parameters α, β and γ. In section 5.4, we study the variation of the cost function with respect to these regularization parameters.

As an alternative to the optimization problem Π, one can express the regularization conditions explicitly through a set of constraints as follows

$$\Pi' \quad : \quad \begin{cases} \text{minimize} & \left\{\lambda_{\mathbf{X}}\left\|\mathbf{X}-[\mathbf{W}\mid\boldsymbol{U}]\mathbf{H}\right\|_F^2+\lambda_{\boldsymbol{Y}}\left\|\boldsymbol{Y}-[\mathbf{W}\mid\boldsymbol{V}]\mathbf{L}\right\|_F^2\right\} \\ \text{subject to} & \mathbf{W},\boldsymbol{U},\boldsymbol{V},\mathbf{H},\mathbf{L}\geq 0\ ,\ \left\|\mathbf{W}^{\mathsf{T}}\boldsymbol{U}\right\|_F^2\leq T_\alpha\ , \\ & \left\|\mathbf{W}^{\mathsf{T}}\boldsymbol{V}\right\|_F^2\leq T_\beta\ ,\ \left\|\boldsymbol{U}^{\mathsf{T}}\boldsymbol{V}\right\|_F^2\leq T_\gamma \end{cases}$$

However, it can be verified that, for every solution $\{\mathbf{W},\boldsymbol{U},\boldsymbol{V},\mathbf{H},\mathbf{L}\}$ to the optimization problem Π using particular values of parameters α, β and γ, there are equivalent parameters T_α, T_β and T_γ in the optimization problem Π' that lead to the same solution

[2] Note that taking the modulus of the entries of matrix $\mathbf{W}^{\mathsf{T}}\boldsymbol{U}$ is not needed as matrices $\mathbf{W}$ and $\boldsymbol{U}$ are already constrained to be nonnegative.

[3] Assuming that matrices $\mathbf{W},\boldsymbol{U},\boldsymbol{V}$ have their columns normalized to norm one (which is a usually done to reduce the ill-posed nature or non-uniqueness of the problem in NMF based factorizations), in practice, it is considered small enough if $\left(\mathbf{W}^{\mathsf{T}}\boldsymbol{U}\right)_{ij}\leq 0.1$.

$\{\mathbf{W}, \boldsymbol{U}, \boldsymbol{V}, \mathbf{H}, \mathbf{L}\}$ for Π'. Thus, Π and Π' are equivalent and it is sufficient to solve the optimization problem Π. **From this point onwards, we refer to this regularized extension as RS-NMF.**

5.1.2 Optimization and Algorithm

Note that the optimization problem Π is not convex in the combined space of variables $\{\mathbf{W}, \boldsymbol{U}, \boldsymbol{V}, \mathbf{H}, \mathbf{L}\}$. However, considering one variable at a time, the cost function turns out to be convex (for details, see Appendix A). For example, given $\{\boldsymbol{U}, \boldsymbol{V}, \mathbf{H}, \mathbf{L}\}$, the cost function is a convex function w.r.t. $\mathbf{W}$. Therefore, although we do not expect to get a global minimum for the above problem, we shall develop an algorithm that is not only simple and efficient but whose convergence can also be guaranteed.

Problem Π is a constrained optimization problem due to the nonnegative constraints on the factorization matrices, and can be solved using the Lagrange multiplier method. Let A^w_{ij} be the Lagrangian multiplier for constraints $\mathbf{W}_{ij} \geq 0$, i.e. for (i, j)-th element of matrix $\mathbf{W}$ and let $\mathbf{A}^w = \left[A^w_{ij}\right]$. Similarly if $\mathbf{A}^u, \mathbf{A}^v, \mathbf{A}^h, \mathbf{A}^l$ are the Lagrangian multiplier matrices for nonnegative constraints of matrices $\boldsymbol{U}, \boldsymbol{V}, \mathbf{H}, \mathbf{L}$, then the above cost function in an unconstrained form (denoted by C) can be written as

$$C = \lambda_{\mathbf{X}} \left\| \mathbf{X} - [\mathbf{W} \mid \boldsymbol{U}]\mathbf{H} \right\|_F^2 + \lambda_{\boldsymbol{Y}} \left\| \boldsymbol{Y} - [\mathbf{W} \mid \boldsymbol{V}]\mathbf{L} \right\|_F^2 + D\left(\mathbf{W}, \boldsymbol{U}, \boldsymbol{V}, \mathbf{H}, \mathbf{L}\right) \tag{5.3}$$

where $D\left(\mathbf{W}, \boldsymbol{U}, \boldsymbol{V}, \mathbf{H}, \mathbf{L}\right)$ is defined as

$$\begin{aligned} D\left(\mathbf{W}, \boldsymbol{U}, \boldsymbol{V}, \mathbf{H}, \mathbf{L}\right) \triangleq \alpha \left\| \mathbf{W}^{\mathsf{T}}\boldsymbol{U} \right\|_F^2 + \beta \left\| \mathbf{W}^{\mathsf{T}}\boldsymbol{V} \right\|_F^2 + \gamma \left\| \boldsymbol{U}^{\mathsf{T}}\boldsymbol{V} \right\|_F^2 + \mathrm{Tr}\left(\mathbf{A}^w\mathbf{W}^{\mathsf{T}}\right) \\ + \mathrm{Tr}\left(\mathbf{A}^u\boldsymbol{U}^{\mathsf{T}}\right) + \mathrm{Tr}\left(\mathbf{A}^v\boldsymbol{V}^{\mathsf{T}}\right) + \mathrm{Tr}\left(\mathbf{A}^h\mathbf{H}^{\mathsf{T}}\right) + \mathrm{Tr}\left(\mathbf{A}^l\mathbf{L}^{\mathsf{T}}\right) \end{aligned}$$

Optimize $\mathbf{W}$ given $\{\boldsymbol{U}, \boldsymbol{V}, \mathbf{H}, \mathbf{L}\}$

The first derivative of the cost function C with respect to $\mathbf{W}$ is given by

$$\nabla_{\mathbf{W}} C = 2\left[\lambda_{\mathbf{X}}\left(\mathbf{X}^{(t)} - \mathbf{X}\right)\mathbf{H}_w^{\mathsf{T}} + \lambda_{\boldsymbol{Y}}\left(\boldsymbol{Y}^{(t)} - \boldsymbol{Y}\right)\mathbf{L}_w^{\mathsf{T}} + \left(\alpha\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}} + \beta\boldsymbol{V}\boldsymbol{V}^{\mathsf{T}}\right)\mathbf{W} + \mathbf{A}^w\right]$$

where $\mathbf{X}^{(t)} \triangleq [\mathbf{W}^{(t)} \mid \boldsymbol{U}^{(t)}]\mathbf{H}^{(t)}$ and $\boldsymbol{Y}^{(t)} \triangleq [\mathbf{W}^{(t)} \mid \boldsymbol{V}^{(t)}]\mathbf{L}^{(t)}$. Using Karush–Kuhn–Tucker (KKT) conditions $\mathbf{A}^w_{ij}\mathbf{W}_{ij} = 0$ with the expression of the gradient $\nabla_{\mathbf{W}} C$, for any stationary point, we get the following

$$\left[\lambda_{\mathbf{X}}\left(\mathbf{X}^{(t)} - \mathbf{X}\right)\mathbf{H}_w^{\mathsf{T}} + \lambda_{\boldsymbol{Y}}\left(\boldsymbol{Y}^{(t)} - \boldsymbol{Y}\right)\mathbf{L}_w^{\mathsf{T}} + \left(\alpha\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}} + \beta\boldsymbol{V}\boldsymbol{V}^{\mathsf{T}}\right)\mathbf{W}\right]_{ij}\mathbf{W}_{ij} = 0$$

which leads to the following update equation

$$\mathbf{W}_{ij} \leftarrow \mathbf{W}_{ij} \frac{\left[\lambda_{\mathbf{X}} \mathbf{X} \mathbf{H}_w^{\mathsf{T}} + \lambda_{\boldsymbol{Y}} \boldsymbol{Y} \mathbf{L}_w^{\mathsf{T}}\right]_{ij}}{\left[\lambda_{\mathbf{X}} \mathbf{X}^{(t)} \mathbf{H}_w^{\mathsf{T}} + \lambda_{\boldsymbol{Y}} \boldsymbol{Y}^{(t)} \mathbf{L}_w^{\mathsf{T}} + \left(\alpha \boldsymbol{U} \boldsymbol{U}^{\mathsf{T}} + \beta \boldsymbol{V} \boldsymbol{V}^{\mathsf{T}}\right) \mathbf{W}\right]_{ij}} \tag{5.4}$$

Optimize $\boldsymbol{U}$ given $\{\mathbf{W}, \boldsymbol{V}, \mathbf{H}, \mathbf{L}\}$

The first derivative of the cost function C with respect to $\boldsymbol{U}$ is given by

$$\nabla_{\boldsymbol{U}} C = 2\left[\lambda_{\mathbf{X}} \left(\mathbf{X}^{(t)} - \mathbf{X}\right) \mathbf{H}_u^{\mathsf{T}} + \left(\alpha \mathbf{W} \mathbf{W}^{\mathsf{T}} + \gamma \boldsymbol{V} \boldsymbol{V}^{\mathsf{T}}\right) \boldsymbol{U} + \mathbf{A}^u\right]$$

As above, using KKT conditions $\mathbf{A}_{ij}^u \boldsymbol{U}_{ij} = 0$ with the expression of the gradient $\nabla_{\boldsymbol{U}} C$, for any stationary point, we get the following update equation

$$\boldsymbol{U}_{ij} \leftarrow \boldsymbol{U}_{ij} \frac{\left[\lambda_{\mathbf{X}} \mathbf{X} \mathbf{H}_u^{\mathsf{T}}\right]_{ij}}{\left[\lambda_{\mathbf{X}} \mathbf{X}^{(t)} \mathbf{H}_u^{\mathsf{T}} + \left(\alpha \mathbf{W} \mathbf{W}^{\mathsf{T}} + \gamma \boldsymbol{V} \boldsymbol{V}^{\mathsf{T}}\right) \boldsymbol{U}\right]_{ij}} \tag{5.5}$$

Similarly, optimizing for $\boldsymbol{V}$ given $\{\mathbf{W}, \boldsymbol{U}, \mathbf{H}, \mathbf{L}\}$, we get the following update equation

$$\boldsymbol{V}_{ij} \leftarrow \boldsymbol{V}_{ij} \frac{\left[\lambda_{\boldsymbol{Y}} \boldsymbol{Y} \mathbf{L}_w^{\mathsf{T}}\right]_{ij}}{\left[\lambda_{\boldsymbol{Y}} \boldsymbol{Y}^{(t)} \mathbf{L}_w^{\mathsf{T}} + \left(\beta \mathbf{W} \mathbf{W}^{\mathsf{T}} + \gamma \boldsymbol{U} \boldsymbol{U}^{\mathsf{T}}\right) \boldsymbol{V}\right]_{ij}} \tag{5.6}$$

Update equations for matrices $\mathbf{H}$ and $\mathbf{L}$ given $\{\mathbf{W}, \boldsymbol{U}, \boldsymbol{V}, \mathbf{L}\}$ and $\{\mathbf{W}, \boldsymbol{U}, \boldsymbol{V}, \mathbf{H}\}$ respectively, are similar to standard NMF and given by

$$\mathbf{H}_{ij} \leftarrow \mathbf{H}_{ij} \frac{\left[\mathbf{F}^{\mathsf{T}} \mathbf{X}\right]_{ij}}{\left[\mathbf{F}^{\mathsf{T}} \mathbf{F} \mathbf{H}\right]_{ij}}, \; \mathbf{L}_{ij} \leftarrow \mathbf{L}_{ij} \frac{\left[\mathbf{G}^{\mathsf{T}} \boldsymbol{Y}\right]_{ij}}{\left[\mathbf{G}^{\mathsf{T}} \mathbf{G} \mathbf{L}\right]_{ij}} \tag{5.7}$$

We note that multiplicative updates given by Eqs (5.4–5.7) are obtained by extending the updates of standard NMF [Lee and Seung, 2001]. There are alternative ways of optimizing the objective function Π such as alternating least squares and the active set method [Kim and Park, 2008] or the projected gradients approach [Lin, 2007], which often have better convergence behavior. Nonetheless, the multiplicative updates derived in this work have reasonably fast convergence behavior as shown empirically in section 5.4.3. Similarly, a good initialization of matrices $\mathbf{W}$, $\boldsymbol{U}$, $\boldsymbol{V}$, $\mathbf{H}$ and $\mathbf{L}$ may lead to quicker convergence of the proposed algorithm. Several methods have been proposed for NMF in literature [Wild et al., 2004, Langville et al., 2006, Boutsidis and Gallopoulos, 2008]. However, it is not obvious how to use them for initializing the shared subspace matrix $\mathbf{W}$ and the corresponding coefficient matrices $\mathbf{H}_w$ and $\mathbf{L}_w$. Other matrices such as $\boldsymbol{U}$, $\boldsymbol{V}$ also depend on $\mathbf{W}$ given the data. Therefore, we confine ourselves to the random initialization of these matrices.

Algorithm 5.1 Regularized Nonnegative Shared Subspace Learning using RS-NMF.

```
Input: Data matrices $\mathbf{X}$, $\boldsymbol{Y}$, parameter set $\Psi$, convergence threshold $\epsilon$ or maximum number of iteration ($Maxiter$).
compute $\lambda_{\mathbf{X}}$ and $\lambda_{\boldsymbol{Y}}$ as $\lambda_{\mathbf{X}} = \|\mathbf{X}\|_F^{-2}$ and $\lambda_{\boldsymbol{Y}} = \|\boldsymbol{Y}\|_F^{-2}$.
initialize matrices $\mathbf{W}$, $\boldsymbol{U}$, $\boldsymbol{V}$, $\mathbf{H}$ and $\mathbf{L}$ with random nonnegative values.
$r = 1$.
while ($r < Maxiter$) or ($C > \epsilon$) do
    update matrices $\mathbf{W}$, $\boldsymbol{U}$, $\boldsymbol{V}$, $\mathbf{H}$ and $\mathbf{L}$ using Eqs (5.4–5.7).
    normalize each column of matrices $\mathbf{W}$, $\boldsymbol{U}$ and $\boldsymbol{V}$ to norm one.
    compute the cost function ($C$) of optimization problem $\Pi$.
    $r = r + 1$.
end while
Output: Return subspace matrices $\mathbf{W}$, $\boldsymbol{U}$, $\boldsymbol{V}$, $\mathbf{H}$ and $\mathbf{L}$.
```

For future references, we denote the parameters required for RS-NMF collectively as a set $\Psi = \{R_1, R_2, K, \alpha, \beta, \gamma\}$. Algorithm 5.1 provides the details of learning the subspace matrices $\mathbf{W}$, $\boldsymbol{U}$, $\boldsymbol{V}$, $\mathbf{H}$ and $\mathbf{L}$ given the data matrices $\mathbf{X}$ and $\boldsymbol{Y}$.

5.1.3 Convergence Analysis

In this section, we prove the convergence of multiplicative updates given by Eqs (5.4–5.7) and analyze the computational complexity of Algorithm 5.1. We first prove some intermediate results (stated in Lemmas 5.1-5.4) and use them later to prove the main convergence result (stated in Theorem 5.1). The proof follows similar lines as the convergence proof of Expectation-Maximization (EM) algorithm [Dempster et al., 1977] and NMF [Lee and Seung, 2001] where the desired cost function is minimized indirectly by minimizing an upper bound function to the cost function.

Definition 5.1. $J(w, w')$ is an auxiliary function for $C(w)$ if $C(w) \leq J(w, w')$ and equality holds if and only if $w = w'$.

Lemma 5.1. *[Lee and Seung, 2001] If J is an auxiliary function for C, C is non-increasing under the update*

$$w^{t+1} = \underset{w}{argmin}\, J\left(w, w^t\right)$$

Proof. By definition, $C\left(w^{t+1}\right) \leq J\left(w^{t+1}, w^t\right) \leq J\left(w^t, w^t\right) = C\left(w^t\right)$. □

Lemma 5.2. *If* $C_{ij}\left(\mathbf{W}_{ij}^{(t)}\right)$ *consists of all the terms of cost function* $C\left(\mathbf{W}^{(t)}\right)$ *involving* $\mathbf{W}_{ij}^{(t)}$ *and if* $S_{ij}\left(\mathbf{W}^{(t)}\right)=\frac{\left[\lambda_{\mathbf{X}}\mathbf{X}^{(t)}\mathbf{H}_w^{\mathsf{T}}+\lambda_{\boldsymbol{Y}}\boldsymbol{Y}^{(t)}\mathbf{L}_w^{\mathsf{T}}+\left(\alpha\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}+\beta\boldsymbol{V}\boldsymbol{V}^{\mathsf{T}}\right)\mathbf{W}^{(t)}\right]_{ij}}{\mathbf{W}_{ij}^{(t)}}$ *then*

$$J\left(\mathbf{W}_{ij},\mathbf{W}_{ij}^{(t)}\right)=C_{ij}\left(\mathbf{W}_{ij}^{(t)}\right)+\left(\mathbf{W}_{ij}-\mathbf{W}_{ij}^{(t)}\right)\nabla C_{ij}\left(\mathbf{W}_{ij}^{(t)}\right)+\frac{1}{2}\left(\mathbf{W}_{ij}-\mathbf{W}_{ij}^{(t)}\right)^2 S_{ij}\left(\mathbf{W}^{(t)}\right)$$

is an auxiliary function for $C_{ij}\left(\mathbf{W}_{ij}\right)$.

Proof. When $\mathbf{W}_{ij}^{(t)}=\mathbf{W}_{ij}$, we clearly have $J\left(\mathbf{W}_{ij},\mathbf{W}_{ij}\right)=C_{ij}\left(\mathbf{W}_{ij}\right)$, therefore, for all other values of $\mathbf{W}_{ij}^{(t)}$, we need to show that $J\left(\mathbf{W}_{ij},\mathbf{W}_{ij}\right)\geq C_{ij}\left(\mathbf{W}_{ij}\right)$. Consider the Taylor expansion of $C_{ij}\left(\mathbf{W}_{ij}\right)$ around $\mathbf{W}_{ij}^{(t)}$, which is given as

$$C_{ij}\left(\mathbf{W}_{ij}\right)=C_{ij}\left(\mathbf{W}_{ij}^{(t)}\right)+\left(\mathbf{W}_{ij}-\mathbf{W}_{ij}^{(t)}\right)\nabla C_{ij}\left(\mathbf{W}_{ij}^{(t)}\right)+\frac{1}{2}\left(\mathbf{W}_{ij}-\mathbf{W}_{ij}^{(t)}\right)^2\nabla^2 C_{ij}\left(\mathbf{W}_{ij}^{(t)}\right)$$

The second derivative of $C_{ij}\left(\mathbf{W}_{ij}^{(t)}\right)$ is given as

$$\nabla^2 C_{ij}\left(\mathbf{W}_{ij}^{(t)}\right)=\left[\lambda_{\mathbf{X}}\mathbf{H}_w\mathbf{H}_w^{\mathsf{T}}+\lambda_{\boldsymbol{Y}}\mathbf{L}_w\mathbf{L}_w^{\mathsf{T}}\right]_{jj}+\left[\alpha\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}+\beta\boldsymbol{V}\boldsymbol{V}^{\mathsf{T}}\right]_{ii}$$

Now, consider the difference between the auxiliary and cost function as

$$J\left(\mathbf{W}_{ij},\mathbf{W}_{ij}^{(t)}\right)-C_{ij}\left(\mathbf{W}_{ij}\right)=\frac{1}{2}\left(\mathbf{W}_{ij}-\mathbf{W}_{ij}^{(t)}\right)^2\left[S_{ij}\left(\mathbf{W}^{(t)}\right)-\nabla^2 C_{ij}\left(\mathbf{W}_{ij}^{(t)}\right)\right] \tag{5.8}$$

and note that

$$\lambda_{\mathbf{X}}\mathbf{X}^{(t)}\mathbf{H}_w^{\mathsf{T}}+\lambda_{\boldsymbol{Y}}\boldsymbol{Y}^{(t)}\mathbf{L}_w^{\mathsf{T}}=\lambda_{\mathbf{X}}\left(\mathbf{W}^{(t)}\mathbf{H}_w+\boldsymbol{U}\mathbf{H}_u\right)\mathbf{H}_w^{\mathsf{T}}+\lambda_{\boldsymbol{Y}}\left(\mathbf{W}^{(t)}\mathbf{L}_w+\boldsymbol{V}\mathbf{L}_v\right)\mathbf{L}_w^{\mathsf{T}}$$

Writing the above expression elementwise, we have the following inequality

$$\begin{aligned}\left[\lambda_{\mathbf{X}}\mathbf{X}^{(t)}\mathbf{H}_w^{\mathsf{T}}+\lambda_{\boldsymbol{Y}}\boldsymbol{Y}^{(t)}\mathbf{L}_w^{\mathsf{T}}\right]_{ij} &\geq \lambda_{\mathbf{X}}\sum_k\mathbf{W}_{ik}^{(t)}\left[\mathbf{H}_w\mathbf{H}_w^{\mathsf{T}}\right]_{kj}+\lambda_{\boldsymbol{Y}}\sum_l\mathbf{W}_{il}^{(t)}\left[\mathbf{L}_w\mathbf{L}_w^{\mathsf{T}}\right]_{lj}\\ &\geq \mathbf{W}_{ij}^{(t)}\left[\lambda_{\mathbf{X}}\mathbf{H}_w\mathbf{H}_w^{\mathsf{T}}+\lambda_{\boldsymbol{Y}}\mathbf{L}_w\mathbf{L}_w^{\mathsf{T}}\right]_{jj}\end{aligned} \tag{5.9}$$

and similarly

$$\begin{aligned}\left[\left(\alpha\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}+\beta\boldsymbol{V}\boldsymbol{V}^{\mathsf{T}}\right)\mathbf{W}^{(t)}\right]_{ij} &= \alpha\sum_k\left[\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}\right]_{ik}\mathbf{W}_{kj}^{(t)}+\beta\sum_l\left[\boldsymbol{V}\boldsymbol{V}^{\mathsf{T}}\right]_{il}\mathbf{W}_{lj}^{(t)}\\ &\geq \mathbf{W}_{ij}^{(t)}\left[\alpha\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}+\beta\boldsymbol{V}\boldsymbol{V}^{\mathsf{T}}\right]_{ii}\end{aligned} \tag{5.10}$$

Therefore, from Eqs (5.9) and (5.10), we have

$$\left[\lambda_{\mathbf{X}}\mathbf{X}^{(t)}\mathbf{H}_w^{\mathsf{T}} + \lambda_{\boldsymbol{Y}}\boldsymbol{Y}^{(t)}\mathbf{L}_w^{\mathsf{T}} + \left(\alpha\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}} + \beta\boldsymbol{V}\boldsymbol{V}^{\mathsf{T}}\right)\mathbf{W}^{(t)}\right]_{ij} \geq \mathbf{W}_{ij}^{(t)}\left(\left[\lambda_{\mathbf{X}}\mathbf{H}_w\mathbf{H}_w^{\mathsf{T}} + \lambda_{\boldsymbol{Y}}\mathbf{L}_w\mathbf{L}_w^{\mathsf{T}}\right]_{jj} + \left[\alpha\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}} + \beta\boldsymbol{V}\boldsymbol{V}^{\mathsf{T}}\right]_{ii}\right) \tag{5.11}$$

Finally, from Eqs (5.8) and (5.11), we note that $J\left(\mathbf{W}_{ij}, \mathbf{W}_{ij}^{(t)}\right) \geq C_{ij}\left(\mathbf{W}_{ij}\right)$. □

Lemma 5.3. *If* $C_{ij}\left(\boldsymbol{U}_{ij}^{(t)}\right)$ *consists of all the terms of cost function* $C\left(\boldsymbol{U}^{(t)}\right)$ *involving* $\boldsymbol{U}_{ij}^{(t)}$ *and if* $S_{ij}\left(\boldsymbol{U}^{(t)}\right) = \frac{\left[\lambda_{\mathbf{X}}\mathbf{X}^{(t)}\mathbf{H}_u^{\mathsf{T}} + \left(\alpha\mathbf{W}\mathbf{W}^{\mathsf{T}} + \gamma\boldsymbol{V}\boldsymbol{V}^{\mathsf{T}}\right)\boldsymbol{U}^{(t)}\right]_{ij}}{\boldsymbol{U}_{ij}^{(t)}}$ *then*

$$J\left(\boldsymbol{U}_{ij}, \boldsymbol{U}_{ij}^{(t)}\right) = C_{ij}\left(\boldsymbol{U}_{ij}^{(t)}\right) + \left(\boldsymbol{U}_{ij} - \boldsymbol{U}_{ij}^{(t)}\right)\nabla C_{ij}\left(\boldsymbol{U}_{ij}^{(t)}\right) + \frac{1}{2}\left(\boldsymbol{U}_{ij} - \boldsymbol{U}_{ij}^{(t)}\right)^2 S_{ij}\left(\boldsymbol{U}^{(t)}\right)$$

is an auxiliary function for $C_{ij}\left(\boldsymbol{U}_{ij}\right)$*.*

Proof. The proof is similar to the proof of the Lemma 5.2. □

Lemma 5.4. *If* $C_{ij}\left(\boldsymbol{V}_{ij}^{(t)}\right)$ *consists of all the terms of cost function* $C\left(\boldsymbol{V}^{(t)}\right)$ *involving* $\boldsymbol{V}_{ij}^{(t)}$ *and if* $S_{ij}\left(\boldsymbol{V}^{(t)}\right) = \frac{\left[\lambda_{\boldsymbol{Y}}\boldsymbol{Y}^{(t)}\mathbf{L}_v^{\mathsf{T}} + \left(\beta\mathbf{W}\mathbf{W}^{\mathsf{T}} + \gamma\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}\right)\boldsymbol{V}^{(t)}\right]_{ij}}{\boldsymbol{V}_{ij}^{(t)}}$ *then*

$$J\left(\boldsymbol{V}_{ij}, \boldsymbol{V}_{ij}^{(t)}\right) = C_{ij}\left(\boldsymbol{V}_{ij}^{(t)}\right) + \left(\boldsymbol{V}_{ij} - \boldsymbol{V}_{ij}^{(t)}\right)\nabla C_{ij}\left(\boldsymbol{V}_{ij}^{(t)}\right) + \frac{1}{2}\left(\boldsymbol{V}_{ij} - \boldsymbol{V}_{ij}^{(t)}\right)^2 S_{ij}\left(\boldsymbol{V}^{(t)}\right)$$

is an auxiliary function for $C_{ij}\left(\boldsymbol{V}_{ij}\right)$*.*

Proof. Again, the proof is similar to the proof of the lemma 5.2. □

We do not provide any Lemma for proving update equation for matrix $\mathbf{H}$ as it can be seen from the optimization problem Π that the solution for $\mathbf{H}$ can be obtained using the results of standard NMF [Lee and Seung, 2001]. By considering the definitions $\mathbf{F} \triangleq [\mathbf{W} \mid \boldsymbol{U}]$ and $\mathbf{G} \triangleq [\mathbf{W} \mid \boldsymbol{V}]$ and noting that the regularization terms have no effect (as $\mathbf{W}, \boldsymbol{U}, \boldsymbol{V}, \mathbf{L}$ are fixed), this intuition becomes clear. Similar arguments hold for the update equation of matrix $\mathbf{L}$.

Theorem 5.1. *The cost function* $C\left(\mathbf{W}, \boldsymbol{U}, \boldsymbol{V}, \mathbf{H}, \mathbf{L}\right)$ *is non-increasing under the alternating multiplicative update rules of Eqs (5.4-5.7).*

Proof. We are minimizing $C\left(\mathbf{W}_{ij}\right)$ indirectly through the auxiliary function $J\left(\mathbf{W}_{ij}, \mathbf{W}_{ij}^{(t)}\right)$. Therefore, evaluating $\nabla J\left(\mathbf{W}_{ij}, \mathbf{W}_{ij}^{(t)}\right) = 0$ and utilizing the results of Lemma 5.1 and 5.2, we get the following update equation

$$\mathbf{W}_{ij}^{(t+1)} = \mathbf{W}_{ij}^{(t)} - \left[\nabla C_{ij}\left(\mathbf{W}_{ij}^{(t)}\right)/S_{ij}\left(\mathbf{W}_{ij}^{(t)}\right)\right]$$

Noting that

$$\nabla C_{ij}\left(\mathbf{W}_{ij}^{(t)}\right) = \left[\lambda_{\mathbf{X}}\left(\mathbf{X}^{(t)} - \mathbf{X}\right)\mathbf{H}_w^{\mathsf{T}} + \lambda_{\boldsymbol{Y}}\left(\boldsymbol{Y}^{(t)} - \boldsymbol{Y}\right)\mathbf{L}_w^{\mathsf{T}} + \alpha\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}\mathbf{W}^{(t)} + \beta\boldsymbol{V}\boldsymbol{V}^{\mathsf{T}}\mathbf{W}^{(t)}\right]_{ij}$$

and substituting $S_{ij}\left(\mathbf{W}_{ij}^{(t)}\right)$ from Lemma 5.2, we get the desired update of Eq (5.4). □

5.1.4 Shared Subspace Dimensionality and Algorithm Complexity

Ideally, the shared subspace dimensionality should be learnt automatically using the two data sources. However, this presents a general model selection problem which is known to be hard. Nonetheless, since RS-NMF framework learns mutually orthogonal subspaces, a rough estimate of subspace dimensionalities can be easily obtained from the data. Since, we have $\mathbf{W}^{\mathsf{T}}\boldsymbol{U} \approx \mathbf{0}$, $\mathbf{W}^{\mathsf{T}}\boldsymbol{V} \approx \mathbf{0}$ and $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{V} \approx \mathbf{0}$, shared subspace dimensionality is given by the rank of matrix $\mathbf{X}^{\mathsf{T}}\boldsymbol{Y}$, i.e. $K \approx \text{rank}\left(\mathbf{X}^{\mathsf{T}}\boldsymbol{Y}\right)$. Similarly, rough estimates of R_1 and R_2 are given by $\text{rank}(\mathbf{X})$ and $\text{rank}(\boldsymbol{Y})$ respectively. A detailed procedure to estimate these parameters is demonstrated in section 5.4.

When analyzing the computational complexity of RS-NMF algorithm, we refer to Eqs (5.4-5.7). Computational complexity of learning matrix $\mathbf{W}$ is $\mathcal{O}(M \times N \times K)$ per iteration where $N = \max(N_1, N_2)$. Similarly, for each iteration, learning matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ takes $\mathcal{O}(M \times N_1 \times K_u)$ and $\mathcal{O}(M \times N_2 \times K_v)$ respectively where $K_u = R_1 - K$ and $K_v = R_2 - K$. Learning matrices $\mathbf{H}$ and $\mathbf{L}$ takes $\mathcal{O}(M \times N_1 \times R_1)$ and $\mathcal{O}(M \times N_2 \times R_2)$ operations per iteration. Therefore, the overall complexity of the algorithm is dominated by computation of matrices $\mathbf{H}$ and $\mathbf{L}$ and is $\mathcal{O}(M \times N \times R)$ where $R = \max(R_1, R_2)$. This is also the order of complexity for S-NMF algorithm (presented in chapter 3) and NMF algorithm [Lee and Seung, 2001] for each iteration. Note that, for implementation efficiency, when computing matrices such as $\mathbf{F}^{\mathsf{T}}\mathbf{F}\mathbf{H}$, we compute $\mathbf{F}^{\mathsf{T}}\mathbf{F}$ first and then multiply with matrix $\mathbf{H}$. Similarly, when computing $\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}\mathbf{W}$, we compute $\boldsymbol{U}^{\mathsf{T}}\mathbf{W}$ first and pre-multiply by $\boldsymbol{U}$. The main idea is to group the matrix multiplications by their inner products rather than outer products.

5.2 Social Media Applications

Given related data sources, RS-NMF learns both the shared and the individual structures present in the data. This is desired in many applications such as multimedia indexing, text mining, computer vision and social media. For example, it might be of interest to determine the commonality among documents from computer science and biological science, and at the same time find the differences. The commonality among the multiple data sources often implies basic features across the data sources while discriminating features can be utilized for classifying documents from one source against other sources. Another application is in social media, where there exist many popular social networking sites. Users of these websites upload and share content with one another. As an example, YouTube users upload videos related to some topics and Flickr users upload photos which may be similar to and/or different from YouTube videos. The data from different social media sources are semantically related due to the common cause of their creation (for example, in response to real world events such as travel, oscars, olympics, wedding receptions, earthquakes etc). When modeled by RS-NMF, the shared features can be used to extract the basic tag structures present across the two media (YouTube and Flickr) accurately by making it less subjective to either medium, and thus improve the retrieval and clustering performance for each medium. In addition, the shared features can also be used to relate items from the two media using the shared subspace representation. While shared subspace across the related sources can boost the performance, the provision to maintain individual variations of each source ensures that domain-specific information is not lost.

In the rest of this section, we focus on the social media domain and show the usefulness of RS-NMF for two tasks (1) Improved social media retrieval by leveraging tags from auxiliary data sources. (2) Improved simultaneous clustering of related data sources and discovery of shared and individual clusters.

5.2.1 Improving Social Media Retrieval using Auxiliary Sources

In this application, we improve the social media retrieval by using tag data from another related auxiliary source. This application is already described in section 3.2 of chapter 3. The retrieval algorithm based on RS-NMF remains similar to Algorithm 3.2 presented in chapter 3 except that the subspace learning is carried out using the RS-NMF framework instead of S-NMF framework.

5.2.2 Joint Clustering of Related Data Sources

Clustering is one of the most important problems in data mining and is considered to be an unsupervised task. Often, in real world applications, related data may arise from different sources and have significant variations in their distributions. Although these data sources have a great degree of commonality, direct clustering of the data combined together from these sources results in poor performance due to their individual differences. In such scenarios, the clustering method needs to deal with both shared and individual characteristics. RS-NMF is suitable for this joint clustering task.

Given two data sources, let us denote their data corpora by $\mathbf{X}$ and $\boldsymbol{Y}$. Using RS-NMF,

Algorithm 5.2 RS-NMF based Clustering.

```
Input: Target $\mathbf{X}$, auxiliary $\boldsymbol{Y}$, parameter set $\Psi$, convergence threshold $\epsilon$ and number of clusters $P$.
learn subspace matrices $\mathbf{W}$, $\boldsymbol{U}$, $\boldsymbol{V}$, $\mathbf{H}$ and $\mathbf{L}$ using Algorithm 5.1.
perform K-means clustering on matrix $\mathbf{H}$ considering its columns as subspace features for data $\mathbf{X}$ and denote the cluster membership vector as $m_{\mathbf{X}}$.
perform K-means clustering on matrix $\mathbf{L}$ considering its columns as subspace features for data $\boldsymbol{Y}$ and denote the cluster membership vector as $m_{\boldsymbol{Y}}$.
Output: Return cluster memebership vectors $m_{\mathbf{X}}$, $m_{\boldsymbol{Y}}$ for data items in $\mathbf{X}$ and $\boldsymbol{Y}$ respectively.
```

we learn the common subspace $\mathbf{W}$ and individual subspaces $\boldsymbol{U}$ and $\boldsymbol{V}$. Representation of the two data matrices $\mathbf{X}$ and $\boldsymbol{Y}$ in the subspaces spanned by the columns of $[\mathbf{W} \mid \boldsymbol{U}]$ and $[\mathbf{W} \mid \boldsymbol{V}]$ is given by $\mathbf{H}$ and $\mathbf{L}$ respectively. Note that, originally, the data points of $\mathbf{X}$ and $\boldsymbol{Y}$ are in M-dimensional space but RS-NMF projects them to a R-dimensional (note that dimensionality is reduced as $R < M$) space thus, partially avoiding the problems of high dimensionality. After projecting the data points in the combined shared and individual subspaces, we use standard single task clustering methods such as K-means (or any other clustering method), for clustering the data in the regularized subspaces. We choose basic K-means with the purpose of emphasizing the clustering strength of subspaces and not of the clustering method *per se*. Algorithm 5.2 provides details of the RS-NMF based clustering.

5.3 Experiments

In this section, we demonstrate the effectiveness of RS-NMF for the social media retrieval and clustering applications. We conduct two sets of experiments. In the first experiment set, we show the performance improvement for social media retrieval using auxiliary sources, clearly demonstrating the superiority of RS-NMF compared to other appropriate baselines. In the second set of experiments, we perform simultaneous clustering of related data sources. Through these experiments, we demonstrate the improvement in clustering performance achieved by simultaneous clustering and show that RS-NMF significantly outperforms other recently proposed single and multi-task clustering methods.

5.3.1 Experiment-I : Social Media Retrieval

5.3.1.1 Data Set

For social media retrieval, we conduct our experiments on the same social media data set (created from Blogspot, Flickr and YouTube websites) that was utilized for demonstrating MS-NMF in chapter 4 and is described in section 4.3.1.

5.3.1.2 Baseline Methods

To compare the performance with other methods, we choose three baselines :

- The first baseline performs retrieval by matching the query keyword with the tag-lists of each YouTube video without learning any subspace. To perform this matching, we use Jaccard[4] coefficient and rank the results based on these scores. This baseline is referred to as "Tag-based matching".

- In the second baseline, we take the tags of YouTube only (no auxiliary source is used) and apply standard NMF for retrieval. The number of basis vectors for NMF is set to 56. This baseline is referred to as "Standard NMF".

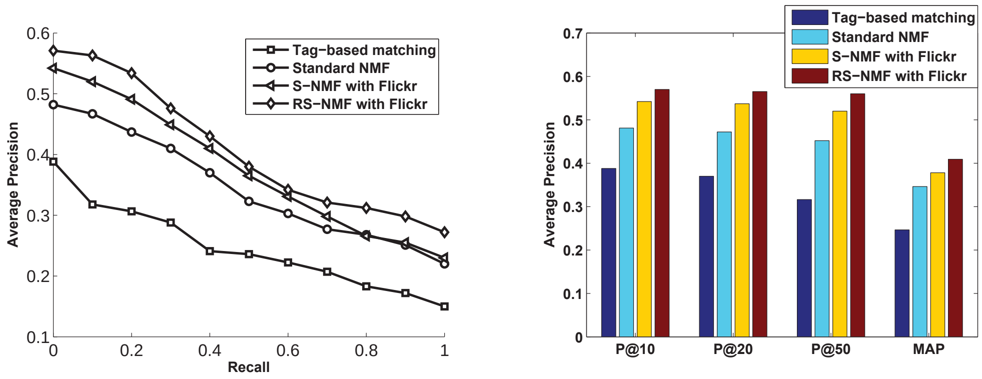
- As a third baseline, we use the S-NMF framework (described in chapter 3) which is a special case of RS-NMF without any mutual orthogonality constraints on the learnt subspaces. This baseline is referred to as "S-NMF".

To learn the subspace dimensionalities required for S-NMF and RS-NMF, we follow the procedure described in section 5.4 and do cross-validation based on retrieval performance. The best performance was achieved by setting $R_Y = 56, R_F = 65, R_B = 62$ and $K_{YB} = 37, K_{YF} = 40$ where R_Y, R_F, R_B are total subspace dimensionalities of YouTube, Flickr and Blogspot data respectively and K_{YB}, K_{YF} are the shared subspace dimensionalities.

5.3.1.3 Evaluation Metrics

For the purpose of evaluation, we define a query set {'beach', 'america', 'bomb', 'animal', 'bank', 'movie', 'river', 'cable', 'climate', 'federer', 'disaster', 'elephant', 'europe', 'fire', 'festival', 'ice', 'obama', 'phone', 'santa', 'tsunami'} and denote it as Q. To evaluate retrieval methods, we use *11-point precision recall curve* and *mean average precision* (*MAP*) [Baeza-Yates and Ribeiro-Neto, 1999]. For social media retrieval, items relevant to user queries should be ranked high, as users would typically like every result on the first few pages to be as relevant as possible. *Precision-scope (P@N)* curve has been used as a ranking measure by many researchers [Rui and Huang, 2000, Cai et al., 2007]. Therefore, to evaluate the ranking performance of different methods clearly, we use the *P@N* curve.

[4] Jaccard $(A, B) = |A \cap B| \, / \, |A \cup B|$.

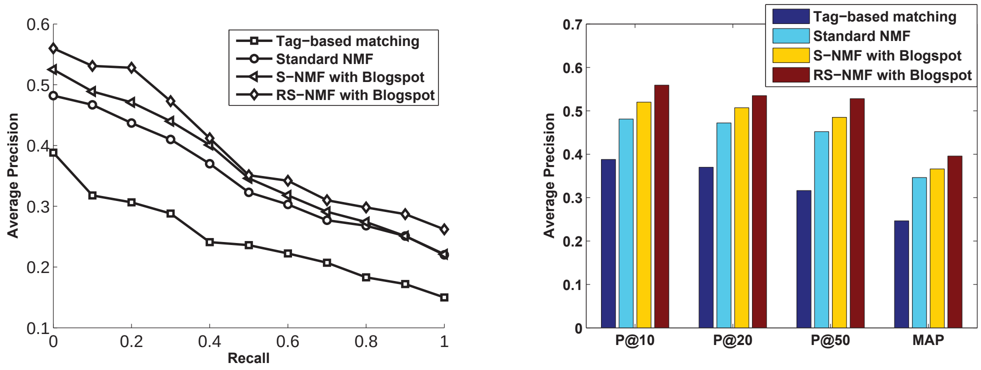


(a) (b)

Figure 5.1: YouTube video retrieval results using tags of Flickr as auxiliary source (a) 11-point interpolated precision-recall (b) precision-scope (P@N) and MAP; for tag-based matching (baseline 1), standard NMF (baseline 2), S-NMF (baseline 3) and the proposed RS-NMF.

5.3.1.4 Experimental Results

YouTube-Flickr Retrieval Results Figure 5.1 depicts the YouTube video retrieval results and compares the performance of RS-NMF with the three baselines in terms of *11-point precision recall* curve, *P@N* curve and *MAP* evaluation criteria. It can be seen from the figure that RS-NMF clearly outperforms all the baselines in all three evaluation criteria. Performance of the method based on tag matching using Jaccard similarity is poor due to the high dimensionality of input tag space and the ambiguities caused by polysemy and synonymy. Subspace learning methods (NMF, S-NMF and RS-NMF) overcome these problems to some extent by learning a reduced dimensional representation in latent space. NMF, being a subspace learning method, performs clearly better than tag-based matching but falls short when compared with S-NMF and RS-NMF. This is mainly due to the additional knowledge acquired from the auxiliary source, which helps in disambiguating tag co-occurrences. When we compare S-NMF with RS-NMF, we see that RS-NMF clearly performs better than S-NMF in terms of all evaluation criteria. This gain in performance can be attributed to the segregation of shared and individual subspaces which ensures the transference of useful knowledge only. Looking at the top 10 results, we see that RS-NMF achieves around 57% precision as compared with 54%, 48% and 39% precision achieved by S-NMF, NMF and tag-based matching methods respectively.



(a) (b)

Figure 5.2: YouTube video retrieval results using tags of Blogspot as auxiliary source (a) 11-point interpolated precision-recall (b) precision-scope (P@N) and MAP; for tag-based matching (baseline 1), standard NMF (baseline 2), S-NMF (baseline 3) and the proposed RS-NMF.

YouTube-Blogspot Retrieval Results When we use Blogspot data as the auxiliary source, the video retrieval results from YouTube follow similar trends. RS-NMF performs the best in terms of all three evaluation criteria followed by S-NMF, NMF and tag-based matching respectively. When looking at the top 10 results, we see that RS-NMF achieves around 56% precision as opposed to 52% precision achieved by S-NMF. The performance of NMF and tag-based matching methods remains the same as they do not use any auxiliary data.

It is interesting to note from Figures 5.1 and 5.2 that YouTube retrieval performance benefits slightly more from Flickr data than Blogspot data. This gain in performance could be due to the fact that the Flickr data set has more data points than the Blogspot data set. The second reason which may explain this performance gain is that the tags attached to Flickr images may be closer in semantics to the YouTube video tags than the Blogspot tags.

5.3.2 Experiment-II : Clustering Social Media and News Articles

In this subsection, we further demonstrate the usefulness of our framework for clustering applications. Beyond the traditional setting of a standard clustering algorithm, where often only one cluster partition is returned, our proposed framework performs simultaneous clustering of the related data sources and provides common and individual clusters. This directly implies the automatic discovery of common and individual topics[5] (or stories)

[5] Our shared and individual subspaces (represented by matrices $\mathbf{W}$, $\boldsymbol{U}$ and $\boldsymbol{V}$) directly provide common and individual topics. Alternatively, one can use standard topic models, e.g. Latent Dirichlet Allocation (LDA) on the documents of common and individual clusters.

Table 5.1: Description of the CNN-BBC data set : The second column shows the news feeds crawled from websites of the two news channels. The third column shows the total number of articles downloaded from each channel.

News Channel	*Feeds Used for Creating Data Set*	*No. of Articles*
CNN	‘Politics’, ‘Crime’, ‘Health’, ‘Living’, ‘Showbiz’, ‘Tech’, ‘Top Stories’, ‘Travel’, ‘US’, ‘World’, ‘Money Latest’, ‘Sports Illustrated (SI)’	4593
BBC	‘Business’, ‘Entertainment/Arts’, ‘Health’, ‘Miscellaneous’, ‘Science/Environment’, ‘Technology’, ‘Sport’, ‘World-US-Canada’	7612

across related data sources.

5.3.2.1 Data Sets

For the clustering experiments, we use *three* data sets. The *first* data set is the same social media data set that we used for the retrieval experiments described in subsection 5.3.1. The *second* data set is created by crawling news articles from CNN and BBC feeds[6]. Both news sources cover many of the same high profile real world events, but they also cover events relevant to their particular focus. Table 5.1 provides details of the selected feeds and the number of news articles for each news channel. We refer to this data set as CNN-BBC data set. For each channel, using their news articles, we generate a *tf-idf* weighted term-document matrix.

Our *third* data set is the 20 Newsgroup benchmark data set widely used for evaluating clustering methods. We mainly use this data set to compare the performance of RS-NMF with other state-of-the-art techniques in single as well as multi-task clustering.

5.3.2.2 Baseline Methods

To demonstrate the advantages of simultaneous clustering, we compare our proposed RS-NMF with the following baseline methods

[6] All CNN and BBC feeds were accessed in August 2010.

- The first baseline is a subspace based clustering which uses NMF in conjunction with K-means (similar to the work in Xu et al. [2003]) . We refer to it as "NMF+KM". Since RS-NMF is based on NMF, improvement on this baseline directly shows the benefits of simultaneous clustering using related auxiliary sources.

- The second baseline is the S-NMF algorithm which can be combined with K-means similar to the proposed RS-NMF. This baseline is referred to as "S-NMF+KM". This baseline is chosen to show how additional mutual orthogonality constraints of RS-NMF clearly segregate the common and individual subspaces and help in avoiding negative transfer learning.

- The third baseline is another subspace based clustering which uses PCA in conjunction with K-means. We refer to it as "PCA+KM". This baseline is chosen to show the benefits of nonnegative shared subspace learning over mixed-sign subspace techniques for document clustering. We also adapt PCA to perform joint clustering by augmenting the data from both sources together and use Kmeans for clustering in the learnt subspace. We refer to this baseline as "Aug-PCA+KM".

- The fourth baseline is the state-of-the-art multi-task clustering technique proposed in Gu and Zhou [2009a] which performs joint clustering of multiple data sources as a linear combination of both original input space and a shared subspace. The linear combination is controlled using a parameter λ. We refer to this method as "LSSMTC". When the parameter $\lambda = 0$, this method reduces to Adaptive Subspace Iteration (ASI) proposed in Li et al. [2004]. ASI can be used in two versions - separately clustering each source or jointly clustering the two sources by augmentine the data from the two sources. The two versions are referred to as "ASI" and "Aug-ASI" respectively.

5.3.2.3 Evaluation Metrics

We evaluate the clustering performance using four well-known metrics : accuracy (AC), normalized mutual information (NMI), average cluster entropy (ACE), cluster purity (CP).

Accuracy Given a data point y_k, let c_k and z_k denote the induced cluster label and the ground-truth category labels respectively. The accuracy (AC) is defined as

$$AC = \frac{\sum_{k=1}^{n} \delta\left(\mathrm{map}\left(c_k\right), z_k\right)}{n} \tag{5.12}$$

where n denotes the total number of data points, $\delta\left(a, b\right)$ is the delta function that equals one if $a = b$ and zero otherwise and $\mathrm{map}\left(c_k\right)$ is the mapping function that maps each cluster label c_k to the most likely category from the data set. In our experiments, we used the popular *Hungarian* algorithm [Lovász and Plummer, 1986] to implement the mapping function similar to other clustering works [Xu et al., 2003, Gu and Zhou, 2009a]. The higher the AC, the better is the clustering result.

Normalized Mutual Information Given induced partition $\mathcal{C}$ with P labels $c_1, \ldots, c_P$ and true partition $\mathcal{T}$ with Q ground-truth category labels $z_1, \ldots, z_Q$, normalized mutual information (NMI) is defined as

$$NMI = \frac{\sum_{p,q} n_{p,q} \log \frac{n_{p,q}}{n_p^{\mathcal{C}} n_q^{\mathcal{T}}}}{\sqrt{\left(\sum_p n_p^{\mathcal{C}} \log \frac{n_p^{\mathcal{C}}}{n^{\mathcal{C}}}\right)\left(\sum_q n_q^{\mathcal{T}} \log \frac{n_q^{\mathcal{T}}}{n^{\mathcal{T}}}\right)}} \tag{5.13}$$

where $n_p^{\mathcal{C}}$, $n_q^{\mathcal{T}}$ denote the number of data points in p-th cluster from partition $\mathcal{C}$ and q-th cluster from partition $\mathcal{T}$ respectively and $n_{p,q}$ denotes the number of common data points between p-th cluster from partition $\mathcal{C}$ and q-th cluster from partition $\mathcal{T}$. The higher the NMI, the better is the clustering result.

Average Cluster Entropy Given the induced partition $\mathcal{C}$ and the true partition $\mathcal{T}$, the entropy of j-th induced cluster is defined as

$$E_j = -\sum_i p_{ij} \log p_{ij}$$

where p_{ij} is the probability that a member in j-th induced cluster belongs to the ground-truth category i. Given entropy of clusters (indexed by $j = 1, \ldots, P$) in induced partition $\mathcal{C}$, average cluster entropy (ACE) is defined as follows

$$ACE = \frac{\sum_{j \in \mathcal{C}} n_j E_j}{n}$$

where n_j is the number of data points in j-th induced cluster and n is the total number of data points. To define the average cluster entropy for ground-truth category i, let $\mathcal{C}_i$ denote the subset of clusters in partition $\mathcal{C}$ such that the majority of data points belong to category i; formally, define

$$\mathcal{C}_i \triangleq \{c_j \mid p_{ij} \geq p_{kj}, k \neq i\}$$

and $N_i \triangleq \sum_{j \in \mathcal{C}_i} n_j$ then ACE_i is given as

$$ACE_i = \frac{\sum_{j \in \mathcal{C}_i} n_j E_j}{N_i}$$

Lower the ACE and ACE_i, better is the clustering result.

Cluster Purity Similar to the entropy definition, we define purity of clusters as a whole and also for each ground-truth category. As a definition, for j-th induced cluster, the major category is i if $p_{ij} \geq p_{kj}, \forall k \neq i$. Now, if, for j-th induced cluster, m_j denotes the number of points from the major category, the purity of cluster partition $\mathcal{C}$ is defined as [Manning et al., 2008]

$$CP = \frac{\sum_{j \in \mathcal{C}} m_j}{n}$$

Again, n is the total number of data points. Cluster purity for ground-truth category i is defined as

$$CP_i = \frac{\sum_{j \in \mathcal{C}_i} m_j}{N_i}$$

where definition of $\mathcal{C}_i$ and N_i remains same as above. Higher the CP and CP_i, better is the clustering result.

We follow an arrow notation such that symbol ↑ denotes that the performance in terms of an evaluation measure is *better* if its value is *higher*. Similarly, the symbol ↓ denotes that the performance in terms of an evaluation measure is *better* if its value is *lower*.

5.3.2.4 Experimental Results

Clustering based on the proposed RS-NMF first uses Algorithm 5.1 for learning shared subspaces and then follows Algorithm 5.2. Clustering based on NMF (or S-NMF) can be carried out in a similar manner except that we replace Algorithm 5.1 with the subspace learning algorithm of NMF (or S-NMF).

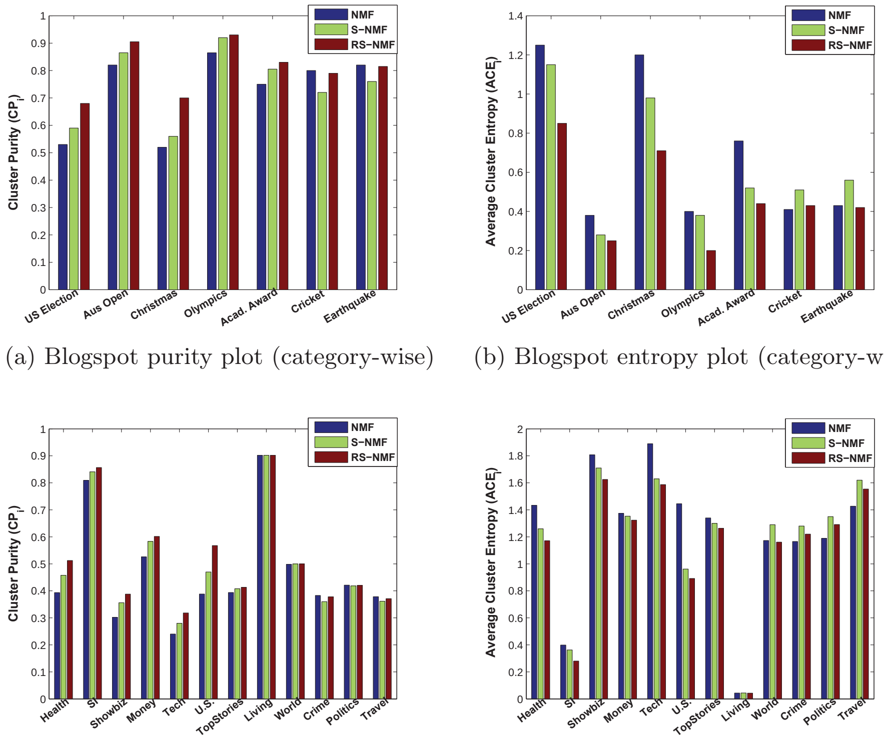
Blogspot-Flickr Clustering Results : Using the clustering results on the Blogspot-Flickr data set, we provide a detailed analysis of how shared subspace representing the common knowledge of the two data sources helps improving the clustering task (S-NMF vs NMF and RS-NMF vs NMF). We also present the comparison between S-NMF and RS-NMF by performing clustering using only shared, only individual and combined subspaces. This provides some intuition on how the regularization scheme applied in RS-NMF helps in segregating common and individual subspaces, and thus provides significant improvements in clustering performance. Following the strategy explained in section 5.4, subspace dimensionalities were set to the following values : $R_B = 62$, $R_F = 65$ and $K_{BF} = 43$ where R_B and R_F are the total subspace dimensionalities of Blogspot and Flickr data respectively and K_{BF} is the dimensionality of their shared subspace. The total number of clusters were set to 7 (equal to the true number of categories).

Since the subspace dimensionailties are chosen based on rank estimation, the same dimensionalities were also used for PCA+KM. The best performance for Aug-PCA+KM method was found by setting the number of basis vectors to 93. Since LSSMTC, Aug-PCA+KM and Aug-ASI require an equal number of clusters for each source, we set this value to 9, as the total number of different categories across the two data sets is equal to 9. The best performance of LSSMTC was achieved at $\lambda = 0.2$ and $l = 12$ where l denotes the dimensionality of shared subspace used in this method.

Table 5.2: Summary of entropy, purity, accuracy and NMI values for Blogspot clusters with Flickr used as the auxiliary source.

Blogspot (Task-1)				
Method	Avg. Cluster Entropy (↓)	Cluster Purity (↑)	Accuracy (↑)	NMI (↑)
NMF+KM	0.69 ± 0.01	0.73 ± 0.08	0.68 ± 0.01	0.40 ± 0.02
PCA+KM	0.78 ± 0.07	0.62 ± 0.11	0.63 ± 0.08	0.38 ± 0.06
Aug-PCA+KM	0.81 ± 0.13	0.58 ± 0.05	0.61 ± 0.12	0.37 ± 0.08
ASI	0.67 ± 0.06	0.74 ± 0.03	0.69 ± 0.05	0.41 ± 0.03
Aug-ASI	0.79 ± 0.06	0.63 ± 0.03	0.62 ± 0.01	0.37 ± 0.05
LSSMTC	0.66 ± 0.07	0.78 ± 0.03	0.71 ± 0.08	0.43 ± 0.02
S-NMF+KM	0.63 ± 0.08	0.75 ± 0.07	0.70 ± 0.00	0.41 ± 0.01
RS-NMF+KM	**0.47 ± 0.03**	**0.81 ± 0.05**	**0.78 ± 0.02**	**0.44 ± 0.05**

Table 5.2 provides clustering results for Blogspot data using Flickr as the auxiliary source and presents a comparison of RS-NMF+KM with the baselines in terms of average clustering entropy (ACE), cluster purity (CP), cluster accuracy (AC) and normalized mutual information (NMI). When looking at the results obtained using single source clustering methods (NMF+KM, PCA+KM and ASI+KM), we see that NMF performs better than both PCA and ASI. Similar superiority of NMF has also been reported in Xu et al. [2003]. When comparing the joint clustering methods, Aug-ASI performs slightly better than Aug-PCA+KM. Nonetheless, the performance of both the augmented methods fall short compared to the shared subspace learning methods. Out of the shared subspace methods, we see that although LSSMTC peforms better than S-NMF+KM, it is outperformed by RS-NMF+KM in terms of all evaluation metrics.



(a) Blogspot purity plot (category-wise)

(b) Blogspot entropy plot (category-wise)

(c) CNN purity plot (category-wise)

(d) CNN entropy plot (category-wise)

Figure 5.3: Category-wise purity/entropy plots depicting cluster quality for Blogspot-Flickr and CNN-BBC data sets. The first row (a and b) depicts the clustering results of Blogspot articles using Flickr as the auxiliary source. The second row (c and d) depicts the clustering results of CNN news articles using BBC as the auxiliary source. The proposed RS-NMF is compared with NMF (baseline 1) and S-NMF (baseline 2) to show the benefits of transfer learning using the cluster purity (CP) and the average cluster entropy (ACE) measures.

To investigate this performance improvement further, we compute the ACE and CP values at the cluster category level. Figure 5.3 (the first row) depicts ACE and CP values for each category. It is interesting to note from Figures 5.3a and 5.3b that RS-NMF+KM performs significantly better than NMF+KM for all but ‘Cricket World Cup’ and ‘Earthquake’ concepts (or categories) which are *individual* to the Blogspot data set (refer Table 4.1). The concepts ‘Cricket World Cup’ and ‘Earthquake’ are not available in the Flickr data set and therefore, combined learning with Flickr data does not help in improving clustering performance for these concepts. Interestingly, there is not much negative transfer in the RS-NMF+KM case which happens to occur in the case of S-NMF+KM wherein the performance clearly degrades for these two concepts. At the same time, performance improvement achieved by S-NMF+KM over NMF+KM is minimal as it is unable to clearly segregate the common and individual subspaces. As a result, S-NMF+KM is neither able to exploit the mutual knowledge for related concepts nor is it able to avoid negative transfer for domain specific concepts.

Table 5.3: Summary of entropy, purity, accuracy and NMI values for Flickr clusters with Blogspot used as the auxiliary source.

Flickr (Task-2)				
Method	Avg. Cluster Entropy (↓)	Cluster Purity (↑)	Accuracy (↑)	NMI (↑)
NMF+KM	0.74 ± 0.01	0.66 ± 0.03	0.64 ± 0.01	0.39 ± 0.03
PCA+KM	0.78 ± 0.08	0.65 ± 0.07	0.63 ± 0.05	0.37 ± 0.06
Aug-PCA+KM	0.79 ± 0.10	0.64 ± 0.09	0.64 ± 0.08	0.37 ± 0.09
ASI	0.83 ± 0.06	0.58 ± 0.04	0.53 ± 0.05	0.33 ± 0.02
Aug-ASI	0.87 ± 0.05	0.59 ± 0.03	0.54 ± 0.07	0.34 ± 0.01
LSSMTC	0.89 ± 0.04	0.54 ± 0.03	0.52 ± 0.06	0.32 ± 0.05
S-NMF+KM	0.72 ± 0.02	0.69 ± 0.06	0.67 ± 0.05	0.41 ± 0.04
RS-NMF+KM	$\mathbf{0.68 \pm 0.03}$	$\mathbf{0.72 \pm 0.05}$	$\mathbf{0.71 \pm 0.01}$	$\mathbf{0.45 \pm 0.07}$

Tables 5.3 presents the clustering results for the Flickr data set (Task-2) using Blogspot as the auxiliary source and provides a comparison of RS-NMF+KM with the two baselines. As it can be seen from the table, RS-NMF+KM remains to be the best method compared to all the baselines. Moreover, in this case, both ASI and LSSMTC are outperformed by all other methods. This degradation in performance can be attributed to the fact that ASI does not use any individual subspace. Similarly, the performance degradations of LSSMTC can be attributed to its limitation of requiring common cluster centroids for both the data sets, and thus unable to capture individual variations.

CNN-BBC News Clustering Results : Table 5.4 and 5.5 present the joint clustering results on CNN (task-1; using BBC data as auxiliary source) and BBC (task-2; using CNN data as auxiliary source) news articles. To perform clustering, the total number of clusters were set to 12 and 8 (equal to the number of categories in the data set) for CNN and BBC respectively. To learn the subspaces, we used the following set of parameters : $R_{CNN} = 50$, $R_{BBC} = 64$ and $K_{CB} = 34$ where R_{CNN} and R_{BBC} are the total subspace dimensionalities of CNN and BBC data respectively, and K_{CB} is the dimensionality of their shared subspace.

Parameter selection for other baselines remains similar to the Blogspot-Flickr experiments. The best performance for the Aug-PCA+KM method, in this case, was found by setting the number of basis vectors to 78. Again, since LSSMTC, Aug-PCA+KM and Aug-ASI require an equal number of clusters for each source, we set this value to 13 as the total number of categories in CNN and BBC are 12 and 8 respectively and most of the BBC categories (except the 'miscellaneous' category) are similar to those in CNN. The best performance for LSSMTC was obtained at $\lambda = 0.3$ and $l = 16$.

It can be seen from Table 5.4 that the clustering results of CNN show similar behavior with respect to the different methods as seen for the Blogspot-Flickr data set. RS-NMF+KM clearly outperforms all other methods including S-NMF+KM, ASI, Aug-ASI, PCA+KM, Aug-PCA+KM and LSSMTC. Figure 5.3 (the second row) depicts the category-wise purity and entropy values for clustering using RS-NMF and S-NMF and compares them with NMF based clustering. Looking at the ACE and CP values in Figure 5.3c and 5.3d, it is interesting to note that RS-NMF clearly exploits the auxiliary data of BBC channel and improves the performance for all categories *except* 'World', 'Crime', 'Politics' and 'Travel'. This is along the lines of our intuition, as out of these four categories, we do not expect the articles of 'Crime' and 'Travel' to be similar for CNN and BBC as 'Crime' related news articles are usually geographically local and 'Travel' related articles are random in nature. Similarly, there seems to be diversity among the categories 'World' and 'Politics' and the two channels need not cover similar stories.

Table 5.4: Summary of entropy, purity, accuracy and NMI values for CNN clusters with BBC used as the auxiliary source.

CNN (Task-1)				
Method	Avg. Cluster Entropy (↓)	Cluster Purity (↑)	Accuracy (↑)	NMI (↑)
NMF+KM	1.22 ± 0.08	0.47 ± 0.02	0.42 ± 0.03	0.27 ± 0.01
PCA+KM	1.26 ± 0.07	0.46 ± 0.06	0.44 ± 0.07	0.28 ± 0.03
Aug-PCA+KM	1.29 ± 0.08	0.45 ± 0.08	0.45 ± 0.03	0.29 ± 0.09
ASI	1.20 ± 0.03	0.49 ± 0.07	0.49 ± 0.05	0.31 ± 0.03
Aug-ASI	1.25 ± 0.06	0.47 ± 0.05	0.48 ± 0.03	0.30 ± 0.02
LSSMTC	1.20 ± 0.09	0.48 ± 0.12	0.47 ± 0.08	0.30 ± 0.06
S-NMF+KM	1.18 ± 0.04	0.49 ± 0.07	0.46 ± 0.02	0.29 ± 0.00
RS-NMF+KM	$\mathbf{1.12 \pm 0.06}$	$\mathbf{0.52 \pm 0.04}$	$\mathbf{0.51 \pm 0.03}$	$\mathbf{0.32 \pm 0.07}$

The clustering results on the BBC data set (task-2) is presented in Table 5.5 in the same fashion as the results of CNN (task-1). However, these results are *more interesting* due to the clear demonstration of S-NMF+KM being affected by *negative transfer learning* while RS-NMF+KM still retains the benefits of auxiliary information. Note from Table 5.5 that the performance of S-NMF+KM degrades compared to NMF+KM in terms of each evaluation criteria except ACE (still slightly better than NMF) which shows that S-NMF+KM is susceptible to negative transfer learning. This happens because the overall clustering performance for BBC data is poor due to wide variations in the data and S-NMF finds it difficult to segregate the commonalities and differences of BBC and CNN articles. Regularization used by RS-NMF stops any CNN specific patterns from entering into the shared subspace and therefore transfers only useful knowledge. In addition, RS-NMF+KM clearly outperforms LSSMTC, PCA+KM, ASI and the augmented methods. Performance of ASI and Aug-ASI has degraded heavily followed by Aug-PCA+KM and PCA+KM. Although LSSMTC performs slightly better than S-NMF+KM, its performance remains lower than RS-NMF+KM.

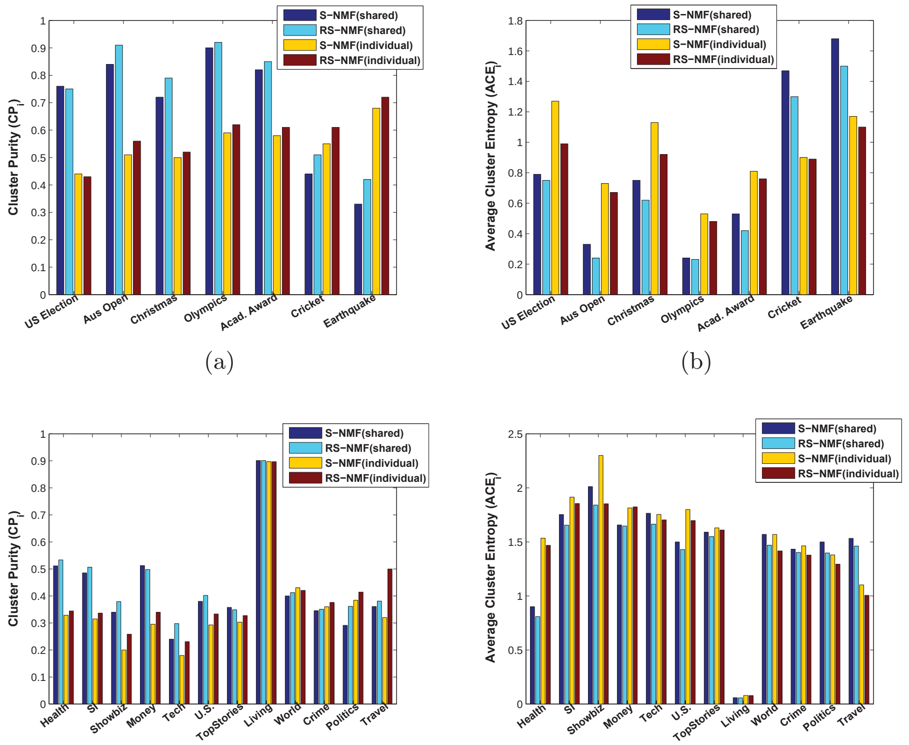
Table 5.5: Summary of entropy, purity, accuracy and NMI values for BBC clusters with CNN used as the auxiliary source.

BBC (Task-2)				
Method	Avg. Cluster Entropy (↓)	Cluster Purity (↑)	Accuracy (↑)	NMI (↑)
NMF+KM	1.23 ± 0.03	0.41 ± 0.00	0.39 ± 0.04	0.18 ± 0.02
PCA+KM	1.28 ± 0.07	0.38 ± 0.05	0.39 ± 0.05	0.17 ± 0.01
Aug-PCA+KM	1.32 ± 0.06	0.36 ± 0.02	0.39 ± 0.03	0.17 ± 0.08
ASI	1.43 ± 0.07	0.34 ± 0.06	0.33 ± 0.04	0.14 ± 0.03
Aug-ASI	1.45 ± 0.09	0.33 ± 0.08	0.34 ± 0.03	0.15 ± 0.05
LSSMTC	1.33 ± 0.05	0.36 ± 0.03	0.38 ± 0.04	0.17 ± 0.03
S-NMF+KM	1.20 ± 0.05	0.40 ± 0.02	0.37 ± 0.01	0.16 ± 0.04
RS-NMF+KM	$\mathbf{1.17 \pm 0.04}$	$\mathbf{0.44 \pm 0.03}$	$\mathbf{0.40 \pm 0.04}$	$\mathbf{0.20 \pm 0.02}$

5.3.2.5 Common and Individual Clusters

Explicit learning of the shared and individual subspaces enables us to distinguish between the *related* or *different* clusters of the two data sources. For clarity, we explain this procedure for Blogspot-Flickr data set only, similar explanation holds for the CNN-BBC data set. However, we provide results for both data sets. After learning the shared and individual subspaces for the Blogspot-Flickr data set, we use them to cluster the Blogspot data. Instead of using the full subspace, we first use only shared subspace representation ($\mathbf{H}_w$)

of the Blogspot data and cluster them to get a partition of the data using K-means. This is referred to as “RS-NMF (shared)” or “S-NMF (shared)”. Then, we use only individual subspace representation ($\mathbf{H}_u$) of Blogspot data and cluster them again to get another partition of the data using K-means. This is referred to as “RS-NMF (individual)” or “S-NMF (individual)”. We compute the purity and entropy values for each induced cluster in both data partitions. Comparing the performance of RS-NMF with S-NMF, these results are shown in Figure 5.4.

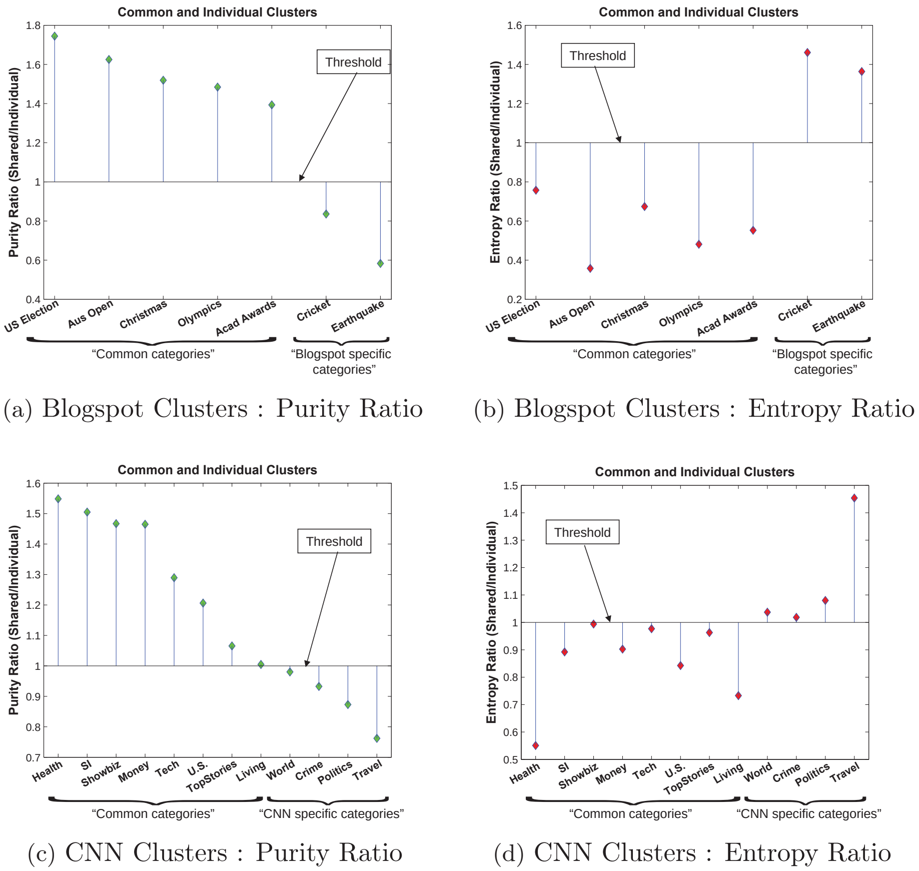


(a) (b)

(c) (d)

Figure 5.4: Comparison of RS-NMF and S-NMF category-wise clustering results using “shared subspace only representation” and “individual subspace only representation”. Using the two representations, clustering results for Blogspot data (with Flickr as auxiliary) are shown in terms of (a) cluster purity (b) average cluster entropy. Similar results for CNN data (with BBC as auxiliary) are shown in (c) and (d).

It can be clearly seen from Figure 5.4 that categories which are similar in the Blogspot and Flickr data sets are clustered well using the shared subspace representation whereas the data from Blogspot specific categories are clustered well using the individual subspace representation. Building upon this idea, we further compute the ratio of the purity values (and similarly, the ratio of entropy values) for the respective clusters in the two partitions.



(a) Blogspot Clusters : Purity Ratio

(b) Blogspot Clusters : Entropy Ratio

(c) CNN Clusters : Purity Ratio

(d) CNN Clusters : Entropy Ratio

Figure 5.5: Plots depicting the ratio of category level purities (and entropies) using "shared subspace only representation" and "individual subspace only representation" (a) For a category, purity ratio greater than 1 indicates that the corresponding news feed is common to both Blogspot and Flickr whereas purity ratio less than 1 indicates that the corresponding news feed is specific to Blogspot only (b) Similar inference can be made using entropy ratio showing opposite behavior. The corresponding results for CNN-BBC data set are shown in (c) and (d).

Figure 5.5a depicts the purity ratio plot for Blogspot clusters. It can be seen from Figure 5.5a that purity ratio clearly remains *higher than one* for the clusters which are of *common* category between Blogspot and Flickr data set. Similarly, purity ratio is always *lower than one* for clusters which are *specific* to Blogspot data set. An inverse behavior can be seen from Figure 5.5b depicting entropy ratio for each category. For common categories, a value of cluster purity ratio higher than one *implies* that clusters obtained using shared subspace representation are purer than those obtained using individual subspace representation. We note that this phenomenon is due to the clear segregation of shared and individual subspaces. As a result, the clusters which are from the common categories of the two data sets are clustered well (with high purity and low entropy) in shared subspace and not so well (with low purity and high entropy) in individual subspace. Exactly the opposite happens to the clusters of specific categories to Blogspot data which are clustered well in the individual rather than the shared subspace. Similar results are obtained for CNN-BBC data set and shown in Figures 5.5c and 5.5d.

5.3.3 Comparison using 20 Newsgroup Benchmark Data Set

5.3.3.1 20 Newsgroup Data Set[7]

20-Newsgroup is one of the most widely used data set for both single domain clustering and cross-domain learning [Xu et al., 2003, Cai et al., 2008, Gu and Zhou, 2009a]. It is a collection of 20000 newsgroup documents, almost evenly divided into 20 categories. We use this data set to compare with a variety of single-domain clustering techniques and a recent multi-task clustering method proposed in Gu and Zhou [2009a]. To have a fair comparison, we create a multi-task clustering (having two tasks) data set from 20 Newsgroup documents in the same fashion as in Gu and Zhou [2009a], and use this data set (we shall refer to it as Rec-Talk as in Gu and Zhou [2009a]) for evaluating multi-task clustering capabilities of the RS-NMF model. For the first task, we include the documents from “rec.autos” and “talk.politics.guns” categories while, for the second task, we use the documents from “rec.sport.baseball” and “talk.politics.mideast” categories. This kind of splitting ensures that the two tasks are related, yet different, due to being constructed from the same categories but with different sub-categories.

5.3.3.2 Baseline Methods

We show the superiority of RS-NMF to several single-task clustering methods such as K-means (KM) in input data space, K-means followed by Principal Component Analysis (PCA+KM), Normalized Cut (Ncut) [Shi and Malik, 2000], K-means followed by Non-negative Matrix Factorization (NMF+KM) [Xu et al., 2003], adaptive subspace iteration (ASI) [Li et al., 2004] and a multi-task clustering method LSSMTC [Gu and Zhou, 2009a]. Single-domain clustering algorithms are applied on the 20 Newsgroup Rec-Talk data set in two ways (1) apply a single-domain clustering algorithm on the two tasks independently (2) apply a single-domain clustering algorithm on the two tasks by merging the data from both the tasks. Therefore, every single-domain clustering task has two versions. For example, the two versions for K-means algorithm are referred to as “KM” and “All KM”. Similarly, K-means followed by PCA are referred to as “PCA+KM” and “All PCA+KM” and so on. LSSMTC is a recent multi-task clustering algorithm closely related with our

[7] http://people.csail.mit.edu/jrennie/20Newsgroups/

Table 5.6: Comparison with the state-of-the-art single and multi-task clustering methods using Rec-Talk (20 Newsgroup) benchmark data set. The clustering results for LSSMTC are taken from Gu and Zhou [2009a] for reference purpose.

	Rec (Task-1)		*Talk (Task-2)*	
Method	Accuracy (↑)	NMI (↑)	Accuracy (↑)	NMI (↑)
NMF+KM	0.6047 ± 0.13	0.4111 ± 0.05	0.8791 ± 0.14	0.6030 ± 0.08
Aug-NMF+KM	0.6421 ± 0.10	0.4232 ± 0.06	0.7636 ± 0.10	0.5812 ± 0.05
LSSMTC	0.8433 ± 0.08	0.4306 ± 0.06	0.7895 ± 0.08	0.3473 ± 0.08
S-NMF+KM	0.8253 ± 0.00	0.4362 ± 0.00	0.8196 ± 0.13	0.4202 ± 0.17
RS-NMF+KM	$\mathbf{0.9674 \pm 0.11}$	$\mathbf{0.7933 \pm 0.14}$	$\mathbf{0.9029 \pm 0.09}$	$\mathbf{0.6763 \pm 0.08}$

work with the difference that learning the shared subspace is carried out in an entirely different manner. Using Rec-Talk data set, we compare our results with LSSMTC only as LSSMTC has already been shown to outperform other baselines (comparison with NMF is not carried out) in Gu and Zhou [2009a]. We use an *identical* setting as used in Gu and Zhou [2009a] so that a comparison can be made with NMF+KM and the proposed RS-NMF. In addition, we also compare RS-NMF with S-NMF (a special case of RS-NMF model without any regularization) to show the benefits of regularization.

5.3.3.3 20 Newsgroup Clustering Results

To generate clustering results based on RS-NMF, we use Algorithms 5.1 (with parameters $R_1 = R_2 = 30$ and $K = 18$; learnt using the procedure described in section 5.4) followed by Algorithm 5.2. For NMF and S-NMF, instead of Algorithm 5.1, we use equivalent subspace learning algorithm and then use K-means as in Algorithm 5.2. Table 5.6 presents the comparison of RS-NMF with the above-mentioned baselines (S-NMF, LSSMTC and various single-task clustering methods) using Rec-Talk data set. For the experiment, the number of clusters for each task were set to 2 as in Gu and Zhou [2009a]. Since the algorithms are iterative, each algorithm was run 50 times and we report the mean value along with standard deviations. The clustering results for LSSMTC were reported in Gu and Zhou [2009a] under identical settings. We use those results here for a comparison with other NMF based methods.

It can be seen from Table 5.6 that all three shared subspace learning methods (LSSMTC, S-NMF+KM and RS-NMF+KM) clearly outperform NMF+KM for Rec data set (Task-1) whereas only RS-NMF outperforms NMF+KM for Talk data set (Task-2). This improvement in performance stems from the ability of RS-NMF to exploit the knowledge available in auxiliary domains and transfer it to the target task appropriately without any negative transfer learning. Note that when single-task clustering algorithms are used on the combined data of Task-1 and Task-2, they do not necessarily perform better since the data distribution differs between the two tasks and simply combining them may even lead to negative knowledge transfer. This can be clearly seen in the case of Aug-NMF+KM results for Talk data set. When comparing shared subspace learning methods, we can see that RS-NMF+KM (and S-NMF+KM except on Task-1) outperforms LSSMTC. This is because, although LSSMTC learns a shared subspace to exploit the common knowledge between the two tasks, it forces common centroids for each task. This is too restrictive to model real world data because, in spite of sharing a subspace, the two data sources would usually have different coordinates in the shared subspace and their clusters need not share the cluster centroids. When comparing the results of RS-NMF and S-NMF, we see that RS-NMF clearly outperforms S-NMF. This is due to the regularization scheme imposed through RS-NMF formulation. Through regularization, RS-NMF tries to get subspaces corresponding to $\mathbf{W}$,$\boldsymbol{U}$ and $\boldsymbol{V}$ matrices such that they are mutually disjoint. This separates the common knowledge between the tasks from their individual knowledge and transfers only the common knowledge through matrix $\mathbf{W}$ whereas S-NMF does not force any such regularization. Due to the lack of regularization in S-NMF, the matrix $\mathbf{W}$ may contain basis vectors representing not only common structures but also some individual structures (usually happens for strong topics). This causes negative knowledge transfer (compare NMF+KM and S-NMF+KM results for Talk data set) and results in suboptimal performance.

5.4 Learning Parameters and Convergence Behavior

In this section, we provide a way to learn the various parameters required for our regularized shared subspace learning framework proposed in section 5.1.

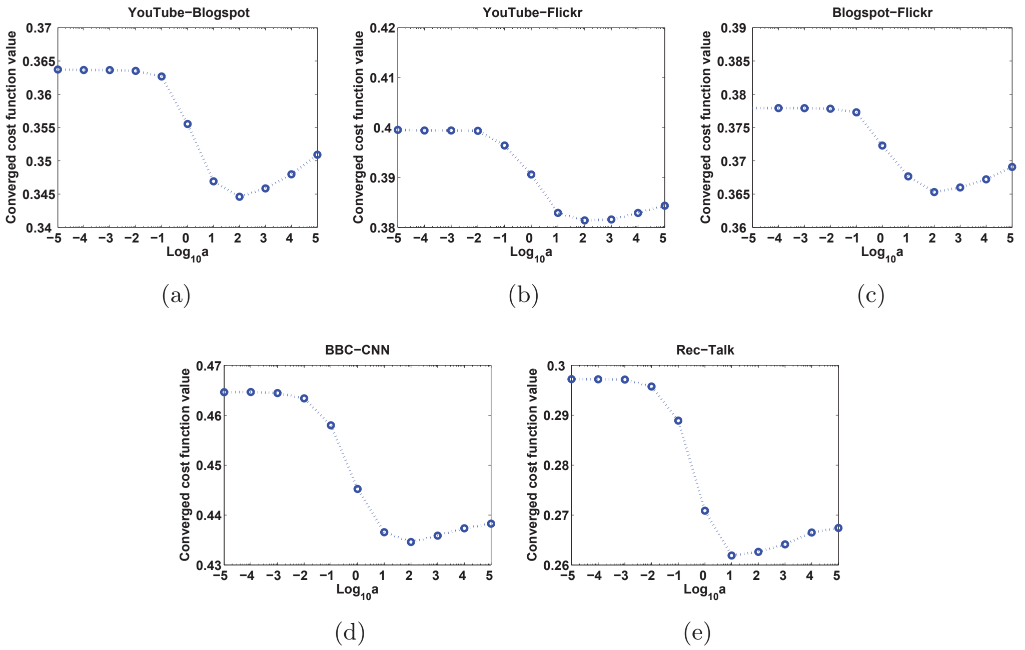
5.4.1 Regularization Parameters

Consider the regularization term used in Eq (5.2)

$$R\left(\mathbf{W},\boldsymbol{U},\boldsymbol{V}\right)=\alpha\left\|\mathbf{W}^{\mathsf{T}}\boldsymbol{U}\right\|_F^2+\beta\left\|\mathbf{W}^{\mathsf{T}}\boldsymbol{V}\right\|_F^2+\gamma\left\|\boldsymbol{U}^{\mathsf{T}}\boldsymbol{V}\right\|_F^2 \tag{5.14}$$

Each element in matrices $\mathbf{W}^{\mathsf{T}}\boldsymbol{U}$, $\mathbf{W}^{\mathsf{T}}\boldsymbol{V}$ and $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{V}$ has a value between 0 and 1 as each column of $\mathbf{W}$, $\boldsymbol{U}$ and $\boldsymbol{V}$ is normalized to L_2-norm 1. Therefore, it is appropriate to normalize the term having $\left\|\mathbf{W}^{\mathsf{T}}\boldsymbol{U}\right\|_F^2$ by K_wK_u since there are $K_w \times K_u$ elements in $\mathbf{W}^{\mathsf{T}}\boldsymbol{U}$.

Similarly, $\left\|\mathbf{W}^{\mathsf{T}}\boldsymbol{V}\right\|_F^2$ and $\left\|\boldsymbol{U}^{\mathsf{T}}\boldsymbol{V}\right\|_F^2$ are normalized by K_wK_v and K_uK_v respectively. Except for these differences in normalization, we treat the regularization for all three terms in Eq (5.14) equally, i.e. weighting them by a common factor a. Explicitly speaking, we use $\alpha = \frac{a}{K_wK_u}$, $\beta = \frac{a}{K_wK_v}$ and $\gamma = \frac{a}{K_uK_v}$ where $K_w \triangleq K$, $K_u = R_1 - K$, $K_v = R_2 - K$ and a is a common regularization factor for each orthogonalization. By increasing the value of a, we obtain solutions which become increasingly mutually orthogonal (i.e. $R(\mathbf{W}, \boldsymbol{U}, \boldsymbol{V})$ moves closer to zero) but at the same time, it also causes an increase in the joint factorization error, i.e. $\lambda_{\mathbf{X}} \|\mathbf{X} - [\mathbf{W} \mid \boldsymbol{U}]\mathbf{H}\|_F^2 + \lambda_{\boldsymbol{Y}} \|\boldsymbol{Y} - [\mathbf{W} \mid \boldsymbol{V}]\mathbf{L}\|_F^2$. This necessitates a trade-off between the data-fitting and mutual orthogonality. Motivated by this, when we look at the combined objective function of optimization problem Π, we find that there exists an optimum value for a. Figure 5.6 shows the variations of combined objective function w.r.t the parameter a. We can see that, for all the data sets except Rec-Talk, the optimum value of a is 100 whereas the optimum value of a for Rec-Talk data set is 10. Even for Rec-Talk, $a = 100$ achieves a value of combined objective function almost equal to the optimum value. Therefore, the parameter a can be fixed at 100 for most of the real-world data sets.



(a) (b) (c)

(d) (e)

Figure 5.6: Variations in cost function w.r.t. the regularization parameter for various target/auxiliary pairs.

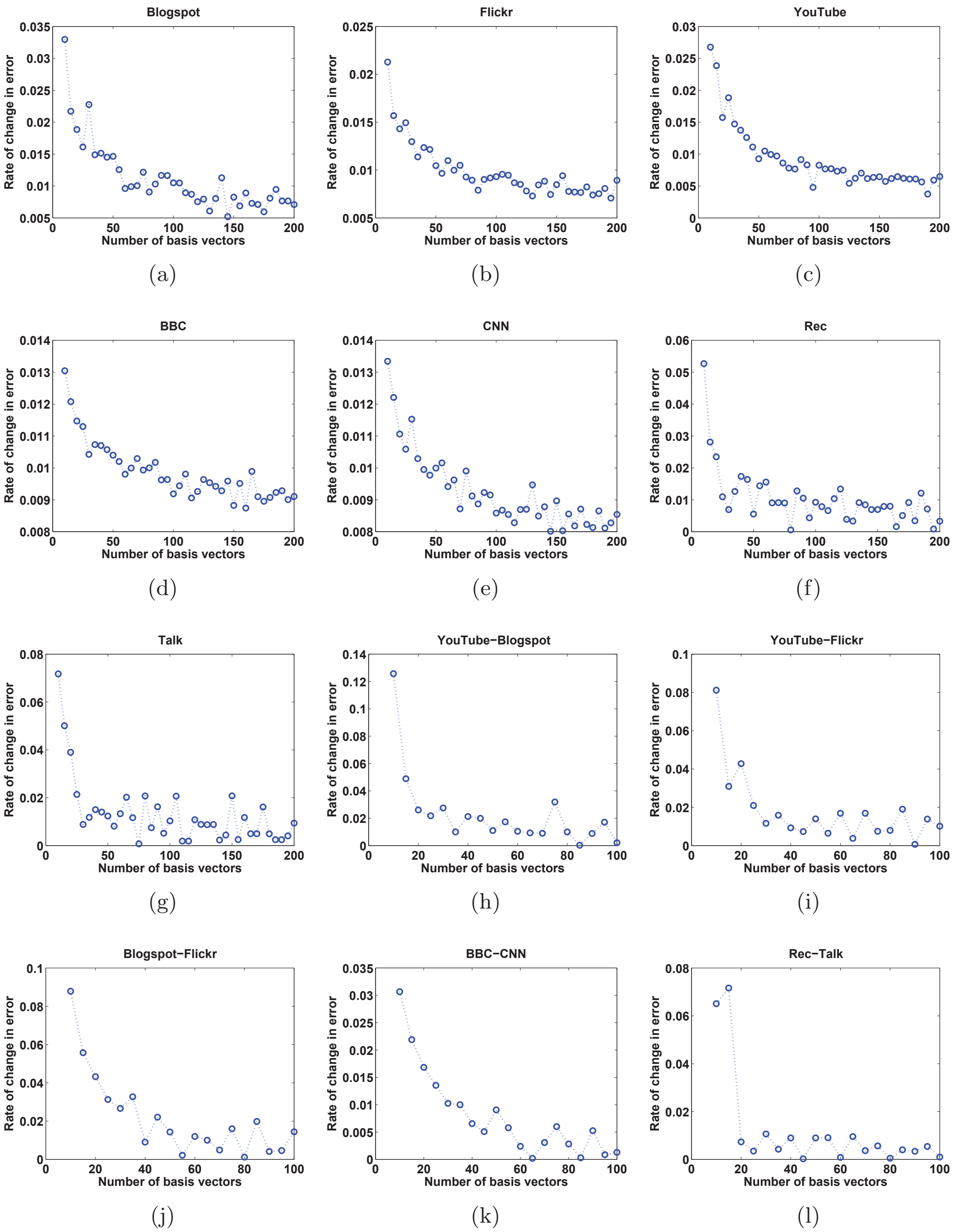
5.4.2 Subspace Dimensionality

The dimensionality of shared and individual subspaces can be determined by estimating the number of significant nonnegative basis vectors used for fitting the data. In particular, the subspace dimensionalities R_1 and R_2 can be estimated from nonnegative matrix factorization of matrices $\mathbf{X}$ and $\boldsymbol{Y}$ respectively with *increasing* value of basis vectors. We plot the rate of change of factorization error (Δ_k) with respect to the number of basis vectors where $\Delta_k \triangleq 1 - \frac{\mathbf{E}_{k+1}}{\mathbf{E}_k}$ and $\mathbf{E}_k$ is the converged NMF factorization error at the number of basis vectors used at k-th index. As can be seen from Figure 5.7, Δ_k decreases sharply in the beginning and then, becomes almost constant. In other words, after a certain value of the number of basis vectors, the value of Δ_k saturates inspite of further increases in the number of basis vectors - indicating that the number of basis vectors have already reached the inherent true dimensionality. As a conservative estimate, the number of basis vectors are set to a value after which Δ_k reduces to a value less than 1%. The first seven plots in Figure 5.7 (a-g) show the variation of Δ_k w. r. t. the number of basis vectors (R) for each data set. Since we have generated these plots at an interval of 5, we choose the best value within the selected interval based on the task performance.

For selecting the shared subspace dimensionality, i.e. K, we use a similar strategy as above. Assuming that $\mathbf{H}_w$ and $\mathbf{L}_w$ are full-rank (which is always the case if $N_1 > K$ and $N_2 > K$) and $\mathbf{W}$, $\boldsymbol{U}$ and $\boldsymbol{V}$ are mutually orthogonal, K can be estimated from the nonnegative matrix factorization of matrix $\mathbf{X}^\mathsf{T}\boldsymbol{Y}$. Consider,

$$\mathbf{X}^\mathsf{T}\boldsymbol{Y} \approx \mathbf{H}^\mathsf{T}\begin{bmatrix} \mathbf{W}^\mathsf{T}\mathbf{W} & \mathbf{W}^\mathsf{T}\boldsymbol{V} \\ \boldsymbol{U}^\mathsf{T}\mathbf{W} & \boldsymbol{U}^\mathsf{T}\boldsymbol{V} \end{bmatrix}\mathbf{L} = \mathbf{H}^\mathsf{T}\begin{bmatrix} \mathbf{W}^\mathsf{T}\mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\mathbf{L} = \left(\mathbf{H}_w^\mathsf{T}\mathbf{W}^\mathsf{T}\right)(\mathbf{W}\mathbf{L}_w)$$

which is a product of two nonnegative matrices. Using above observation, we compute the nonnegative matrix factorization of $\mathbf{X}^\mathsf{T}\boldsymbol{Y}$ with increasing numbers of basis vectors. Once the number of basis vectors reaches the true matrix rank, the rate of change of the cost function (Δ_k) becomes minimal inspite of further increases in the number of basis vectors. Again, we select a value for the number of basis vectors after which Δ_k reduces to a value less than 1% and set this value as K. The last five plots in Figure 5.7 (h-l) show the variation of Δ_k w. r. t. the number of shared basis vectors (K) for data set pairs (target/auxiliary) used in our experiments section. Again, after selecting a particular interval with 1% criteria, the best value within the selected interval is chosen based on the task performance.



(a) (b) (c)

(d) (e) (f)

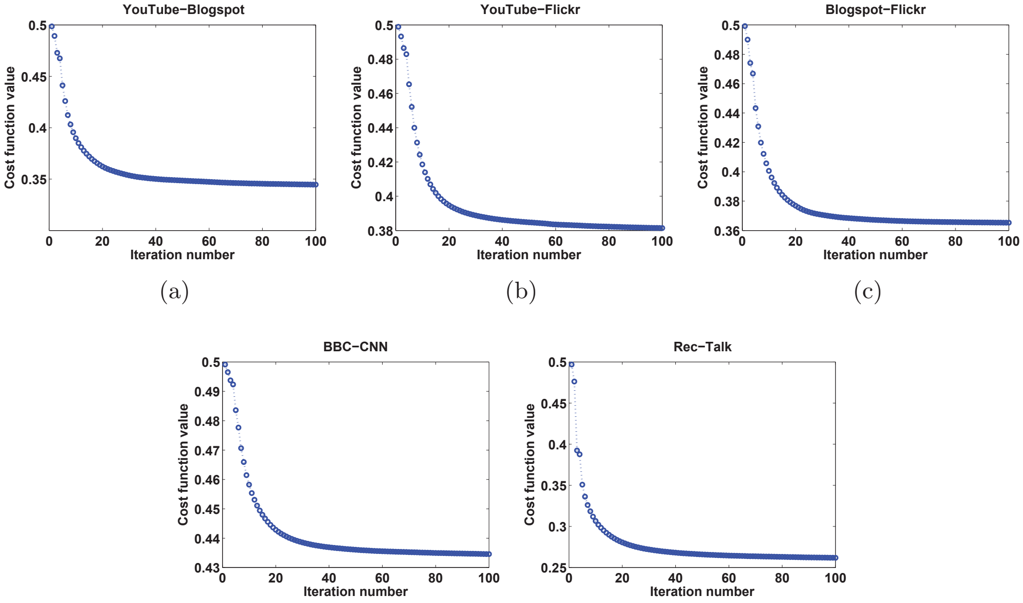
(g) (h) (i)

(j) (k) (l)

Figure 5.7: Subspace dimensionality plots; the variations w.r.t. the total subspace dimensionalities are shown in (a)-(g) whereas those w.r.t. the shared subspace dimensionalities are shown in (h)-(l).

5.4.3 Convergence Behavior

In section 5.1.3, we have shown that the iterative updates in Eqs (5.4–5.7) are convergent. Here, we empirically show the convergence behavior of RS-NMF for all the target/auxiliary pairs (e.g. Blogspot-Flickr, CNN-BBC etc) used for the experiments. Figure 5.8 shows that RS-NMF converges within almost 50 iterations for each target/auxiliary data set pair. Moreover, we note that the time taken per iteration for RS-NMF remains similar to the standard NMF as the order of complexity for both NMF and the RS-NMF is the same.



(a) (b) (c)

(d) (e)

Figure 5.8: Convergence behavior of RS-NMF for various target and auxiliary data set pairs.

5.5 Extension to Multiple Data Sources

In this section, we extend the regularized shared subspace learning framework to multiple data sources having an arbitrary sharing configuration in the same fashion along the lines of MS-NMF. We term this framework as RMS-NMF. Before proceeding further, let us recall some of the notations used in chapter 4. For a set of n data sets having an arbitrary sharing configuration, $S(n)$ denotes the power set of $\{1, 2, \ldots n\}$. For each $i = 1, \ldots, n$, let the index set associated with the i-th data source be defined as $S(n, i) = \{v \in S(n) \mid i \in v\}$. Using this notation, our proposed joint matrix factorization for n data sources can be written as

$$\mathbf{X}_i = \mathbf{W}_i \cdot \mathbf{H}_i = \sum_{v \in S(n,i)} W_v \cdot H_{i,v}$$

If we explicitly list out the subspace indices of i-th data source as $S(n,i) = \{v_1, v_2, \ldots, v_Z\}$ (where $Z = 2^{n-1}$), $\mathbf{W}_i$ and $\mathbf{H}_i$ can then be written as

$$\mathbf{W}_i = \left[W_{v_1} \mid W_{v_2} \mid \ldots \mid W_{v_Z} \right] \ , \mathbf{H}_i = \begin{bmatrix} H_{i,v_1} \\ \vdots \\ H_{i,v_Z} \end{bmatrix}$$

Using the above notation, we note that the subspace matrices W_{v_z} for different z, may capture basis vectors which are redundant and not segregated well. This problem can be addressed by using a regularization scheme as discussed earlier in this chapter. In the case of multiple data sources, each of the W_{v_z} should be mutually orthogonal with other matrices. Formally, the regularization scheme aims to achieve the following relation

$$W_{v_{z_j}}^{\mathsf{T}} W_{v_{z_k}} = 0 \text{ , for every } j,\, k \text{ such that } j \neq k \tag{5.15}$$

Incorporating the above mutual orthogonality constraint in the cost function of Eq (4.5) (used for MS-NMF), the modified cost function can be written as the following

$$J(\mathbf{W}, \mathbf{H}) = \sum_{i=1}^{n} \lambda_i \left\| \mathbf{X}_i - \sum_{v \in S(n,i)} W_v \cdot H_{i,v} \right\|_F^2 + \sum_{\substack{j,k=1 \\ (j \neq k)}}^{Z} \mu_{jk} W_{v_{z_j}}^{\mathsf{T}} W_{v_{z_k}} \tag{5.16}$$

where $\mu_{jk} = \frac{\mu}{K_{v_{z_j}} \times K_{v_{z_k}}}$ and $K_{v_{z_j}}$ comes from the size of $(M \times K_{v_{z_j}})$ dimensional matrix $W_{v_{z_j}}$. Note that the similar regularization parameter is chosen for RS-NMF and is motivated in section 5.4.1. The objective function in Eq (5.16) can be minimized following the same lines of RS-NMF described earlier in this chapter. The multiplicative update for W_v in this case, is given as

$$(W_v)_{lk} \leftarrow (W_v)_{lk} \frac{\left(\sum_{i \in v} \lambda_i \mathbf{X}_i \cdot H_{i,v}^{\mathsf{T}} \right)_{lk}}{\left(\sum_{i \in v} \lambda_i \mathbf{X}_i^{(t)} \cdot H_{i,v}^{\mathsf{T}} + \sum_{u \neq v} \left(\frac{\mu}{K_u K_v} W_u W_u^{\mathsf{T}} W_v \right) \right)_{lk}} \tag{5.17}$$

Update equations for matrices $H_{i,v}$ do not get affected from the regularization. Therefore, these updates remain similar to the MS-NMF case and are given in Eq (4.11). As can be verified by inspection from Eq (5.17) that, for two data source case, i.e. when $n = 2$, Eq (5.17) reduces to Eqs (5.4-5.6). Similarly, when no regularization is imposed, e.g. the problem setting of MS-NMF, Eq (5.17) reduces to Eq (4.10). The convergence of above updates can be proved along the same lines as RS-NMF.

5.6 Discussion

In this chapter, we have presented a regularized shared subspace learning framework which captures the commonalities (through a shared subspace) and differences (through individual subspaces) of the related data sources while ensuring segregation of the constituent subspaces. This segregation is important in dealing with imbalanced real-world data sets, where without such a segregation, negative transfer learning [Rosenstein et al., 2005] may occur when transferring the common knowledge across different sources. To achieve such segregation between the shared and individual subspaces, our framework imposes a set of mutual orthogonality constraints. To solve the resulting constrained-optimization problem, we have derived efficient iterative multiplicative updates and show that convergence of these iterative updates is mathematically guaranteed. Based on the proposed framework, we have developed efficient social media retrieval and clustering algorithms and demonstrated them using three real-world social media and news data sets. Our experiments consistently validate the effectiveness of the proposed regularized model (RS-NMF) by achieving better performance compared to the unregularized counterpart (S-NMF) for both retrieval and clustering applications using three real world data sets. In addition, comparisons made with various state-of-the-art single and multi-task clustering techniques using both social media and news data sets demonstrate the effectiveness of RS-NMF for multi-task clustering. While in the major part of the chapter we have focused on the regularized subspace learning for two data sources, we have shown that the proposed approach can readily be extended to model multiple data sources with arbitrary sharing configurations along the lines of MS-NMF.

In the previous chapters, we have developed various shared subspace learning models for the nonnegative data, the similar models for *mixed-sign* data are highly desirable. Another direction is to extend these models by treating the factorization matrices *probabilistically* allowing us to deal with uncertainties of the real-world data in an effective manner. In fact, we can address both these above limitations simultaneously by proposing probabilistic shared subspace learning without imposing any constraints of nonnegativity. This is the subject of the next chapter.

Chapter 6

Bayesian Shared Subspace Learning

In this chapter, we depart from the nonnegative matrix factorization based models in two ways. First, we aim to develop a framework alternative to MS-NMF (described in chapter 4), which can model the mixed-sign data from multiple sources, not merely the nonnegative data. Second, we would like to have a probabilistic model that can effectively handle the uncertainities of real world data. To achieve these goals in a unified manner, we develop a fully Bayesian framework that can jointly model the data from multiple data sources without assuming nonnegative constraints. Carrying forward the idea of shared subspace learning, the proposed framework allows multiple data sources to exploit their collective and individual strengths through probabilistic shared and individual subspaces. Being a Bayesian framework, this yields a probability distribution over the shared and individual subspaces – a feature that is important for dealing with uncertainties and model overfitting.

The proposed framework is based on the state-of-the-art Bayesian probabilistic matrix factorization (BPMF) model, recently proposed in Salakhutdinov and Mnih [2008]. We extend BPMF to enable joint modeling of multiple data sources through shared and individual subspaces, and derive inference algorithms using *Rao-Blackwellized* Gibbs Sampling (RBGS). In spite of being a Bayesian technique, which is usually conceived as computationally expensive, the BPMF model has been shown to be efficient and applied on large scale data sets. While developing the extension of BPMF for multiple data sources, we ensure that the efficiency of BPMF is retained. To demonstrate the usefulness of our approach, we choose the same applications as in chapter 4, i.e., improving social media retrieval using auxiliary sources, and cross-media retrieval. We use three disparate data sources – Flickr, YouTube and Blogspot – to show the effectiveness of the proposed probabilistic shared subspace learning framework.

This chapter is organized as follows. Section 6.1 presents the Bayesian shared subspace learning (BSSL) formulation and describes Gibbs sampling based inference procedure. Sections 6.2 and 6.3 demonstrate the proposed framework for social media retrieval applications. Finally, a discussion is provided in Section 6.4.

6.1 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 $\mathbf{X}_1, \ldots, \mathbf{X}_n$, which, for example, can be term-document matrices (each row a word and each column a document with *tf-idf* features [Baeza-Yates and Ribeiro-Neto, 1999]) 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 this is not the case, one can always merge the vocabularies of each data source to form a common vocabulary. Following the notation of chapter 4 that was developed for indexing shared and individual subspaces for multiple data sources, our proposed shared subspace learning framework seeks a set of expressions in the following form, for $i = 1, \ldots, n$

$$\mathbf{X}_i = \mathbf{W}_i \cdot \mathbf{H}_i + \mathbf{E}_i = \sum_{v \in S(n,i)} W_v \cdot H_{i,v} + \mathbf{E}_i,\ i = 1, \ldots, n \tag{6.1}$$

It is also clear from Eq (6.1) 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 $S(n,i) = \{v_1, v_2, \ldots, v_Z\}$ then $\mathbf{W}_i$, $\mathbf{H}_i$ are

$$\mathbf{W}_i = [W_{v_1} \mid W_{v_2} \mid \ldots \mid W_{v_Z}] \text{ and } \mathbf{H}_i = \begin{bmatrix} H_{i,v_1} \\ \vdots \\ H_{i,v_Z} \end{bmatrix} \tag{6.2}$$

As an example of the above factorization, 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

$$\mathbf{X}_1 = \underbrace{[W_{12} \mid W_1]}_{\mathbf{W}_1} \underbrace{\begin{bmatrix} H_{1,12} \\ \\ H_{1,1} \end{bmatrix}}_{\mathbf{H}_1} + \mathbf{E}_1 \text{ and } \mathbf{X}_2 = \underbrace{[W_{12} \mid W_2]}_{\mathbf{W}_2} \underbrace{\begin{bmatrix} H_{2,12} \\ \\ H_{2,2} \end{bmatrix}}_{\mathbf{H}_2} + \mathbf{E}_2 \tag{6.3}$$

6.1.1 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

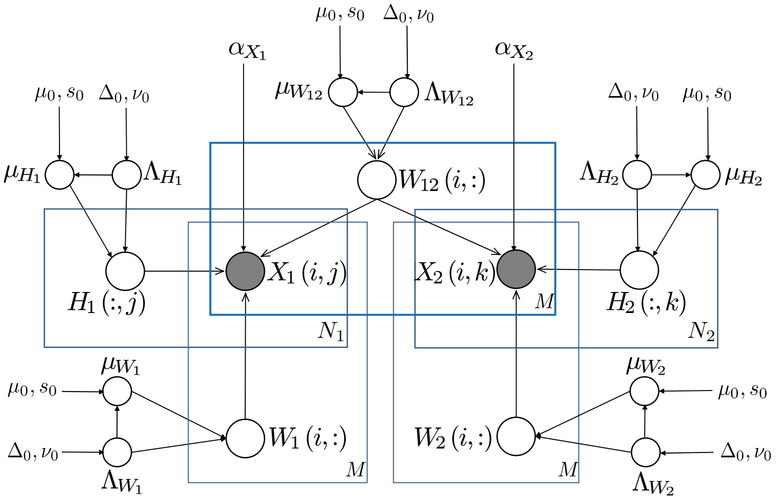


Figure 6.1: Directed graphical representation of BSSL (shown for a special case of two data sources, i.e. $n = 2$).

shared subspace learning for an arbitrary number of data sources, for simplicity, we show the graphical model for the case of two data sources in Figure 6.1. For each $i \in \{1, 2, \dots n\}$ and $v \in S(n, i)$, our probabilistic model is then given as

$$p\left(\mathbf{X}_i(m,l) \mid \mathbf{W}_i, \mathbf{H}_i, \Lambda_{\mathbf{X}_i}\right) = \mathcal{N}\left(\mathbf{X}_i(m,l) \mid \mathbf{W}_i^{(m)} \mathbf{H}_i^{[l]}, \Lambda_{\mathbf{X}_i}^{-1}\right) \tag{6.4}$$

$$p\left(W_v \mid \mu_{W_v}, \Lambda_{W_v}\right) = \prod_{m=1}^{M} \mathcal{N}\left(W_v^{(m)} \mid \mu_{W_v}, \mathit{\Lambda}_{W_v}^{-1}\right) \tag{6.5}$$

$$p\left(\mathbf{H}_i \mid \mu_{\mathbf{H}_i}, \Lambda_{\mathbf{H}_i}\right) = \prod_{l=1}^{N_i} \mathcal{N}\left(\mathbf{H}_i^{[l]} \mid \mu_{\mathbf{H}_i}, \mathit{\Lambda}_{\mathbf{H}_i}^{-1}\right) \tag{6.6}$$

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

$$p\left(\mu_{W_v}, \Lambda_{W_v} \mid \Psi_0\right) = \mathcal{N}\left(\mu_{W_v} \mid \mu_0, \left(s_0 \Lambda_{W_v}\right)^{-1}\right) \mathcal{W}\left(\Lambda_{W_v} \mid \Delta_0, \nu_0\right) \tag{6.7}$$

where $\mathcal{W}(. \mid \Delta_0, \nu_0)$ is Wishart distribution with $K_v \times K_v$ scale matrix Δ_0 and ν_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 $\Psi_0 \triangleq \{\mu_0, s_0, \Delta_0, \nu_0, \alpha_{\mathbf{X}_1}, \dots, \alpha_{\mathbf{X}_n}\}$.

6.1.2 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 $i \in \{1, 2, \ldots n\}$ 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 Ψ_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 auto-covariance of the Gibbs samples and provide better mixing of Markov chain [Liu, 1994]. This scheme is known as collapsed Gibbs sampling or *Rao-Blackwellized* Gibbs Sampling [Casella and Robert, 1996, Sudderth, 2006].

After integrating out these nuisance parameters, our state space reduces to only the factor matrices $\{W_v, \mathbf{H}_i\}$ for all $i \in \{1, 2, \ldots n\}$ 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 6.1 outlines these sampling steps while the rest of this section shall briefly explain how to obtain the Gibbs conditional distributions as in Eqs (6.10-6.13).

Recalling $\mathbf{W}_i$ and $\mathbf{H}_i$ from Eq. (6.2), 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 $W_v^{(-m)}$), the coefficient matrices $\mathbf{H}_i$ for all i, observed data $\mathbf{X}_i$ for all i and the hyperparameters Ψ_0, is given by

$$\begin{aligned} p\left(W_v^{(m)} \mid \mathbf{X}_{1:n}, W_v^{(-m)}, W_{\{u \in S(n,i) \mid u \neq v\}}, \mathbf{H}_{1:n}, \Psi_0\right) \\ \propto \left[\prod_{i=1}^{n}\prod_{l=1}^{N_i} p\left(\mathbf{X}_i(m,l) \mid \mathbf{W}_i^{(m)}\mathbf{H}_i^{[l]}, \Lambda_{\mathbf{X}_i}^{-1}\right)\right] \times p\left(W_v^{(m)} \mid W_v^{(-m)}, \Psi_0\right) \end{aligned} \tag{6.8}$$

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 $W_v^{(-m)}$. The predictive distribution of $W_v^{(m)}$ conditioned on $W_v^{(-m)}$ and Ψ_0 is obtained by integrating over the parameters of the normal–inverse–Wishart posterior distribution and is *multivariate Student–t* [Gelman, 2004]. 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 [Gelman, 2004]. Thus, the predictive distribution is given as

Algorithm 6.1 Rao-Blackwellized Gibbs Sampling (RBGS) for BSSL.

```
Input: Hyperparameters $\Psi_0$, number of samples $L$.
For each $i$, initialize matrices $\mathbf{W}_i$, $\mathbf{H}_i$ randomly.
for $r = 1, \ldots, L$ do
    For each $v$, draw $r$-th sample $[W_v]^r$ from normal distribution parameterized by Eqs.(6.10–6.11).
    For each $i$, draw $r$-th sample $[\mathbf{H}_i]^r$ from normal distribution parameterized by Eqs.(6.12–6.13).
end for
For each $v$ and $i$, get an estimate of $W_v$ and $\mathbf{H}_i$ using the Gibbs samples as $W_v \approxeq \frac{1}{L}\sum_{r=1}^{L}[W_v]^r$, $\mathbf{H}_i \approxeq \frac{1}{L}\sum_{r=1}^{L}[\mathbf{H}_i]^r$.
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 $i$.
```

$$p\left(W_v^{(m)} \mid W_v^{(-m)}, \Psi_0\right) \approx \mathcal{N}\left(W_v^{(m)} \mid \mu^{pred}_{W_v^{(m)}}, \Lambda^{pred}_{W_v^{(m)}}\right) \tag{6.9}$$

$$\text{where} \quad \mu^{pred}_{W_v^{(m)}} = \frac{s_0\mu_0 + \sum_{\substack{l=1\\l\neq m}}^{M} W_v^{(l)}}{s_0 + (M-1)}, \quad \Lambda^{pred}_{W_v^{(m)}} = \frac{(s_m+1)\,\Delta_m^{-1}}{s_m\left(\nu_m - K_{W_v} - 1\right)},$$

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

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

Using Eqs (6.8) and (6.9), the posterior distribution can be written as

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

$$\Lambda^{post}_{W_v^{(m)}} = \Lambda^{pred}_{W_v^{(m)}} + \sum_{i=1}^{n} H_{i,v}\Lambda_{\mathbf{X}_i}H_{i,v}^{\mathsf{T}} \tag{6.10}$$

$$\left(\mu^{post}_{W_v^{(m)}}\right)^{\mathsf{T}} = \left[\Lambda^{post}_{W_v^{(m)}}\right]^{-1}\left[\Lambda^{pred}_{W_v^{(m)}}\left(\mu^{pred}_{W_v^{(m)}}\right)^{\mathsf{T}} + \sum_{i=1}^{n} H_{i,v}\Lambda_{\mathbf{X}_i}\left(\mathbf{A}_{i,v}^{(m)}\right)^{\mathsf{T}}\right] \tag{6.11}$$

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

$$\Lambda^{post}_{\mathbf{H}_i^{[l]}} = \Lambda^{pred}_{\mathbf{H}_i^{[l]}} + \mathbf{W}_i^{\mathsf{T}} \Lambda_{\mathbf{X}_i} \mathbf{W}_i \tag{6.12}$$

$$\mu^{post}_{\mathbf{H}_i^{[l]}} = \left[\Lambda^{post}_{\mathbf{H}_i^{[l]}}\right]^{-1} \left[\Lambda^{pred}_{\mathbf{H}_i^{[l]}} \mu^{pred}_{\mathbf{H}_i^{[l]}} + \mathbf{W}_i^{\mathsf{T}} \Lambda_{\mathbf{X}_i} \mathbf{X}_i^{[l]}\right] \tag{6.13}$$

6.1.3 Subspace Dimensionality and Complexity Analysis

Let the number of rows in $\mathbf{X}_1, \ldots, \mathbf{X}_n$ 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 $\sum_{v \in S(n,i)} K_v = R_i$. Determining the value of K_v is a model selection problem and depends upon the common features among various data sources. Similar to MS-NMF, we use a rule of thumb as $K_v \approx \sqrt{M_v/2}$ where M_v is the number of common features in sharing configuration v. Following the above notation for the dimensionalities of matrices, for each $i \in \{1, 2, \ldots n\}$, assuming $R_i < M$ (generally the case for real world data), the complexity of sampling $\mathbf{H}_i$ matrices from its posterior distributions is $\mathcal{O}\left(M \times N_i \times R_i\right)$ whereas the complexity involved in sampling $\mathbf{W}_i$ matrices from its posterior distributions is $\mathcal{O}\left(M \times N_i \times R_i^2\right)$. Thus the computation complexity of BSSL remains similar to BPMF model [Salakhutdinov and Mnih, 2008] and does not grow any further.

6.2 Social Media Applications

We demonstrate the usefulness of BSSL in two real world applications. (1) Improving social media retrieval in target medium with the help of multiple related auxiliary social media sources. (2) Retrieving items across multiple social media sources. 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.

6.2.1 Improving Social Media Retrieval using Auxiliary Sources

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 6.2 for BSSL based retrieval.

6.2.2 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 6.2.1, we construct a vector q of length M using a set of query keywords S_Q. We proceed similar to Algorithm 6.2 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 6.2 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 $\mathcal{I} = \left\{I_1, I_2, ......., I_{N_j}\right\}$, number of items to be retrieved $N$.
2: Get Gibbs estimates of $\mathbf{W}_j$ and $\mathbf{H}_j$ using Algorithm 6.1.
3: Project q onto the subspace spanned by $\mathbf{W}_j$ to get $h$ as $h = \mathbf{W}_j^{\dagger} q$ where † 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 $h$ : $\text{sim}\left(h, h_m\right) = \frac{h^{\intercal} h_m}{\|h\|_2 \|h_m\|_2}$
5: Output: Return the top $N$ items in decreasing order of similarity.
```

6.3 Experiments

6.3.1 Data Set

For social media retrieval, we conduct our experiments on the same social media data set (created from Blogspot, Flickr and YouTube websites) that was utilized for demonstrating

MS-NMF in chapter 4 and is already described in section 4.3.1.

6.3.2 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 (6.10)–(6.13) 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 Mardia et al. [1979] 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 $\mu_0 = [0, \ldots, 0]^{\mathsf{T}}$, $s_0 = 1$, $\Delta_0 = \mathbf{I}$ 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 data set described 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.

6.3.3 Experiment 1 : Improving Social Media Retrieval using Auxiliary Sources

To carry out our experiments, we choose YouTube as the target data set and Blogspot and Flickr as auxiliary data sets. To perform BSSL based retrieval from YouTube, we first generate samples of basis matrix $\mathbf{W}_1 \triangleq [W_1 \mid W_{12} \mid W_{13} \mid W_{123}]$ and representation matrix $\mathbf{H}_1 \triangleq [H_{1,1} \mid H_{1,12} \mid H_{1,13} \mid H_{1,123}]$ according to Eqs (6.10)–(6.13) and then get an estimate of $\mathbf{W}_1$ and $\mathbf{H}_1$ following Algorithm 6.2.

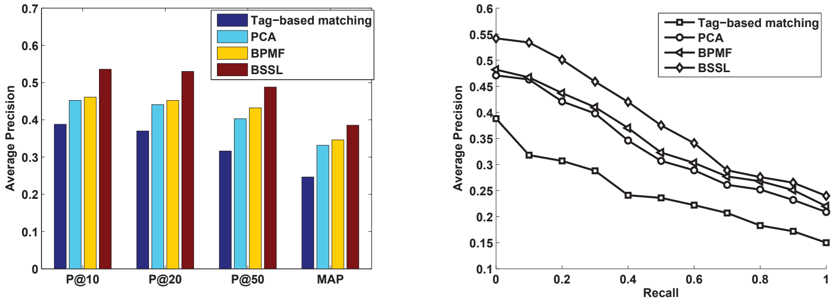
To compare the performance with other methods, we choose *three* baselines. The *first* baseline performs retrieval by matching the tag-lists of videos without any subspace learning. To get the similarity with other items, Jaccard coefficent[1] is used and the results are ranked based on the similarity scores. For the other two baselines, we have chosen models that allow modeling of mixed-signed data for having a comparison based on common setting. The *second* baseline is retrieval 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 Salakhutdinov and Mnih [2008] 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 datasize so as to

[1] Jaccard $(A, B) = |A \cap B| \,/\, |A \cup B|$.

keep the total datasize 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 $\mathbb{Q} = \{$'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 6.2 compares the retrieval performance of BSSL with all the baselines in terms of *precision-scope (P@N)* curve[2], *mean average precision (MAP)* and *11-point precision-recall curve* Baeza-Yates and Ribeiro-Neto [1999]. Figure 6.2 clearly shows that BSSL outperforms the baselines in terms of all three evaluation criteria. Although, BPMF 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.



(a)

(b)

Figure 6.2: 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 [Salakhutdinov and Mnih, 2008] (baseline 3) and the proposed BSSL.

6.3.4 Experiment 2 : Cross-Social Media Retrieval

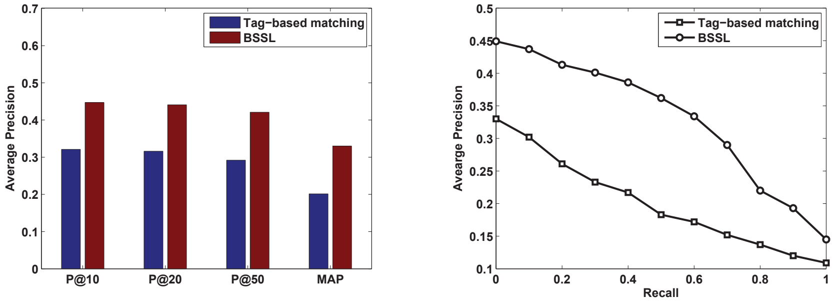
The effectiveness of BSSL for cross-media retrieval is demonstrated using the YouTube-Flickr-Blogspot data set with the subspace learning parameters as described in subsection 6.3.2. For evaluation, we again use *precision-scope (P@N), mean average precision (MAP)*

[2] For example, P@10 is the retrieval precision when considering the top 10 retrieved items.

and *11-point interpolated precision-recall curves.* The precision and recall measures for cross-media retrieval are already defined in Eq (4.16) of chapter 4.

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 [Yang et al., 2009, Yi et al., 2008] 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 6.3 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 6.3.3. It can be seen from Figure 6.3 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.



(a) (b)

Figure 6.3: Cross-media retrieval results: (a) precision-scope, MAP (b) 11-point interpolated precision-recall; for tag-based matching and the proposed BSSL.

6.4 Discussion

We have presented a Bayesian framework to learn individual and shared subspaces from multiple data sources 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) [Salakhutdinov and Mnih, 2008], provides an efficient algorithm to learn the subspace. Inference based on Rao-Blackwellized 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 data sets. More importantly, our solution provides a generic framework to exploit collective strengths from heterogeneous sources with mixed-sign data and useful for various applications involving cross-domain data mining and beyond. Being a probabilistic model, it can be plugged to build more complex hierarchical and nested models.

So far in this thesis, we have developed various shared subspace learning models and all of these models require to specify the shared and individual subspace dimensionalities *a priori*. We have learned these parameters separately either following some heuristics or through a rigorous model selection procedure. Under these circumstances, it is highly desirable to embed the learning of these parameters within the shared subspace modeling framework. The model proposed in this chapter, being an instance of hierarchical Bayesian models, can be extended to learn these dimensionalities automatically using the theory of nonparametric Bayesian priors. We address this extension in the next chapter.

Chapter 7

Nonparametric Bayesian Shared Subspace Learning

In chapter 6, we described a hierarchical Bayesian model that can jointly model related data from multiple sources. This model is generic in the sense that it can handle modeling of both nonnegative and mixed-sign data sources. In addition, being a probabilistic model, it can handle uncertainties of the real-world data. However, performance of this approach depends on the subspace dimensionalities and the optimal sharing needs to be specified *a priori*. Learning subspace dimensionalities involves a separate model selection, which is difficult and time consuming. To address this issue, in this chapter, we propose a Bayesian *nonparametric* shared subspace framework for jointly modeling multiple related data sources. Our model utilizes the hierarchical beta process [Thibaux and Jordan, 2007] (HBP) as a nonparametric prior to automatically infer the dimensionality of shared and individual subspaces. For posterior inference, it is customary to integrate out the random prior distribution to avoid dealing with its infinite support. However, it induces coupling (or dependency) across each source. Therefore, to avoid coupling, we represent the random prior distribution explicitly and use auxiliary variable sampling for inference. The inference based on Gibbs sampling also helps in sampling hyperparameters of the model.

The proposed nonparametric shared subspace learning is developed by extending the beta process factor analysis (BPFA) model proposed in Paisley and Carin [2009] wherein the data matrix is decomposed as a product of two matrices : the factors and their features. The feature matrix is further decomposed into an elementwise product of a binary matrix (say $\mathbf{Z}$) (indicating absence or presence of a feature) and a weight matrix (feature values). The binary matrix $\mathbf{Z}$ is modeled using a Bernoulli process parametrized by a beta process. Our model allows sharing of factors across different sources and learns the number of shared and individual factors automatically from data. We model each data source analogously to Paisley and Carin [2009], but in this case, the binary matrix for each source (say $\mathbf{Z}_j$ for source j) is modeled using a Bernoulli process parametrized by a hierarchical beta process. This allows flexible representation of both the shared and individual factors along with the corresponding features. Unlike previous works [Thibaux and Jordan, 2007, Paisley and Carin, 2009], a Gibbs sampling based inference is presented, which, in addition to sampling the main variables, provides an elegant way to sample the hyperparameters of the hierarchical beta process. Sampling hyperparameters is important as these parameters control the extent of sharing across data sources. We demonstrate the usefulness of the proposed model through its application to two different tasks – transfer learning using text data and image retrieval. The experiments using NIPS 0-12 data set validates the effectiveness of our model for the transfer learning task over NIPS sections. For image retrieval, the proposed method outperforms recent state-of-the-art techniques on the NUS-WIDE animal data set [Chua et al., 2009].

This chapter is organized as follows. Section 7.1 presents a nonparametric model for learning shared and individual subspaces from multiple data sources. Section 7.2 provides model inference based on collapsed Gibbs sampling using auxiliary variable scheme. Section 7.3 describes the experiments conducted on both the synthetic data set and the real world data sets. Finally, a discussion is provided in section 7.4.

7.1 Joint Factor Model using HBP Prior

7.1.1 Nonparametric Joint Factor Analysis (NJFA)

Factor analysis attempts to model a data matrix $\mathbf{X}$ as the product of two matrices Φ and $\mathbf{H}$ plus an error matrix $\mathbf{E}$. The matrix Φ contains the factors, which can also be viewed as basis vectors that spans a subspace and the matrix $\mathbf{H}$ contains the factor loadings or the representation of the data matrix in the subspace. Formally,

$$\mathbf{X} = \Phi\mathbf{H} + \mathbf{E} \tag{7.1}$$

Departing from this setting, we propose a joint factor analysis model, whose goal is to model multiple data matrices $\mathbf{X}_1, ..., \mathbf{X}_J$ using a factor matrix $\Phi = [\phi_1, ..., \phi_K]$ where $\phi \in \mathbb{R}^K$. Some of the factors in Φ may be shared amongst the various data sources whereas other factors would be specific to individual ones. For each $j = 1, \ldots, J$, $\mathbf{X}_j \in \mathbb{R}^{M \times N_j}$ has a representation $\mathbf{H}_j \in \mathbb{R}^{K \times N_j}$ in the subspace spanned by Φ along with the factorization error $\mathbf{E}_j$. We represent our joint factor analysis model as a set of factor models which shall be jointly inferred

$$\Pi : \begin{cases} \mathbf{X}_1 = & \Phi\mathbf{H}_1 + \mathbf{E}_1 \\ \vdots & \vdots \\ \mathbf{X}_J = & \Phi\mathbf{H}_J + \mathbf{E}_J \end{cases} \tag{7.2}$$

For this model, we allow the number of factors (K) to grow as large as needed when more data is observed. When the number of factors are infinitely large, each data point may be using a few factors out of the infinitely large pool and therefore, the representation of a data point is usually sparse. Due to this sparseness, we represent the matrix $\mathbf{H}_j$ as the element-wise multiplication of two matrices $\mathbf{Z}_j$ and $\mathbf{W}_j$, i.e. $\mathbf{H}_j = \mathbf{Z}_j \odot \mathbf{W}_j$, where $\mathbf{Z}_{ji}^k = 1$ implies the presence of factor ϕ_k for i-th data point $\mathbf{X}_{ji}$ from source j and $\mathbf{W}_{ji}^k$ represents the coefficient or weight of the factor ϕ_k.

We use HBP prior [Thibaux and Jordan, 2007] on the collection of matrices $\mathbf{Z}_1, \ldots, \mathbf{Z}_J$. In particular, we model each column of $\mathbf{Z}_j$ as a draw from a Bernoulli process parametrized by a beta process A_j. Since our goal is to share some of the factors ϕ_k's across more than one data source, we require A_j's to have a common support. To do so, we tie A_j's together via a beta process B so that $A_j \sim \text{BP}(\alpha_j, B)$ where base measure B itself is a draw from another beta process, i.e. $B \sim \text{BP}(\gamma_0, B_0)$ (see Figure 7.1a). This gives rise to a HBP prior on $\mathbf{Z}_j$ which ensures that some of the factors are shared across multiple sources while other factors remain specific to a source. This property avoids negative knowledge transfer and is important for transfer learning. The measure of the whole parameter space is denoted as τ_0 (i.e. $B_0(\Omega) = \tau_0$).

Another way of jointly modeling these data sources is by adapting the single source beta process factor analysis [Paisley and Carin, 2009] and considering $[\mathbf{X}_1, \ldots, \mathbf{X}_J] = \Phi[\mathbf{H}_1, \ldots, \mathbf{H}_J] + [\mathbf{E}_1, \ldots, \mathbf{E}_J]$. However, this method does not distinguish among the data from different groups. Through our experiments, we empirically show the superiority of the proposed hierarchical model over this augmented model.

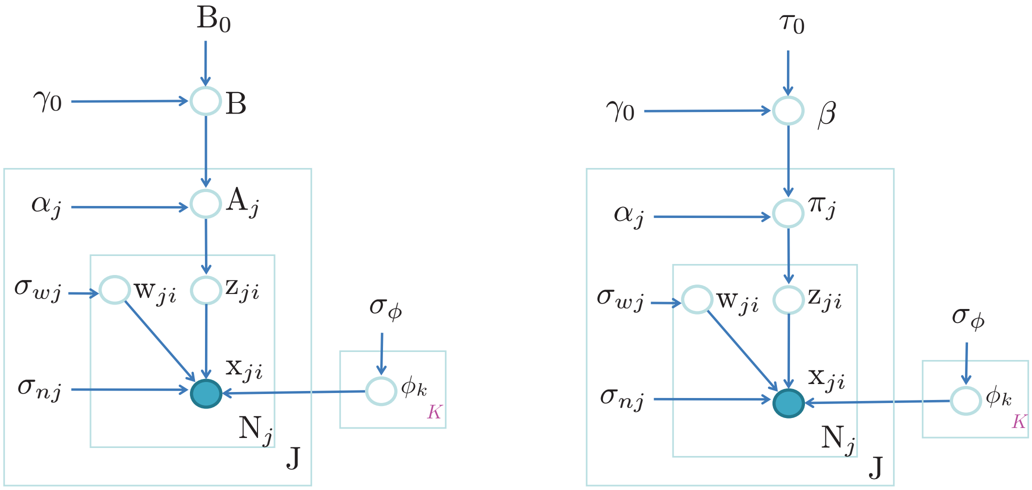
The other matrices such as factors Φ, features $\mathbf{W}_j$ and errors $\mathbf{E}_j$ are assumed to be i.i.d. and normal distributed. Using the representation for beta process [Thibaux and Jordan, 2007], the whole model can be described as (refer section 2.4.4)

$$\beta_k \sim \text{beta}(\gamma_0 b_k, \gamma_0(1 - b_k)), \qquad b_k \triangleq B_0(d\phi_k), \quad \beta_k \triangleq B(d\phi_k) \tag{7.3}$$

$$\pi_{jk} \sim \text{beta}(\alpha_0\beta_k, \alpha_0(1 - \beta_k)), \qquad \phi_k \sim \mathcal{N}\left(\mathbf{0}, \sigma_\phi^2\mathbf{I}\right) \tag{7.4}$$

$$\mathbf{Z}_{ji} \sim \text{BeP}(\pi_j), \qquad \mathbf{W}_{ji} \sim \mathcal{N}\left(\mathbf{0}, \sigma_{wj}^2\mathbf{I}\right) \tag{7.5}$$

$$\mathbf{X}_{ji} \mid \Phi, \mathbf{Z}_{ji}, \mathbf{W}_{ji} \sim \mathcal{N}\left(\Phi(\mathbf{Z}_{ji} \odot \mathbf{W}_{ji}), \sigma_{nj}^2\mathbf{I}\right) \tag{7.6}$$



(a) (b)

Figure 7.1: (a) Graphical representation of NJFA model (stochastic process view) (b) An alternative view.

We note from Eq (7.3) that the distribution for β_k is improper [Thibaux and Jordan, 2007] in case of continuous B_0, as in this case, the measure associated to a single atom ϕ_k is equal to zero. Nonetheless, instead of using a single atom ϕ_k, we can consider infinitesimally small region $d\phi_k$ in Ω around ϕ_k. [Hjort, 1990] has shown that increments of such infinitesimal form for the beta process are independent and approximately follow beta distribution. Notationwise, we use $\beta = \{\beta_k : k = 1, \ldots, K\}$ and $\pi = \{\pi_{jk} : j = 1, \ldots, J \text{ and } k = 1, \ldots, K\}$.

7.2 Model Inference

For inference, we use collapsed Gibbs sampling. It can be seen from the graphical representation in Figure 7.1b that main variables of interest for NJFA model are $\mathbf{Z}$, π, Φ and β. We integrate out π and Gibbs sample only $\mathbf{Z}$, Φ and β. In addition to sampling main variables, we also sample the hyperparameters α_j's and γ_0. The factors ϕ_k's are drawn i.i.d. from B_0 and since B_0 is assumed to be a normal distribution (see Eq (7.4)), $B_0(\Omega) = \tau_0 = 1$. For each ϕ_k, we also sample b_k from its posterior. Notationwise, a superscript attached to a symbol following a '-' sign, e.g. $\mathbf{m}^{-jk}$, $\mathbf{Z}^{-ji}$, $\mathbf{Z}_{ji}^{-k}$ etc, means a set of variables excluding the variable indexed by the superscript.

7.2.1 Collapsed Gibbs Sampling

Sampling β Before proposing an update for β, we integrate out π and write the joint distribution of $\mathbf{Z}$ and β as

$$p\left(\mathbf{Z},\beta\right)=\prod_{k=1}^{K}p\left(\beta_k\right)\prod_{j=1}^{J}\frac{\Gamma\left(\alpha_j\right)}{\Gamma\left(N_j+\alpha_j\right)}\frac{\Gamma\left(\alpha_j\beta_k+n_{jk}\right)}{\Gamma\left(\alpha_j\beta_k\right)}\frac{\Gamma\left(\alpha_j-\alpha_j\beta_k+N_j-n_{jk}\right)}{\Gamma\left(\alpha_j-\alpha_j\beta_k\right)} \tag{7.7}$$

We use auxiliary variable method [Damien et al., 1999, Teh et al., 2006] for sampling from posterior of β. Likelihood expression of Eq. (7.7) involves β in the argument of gamma functions and such ratios of gamma functions lead to independent polynomials in $\alpha_j\beta_k$ and $\alpha_j\left(1-\beta_k\right)$. In particular,

$$\frac{\Gamma\left(\alpha_j\beta_k+n_{jk}\right)}{\Gamma\left(\alpha_j\beta_k\right)} = \sum_{m_{jk}=0}^{n_{jk}}s\left(n_{jk},m_{jk}\right)\left(\alpha_j\beta_k\right)^{m_{jk}} \text{ and} \tag{7.8}$$

$$\frac{\Gamma\left(\alpha_j-\alpha_j\beta_k+N_j-n_{jk}\right)}{\Gamma\left(\alpha_j-\alpha_j\beta_k\right)} = \sum_{l_{jk}=0}^{N_j-n_{jk}}s\left(N_j-n_{jk},l_{jk}\right)\left(\alpha_j-\alpha_j\beta_k\right)^{l_{jk}} \tag{7.9}$$

where $s\left(n,m\right)$ are the unsigned Stirling numbers of the first kind. Next, we define $\mathbf{m}=\left(m_{jk}:\forall j,k\right)$ and $\mathbf{l}=\left(l_{jk}:\forall j,k\right)$ and treat them as auxiliary variables to the model. Using Eqs. (7.7), (7.9) and the prior on β, we can write the joint distribution over $\mathbf{Z}$, $\mathbf{m}$, $\mathbf{l}$ and β as the following

$$p\left(\mathbf{Z},\mathbf{m},\mathbf{l},\beta\right)=\prod_{k=1}^{K}\left(\frac{\Gamma\left(\gamma_0\right)}{\Gamma\left(\gamma_0 b_k\right)\Gamma\left(\gamma_0\left(1-b_k\right)\right)}\right)\beta_k^{\gamma_0 b_k}\left(1-\beta_k\right)^{\gamma_0\left(1-b_k\right)}\times$$
$$\prod_{j=1}^{J}\left(\frac{\Gamma\left(\alpha_j\right)}{\Gamma\left(N_j+\alpha_j\right)}\right)s\left(n_{jk},m_{jk}\right)\left(\alpha_j\beta_k\right)^{m_{jk}}s\left(N_j-n_{jk},l_{jk}\right)\left(\alpha_j\left(1-\beta_k\right)\right)^{l_{jk}} \tag{7.10}$$

To sample from above joint distribution given $\mathbf{Z}$, we iterate sampling between $\mathbf{m}$, $\mathbf{l}$ and β variables. The desired conditional distributions are given by

$$p\left(m_{jk}\mid\mathbf{m}^{-jk},\mathbf{l},\beta,\mathbf{Z}\right) \propto s\left(n_{jk},m_{jk}\right)\left(\alpha_j\beta_k\right)^{m_{jk}} \tag{7.11}$$

$$p\left(l_{jk}\mid\mathbf{l}^{-jk},\mathbf{m},\beta,\mathbf{Z}\right) \propto s\left(N_j-n_{jk},l_{jk}\right)\left(\alpha_j-\alpha_j\beta_k\right)^{l_{jk}} \tag{7.12}$$

$$p\left(\beta_k\mid\mathbf{m},\mathbf{l},\mathbf{Z}\right) \propto \text{beta}\left(\sum_j m_{jk}+\gamma_0 b_k,\sum_j l_{jk}+\gamma_0\left(1-b_k\right)\right) \tag{7.13}$$

Updates of b_k are provided in a later section where we describe the sampling scheme for the set of hyperparameters.

Sampling Z Let us assume that the whole space Ω is partitioned in L equal parts such that $\cup_{k=1}^{L} d\phi_k = \Omega$, where L should be a large number to obtain a good sampling approximation in Eq (7.3). The number of active factors used by data across all J sources is K. Since the random measure B is a draw from BP (γ_0, B_0), it can be written as

$$B = \sum_{k=1}^{L} \beta_k \delta_{d\phi_k} = \sum_{k=1}^{K} \beta_k \delta_{d\phi_k} + \sum_{k=K+1}^{L} \beta_k \delta_{d\phi_k} = \sum_{k=1}^{K} \beta_k \delta_{d\phi_k} + B^u$$

where $B^u \triangleq \sum_{k=K+1}^{L} \beta_k \delta_{d\phi_k} = B\left(\Omega \backslash \{d\phi_1, \ldots, d\phi_K\}\right)$. Since, at any time, we have only K atoms (or factors) which are active, B^u is approximated by its expected value $\mathbb{E}\left[B\left(\Omega \backslash \{d\phi_1, \ldots, d\phi_K\}\right)\right]$ which is equal to $1 - \sum_{k=1}^{K} b_k$ where b_k is defined in Eq (7.3). Given B and $\mathbf{Z}^{-ji}$, we can use the conjugacy property of beta process and sample $\mathbf{Z}_{ji}$ as the following

$$\mathbf{Z}_{ji} \mid \mathbf{Z}^{-ji}, B \sim \text{BeP}\left(\frac{\alpha_j}{\alpha_j + N_j - 1} \sum_{k=1}^{K} \beta_k \delta_{\phi_k} + \sum_{l=1}^{K_j} \frac{n_{j,k_{jl}}^{-i}}{\alpha_j + N_j - 1} \delta_{\phi_{k_{jl}}} + \frac{\alpha_j}{\alpha_j + N_j - 1} B^u\right) \tag{7.14}$$

where $n_{j,k}^{-i} \triangleq \sum_{i' \neq i} \mathbf{Z}_{ji'}^{k}$ and k_{jl} relates a factor indexed locally in source j as l and globally as k. Similar to Indian buffet process (IBP) culinary analogy [Griffiths and Ghahramani, 2006], in hierarchical beta process, i-th customer entering in j-th restaurant chooses a dish (say k-th w.r.t. global index and l w.r.t. local index) already being served in j-th restaurant with probability $\frac{n_{j,k_{jl}}^{-i} + \alpha_j \beta_k}{\alpha_j + N_j - 1}$ and a dish which is not being served in j-th restaurant but available in other restaurants (say q-th w.r.t. global index) with probability $\frac{\alpha_j \beta_q}{\alpha_j + N_j - 1}$. After choosing these dishes, i-th customer entering in j-th restaurant chooses Poi $\left(\frac{\alpha_j}{\alpha_j + N_j - 1} B^u\right)$ number of new dishes where Poi $(.)$ denotes Poisson distribution. Proceeding further, we can write the Gibbs sampling update equations of $\mathbf{Z}_{ji}^{k}$ for $k = 1, \ldots K$ as

$$p\left(\mathbf{Z}_{ji}^{k} \mid \mathbf{Z}_{ji}^{-k}, \mathbf{W}_{ji}, \beta_k, \Phi, x_{ji}\right) \propto p\left(\mathbf{Z}_{ji}^{k} \mid \mathbf{Z}_{ji}^{-k}, \beta_k\right) p\left(\mathbf{X}_{ji} \mid \mathbf{Z}_{ji}^{-k}, \mathbf{W}_{ji}, \Phi, \sigma_{jn}^{2}\right) \tag{7.15}$$

If F_{ji} be the number of new factors chosen by i-th data point from j-th source then from Eq. (7.14), prior on F_{ji} can be written as $F_{ji} \mid \alpha_j, B \sim$ Poi $\left(\frac{\alpha_j B^u}{\alpha_j + N_j - 1}\right)$. Given this prior, the posterior distribution of F_{ji} is given as

$$\begin{aligned} p\left(F_{ji} \mid \text{rest}\right) \quad \propto \quad & \text{Poi}\left(\frac{\alpha_j B^u}{\alpha_j + N_j - 1}\right) \times \\ & \int p\left(\mathbf{X}_{ji} \mid \mathbf{Z}_{ji}, \mathbf{Z}_{ji}^{new}, \Phi, \Phi^{new}, \mathbf{W}_{ji}, \mathbf{W}_{ji}^{new}, \sigma_{jn}^{2}\right) dB_0\left(\Phi^{new}\right) \end{aligned} \tag{7.16}$$

where Φ^{new} is a matrix which accommodates the new factors and $\mathbf{Z}_{ji}^{new}$ is the corresponding indicator vector for which all the elements are equal to one. For each value of F_{ji}, given $\mathbf{W}_{ji}^{new}$, the integral in above equation can be solved in closed form and is given by

$$\int p\left(\mathbf{X}_{ji} \mid \mathbf{Z}_{ji}, \mathbf{Z}_{ji}^{new}, \Phi, \Phi^{new}, \mathbf{W}_{ji}, \mathbf{W}_{ji}^{new}, \sigma_{jn}^{2}\right) dB_0\left(\Phi^{new}\right) = \frac{\left[\det\left(S_F\right)\right]^{M/2}}{\sigma_{\phi}^{M\times F_{ji}}}\times$$
$$\exp\left(\frac{1}{2}\mathrm{Tr}\left[\mu_F S_F^{-1}\mu_F^{\mathsf{T}}\right]\right)\times\mathcal{N}\left(\mathbf{X}_{ji} \mid \Phi\left(\mathbf{Z}_{ji}\odot\mathbf{W}_{ji}\right), \sigma_{nj}^{2}\mathbf{I}\right)$$

where $\mu_F = \left(\mathbf{X}_{ji} - \Phi\left(\mathbf{Z}_{ji}\odot\mathbf{W}_{ji}\right)\right)\left(\boldsymbol{Y}_{ji}^{new}\right)^{\mathsf{T}} S_F$, $S_F^{-1} = \frac{\boldsymbol{Y}_{ji}^{new}\left(\boldsymbol{Y}_{ji}^{new}\right)^{\mathsf{T}}}{\sigma_{nj}^{2}} + \frac{\mathbf{I}_{F_{ji}}}{\sigma_{\phi}^{2}}$ and $\boldsymbol{Y}_{ji}^{new} \triangleq \mathbf{Z}_{ji}^{new}\odot\mathbf{W}_{ji}^{new}$. The symbol $\mathcal{N}(.)$ is used to denote the normal probability distribution function. In our implementation, for Eq (7.16), we truncate the support of Poi (.) at five times its mean value. The samples of β_k for $k > K$ and $\mathbf{W}_{ji}$ can be obtained using their priors.

Sampling Φ and W Gibbs sampling update for Φ conditioned on data $\mathbf{X}_1, \ldots, \mathbf{X}_J$ and latent indicator variables $\mathbf{Z}$ is given by

$$p\left(\Phi \mid \mathbf{Z}_{1:J}, \mathbf{W}_{1:J}, \mathbf{X}_{1:J}\right) \propto p\left(\Phi \mid B_0, \sigma_{\Phi}^{2}\right) p\left(\mathbf{X}_{1:J} \mid \Phi, \mathbf{Z}_{1:J}, \mathbf{W}_{1:J}\right) \tag{7.17}$$

Taking logarithm on both sides, we get

$$\begin{aligned}
-\log p\left(\Phi \mid \mathbf{Z}_{1:J}, \mathbf{W}_{1:J}, \mathbf{X}_{1:J}\right) &= \text{constant} + \frac{1}{2\sigma_{\phi}^{2}}\mathrm{Tr}\left[\Phi\Phi^{\mathsf{T}}\right] + \\
&\quad \sum_{j}\frac{1}{2\sigma_{nj}^{2}}\mathrm{Tr}\left[\left(\mathbf{X}_j - \Phi\left(\mathbf{Z}_j\odot\mathbf{W}_j\right)\right)\left(\mathbf{X}_j - \Phi\left(\mathbf{Z}_j\odot\mathbf{W}_j\right)\right)^{\mathsf{T}}\right] \\
&= \text{constant} + \frac{1}{2}\mathrm{Tr}\left[S_{\Phi}^{-1}\left(\Phi-\mu_{\Phi}\right)^{\mathsf{T}}\left(\Phi-\mu_{\Phi}\right)\right] \\
\Rightarrow p\left(\Phi \mid \mathbf{Z}_{1:J}, \mathbf{W}_{1:J}, \mathbf{X}_{1:J}\right) &= \mathcal{N}\left(\Phi \mid \mu_{\Phi}, S_{\Phi}\right)
\end{aligned}$$

where $\mu_{\Phi} \triangleq \left(\sum_j \frac{\mathbf{X}_j\left(\mathbf{Z}_j\odot\mathbf{W}_j\right)^{\mathsf{T}}}{\sigma_{nj}^{2}}\right) S_{\Phi}$ and $S_{\Phi}^{-1} \triangleq \sum_j\left(\frac{\left(\mathbf{Z}_j\odot\mathbf{W}_j\right)\left(\mathbf{Z}_j\odot\mathbf{W}_j\right)^{\mathsf{T}}}{\sigma_{nj}^{2}}\right) + \frac{\mathbf{I}_K}{\sigma_{\phi}^{2}}$.
Similarly, Gibbs sampling update for $\mathbf{W}_{ji}$ conditioned on $\mathbf{X}_{ji}$, $\mathbf{Z}_{ji}$ and Φ is given by

$$p\left(\mathbf{W}_{ji} \mid \mathbf{X}_{ji}, \mathbf{Z}_{ji}, \Phi\right) \propto p\left(\mathbf{W}_{ji} \mid \sigma_{jw}^{2}\right) p\left(\mathbf{X}_{ji} \mid \Phi, \mathbf{Z}_{ji}, \mathbf{W}_{ji}\right) \tag{7.18}$$

Again, under the assumed model distributions, this reduces to a normal distribution, i.e.

$$p\left(\mathbf{W}_{ji} \mid \mathbf{X}_{ji}, \mathbf{Z}_{ji}, \Phi\right) = \mathcal{N}\left(\mathbf{W}_{ji} \mid \mu_{w_{ji}}, S_{w_{ji}}\right) \tag{7.19}$$

where $\mu_{w_{ji}} \triangleq \frac{S_{w_{ji}} D_{\mathbf{Z}_{ji}}\Phi^{\mathsf{T}}\mathbf{X}_{ji}}{\sigma_{nj}^{2}}$, $S_{w_{ji}}^{-1} \triangleq \frac{D_{\mathbf{Z}_{ji}}\Phi^{\mathsf{T}}\Phi D_{\mathbf{Z}_{ji}}}{\sigma_{nj}^{2}} + \frac{\mathbf{I}_K}{\sigma_{\phi}^{2}}$ and $D_{\mathbf{Z}_{ji}} \triangleq \mathrm{diag}\left(\mathbf{Z}_{ji}\right)$, a diagonal matrix constructed by keeping $\mathbf{Z}_{ji}$ on its diagonal.

Integrating out β To be able to sample for hyperparameters γ_0 and τ_0, we need to integrate out β from the joint distribution expression of Eq (7.24). The required marginal after integrating out β is given as the following (note the emphasis on the hyperparameters)

$$p\left(\mathbf{Z},\mathbf{m},\mathbf{l},\gamma_0,\tau_0,\alpha_{1:J}\right)=\left(\prod_{k=1}^{K}\prod_{j=1}^{J}\frac{\Gamma\left(\alpha_j\right)}{\Gamma\left(N_j+\alpha_j\right)}\alpha_j^{m_{jk}+l_{jk}}s\left(n_{jk},m_{jk}\right)s\left(N_j-n_{jk},l_{jk}\right)\right)\times$$
$$\prod_{k=1}^{K}\left(\frac{\Gamma\left(\gamma_0\right)}{\Gamma\left(\gamma_0+m_k+l_k\right)}\times\frac{\Gamma\left(m_k+\gamma_0 b_k\right)}{\Gamma\left(\gamma_0 b_k\right)}\times\frac{\Gamma\left(l_k+\gamma_0\left(1-b_k\right)\right)}{\Gamma\left(\gamma_0\left(1-b_k\right)\right)}\right)$$

where we define $m_k\triangleq\sum_j m_{jk}$ and $l_k\triangleq\sum_j l_{jk}$. Expanding second and third terms of gamma ratios in the fashion similar to Eq (7.9), and introducing further auxiliary variables $\mathbf{r}=\left(r_k:k=1,\ldots K\right)$ and $\mathbf{r}'=\left(r'_k:k=1,\ldots K\right)$ (where $r_k\in\{0,m_k\}$ and $r'_k\in\{0,l_k\}$) to above joint distribution, we obtain

$$p\left(\mathbf{Z},\mathbf{m},\mathbf{l},\mathbf{r},\mathbf{r}',\gamma_0,\tau_0,\alpha_{1:J}\right)=\prod_{k=1}^{K}\prod_{j=1}^{J}\frac{\Gamma\left(\alpha_j\right)}{\Gamma\left(N_j+\alpha_j\right)}\alpha_j^{\left(m_{jk}+l_{jk}\right)}s\left(n_{jk},m_{jk}\right)s\left(N_j-n_{jk},l_{jk}\right)\times$$
$$\prod_{k=1}^{K}\left(\frac{\Gamma\left(\gamma_0\right)}{\Gamma\left(\gamma_0+m_k+l_k\right)}\times\left\{s\left(m_k,r_k\right)\left(\gamma_0 b_k\right)^{r_k}\right\}\times\left\{s\left(l_k,r'_k\right)\left(\gamma_0\left(1-b_k\right)\right)^{r'_k}\right\}\right)\quad(7.20)$$

7.2.2 Sampling Hyperparameters

Update for σ_ϕ, σ_{nj} and σ_{wj} Assuming an inverse-gamma prior on σ_ϕ^2 with shape and scale parameter a_{σ_ϕ} and b_{σ_ϕ}, the posterior for σ_ϕ^2 is given by

$$\begin{aligned}p\left(\sigma_\phi^2\mid\Phi,a_{\sigma_\phi},b_{\sigma_\phi}\right)&\propto p\left(\sigma_\phi^2\mid a_{\sigma_\phi},b_{\sigma_\phi}\right)p\left(\Phi\mid\sigma_\phi^2\right)\\&\propto\left(\sigma_\phi^2\right)^{-a_{\sigma_\phi}+1}\exp\left(-b_{\sigma_\phi}/\sigma_\phi^2\right)p\left(\Phi\mid\sigma_\phi^2\right)\\&\propto\left(\sigma_\phi^2\right)^{-a_{\sigma_\phi}+1}\exp\left(-b_{\sigma_\phi}/\sigma_\phi^2\right)\left(\sigma_\phi^2\right)^{-MK/2}\exp\left(-\frac{\mathrm{Tr}\left[\Phi\Phi^{\mathsf{T}}\right]}{2\sigma_\phi^2}\right)\\&\propto\left(\sigma_\phi^2\right)^{-\left(a_{\sigma_\phi}+MK/2\right)+1}\exp\left(-\left(b_{\sigma_\phi}+\frac{1}{2}\mathrm{Tr}\left[\Phi\Phi^{\mathsf{T}}\right]\right)/\sigma_\phi^2\right)\quad(7.21)\end{aligned}$$

which is an inverse-gamma distribution with shape and scale parameters $a_{\sigma_\phi}+MK/2$ and $b_{\sigma_\phi}+\frac{1}{2}\mathrm{Tr}\left[\Phi\Phi^{\mathsf{T}}\right]$ respectively. Sampling from posteriors of σ_{nj} and σ_{wj} is similar to the sampling of σ_ϕ and have inverse-gamma form. If $\sigma_{nj}^2\sim\mathrm{invGam}\left(\sigma_{nj}^2\mid a_{\sigma_{nj}},b_{\sigma_{nj}}\right)$, the posterior samples of σ_{nj}^2 are drawn as

$$\sigma_{nj}^2\mid\mathbf{X}_j,\mathbf{W}_j,\mathbf{Z}_j,\Phi\sim\mathrm{invGam}\left(\sigma_{nj}^2\mid a_{\sigma_{nj}}+MN_j/2,b_{\sigma_{nj}}+\frac{1}{2}\mathrm{Tr}\left[\left(\mathbf{X}_j-\Phi\mathbf{H}_j\right)\left(\mathbf{X}_j-\Phi\mathbf{H}_j\right)^{\mathsf{T}}\right]\right)\quad(7.22)$$

where $\mathbf{H}_j \triangleq \mathbf{Z}_j \odot \mathbf{W}_j$. Similarly, the posterior samples of σ_{wj}^2 are drawn as

$$\sigma_{wj}^2 \mid \mathbf{W}_j \sim \text{invGam}\left(\sigma_{wj}^2 \mid a_{\sigma_{wj}} + KN_j/2,\ b_{\sigma_{wj}} + \frac{1}{2}\text{Tr}\left[\mathbf{W}_j\mathbf{W}_j^{\mathsf{T}}\right]\right) \tag{7.23}$$

Sampling α_j We note that the joint distribution of Eq. (7.10) acts as likelihood for α_j. We assume a gamma prior on α_j with parameters a_α and b_α. Before writing the posterior directly, we re-write the likelihood expression emphasizing the hyper-parameters.

$$\begin{aligned} p\left(\mathbf{Z}, \mathbf{m}, \mathbf{l}, \beta\right) &= \prod_{k=1}^{L} p\left(\beta_k\right) \times \\ & \prod_{j=1}^{J} \frac{\Gamma\left(\alpha_j\right)}{\Gamma\left(N_j + \alpha_j\right)} s\left(n_{jk}, m_{jk}\right)\left(\alpha_j\beta_k\right)^{m_{jk}} s\left(N_j - n_{jk}, l_{jk}\right)\left(\alpha_j - \alpha_j\beta_k\right)^{l_{jk}} \end{aligned}$$

The above can be expanded and written as

$$\begin{aligned} p\left(\mathbf{Z}, \beta, \mathbf{m}, \mathbf{l}, \gamma_0, \tau_0 \mid \alpha_{1:J}\right) = \prod_{k=1}^{K}\prod_{j=1}^{J} \frac{\Gamma\left(\alpha_j\right)}{\Gamma\left(N_j + \alpha_j\right)} \alpha_j^{\left(m_{jk} + l_{jk}\right)} \times \\ \left(\prod_{k=1}^{K} p\left(\beta_k \mid \tau_0, \gamma_0\right) \prod_{j=1}^{J} s\left(n_{jk}, m_{jk}\right) \beta_k^{m_{jk}} s\left(N_j - n_{jk}, l_{jk}\right)\left(1 - \beta_k\right)^{l_{jk}}\right) \prod_{k'=K+1}^{L} p\left(\beta_{k'} \mid \tau_0, \gamma_0\right) \end{aligned} \tag{7.24}$$

Before proceeding further, we note that α_j is present in the argument of gamma functions and use an identity from Teh et al. [2006] to represent the ratio of gamma functions as

$$\frac{\Gamma\left(\alpha_j\right)}{\Gamma\left(N_j + \alpha_j\right)} = \int_0^1 u_j^{\alpha_0}\left(1 - u_j\right)^{N_j - 1}\left(1 + \frac{N_j}{\alpha_j}\right) du_j \tag{7.25}$$

Similar to Teh et al. [2006], Escobar and West [1995], we introduce further auxiliary variables $\mathbf{u} = (u_{jk} : j = 1, \ldots J \text{ and } k = 1, \ldots K)$ and $\mathbf{s} = (s_{jk} : j = 1, \ldots J \text{ and } k = 1, \ldots K)$ where $u_{jk} \in [0, 1]$ and $s_{jk} \in \{0, 1\}$. Using these auxiliary variables,

$$\begin{aligned} p\left(\mathbf{Z}, \beta, \mathbf{m}, \mathbf{l}, \mathbf{u}, \mathbf{s}, \gamma_0, \tau_0 \mid \alpha_{1:J}\right) = \left(\prod_{k=1}^{K}\prod_{j=1}^{J} u_{jk}^{\alpha_j}\left(1 - u_{jk}\right)^{N_j - 1}\left(\frac{N_j}{\alpha_j}\right)^{s_{jk}} \alpha_j^{\left(m_{jk} + l_{jk}\right)}\right) \times \\ \prod_{k=1}^{K} p\left(\beta_k \mid \tau_0, \gamma_0\right) \prod_{j=1}^{J} s\left(n_{jk}, m_{jk}\right) \beta_k^{m_{jk}} s\left(N_j - n_{jk}, l_{jk}\right)\left(1 - \beta_k\right)^{l_{jk}} \times \left(\prod_{k=K+1}^{L} p\left(\beta_k \mid \tau_0, \gamma_0\right)\right) \end{aligned}$$

Now, conditioned on auxiliary variables $\mathbf{m}$, $\mathbf{l}$, $\mathbf{u}$ and $\mathbf{s}$, we can write the posterior of α_j as

$$p\left(\alpha_j \mid \mathbf{Z}, \beta, \mathbf{m}, \mathbf{l}, \mathbf{u}, \mathbf{s}, \gamma_0, \tau_0, a_\alpha, b_\alpha\right) \propto p\left(\mathbf{Z}, \beta, \mathbf{m}, \mathbf{l}, \mathbf{u}, \mathbf{s}, \gamma_0, \tau_0 \mid \alpha_j\right) p\left(\alpha_j \mid a_\alpha, b_\alpha\right)$$

which is a gamma distribution with shape and scale parameters $a_\alpha + \sum_k \left(m_{jk} + l_{jk} - s_{jk}\right)$ and $b_\alpha - \sum_k \log u_{jk}$ respectively. The auxiliary variables u_{jk} and s_{jk} can be sampled as

$$p\left(u_{jk} \mid \alpha_j\right) \propto u_{jk}^{\alpha_j}\left(1 - u_{jk}\right)^{N_j - 1} \text{ and } p\left(s_{jk} = 1 \mid \alpha_j\right) = \frac{N_j}{\alpha_j + N_j} \tag{7.26}$$

Sampling γ_0 We note that the expression in Eq (7.20) is nothing but the likelihood of γ_0. Assuming $\gamma_0 \sim \text{Gamma}\left(\gamma_0 \mid a_\gamma, b_\gamma\right)$, the posterior of γ_0 is given as

$$\begin{aligned} p\left(\gamma_0 \mid \text{rest}\right) &\propto p\left(\gamma_0 \mid a_\gamma, b_\gamma\right) \times p\left(\mathbf{Z}, \mathbf{m}, \mathbf{l}, \mathbf{r}, \mathbf{r}', \tau_0, \alpha_{1:J} \mid \gamma_0\right) \\ &\propto \gamma_0^{a_\gamma} \exp\left(-b_\gamma \gamma_0\right) \times \prod_{k=1}^{K}\left(\frac{\Gamma\left(\gamma_0\right)}{\Gamma\left(\gamma_0 + m_k + l_k\right)} \times \left[\gamma_0 b_k\right]^{r_k} \times \left[\gamma_0\left(1 - b_k\right)\right]^{r'_k}\right) \end{aligned}$$

Expanding the term involving the ratio of the gamma functions similar to Eq (7.25) and introducing auxiliary variables $\mathbf{v} = \left(v_k : k = 1, \ldots K\right)$ and $\mathbf{c} = \left(c_k : k = 1, \ldots K\right)$ where $v_k \in [0,1]$ and $c_k \in \{0,1\}$ to the above distribution, we obtain

$$\begin{aligned} p\left(\gamma_0 \mid \mathbf{v}, \mathbf{c} \text{ and rest}\right) &\propto \gamma_0^{a_\gamma - 1} \exp\left(-b_\gamma \gamma_0\right) \times \\ &\prod_{k=1}^{K}\left(v_k^{\gamma_0}\left(1 - v_k\right)^{p_k - 1}\left(\frac{p_k}{\gamma_0}\right)^{c_k} \times \left[\gamma_0 b_k\right]^{r_k} \times \left[\gamma_0\left(1 - b_k\right)\right]^{r'_k}\right) \\ &= \text{gamma}\left(a_\gamma + \sum_k\left(r_k + r'_k - c_k\right), b_\gamma - \sum_k \log v_k\right) \end{aligned} \tag{7.27}$$

where $p_k \triangleq m_k + l_k$. The auxiliary variables $\{v_k\}$ and $\{c_k\}$ can be sampled similar to the scheme as described in Eq (7.26).

Sampling b_k Likelihood of b_k can be obtained from Eq (7.20). Assuming a beta prior on each b_k, i.e. $b_k \sim \text{beta}\left(\mu_a, \mu_b\right)$, the posterior of b_k is given as

$$\begin{aligned} p\left(b_k \mid r_k, r'_k \text{ and rest}\right) &\propto p\left(b_k \mid \mu_a, \mu_b\right) \times p\left(\mathbf{Z}, \mathbf{m}, \mathbf{l}, \mathbf{r}, \mathbf{r}', \tau_0, \gamma_0, \alpha_{1:J} \mid b_k\right) \\ &\propto b_k^{\mu_a - 1}\left(1 - b_k\right)^{\mu_b - 1} \times b_k^{r_k}\left(1 - b_k\right)^{r'_k} \\ &= \text{beta}\left(\mu_a + r_k, \mu_b + r'_k\right) \end{aligned} \tag{7.28}$$

Parameter Settings for Sampling Hyperparameters We use $\mu_a = \tau_0$ (which is equal to 1 for our model) and $\mu_b = L$ where L should be a large value such as $L \gg K$. For all our experiments, we have used $L = 1000$. For all other vague gamma priors used for γ_0, α_j, σ_{nj}, σ_{wj} and σ_ϕ, both the shape and the scale parameters are set to 1.

7.3 Experiments

Given the Gibbs samples of $\{\Phi\}_{r=1}^{R}$ and $\{\beta\}_{r=1}^{R}$ and the test data from j-th source, i.e. $\mathbf{X}_j^{new}$, we infer $\mathbf{Z}_j^{new}$ and $\mathbf{W}_j^{new}$ by iterating between Gibbs sampling updates of Eq (7.15) and (7.18). This provides us the required factorization for the test data matrix in the form : $\mathbf{X}_j^{new} = \Phi\mathbf{H}_j^{new} + \mathbf{E}_j^{new}$ where $\mathbf{H}_j^{new} \triangleq \mathbf{Z}_j^{new} \odot \mathbf{W}_j^{new}$. Monte-Carlo approximation of marginal predictive distribution $p\left(\mathbf{X}_j^{new} \mid \mathbf{X}_{1:J}\right)$ is computed by averaging the predictive distribution over $\{\Phi\}_r$ and $\{\beta\}_r$ samples. This is used to compute the perplexity for the test data.

In the following, to demonstrate our model and its typical behavior, we first perform experiments with a synthetic data set. Then, we show the application of our model for two real-world tasks. For both synthetic and real-world experiments, the priors for hyperparameters were chosen as : $\gamma_0 \sim \text{gamma}(1,1)$, $\alpha_j \sim \text{gamma}(1,1)$ and $b_k \sim \text{beta}(1,1000)$, and both the shape and the scale parameters of gamma priors for σ_ϕ, σ_{nj} and σ_{wj} were set to 1.

7.3.1 Experiments-I : Synthetic Data

For experiments with synthetic data, we create twelve factors (represented by matrix Φ) and distribute them across two sources ($J = 2$). Out of these factors, first four factors (vertical bars) are used in source-1 only, the last four factors (horizontal bars) are used in source-2 only and the remaining four factors (diagonal bars) are shared between both the sources. By distributing the factors in this way, our aim is to see whether the proposed model can discover the factors used for data generation and their correct mixings. $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are chosen as random binary matrices and used for selecting the factors. The entries of weight matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ are chosen to be normally distributed with zero mean and standard deviation equal to 0.1.

We perform the inference based on Gibbs sampling as described in section 7.2 and use only one sample after rejecting 1500 “burn-in” samples. Although the model converges around 1000 iterations but we run the model long to illustrate the point that the number of factors does not change further. Figure 7.2c shows that the model correctly learns the number of factors *automatically* from the data. In addition, it can be seen from Figure 7.2 that indeed, our model is able to learn the factors used for generating the data and their mixing

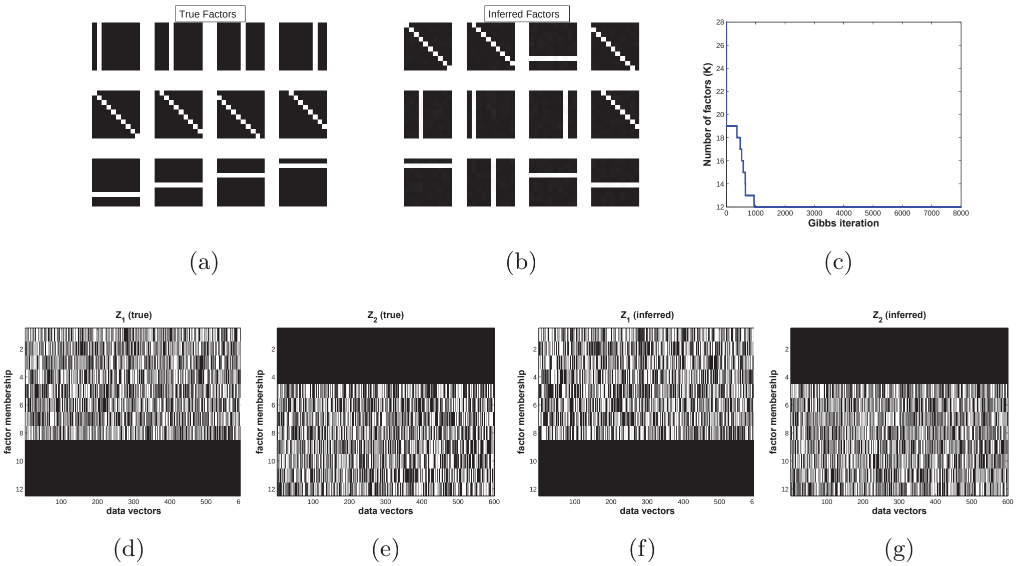


Figure 7.2: Experiments with synthetic data (a) true factors (both shared and individuals) (b) inferred factors (both shared and individuals); note the exact recovery upto a permutation (c) convergence of number of factors (d) true factor mixings of source-1 (e) true factor mixings of source-2 (f) inferred factor mixings of source-1 (g) inferred factor mixings of source-2; rows of (f) and (g) are manually permuted for easy comparison.

combinations correctly.

7.3.2 Experiments-II : Real Data

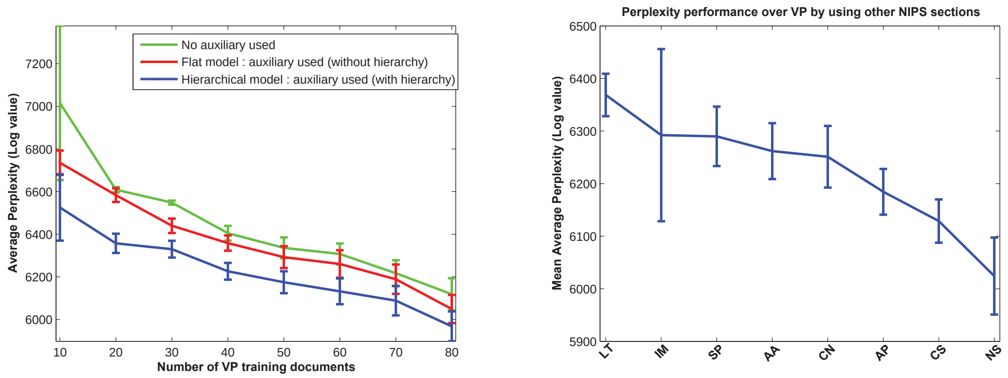
7.3.2.1 Results using NIPS 0-12 Data Set

We train our proposed NJFA model on the NIPS 0-12 data set, which contains the articles from *Neural Information Processing Systems* (NIPS) conference published between 12 years (during 1988 – 1999). In this data set[1], articles are categorized into nine main conference sections and one miscellaneous section. The main conference sections are Cognitive Science (CS), Neuroscience (NS), Learning Theory (LT), Algorithms and Architecture (AA), Implementations (IM), Speech and Signal Processing (SSP), Visual Processing (VP), Applications (AP) and Control, Navigation and Planning (CNP). We ignore articles from the miscellaneous section. Our model treats these sections as individual corpora and learns the factor matrix such as some of the factors are shared among different sections whereas other factors are specific to particular sections.

[1] We downloaded the version available at http://www.gatsby.ucl.ac.uk/~ywteh/research/data.html.

We use the above data set in a transfer learning setting and investigate whether combined learning of one section with others provides benefits. We use an experimental setting similar to Teh et al. [2006] and treat the section VP as the target and other sections as auxiliary. There are a total of 1564 articles combined across nine main sections. We select 80 articles from each section randomly and use them for training. We combine various auxiliary sections, one at a time with the target and compute the perplexity on a test set of articles from VP section. The test set contains 44 articles and is fixed throughout our experimentation.

Evaluation Measures and Baselines To evaluate the proposed method, we use perplexity per document, a widely used measure in language models (LM). Perplexity of a new document indicates the degree of surprise that a model expresses when modeling new documents. Therefore, a low value of perplexity over test set indicates better prediction of test data.



(a)

(b)

Figure 7.3: Perplexity results using NIPS 0-12 data set (a) perplexity for the test data from VP section versus the number of VP training documents (averaged over 10 runs) shown for three different models (b) mean average perplexity for the test data from VP section (averaged over 10 runs) versus other NIPS auxiliary sections using the proposed hierarchical joint factor analysis.

To compare the performance benefits, we use two baselines. The first baseline is a model that does not use any auxiliary data and relies completely on target data. This model is equivalent to the factor analysis model proposed in Paisley and Carin [2009]. The second baseline is a model which uses auxiliary data but does not use them in hierarchical fashion. Instead, it simply combines the auxiliary data with the target data as if they were from the same section. We refer to this model as the 'flat' model. This model can be thought of adaptation of Paisley and Carin [2009] for jointly modeling the data from multiple sources. Also, both of these baseline models are a special case of the proposed model. For this experiment, our Gibbs sampler typically converges in 25 iterations, however, we run the sampler for 100 iterations.

Experimental Results For each target-auxiliary pair, we average the perplexity values over 10 trials and plot them as a function of the number of target (VP section) training articles, varied from 10 to 80 with a step of 10. Figure 7.3 depicts the perplexity values for the proposed model in comparison with the baseline models. For each graph, the perplexity values shown are averaged across all the eight auxiliary sections and 10 trials. It can be clearly seen from Figure 7.3a that both the 'flat model' and the 'hierarchical model' perform better than the 'no auxiliary model', empirically proving the point that use of auxiliary data improves the performance. We can also see that the 'hierarchical model' performs significantly better than the 'flat model'. This improvement in performance is mainly due to the ability of the hierarchical model to treat each section differently whilst allowing sharing. On the contrary, the 'flat model' does not model the variations of each section separately and treats them identically, resulting in performance degradation. To evaluate the benefits obtained from each auxiliary section, Figure 7.3b plots the "mean average perplexity" values (a single point indicator obtained by averaging the perplexity values across increasing number of training documents). We see that the three sections, which provide maximum benefits are NS, CS and AP in decreasing order of performance.

7.3.2.2 Results using NUS-WIDE Data Set

Our second data set is based on the NUS-WIDE [Chua et al., 2009] data set, which is a large collection of Flickr images. We select a subset[2] comprising of 3411 images involving 13 animals[3]. This data set provides six different low-level features [Chua et al., 2009] (64-D color histogram, 144-D color correlogram, 73-D edge direction histogram, 128-D wavelet texture, 225-D block-wise color moments) and 500-D SIFT descriptions along with their ground-truths. We ignore SIFT features and use low-level features only. Our model treats each animal category as an individual group and learns the factor matrix such as some of the factors are shared among different groups whereas other factors are specific to a particular group. We use a set of 2054 images for training and the remaining for testing – an *identical* training and test settings as used in Chen et al. [2010] – so that a comparison

[2] This data set has already been used in Chen et al. [2010] and available at http://www.cs.cmu.edu/~junzhu/data.htm.

[3] In particular, 13 animals are 'squirrel', 'cow', 'cat', 'zebra', 'tiger', 'lion', 'elephant', 'whales', 'rabbit', 'snake', 'antlers', 'hawk' and 'wolf'.

Table 7.1: Comparison with the state-of-the-art image retrieval techniques using NUS-WIDE animal data set.

Method	*Mean Average Precision (MAP)*
DWH [Xing et al., 2005]	0.153
TWH [Yang et al., 2007]	0.158
MMH [Chen et al., 2010]	0.163
NJFA (proposed method)	0.1789 ± 0.0128

can be made with our work.

Using the features of test data in matrix $\mathbf{X}_j^{test}$, we sample the factorization matrices $\mathbf{W}_j^{test}$ and $\mathbf{Z}_j^{test}$ and define $\mathbf{H}_j^{test} = \mathbf{Z}_j^{test} \odot \mathbf{W}_j^{test}$. Next, to find the similar images for a test image (represented by $\mathbf{X}_{ji}^{test}$), we use *cosine similarity* between its feature $\mathbf{H}_{ji}^{test}$ and the feature matrices of the training data (i.e. $\mathbf{H}_j$ for each j). The retrieved images are ranked in decreasing order of these similarities.

Evaluation Measures and Baselines To evaluate the performance of the proposed method, we use mean average precision (MAP). We compare the result of the proposed model with the recent state-of-the-art techniques [Chen et al., 2010, Xing et al., 2005, Yang et al., 2007], categorized under multi-view learning algorithms.

Experimental Results Table 7.1 compares the proposed NJFA with above-mentioned baselines for image retrieval using the animal data set. It can be seen from the table that NJFA outperforms all three baseline models. This comparison is based on the MAP values presented in Chen et al. [2010]. We note that the data set used to generate these results (including the test set) is *identical*. The MAP results of baseline models are reported using 60 topics. Our work, being a nonparametric model, learns these dimensionalities automatically from the data and avoids any cross-validation for the model selection. We report our results averaged across 20 runs along with the standard deviation. Total number of factors (both shared and individual) for this data set varied between 5 to 8. In our experiments, the Gibbs sampler typically converges between 15-20 iterations, however, we run the sampler for 100 iterations.

7.4 Discussion

We develop a nonparametric joint factor analysis technique for modeling multiple related data sources. Our technique learns shared factors to exploit common statistical strengths and individual factors to model the variations of each source. To infer the number of shared and individual factors automatically from the data, we use hierarchical beta process (HBP) prior [Thibaux and Jordan, 2007]. The auxiliary variable Gibbs sampling provided for hierarchical beta process is general and can be utilized for other matrix factorizations. Our experiments using NIPS 0-12 data set show the usefulness of the proposed model for transfer learning applications. Automatically learning the extent of sharing across data sources avoids the possibility of negative knowledge transfer. For retrieval application, the proposed method outperforms the recent state-of-the-art methods using NUS-WIDE animal data set. This model is complete in the sense that it is probabilistic, it can model both nonnegative and mixed-sign data, it learns the subspace dimensionalities automatically and avoids learning the large number of subsapce dimensionalities, especially when applied for modeling more than two data sources. However, this model assumes an i.i.d. distribution for the basis vectors which are sampled from the base distribution of the beta process. This was mainly done for mathematical convenience as a multivariate covariance model brings the model intractable. One possible direction is the use of variational inference [Wainwright and Jordan, 2008].

Chapter 8

Concluding Remarks

8.1 Summary

This thesis has focused on developing formal frameworks to model data from multiple related sources (e.g. different social media sharing systems such as Flickr, YouTube, Blogspot) in an unsupervised setting. These have implications in unsupervised transfer learning applications where the aim is to improve task performance for a target source leveraging the knowledge present across the auxiliary sources without needing supervised data. To realize this aim, this thesis took a latent subspace approach and presented a shared subspace learning framework for jointly modeling multiple data sources.

In chapter 3, a novel shared nonnegative matrix factorization (S-NMF) is developed for jointly modeling two data sources. The joint factorization leads to a non-convex optimization function that is solved using iterative multiplicative updates. A formal mathematical proof is provided to show the convergence of the iterative algorithm. The utility of S-NMF has been demonstrated through a transfer learning task by improving tag-based image (Flickr) and video (YouTube) retrieval, leveraging information from an auxiliary source (LabelMe).

To allow modeling of multiple data sources, an extension of S-NMF (termed as MS-NMF) is developed in chapter 4. This model allows modeling multiple data sources with arbitrary sharing configurations. The optimization function for the joint factorization is again solved using iterative multiplicative updates with guaranteed convergence. Comparisons with S-NMF, have shown that using multiple auxiliary sources may further improve the retrieval performance on the target source. Additionally, MS-NMF was applied for jointly retrieving items from multiple disparate social media. This application was referred to as *cross-social media retrieval.* The effectiveness of MS-NMF for cross-social media retrieval is demonstrated using three real world representative social media sources namely Blogspot.com, Flickr and YouTube.

In many situations, both S-NMF and MS-NMF learn shared and individual subspaces that are not properly segregated causing negative knowledge transfer. To deal with negative transfer learning, a regularized extension of S-NMF (termed as RS-NMF) is developed in chapter 5. In deriving RS-NMF, the segregation among the shared and individual subspaces is ensured through a set of mutual orthogonality constraints. The use of additional constraints give rise to a modified set of iterative updates, however, we have shown that the convergence of these updates can still be guaranteed. In addition to the modeling, this chapter provides a systematic model selection procedure for learning the model parameters such as regularization parameters, subspace dimensionalities etc. The utility of this model is shown through retrieval and clustering applications using a variety of data sets. The key-point of immunity to negative transfer learning is successfully demonstrated.

Similar to the above nonnegative models, it is desirable to have a model that can deal with mixed-sign data sources. To fulfill this requirement, a hierarchical, fully Bayesian shared subspace learning (BSSL) model is developed under a linear-Gaussian framework. Similar to MS-NMF, this model allows modeling multiple data sources with arbitrary sharing configurations. For model inference, an efficient Rao-Blackwellized Gibbs sampler is derived. The effectiveness of this model is shown on social media retrieval using auxiliary sources and cross-social media retrieval. The proposed BSSL outperforms many non-sharing baselines such as PCA and Bayesian probabilistic matrix factorization (BPMF) Salakhutdinov and Mnih [2008].

In chapter 3 and 4, some heuristics were used to determine the shared subspace dimensionalities. A more systematic model selection procedure based on rank estimation was adopted in chapter 5. However, learning the model parameters through a separate model selection procedure is computationally expensive and time consuming. In chapter 7, we address this problem using the theory of Bayesian nonparametrics. Using hierarchical beta process prior, we have developed a nonparametric Bayesian model (termed as NJFA), which can learn the subspace dimensionalities automatically from the data. The model is general in that it allows modeling any number of data sources with arbitrary sharing configurations. For model inference, an efficient Gibbs sampler is derived using auxiliary variable scheme by directly sampling from the posterior of the underlying beta process. The effectiveness of the model is shown through language model perplexity improvement over non-sharing counterpart and other related models using NIPS 0-12 data set. In addition, NJFA also outperforms many recently proposed state-of-the-art techniques for content-based image retrieval using NUS-WIDE animal data set.

8.2 Future Works

There are several potential extensions from the work taken up in this thesis which have not been addressed here and are left for future investigation. We list them as the following:

- In chapter 7, we have addressed the model selection of joint matrix factorization for mixed-sign data sources. Although this framework can model nonnegative data sources as a special case, there are cases when nonnegative constraints are essential to provide useful and intuitive interpretation of data. This requires the similar Bayesian nonparametric modeling in case of nonnegative models such as S-NMF. This is possible by changing the model distributions in chapter 7 to distributions with nonnegative support while ensuring that inference is tractable.

- In the current nonparametric joint factor analysis (NJFA) mode, we have not exploited the correlation within the factor matrix. To capture this correlation, an extension is required where the factors are a draw from a multivariate distribution. However, at first instance, this extension appears intractable. One fruitful direction to approach this problem is through using variational inference.

- In this thesis, we have mainly focused on developing various shared subspace learning frameworks by treating each data source as a collection of data examples. One promising direction is to extend these models to allow modeling of data streams where data examples appear on a time-line. This would lead to more realistic models, as in the real world, most of the data sources are evolving with time.

- In this thesis, we have confined ourselves to the linear models. Extensions to nonlinear cases may be desirable. Possible approaches are to introduce features based on kernels (akin to kernel PCA or kernel NMF). In fact, a nonlinear kernel version of NMF is recently proposed in Zafeiriou and Petrou [2009]. Another fruitful direction is to proceed from Gaussian process based latent variable models such as GPLVM [Lawrence, 2003], Bayesian GPLVM [Titsias and Lawrence, 2010] etc. This opens connections to manifold learning based approaches. Extension to this end is also possible.

- Another useful direction is to develop supervised or semi-supervised models. We have considered subspace learning to solve mainly unsupervised tasks. Subspace learning models such as linear discriminant analysis (LDA) are quite popular for supervised applications such as face recognition. The main idea is to develop predictive subspace learning techniques similar to Chen et al. [2010]. The subspace learning has also been used in "learning to learn" problems mainly to transfer knowledge but this line of research learns a subspace for the classification/regression parameters whereas less attempts have been made towards modeling the data directly. Single source models for this task are already available, e.g. LDA and probabilistic LDA [Ioffe, 2006] but these models are prone to over-fitting especially in the domains where there are not enough examples. Bayesian extension of probabilistic LDA may help in dealing with these problems. When there are not enough examples for learning the subspace or if the domain is very noisy, it might be a good idea to consider extending these techniques in the paradigm of shared subspace learning. This gives rise to predictive shared subspace learning models. A promising starting point is the work of Chen et al. [2010] which focuses on multi-view learning by relying on undirected graphical models. An extension is required for general settings and one can also use directed graphical models such as Bayesian networks. Since these techniques would require model selection to infer the subspace dimensionalities, nonparametric Bayesian priors can play a major role while developing these models.

- For the retrieval algorithms used in this thesis, the emphasis was on demonstrating the learning capabilities of various proposed models. The data sets comprising of social media tags are generally sparse and thus data structures from sparse linear algebra such as those suggested in Witten et al. [1999] can further be utilized to speed up the retrieval algorithm. In addition, extensions can be developed to incorporate relevance feedback and automatic query expansion schemes to reduce the query ambiguities. There exists huge literature around such techniques. Some of the useful references are the works of Salton and Buckley [1990], Riezler et al. [2007] (a statistical language processing approach to relevance feedback) and the survey in Ruthven and Lalmas [2003].

Through above listing, we have outlined several potential research directions. However, research on unsupervised modeling of multiple data sources is a relatively new area and there are many broad questions to answer. For example, can we develop joint models which keeps refining itself efficiently to leverage continuously growing data sources in an incremental fashion? Can we develop frameworks for joint modeling of multiple sources, which although related at higher level, do not have any lower level commonalities? Can we exploit structured sparsity to learn the higher level objects or structures directly? We believe that attempting to answer many such questions will allow us to build systems which could break the existing barriers in machine learning and exploit the strength of multiple sources in a true sense.

Appendix A

Convexity Analysis

Convex Set

Definition A.1. Let V be a vector space over the set of real numbers $\mathbb{R}$, then a set $Y \in V$ is said to be convex if, for each $y_1, y_2 \in Y$ and a constant $\alpha \in [0,1]$, the point $\bar{y}$, defined as $\bar{y} \triangleq \alpha y_1 + (1-\alpha)\, y_2$ belongs to the set Y.

The above definition of convex sets implies that every point on the line segment connecting y_1 and y_2 belongs to the set Y.

Convex Function

A real-valued function $f(y)$ defined on an interval (a,b) is convex if the graph of the function lies below the line segment joining any two points of the graph.

Definition A.2. Let Y be a convex set defined in a vector space, then a real-valued function $f : Y \to \mathbb{R}$ is defined as convex if, for any two points $y_1, y_2 \in Y$ and a constant $\alpha \in [0,1]$

$$f\left(\alpha y_1 + (1-\alpha)\, y_2\right) \leq \alpha f\left(y_1\right) + (1-\alpha)\, f\left(y_2\right) \tag{A.1}$$

Concave functions can be defined in a similar manner. In particular, a function g is said to be concave if $-g$ is convex.

Convexity Analysis of NMF Objective Function

Consider the least-squares objective function for nonnegative matrix factorization (NMF), which is given as

$$C\left(\mathbf{W},\mathbf{H}\right)=\|\mathbf{X}-\mathbf{W}\mathbf{H}\|_F^2=\mathrm{Tr}\left[\left(\mathbf{X}-\mathbf{W}\mathbf{H}\right)\left(\mathbf{X}-\mathbf{W}\mathbf{H}\right)^{\mathsf{T}}\right] \tag{A.2}$$

where $\mathbf{W}$ and $\mathbf{H}$ are assumed to be nonnegative matrices, i.e., $\mathbf{W},\mathbf{H}\geq 0$. In chapters 3, 4 and 5, we have derived multiplicative updates for various joint factorization models adapted from NMF. In these chapters, we stated that the objective functions of the form as in Eq (A.2) is not convex in general. However, when considering one optimization variable at a time, it is convex. For example, the objective function of Eq (A.2) is convex when considered as a function of $\mathbf{W}$ (or $\mathbf{H}$) alone while fixing $\mathbf{H}$ (or $\mathbf{W}$). Following the definition of convex functions as given in Eq (A.1), we can easily verify this for objective function of Eq (A.2).

To verify the convexity of objective function of Eq (A.2) when the matrix $\mathbf{H}$ is fixed, for any two matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ and any $\alpha\in[0,1]$, consider

$$\begin{aligned}\alpha C\left(\mathbf{W}_1,\mathbf{H}\right)+\left(1-\alpha\right)C\left(\mathbf{W}_2,\mathbf{H}\right) &= \alpha\mathrm{Tr}\left[\left(\mathbf{X}-\mathbf{W}_1\mathbf{H}\right)\left(\mathbf{X}-\mathbf{W}_1\mathbf{H}\right)^{\mathsf{T}}\right]\\ &\quad+\left(1-\alpha\right)\mathrm{Tr}\left[\left(\mathbf{X}-\mathbf{W}_2\mathbf{H}\right)\left(\mathbf{X}-\mathbf{W}_2\mathbf{H}\right)^{\mathsf{T}}\right]\\ &= \mathrm{Tr}\left[\mathbf{X}\mathbf{X}^{\mathsf{T}}\right]-\mathrm{Tr}\left[2\alpha\mathbf{W}_1\mathbf{H}\mathbf{X}^{\mathsf{T}}+2\left(1-\alpha\right)\mathbf{W}_2\mathbf{H}\mathbf{X}^{\mathsf{T}}\right]\\ &\quad+\mathrm{Tr}\left[\alpha\mathbf{W}_1\mathbf{H}\mathbf{H}^{\mathsf{T}}\mathbf{W}_1^{\mathsf{T}}+\left(1-\alpha\right)\mathbf{W}_2\mathbf{H}\mathbf{H}^{\mathsf{T}}\mathbf{W}_2^{\mathsf{T}}\right] \end{aligned}\tag{A.3}$$

Defining $\bar{\mathbf{W}}\triangleq\alpha\mathbf{W}_1+\left(1-\alpha\right)\mathbf{W}_2$, we can write

$$\begin{aligned}C\left(\bar{\mathbf{W}},\mathbf{H}\right) &= \mathrm{Tr}\left[\left(\mathbf{X}-\bar{\mathbf{W}}\mathbf{H}\right)\left(\mathbf{X}-\bar{\mathbf{W}}\mathbf{H}\right)^{\mathsf{T}}\right]\\ &= \mathrm{Tr}\left[\mathbf{X}\mathbf{X}^{\mathsf{T}}\right]-\mathrm{Tr}\left[2\alpha\mathbf{W}_1\mathbf{H}\mathbf{X}^{\mathsf{T}}+2\left(1-\alpha\right)\mathbf{W}_2\mathbf{H}\mathbf{X}^{\mathsf{T}}\right]\\ &\quad+\mathrm{Tr}\left[\alpha\mathbf{W}_1\mathbf{H}\mathbf{H}^{\mathsf{T}}\bar{\mathbf{W}}^{\mathsf{T}}+\left(1-\alpha\right)\mathbf{W}_2\mathbf{H}\mathbf{H}^{\mathsf{T}}\bar{\mathbf{W}}^{\mathsf{T}}\right]\end{aligned}\tag{A.4}$$

Comparing Eqs (A.3) and (A.4), we get

$$\begin{aligned}\alpha C\left(\mathbf{W}_1,\mathbf{H}\right)+\left(1-\alpha\right)C\left(\mathbf{W}_2,\mathbf{H}\right)-C\left(\bar{\mathbf{W}},\mathbf{H}\right) &= \mathrm{Tr}\left[\alpha\mathbf{W}_1\mathbf{H}\mathbf{H}^{\mathsf{T}}\left(\mathbf{W}_1-\bar{\mathbf{W}}\right)^{\mathsf{T}}\right]\\ &\quad+\mathrm{Tr}\left[\left(1-\alpha\right)\mathbf{W}_2\mathbf{H}\mathbf{H}^{\mathsf{T}}\left(\mathbf{W}_2-\bar{\mathbf{W}}\right)^{\mathsf{T}}\right]\\ &= \mathrm{Tr}[\alpha\left(1-\alpha\right)\mathbf{W}_1\mathbf{H}\mathbf{H}^{\mathsf{T}}\left(\mathbf{W}_1-\mathbf{W}_2\right)^{\mathsf{T}}\\ &\quad-\alpha\left(1-\alpha\right)\mathbf{W}_2\mathbf{H}\mathbf{H}^{\mathsf{T}}\left(\mathbf{W}_1-\mathbf{W}_2\right)^{\mathsf{T}}]\\ &= \alpha\left(1-\alpha\right)\|\left(\mathbf{W}_1-\mathbf{W}_2\right)\mathbf{H}\|_F^2\\ &\geq 0\end{aligned}$$

which satistisfy the convexity condition of Eq (A.1). By symmetry between $\mathbf{W}$ and $\mathbf{H}$, we can show the convexity of C in $\mathbf{H}$ when $\mathbf{W}$ is fixed.

Convexity Analysis of S-NMF Objective Function

Before showing the convexity of S-NMF similar to the above, we prove the following lemma.

Lemma A.1. *Given two convex functions f_1 and f_2 and nonnegative constants λ_1 and λ_2, the linear combination $\lambda_1 f_1 + \lambda_2 f_2$ is also convex.*

Proof. Since the functions f_1 and f_2 are convex, for any two points y_1, y_2 and a constant α, define $\bar{y} \triangleq \alpha y_1 + (1-\alpha)\, y_2$, then the following inequalities hold

$$f_1(\bar{y}) \le \alpha f_1(y_1) + (1-\alpha)\, f_1(y_2) \tag{A.5}$$

$$f_2(\bar{y}) \le \alpha f_2(y_1) + (1-\alpha)\, f_2(y_2) \tag{A.6}$$

Let us define the function arising from the linear combination $\lambda_1 f_1 + \lambda_2 f_2$ as h, i.e., $h \triangleq \lambda_1 f_1 + \lambda_2 f_2$, then to show the convexity of the function h, we need to establish the following inequality

$$h(\bar{y}) \le \alpha h(y_1) + (1-\alpha)\, h(y_2)$$

The above can be easily shown by using Eqs (A.5) and (A.6) and writing

$$\begin{aligned} h(\bar{y}) &= \lambda_1 f_1(\bar{y}) + \lambda_2 f_2(\bar{y}) \\ &\le \lambda_1 (\alpha f_1(y_1) + (1-\alpha)\, f_1(y_2)) + \lambda_2 (\alpha f_2(y_1) + (1-\alpha)\, f_2(y_2)) \\ &= \alpha (\lambda_1 f_1(y_1) + \lambda_2 f_2(y_1)) + (1-\alpha)(\lambda_1 f_1(y_2) + \lambda_2 f_2(y_2)) \\ &= \alpha h(y_1) + (1-\alpha)\, h(y_2) \end{aligned}$$

□

Theorem A.1. *The objective function of Eq (3.4) used for S-NMF is convex in $\mathbf{W}$ when the matrices $\boldsymbol{U}$, $\boldsymbol{V}$, $\mathbf{H}$ and $\mathbf{L}$ are fixed.*

Proof. The theorem can be easily proven by applying the result in Lemma A.1 to S-NMF objective function. □

Although, above result is stated for convexity in $\mathbf{W}$, similar convexity folds for other variables provided remaining variables are fixed. In the following, we state equivalent theorems for the models MS-NMF and RS-NMF described in chapters 4 and 5 repectively.

Theorem A.2. *The objective function of Eq (4.5) used for MS-NMF is convex in W_v when other optimization variables (other matrices) are fixed.*

Proof. The proof follows from the application of Lemma A.1 extended from the nonnegative linear combination of two functions case to multiple functions. □

Theorem A.3. *The objective function of Eq (5.1) used for RS-NMF is convex in* $\mathbf{W}$ *when the matrices* $\boldsymbol{U}$, $\boldsymbol{V}$, $\mathbf{H}$ *and* $\mathbf{L}$ *are fixed.*

Proof. The proof follows from the direct application of Lemma A.1 in the context of RS-NMF. □

Bibliography

H.D. Abdulla, M. Polovincak, and V. Snasel. Search results clustering using nonnegative matrix factorization (nmf). *International Conference on Advances in Social Networks Analysis and Mining*, pages 320–323, 2009.

R.K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research*, 6:1817–1853, 2005.

C. Andrieu, N. De Freitas, A. Doucet, and M.I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50(1):5–43, 2003.

C.E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. *Advances in Neural Information Processing Systems*, 19:41–48, 2007.

F.R. Bach and M.I. Jordan. A probabilistic interpretation of canonical correlation analysis. *Technical Report, University of California, Berkeley*, 2005.

E. Bae and J. Bailey. Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. *IEEE International Conference on Data Mining*, pages 53–62, 2006.

R. Baeza-Yates and B. Ribeiro-Neto. *Modern information retrieval*. Addison-Wesley, 1999.

M.F. Balcan and A. Blum. A PAC-style model for learning from labeled and unlabeled data. In *Proceedings of the 18th Annual Conference on Learning Theory*, pages 111–126. Springer, 2005.

S. Basu, M. Bilenko, and R.J. Mooney. A probabilistic framework for semi-supervised clustering. In *Proceedings of the 10th International Conference on Knowledge discovery and data mining*, pages 59–68. ACM, 2004.

J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

G. Begelman, P. Keller, and F. Smadja. Automated tag clustering: improving search and exploration in the tag space. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, pages 15–33. Citeseer, 2006.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

S. Ben-David and R. Schuller. Exploiting task relatedness for multiple task learning. *Computational Learning Theory and Kernel Machines, 16th Annual Conference on Computational Learning Theory*, 2777:567–580, 2003.

M.W. Berry and M. Browne. Email surveillance using non-negative matrix factorization. *Computational and Mathematical Organization Theory*, 11(3):249–264, 2005.

M.W. Berry, M. Browne, A.N. Langville, V.P. Pauca, and R.J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173, 2007.

J. Besag and P.J. Green. Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(1):25–37, 1993.

S. Bickel and T. Scheffer. Multi-view clustering. *Proceedings of the IEEE International Conference on Data Mining*, pages 19–26, 2004.

C.M. Bishop. Bayesian pca. In *Advances in Neural Information Processing Systems*, volume 11, pages 382–388. The MIT Press, 1999.

C.M. Bishop. *Pattern recognition and machine learning*, volume 4. Springer New York, 2006.

D.M. Blei and M.I. Jordan. Modeling annotated data. In *Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 127–134. ACM, 2003.

D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

C. Boutsidis and E. Gallopoulos. SVD based initialization: a head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362, 2008.

G. Buchsbaum and O. Bloch. Color categories revealed by non-negative matrix factorization of Munsell color spectra. *Vision Research*, 42(5):559–563, 2002.

I. Buciu and I. Pitas. Application of non-negative and local non negative matrix factorization to facial expression recognition. *Pattern Recognition*, 1:288–291, 2004.

D. Cai, X. He, and J. Han. Semi-supervised discriminant analysis. *International Conference on Computer Vision*, pages 1–7, 2007.

D. Cai, X. He, X. Wu, and J. Han. Non-negative matrix factorization on manifold. *IEEE International Conference on Data Mining*, pages 63–72, 2008.

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

R.L. Calibrasi, P.M.B. Vitanyi, and A. CWI. The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007.

R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

G. Casella and C.P. Robert. Rao-Blackwellisation of sampling schemes. *Biometrika*, 83 (1):81–94, 1996.

K. Chaudhuri, S.M. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. *Proceedings of the 26th International Conference on Machine Learning*, pages 129–136, 2009.

G. Chechik and N. Tishby. Extracting relevant structures with side information. *Advances in Neural Information Processing Systems*, pages 881–888, 2003.

B. Chen, G. Polatkan, G. Sapiro, D.B. Dunson, and L. Carin. The hierarchical beta process for convolutional factor analysis and deep learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 361–368, 2011.

J. Chen, L. Tang, J. Liu, and J. Ye. A convex formulation for learning shared structures from multiple tasks. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 137–144. ACM, 2009.

N. Chen, J. Zhu, and E.P. Xing. Predictive subspace learning for multi-view data: a large margin approach. *Advances in Neural Information Processing Systems*, pages 361–369, 2010.

X. Chen, L. Gu, S.Z. Li, and H.J. Zhang. Learning representative local features for face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1126–1131. IEEE, 2001.

S. Choi. Algorithms for orthogonal nonnegative matrix factorization. *Proceedings of the International Joint Conference on Neural Networks*, pages 1828–1832, 2008.

T.S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. NUS-WIDE: A real-world web image database from national university of singapore. *Proceeding of the ACM International Conference on Image and Video Retrieval*, pages 48:1–48:9, 2009.

T.F. Cox and M.A.A. Cox. *Multidimensional scaling*, volume 1. CRC Press, 2001.

K. Crammer, M. Kearns, and J. Wortman. Learning from data of variable quality. *Advances in Neural Information Processing Systems*, 18:219, 2006.

Y. Cui, X.Z. Fern, and J.G. Dy. Non-redundant multi-view clustering via orthogonalization. In *IEEE International Conference on Data Mining*, pages 133–142. IEEE, 2007.

W. Dai, G.R. Xue, Q. Yang, and Y. Yu. Co-clustering based classification for out-of-domain documents. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 210–219. ACM, 2007a.

W. Dai, Q. Yang, G.R. Xue, and Y. Yu. Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 193–200. ACM, 2007b.

W. Dai, Q. Yang, G.R. Xue, and Y. Yu. Self-taught clustering. In *Proceedings of the 25th International Conference on Machine Learning*, pages 200–207. ACM, 2008.

P. Damien, J. Wakefield, and S. Walker. Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):331–344, 1999.

R. Datta, D. Joshi, J. Li, and J.Z. Wang. Image retrieval: ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, 2008.

B. De Finetti. *Theory of probability: A critical introductory treatment. Vol. 2.* Wiley, 1990.

A.P. Dempster, N.M. Laird, D.B. Rubin, et al. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

I.S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 269–274. ACM, 2001.

T. Diethe, D.R. Hardoon, and J. Shawe-Taylor. Multiview Fisher discriminant analysis. *NIPS Workshop on Learning from Multiple Sources*, 2008.

C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 126–135. ACM, 2006.

D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts. *Advances in Neural Information Processing Systems*, 16:1141–1148, 2004.

F. Doshi-Velez and Z. Ghahramani. Accelerated sampling for the Indian buffet process. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 273–280. ACM, 2009a.

F. Doshi-Velez and Z. Ghahramani. Correlated non-parametric latent feature models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 143–150. AUAI Press, 2009b.

F. Doshi-Velez, D. Knowles, S. Mohamed, and Z. Ghahramani. Large scale nonparametric Bayesian inference: data parallelisation in the Indian buffet process. In *Advances in Neural Information Processing Systems*, 2009a.

F. Doshi-Velez, K.T. Miller, J. Van Gael, and Y.W. Teh. Variational inference for the Indian buffet process. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 12, pages 137–144. Citeseer, 2009b.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern classification*, volume 2. Citeseer, 2001.

R.G. Edwards and A.D. Sokal. Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm. *Physical Review D*, 38:2009–2012, 1988.

M.D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995. ISSN 0162-1459.

T. Evgeniou and M. Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117. ACM, 2004.

L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

T.S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

X. Fern, C.E. Brodley, and M.A. Friedl. Correlation clustering for learning mixtures of canonical correlation models. In *Proceedings of the 5th SIAM International Conference on Data Mining*, volume 119, pages 439–448. Society for Industrial Mathematics, 2005.

D.P. Foster, S.M. Kakade, and T. Zhang. Multi-view dimensionality reduction via canonical correlation analysis. *Technical Report, TTI-C Tech Report, TTI-TR-2008-4*, 2008.

E.B. Fox, E.B. Sudderth, M.I. Jordan, and A.S. Willsky. Sharing features among dynamical systems with beta processes. *Advances in Neural Information Processing Systems*, 22: 549–557, 2009.

J. Friedman, R. Tibshirani, and T. Hastie. *The elements of statistical learning: data mining, inference, and prediction*. Springer-Verlag New York, 2009.

J. Gao, W. Fan, J. Jiang, and J. Han. Knowledge transfer via multiple model local structure mapping. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 283–291. ACM, 2008.

A. Gelman. *Bayesian data analysis*. CRC press, 2004.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(2):721–741, 1984.

W.R. Gilks, A. Thomas, and D.J. Spiegelhalter. A language and program for complex Bayesian modelling. *The Statistician*, 43(1):169–177, 1994.

A. Globerson, G. Chechik, F. Pereira, and N. Tishby. Euclidean embedding of co-occurrence data. In *Advances in Neural Information Processing Systems 17*. Citeseer, 2005.

S.A. Golder and B.A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

T. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. *Advances in Neural Information Processing Systems*, 18:475, 2006. ISSN 1049-5258.

Q. Gu and J. Zhou. Learning the shared subspace for multi-task clustering and transductive transfer classification. *IEEE International Conference on Data Mining*, pages 159–168, 2009a.

Q. Gu and J. Zhou. Local learning regularized nonnegative matrix factorization. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 1046–1051, 2009b.

D. Guillamet, M. Bressan, and J. Vitrià. A weighted non-negative matrix factorization for local representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 942–947. IEEE, 2001.

D. Guillamet, J. Vitrià, and B. Schiele. Introducing a weighted non-negative matrix factorization for image classification. *Pattern Recognition Letters*, 24(14):2447–2454, 2003.

H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *Proceedings of the 16th International Conference on World Wide Web*, pages 211–220. ACM, 2007.

D.R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: an overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

H.H. Harman. *Modern factor analysis*. University of Chicago Press, 1976.

W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

X. He and P. Niyogi. Locality preserving projections. In *Advances in Neural Information Processing Systems 16*, Cambridge, MA, 2004. MIT Press.

N.L. Hjort. Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, 18(3):1259–1294, 1990.

P.O. Hoyer. Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research*, 5:1457–1469, 2004.

M.H. Hsu and H.H. Chen. Tag normalization and prediction for effective social media retrieval. In *IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 770–774. IEEE, 2008.

S. Ioffe. Probabilistic linear discriminant analysis. *European Conference on Computer Vision*, pages 531–542, 2006.

S. Ji, L. Tang, S. Yu, and J. Ye. Extracting shared subspace for multi-label classification. *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 381–389, 2008.

Y. Jia, M. Salzmann, and T. Darrell. Factorized latent spaces with structured sparsity. *Technical Report, EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2010-99*, 2010.

L.O. Jimenez and D.A. Landgrebe. Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 28(1): 39–54, 1998.

I.T. Jolliffe. *Principle component analysis*. Springer, Heidelberg, 2002.

K. Kailing, H.P. Kriegel, A. Pryakhin, and M. Schubert. Clustering multi-represented objects with noise. *Advances in Knowledge Discovery and Data Mining, 8th Pacific-Asia Conference (PAKDD)*, pages 394–403, 2004.

S.M. Kakade and P. Foster. Multi-view regression via canonical correlation analysis. *Proceedings of the 20th Annual Conference on Learning Theory*, 4539:82–96, 2007.

M.S. Kankanhalli and Y. Rui. Application potential of multimedia information retrieval. *Proceedings of the IEEE*, 96(4):712–720, 2008.

H. Kim and H. Park. Non-negative matrix factorization based on alternating non-negativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications*, 30(2):713–730, 2008.

Y. Kim. Nonparametric Bayesian estimators for counting processes. *The Annals of Statistics*, 27(2):562–588, 1999.

A. Klami and S. Kaski. Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, 72(1-3):39–46, 2008.

D. Knowles and Z. Ghahramani. Infinite sparse factor analysis and infinite independent components analysis. *Independent Component Analysis and Signal Separation*, pages 381–388, 2007.

I. Konstas, V. Stathopoulos, and J.M. Jose. On social networks and collaborative recommendation. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 195–202. ACM, 2009.

B. Kulis, S. Basu, I. Dhillon, and R. Mooney. Semi-supervised graph clustering: a kernel approach. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 457–464. ACM, 2005.

M. Kuss and T. Graepel. The geometry of kernel canonical correlation analysis. *Technical Report TR-108, Max Planck Institute of Biological Cybernetics, Germany*, 2003.

A.N. Langville, C.D. Meyer, R. Albright, J. Cox, and D. Duling. Initializations for the nonnegative matrix factorization. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Citeseer, 2006.

N. D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in Neural Information Processing Systems 16*, Cambridge, MA, 2003. MIT Press.

D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13:556–562, 2001.

J. Lee and Y. Kim. A new algorithm to generate beta processes. *Computational statistics & data analysis*, 47(3):441–453, 2004.

G. Leen and C. Fyfe. Dirichlet process mixture models for finding shared structure between two related data sets. In *Proceedings of the 7th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Databases*, pages 31–40. World Scientific and Engineering Academy and Society (WSEAS), 2008a.

G. Leen and C. Fyfe. Learning shared and separate features from two related data sets using GPLVM's. In *Proceedings of NIPS Workshop on Learning from Multiple Sources*, 2008b.

S.Z. Li, X.W. Hou, H.J. Zhang, and Q.S. Cheng. Learning spatially localized, parts-based representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 207–212. IEEE, 2001.

T. Li, S. Ma, and M. Ogihara. Document clustering via adaptive subspace iteration. *Proceedings of the 27th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 218–225, 2004.

T. Li, C. Ding, and M.I. Jordan. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In *IEEE International Conference on Data Mining*, pages 577–582. IEEE, 2007.

X. Li, C. G. M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, 11(7):1310–1322, 2009a.

X. Li, C.G.M. Snoek, and M. Worring. Annotating images by harnessing worldwide user-tagged photos. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3717–3720, 2009b.

X. Liao, Y. Xue, and L. Carin. Logistic regression with an auxiliary data source. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 505–512. ACM, 2005.

C.J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19(10):2756–2779, 2007.

J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

Y.R. Lin, H. Sundaram, M. De Choudhury, and A. Kelliher. Temporal patterns in social media streams: theme discovery and evolution using joint analysis of content and context. In *IEEE International Conference on Multimedia and Expo*, pages 1456–1459, 2009.

X. Ling, W. Dai, G.R. Xue, Q. Yang, and Y. Yu. Spectral domain-transfer learning. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 488–496. ACM, 2008.

D. Liu, X.S. Hua, L. Yang, M. Wang, and H.J. Zhang. Tag ranking. In *Proceedings of the 18th International Conference on World Wide Web*, pages 351–360. ACM, 2009.

J.S. Liu. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427): 958–966, 1994.

W. Liu and N. Zheng. Learning sparse features for classification by mixture models. *Pattern Recognition Letters*, 25(2):155–161, 2004a.

W. Liu and N. Zheng. Non-negative matrix factorization based methods for object recognition. *Pattern Recognition Letters*, 25(8):893–897, 2004b.

L. Lovász and M.D. Plummer. *Matching theory*. Elsevier Science Ltd, 1986.

S.N. MacEachern. Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science, American Statistical Association*, pages 50–55, 1999.

C. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

C.D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, 1999.

K. V. Mardia, J. M. Bibby, and J. T. Kent. *Multivariate analysis*. Academic Press, New York, 1979.

C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, toread. *Proceedings of the 17th Conference on Hypertext and Hypermedia*, pages 31–40, 2006.

M.E. Maron and J.L. Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM)*, 7(3):216–244, 1960.

N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21 (6):1087–1091, 1953.

C.A. Micchelli and M. Pontil. Kernels for multi-task learning. *Advances in Neural Information Processing Systems*, 17:921–928, 2005.

A. Mira and L. Tierney. On the use of auxiliary variables in Markov chain Monte Carlo sampling. In *Scandinavian Journal of Statistics*, volume 29, pages 1–12. Citeseer, 1997.

R. M. Neal. Slice sampling. *Annals of Statistics*, 31(3):705–767, 2003.

D. Niu, J.G. Dy, and M.I. Jordan. Multiple non-redundant spectral clustering views. *Proceedings of the 27th International Conference on Machine Learning*, pages 831–838, 2010.

J. Paisley and L. Carin. Nonparametric factor analysis with beta process priors. *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 777–784, 2009.

J. Paisley, A. Zaas, C.W. Woods, G.S. Ginsburg, and L. Carin. A stick-breaking construction of the beta process. In *Proceedings of the 27th International Conference on Machine Learning*, pages 847–854. Citeseer, 2010.

S.J. Pan and Q. Yang. A survey on transfer learning. *Technical Report HKUST-CS08-08, Department of Computer Science and Engineering, HKUST, Hong Kong, China*, 2008.

S.J. Pan, J.T. Kwok, and Q. Yang. Transfer learning via dimensionality reduction. In *Proceedings of the 23rd national conference on Artificial intelligence*, pages 677–682, 2008.

L. A. F. Park and K. Ramamohanarao. Multiresolution web link analysis using generalized link relations. *IEEE Transactions on Knowledge and Data Engineering*, 23(11):1691–1703, 2011.

I. Peters. *Folksonomies: indexing and retrieval in Web 2.0*, volume 1. KG Saur Verlag Gmbh & Co, 2009.

J. Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006.

Z.J. Qi and I. Davidson. A principled and flexible framework for finding alternative clusterings. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 717–726, 2009.

V.V. Raghavan and SKM Wong. A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science*, 37(5):279–287, 1986.

P. Rai and H. Daumé III. Multi-label prediction via sparse infinite cca. In *Advances in Neural Information Processing Systems*, volume 22, pages 1518–1526, 2009.

P. Rai and H. Daumé III. Infinite predictor subspace models for multitask learning. *Journal of Machine Learning Research - Proceedings Track*, 9:613–620, 2010.

R. Raina, A. Battle, H. Lee, B. Packer, and A.Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th International Conference on Machine Learning*, pages 759–766. ACM, 2007.

S. Riezler, A. Vasserman, I. Tsochantaridis, V. Mittal, and Y. Liu. Statistical machine translation for query expansion in answer retrieval. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, volume 45, page 464, 2007.

S. Rogers, A. Klami, J. Sinkkonen, M. Girolami, and S. Kaski. Infinite factorization of multiple non-parametric views. *Machine Learning*, 79(1):201–226, 2010.

M.T. Rosenstein, Z. Marx, L.P. Kaelbling, and T.G. Dietterich. To transfer or not to transfer. In *Proceedings of NIPS Workshop on Inductive Transfer 10 Years Later*. Citeseer, 2005.

S. Roweis. EM algorithms for PCA and SPCA. *Advances in Neural Information Processing Systems*, pages 626–632, 1998.

S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Y. Rui and T. Huang. Optimizing learning in image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 236–243. Published by the IEEE Computer Society, 2000.

B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman. LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1): 157–173, 2008.

I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review*, 18(02):95–145, 2003.

R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. *Proceedings of the 25th International Conference on Machine Learning*, pages 880–887, 2008.

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

M. Salzmann, C. H. Ek, R. Urtasun, and T. Darrell. Factorized orthogonal latent spaces. *Journal of Machine Learning Research - Proceedings Track*, 9:701–708, 2010.

S. Saria, D. Koller, and A. Penn. Discovering shared and individual latent structure in multiple time series. *Technical Report, Arxiv preprint arXiv:1008.2028*, 2010.

B. Schölkopf, A. Smola, and K.R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

F. Shahnaz, M.W. Berry, V.P. Pauca, and R.J. Plemmons. Document clustering using nonnegative matrix factorization. *Information Processing and Management*, 42(2):373–386, 2006.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

S. Si, D. Tao, and B. Geng. Bregman divergence based regularization for transfer subspace learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(7):929–942, 2009.

B. Sigurbjörnsson and R. Van Zwol. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th International Conference on World Wide Web*, pages 327–336. ACM New York, NY, USA, 2008.

K. Sridharan and S.M. Kakade. An information theoretic framework for multi-view learning. In *Proceedings of the 21st Annual Conference on Learning Theory*, volume 2, pages 2–6. Citeseer, 2008.

E.B. Sudderth. *Graphical models for visual object recognition and tracking*. PhD thesis, Massachusetts Institute of Technology, 2006.

R.H. Swendsen and J.S. Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58(2):86–88, 1987.

P.N. Tan, M. Steinbach, and V. Kumar. *Introduction to data mining*. Pearson Addison Wesley Boston, 2006.

Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

Y.W. Teh, D. Görür, and Z. Ghahramani. Stick-breaking construction for the Indian buffet process. *Journal of Machine Learning Research - Proceedings Track*, 2:556–563, 2007.

J.B. Tenenbaum, V. Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

R. Thibaux and M.I. Jordan. Hierarchical beta processes and the Indian buffet process. *Journal of Machine Learning Research - Proceedings Track*, 2:564–571, 2007.

S. Thrun. Is learning the n-th thing any easier than learning the first? *Advances in Neural Information Processing Systems*, pages 640–646, 1996.

M.E. Tipping and C.M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):611–622, 1999.

N. Tishby, F.C. Pereira, and W. Bialek. The information bottleneck method. *Technical Report, Computing Research Repository, Arxiv preprint physics/0004057*, 2000.

M. Titsias and N. Lawrence. Bayesian Gaussian process latent variable model. *Journal of Machine Learning Research - Proceedings Track*, 9:844–851, 2010.

A. Tripathi, A. Klami, M. Orešič, and S. Kaski. Matching samples of multiple views. *Data Mining and Knowledge Discovery*, 23(2):300–321, 2011.

T. Vanderwal. Folksonomy coinage and definition. *http://vanderwal.net/folksonomy.html (Accessed in oct 2009)*, 2005.

J. Viinikanoja, A. Klami, and S. Kaski. Variational Bayesian mixture of robust CCA models. *Machine Learning and Knowledge Discovery in Databases*, 6323:370–385, 2010.

S. Virtanen, A. Klami, and S. Kaski. Bayesian CCA via group sparsity. In *Proceedings of the 28th International Conference on Machine Learning*, pages 457–464, 2011.

K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl. Constrained K-means clustering with background knowledge. In *Proceedings of the 18th International Conference on Machine Learning*, pages 577–584. Citeseer, 2001.

M.J. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

C. Wang, F. Jing, L. Zhang, and H.J. Zhang. Scalable search-based image annotation. *Multimedia Systems*, 14(4):205–220, 2008a.

F. Wang, T. Li, and C. Zhang. Semi-supervised clustering via matrix factorization. In *Proceedings of the 8th SIAM Conference on Data Mining*, pages 1–12. Citeseer, 2008b.

S. Wild, J. Curry, and A. Dougherty. Improving non-negative matrix factorizations through structured initialization. *Pattern Recognition*, 37(11):2217–2232, 2004.

B. Wiswedel, F. Höppner, and M.R. Berthold. Learning in parallel universes. *Data Mining and Knowledge Discovery*, 21(1):130–152, 2010.

I.H. Witten, A. Moffat, and T.C. Bell. *Managing gigabytes: compressing and indexing documents and images*. Morgan Kaufmann, 1999.

R.L. Wolpert and K. Ickstadt. Simulation of Lévy random fields. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, volume 133, pages 227–242. New York: Springer-Verlag, 1998.

L. Wu, L. Yang, N. Yu, and X.S. Hua. Learning to tag. *Proceedings of the 18th International Conference on World Wide Web*, pages 361–370, 2009.

P. Wu and T.G. Dietterich. Improving SVM accuracy by training on auxiliary data sources. In *Proceedings of the 21st International Conference on Machine Learning*, pages 110–117. ACM, 2004.

Z. Wu, C.W. Cheng, and C. Li. Social and semantics analysis via non-negative matrix factorization. *Proceedings of the 17th International Conference on World Wide Web*, pages 1245–1246, 2008.

E. Xing, R. Yan, and A.G. Hauptmann. Mining associated text and images with dual-wing harmoniums. *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, pages 633–641, 2005.

W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 267–273, New York, NY, USA, 2003. ACM.

R. Yan, J. Tesic, and J.R. Smith. Model-shared subspace boosting for multi-label classification. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 834–843. ACM, 2007.

J. Yang, Y. Liu, E.X. Ping, and A.G. Hauptmann. Harmonium models for semantic video representation and classification. *SIAM Conf. on Data Mining*, pages 1–12, 2007.

T. Yang, R. Jin, A.K. Jain, Y. Zhou, and W. Tong. Unsupervised transfer classification: application to text categorization. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1159–1168. ACM, 2010.

Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang. Ranking with local regression and global alignment for cross media retrieval. In *Proceedings of the 17th ACM International Conference on Multimedia*, pages 175–184. ACM, 2009.

Y. Yi, Y.T. Zhuang, F. Wu, and Y.H. Pan. Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. *IEEE Transactions on Multimedia*, 10(3):437 –446, 2008.

S. Zafeiriou and M. Petrou. Nonlinear nonnegative component analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2860–2865, 2009.

D. Zhang, S. Chen, and Z.H. Zhou. Two-dimensional non-negative matrix factorization for face representation and recognition. In *Proceedings of 2nd International Workshop on Analysis and Modelling of Faces and Gestures (AMFG)*, pages 350–363. Springer, 2005.

J. Zhang and C. Zhang. Multitask Bregman clustering. *Neurocomputing*, 74(10):1720–1734, 2011.

Z. Zhang and H. Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal of Scientific Computing*, 26(1):313–338, 2004.

W.S. Zheng, S. Gong, and T. Xiang. Unsupervised selective transfer learning for object recognition. In *Proceedings of the 10th Asian Conference on Computer Vision*, pages 527–541. Springer, 2011.

D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. *Advances in Neural Information Processing Systems*, 16:595–602, 2004.

M. Zhou, H. Yang, G. Sapiro, D.B. Dunson, and L. Carin. Dependent hierarchical beta process for image interpolation and denoising. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, volume 15, pages 883–891, 2011.

Y.T. Zhuang, Y. Yang, and F. Wu. Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. *IEEE Transactions on Multimedia*, 10(2):221–229, 2008.

Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.

ASSOCIATION FOR COMPUTING MACHINERY, INC. LICENSE TERMS AND CONDITIONS

Aug 19, 2011

This is a License Agreement between Sunil Kumar Gupta ("You") and Association for Computing Machinery, Inc. ("Association for Computing Machinery, Inc.") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Association for Computing Machinery, Inc., and the payment terms and conditions.

License Number	2732300146066
License date	Aug 19, 2011
Licensed content publisher	Association for Computing Machinery, Inc.
Licensed content publication	Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining
Licensed content title	Nonnegative shared subspace learning and its application to social media retrieval
Licensed content author	Sunil Kumar Gupta, et al
Licensed content date	Jul 25, 2010
Type of Use	Thesis/Dissertation
Requestor type	Author of this ACM article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	Unsupervised Modeling of Multiple Data Sources : A Latent Shared Subspace Approach
Expected completion date	Sep 2011
Estimated size (pages)	150
Billing Type	Credit Card
Credit Card Info	Visa ending in 9092
Credit Card Expiration	06/2013
Total	**18.75 USD**
Terms and Conditions	

Rightslink Terms and Conditions for ACM Material

1. The publisher of this copyrighted material is Association for Computing Machinery, Inc. (ACM). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at).

2. ACM reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

3. ACM hereby grants to licensee a non-exclusive license to use or republish this ACM-copyrighted material* in secondary works (especially for commercial distribution) with the stipulation that consent of the lead author has been obtained independently. Licenses are for one-time use in a single edition of the work, only with a maximum distribution equal to the number that you identified in the licensing process. Any additional form of republication must be specified according to the terms included at the time of licensing.

*Please note that ACM cannot grant republication or distribution licenses for embedded third-party material. You must confirm the ownership of figures, drawings and artwork prior to use.

4. Any form of republication or redistribution must be used within 180 days from the date stated on the license and any electronic posting is limited to a period of six months unless an extended term is selected during the licensing process. Separate subsidiary and subsequent republication licenses must be purchased to redistribute copyrighted material on an extranet. These licenses may be exercised anywhere in the world.

5. Licensee may not alter or modify the material in any manner (except that you may use, within the scope of the license granted, one or more excerpts from the copyrighted material, provided that the process of excerpting does not alter the meaning of the material or in any way reflect negatively on the publisher or any writer of the material).
6. Licensee must include the following copyright and permission notice in connection with any reproduction of the licensed material: "[Citation] © YEAR Association for Computing Machinery, Inc. Reprinted by permission." Include the article DOI as a link to the definitive version in the ACM Digital Library. Example: Charles, L. "How to Improve Digital Rights Management," Communications of the ACM, Vol. 51:12, © 2008 ACM, Inc. http://doi.acm.org/10.1145/nnnnnn.nnnnnn (where nnnnnn.nnnnnn is replaced by the actual number).
7. Translation of the material in any language requires an explicit license identified during the licensing process. Due to the error-prone nature of language translations, Licensee must include the following copyright and permission notice and disclaimer in connection with any reproduction of the licensed material in translation: "This translation is a derivative of ACM-copyrighted material. ACM did not prepare this translation and does not guarantee that it is an accurate copy of the originally published work. The original intellectual property contained in this work remains the property of ACM."
8. You may exercise the rights licensed immediately upon issuance of the license at the end of the licensing transaction, provided that you have disclosed complete and accurate details of your proposed use. No license is finally effective unless and until full payment is received from you (either by CCC or ACM) as provided in CCC's Billing and Payment terms and conditions.
9. If full payment is not received within 90 days from the grant of license transaction, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted.
10. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.
11. ACM makes no representations or warranties with respect to the licensed material and adopts on its own behalf the limitations and disclaimers established by CCC on its behalf in its Billing and Payment terms and conditions for this licensing transaction.
12. You hereby indemnify and agree to hold harmless ACM and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
13. This license is personal to the requestor and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.
14. This license may not be amended except in a writing signed by both parties (or, in the case of ACM, by CCC on its behalf).
15. ACM hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and ACM (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.
16. This license transaction shall be governed by and construed in accordance with the laws of New York State. You hereby agree to submit to the jurisdiction of the federal and state courts located in New York for purposes of resolving any disputes that may arise in connection with this licensing transaction.
17. There are additional terms and conditions, established by Copyright Clearance Center, Inc. ("CCC") as the administrator of this licensing service that relate to billing and payment for licenses provided through this service. Those terms and conditions apply to each transaction as if they were restated here. As a user of this service, you agreed to those terms and conditions at the time that you established your account, and you may see them again at any time at http://myaccount.copyright.com
18. Thesis/Dissertation: This type of use requires only the minimum administrative fee. It is not a fee for permission. Further reuse of ACM content, by ProQuest/UMI or other document delivery providers, or in republication requires a separate permission license and fee. Commercial resellers of your dissertation containing this article must acquire a separate license.
Special Terms:

SPRINGER LICENSE
TERMS AND CONDITIONS

Aug 09, 2011

This is a License Agreement between Sunil Kumar Gupta ("You") and Springer ("Springer") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Springer, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	2724630088942
License date	Aug 09, 2011
Licensed content publisher	Springer
Licensed content publication	Springer eBook
Licensed content title	A Bayesian Framework for Learning Shared and Individual Subspaces from Multiple Data Sources
Licensed content author	Sunil Kumar Gupta
Licensed content date	May 27, 2011
Type of Use	Thesis/Dissertation
Portion	Full text
Number of copies	3
Author of this Springer article	Yes and you are a contributor of the new work
Order reference number	
Title of your thesis / dissertation	Unsupervised Modeling of Multiple Data Sources : A Latent Shared Subspace Approach
Expected completion date	Sep 2011
Estimated size(pages)	150
Total	**0.00 USD**

Terms and Conditions

Introduction

The publisher for this copyrighted material is Springer Science + Business Media. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

Limited License

With reference to your request to reprint in your thesis material on which Springer Science and Business Media control the copyright, permission is granted, free of charge, for the use indicated in your enquiry. Licenses are for one-time use only with a maximum distribution equal to the number that you identified in the licensing process.

This License includes use in an electronic form, provided it is password protected or on the university's intranet, destined to microfilming by UMI and University repository. For any other electronic use, please contact Springer at (permissions.dordrecht@springer.com or permissions.heidelberg@springer.com)

The material can only be used for the purpose of defending your thesis, and with a maximum of 100 extra copies in paper.

Although Springer holds copyright to the material and is entitled to negotiate on rights, this license is only valid, provided permission is also obtained from the (co) author (address is given with the article/chapter) and provided it concerns original material which does not carry references to other sources (if material in question appears with credit to another source, authorization from that source is required as well). Permission free of charge on this occasion does not prejudice any rights we might have to charge for reproduction of our copyrighted material in the future.

Altering/Modifying Material: Not Permitted

However figures and illustrations may be altered minimally to serve your work. Any other abbreviations, additions, deletions

and/or any other alterations shall be made only with prior written authorization of the author(s) and/or Springer Science + Business Media. (Please contact Springer at permissions.dordrecht@springer.com or permissions.heidelberg@springer.com)
Reservation of Rights
Springer Science + Business Media reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
Copyright Notice:
Please include the following copyright citation referencing the publication in which the material was originally published. Where wording is within brackets, please include verbatim.
"With kind permission from Springer Science+Business Media: <book/journal title, chapter/article title, volume, year of publication, page, name(s) of author(s), figure number(s), and any original (first) copyright notice displayed with material>."
Warranties: Springer Science + Business Media makes no representations or warranties with respect to the licensed material.
Indemnity
You hereby indemnify and agree to hold harmless Springer Science + Business Media and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
No Transfer of License
This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without Springer Science + Business Media's written permission.
No Amendment Except in Writing
This license may not be amended except in a writing signed by both parties (or, in the case of Springer Science + Business Media, by CCC on Springer Science + Business Media's behalf).
Objection to Contrary Terms
Springer Science + Business Media hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and Springer Science + Business Media (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.
Jurisdiction
All disputes that may arise in connection with this present License, or the breach thereof, shall be settled exclusively by the country's law in which the work was originally published.
Other terms and conditions:
v1.2

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK0.
Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

For suggestions or comments regarding this order, contact Rightslink Customer Support: customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

SPRINGER LICENSE
TERMS AND CONDITIONS

Dec 05, 2011

This is a License Agreement between Sunil Kumar Gupta ("You") and Springer ("Springer") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Springer, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	2802820567238
License date	Dec 05, 2011
Licensed content publisher	Springer
Licensed content publication	Data Mining and Knowledge Discovery
Licensed content title	Regularized nonnegative shared subspace learning
Licensed content author	Sunil Kumar Gupta
Licensed content date	Jan 1, 2011
Type of Use	Thesis/Dissertation
Portion	Full text
Number of copies	4
Author of this Springer article	Yes and you are a contributor of the new work
Order reference number	
Title of your thesis / dissertation	Unsupervised Modeling of Multiple Data Sources : A Latent Shared Subspace Approach
Expected completion date	Dec 2011
Estimated size(pages)	180
Total	0.00 USD
Terms and Conditions	

Introduction
The publisher for this copyrighted material is Springer Science + Business Media. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

Limited License
With reference to your request to reprint in your thesis material on which Springer Science and Business Media control the copyright, permission is granted, free of charge, for the use indicated in your enquiry. Licenses are for one-time use only with a maximum distribution equal to the number that you identified in the licensing process.

This License includes use in an electronic form, provided it is password protected or on the university's intranet, destined to microfilming by UMI and University repository. For any other electronic use, please contact Springer at (permissions.dordrecht@springer.com or permissions.heidelberg@springer.com)

The material can only be used for the purpose of defending your thesis, and with a maximum of 100 extra copies in paper.

Although Springer holds copyright to the material and is entitled to negotiate on rights, this license is only valid, provided permission is also obtained from the (co) author (address is given with the article/chapter) and provided it concerns original material which does not carry references to other sources (if material in question appears with credit to another source, authorization from that source is required as well). Permission free of charge on this occasion does not prejudice any rights we might have to charge for reproduction of our copyrighted material in the future.

Altering/Modifying Material: Not Permitted
However figures and illustrations may be altered minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of the author(s) and/or Springer Science + Business Media. (Please contact Springer at permissions.dordrecht@springer.com or permissions.heidelberg@springer.com)

Reservation of Rights
Springer Science + Business Media reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

Copyright Notice:
Please include the following copyright citation referencing the publication in which the material was originally published. Where wording is within brackets, please include verbatim.
"With kind permission from Springer Science+Business Media: <book/journal title, chapter/article title, volume, year of publication, page, name(s) of author(s), figure number(s), and any original (first) copyright notice displayed with material>."

Warranties: Springer Science + Business Media makes no representations or warranties with respect to the licensed material.

Indemnity
You hereby indemnify and agree to hold harmless Springer Science + Business Media and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

No Transfer of License
This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without Springer Science + Business Media's written permission.

No Amendment Except in Writing
This license may not be amended except in a writing signed by both parties (or, in the case of Springer Science + Business Media, by CCC on Springer Science + Business Media's behalf).

Objection to Contrary Terms
Springer Science + Business Media hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and