UNIVERSITY OF CALIFORNIA, SAN DIEGO

A Family of Statistical Topic Models for Text and Multimedia Documents

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Signal and Image Processing)

by

Duangmanee (Pew) Putthividhya

Committee in charge:

Kenneth Kreutz-Delgado, Chair
Sanjoy Dasgupta
Gert Lanckriet
Te-Won Lee
Lawrence Saul
Nuno Vasconcelos

2010

The dissertation of Duangmanee (Pew) Putthividhya is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2010

DEDICATION

In the memory of my grandparents—Li Kae-ngo and Ngo Ngek-lang.

EPIGRAPH

I know the price of success:
dedication, hard work, and unremitting
devotion to the things you want to see happen.
—Frank Lloyd Wright.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my family, friends, colleagues, and mentors who have helped made the completion of this dissertation possible.

First and foremost, I would like to thank my family whose unconditional love and support has been a bright light that shines on me even in the darkest hours. Both of my sisters—Dr. Wanida Putthividhya and Dr. Aksara Putthividhya—have been the greatest source of inspiration and I could not have been more proud of their accomplishments. I am extremely happy to have followed in their footsteps and finally be able to cross the finish line with the completion of this dissertation. Their personal advice and encouragement has made me a better researcher, a better sister, and a better person. My parents have been the biggest cheer leaders over the past decade in my pursuit of many degrees. Through their many prayers, unyielding love, and selfless sacrifices, I have been able to find strength and work through problems and obstacles that come my way. I could not have asked for a better support system in my family than what I already have.

My most sincere gratitude is reserved for the best mentor—Dr. Hagai T. Attias of Golden Metallics Inc. Without his guidance, insights, wisdom, and most importantly his understanding, much of the work that makes up this thesis would not have been remotely possible. His enthusiasm and passion has been a great source of motivation and moved me to accomplish many things I could not have dreamed possible. His many kind words and encouragement have filled me with much needed confidence and have made me work harder to reach my potentials.

I would like to thank my advisor—Dr. Te-Won Lee—who has put me in the right path and has sparked my interests in the field of machine learning and data mining. Without his financial support and belief in my potentials, my PhD journey in this interesting field would not have even been launched. I am indebted to Professor Srikantan S. Nagarajan of UCSF whose lab I visited as often as my own. His keen intellect and brilliant mind is always open to opportunities in the new fields and applications, and I have gained invaluable perspectives from many discussions we have had.

I would like to thank all my committee members for their advice and constructive criticism. Professor Kenneth Kreutz-Delgado has been nothing but supportive for the past many years. I learned a great deal from many insightful discussions with Professor Sanjoy Dasgupta, Professor Nuno Vasconcelos, Professor Lawrence Saul, Professor Gert Lanckriet, and Professor Charles Elkan.

I would like to acknowledge my colleagues and friends at the Lee lab. Specifically, I would like to thank Jiucang Hao for many insightful and interesting discussions we shared over many years. My special thanks to Hyun-jin Park for getting me started on the work on ICA for image and video representation. I am grateful for all the fruitful discussions with Gil-jin Jang, Tae-su Kim, Alice (Eunjung) Lee and Professor Torbjorn Eltoft of University of Tromso, Norway.

I am grateful for the support and friendships of so many good friends—Samerkhae (Nok) Jongthammanurak, Piyada (Kaew) Phanaphat, Rojana (E) Pornprasertsuk, Nirattaya (Ning) Khamsemanan, Pakorn (Tony) Kanchanawong, Siraprapha (Yui) Sanchatjate, Bunpote (A) Siridechadilok—who have been relentless cheer leaders from the start of my PhD journey. I am grateful to call them my friends with whom I can share not only laughters but tears. Special thanks to my friends from MIT—Xuemin Chi, Ying A. Cao, and Xixi D'Moon—who have stayed in touch and been the greatest supporters. I am also grateful for the friendship of Marc Freese and my mentor at Toshiba Research and Development Center—Hiroaki Nakai—who has shown me nothing but kindness.

Many friends I met in San Diego and San Francisco have been the greatest comfort throughout the turbulent years. They are the people who keep me grounded and sane. These great people are Douglas and Karen Fidaleo, Emelia Marapao, Thawee Techathamnukool, Sutharin Kampuntip, Nut Taesombut, Somsak Kittipiyakul, Sataporn Pornpromlikit, Dan Liu, Dashan Gao, Honghao Shan, Wen-yi Zhang, Paul Hammon, Jun-wen Wu, Stephen Krotosky, David Chi, Gwen Littlewort, Aaron Jow, Saeko Nomura, Tomomi Ushii, Sandeep Manyam, and Kelly Westlake.

Last but most importantly, I am forever indebted to my best friend and love of my life—Shinko Cheng. I am grateful to his companionship, support, and his utmost understanding throughout the years. I have enjoyed our technical discussions and without his insights and help in small and large ways, the completion of this dissertation would not have been possible. His companionship and kindness has made this bumpy journey with twists and turns a truly enjoyable and worthwhile ride.

Chapter 2, in part, is a reprint of the material as it appears in: D. Putthividhya, H. T. Attias, S. Nagarajan, "Independent Factor Topic Models," in International Conference on Machine Learning (ICML) 2009. I was the primary researcher of the cited materials and the co-author listed in these publications supervised the work which forms the basis of this chapter.

Chapter 4, in part, is a reprint of the material as it appears in: D. Putthividhya, H. T. Attias, S. Nagarajan. "Topic-Regression Multi-Modal Latent Dirichlet Allocation for Auto-Annotation," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, and D. Putthividhya, H. T. Attias, S. Nagarajan, T.-W. Lee, "Probabilistic Graphical Model for Auto-Annotation, Content-based Retrieval and Classification of TV clips containing Audio, Video, and Text," in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2007. I was the primary researcher of the cited materials and the co-author listed in these publications supervised the work which forms the basis of this chapter.

Chapter 5, in part, is a reprint of the material as it appears in: D. Putthividhya, H. T. Attias, S. Nagarajan. "Supervised Topic Model for Automatic Image Annotation", in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2010. I was the primary researcher of the cited materials and the co-author listed in these publications supervised the work which forms the basis of this chapter.

Chapter 7, in part, is a reprint of the material as it appears in: D. Putthividhya, T.-W. Lee, “Motion Patterns: High-level Representations of Natural Video Sequences,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006. I was the primary researcher of the cited materials and the co-author listed in these publications supervised the work which forms the basis of this chapter.

VITA

2001	B. A. in Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge.
2002	M. E. in Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge.
2003-2009	Graduate Student Researcher, University of California, San Diego.
2010	Ph. D. in Electrical and Computer Engineering, University of California, San Diego.

PUBLICATIONS

D. Putthividhya, T.-W. Lee, "Motion Patterns: High-level Representations of Natural Video Sequences", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

D. Putthividhya, H. T. Attias, S. Nagarajan, T.-W. Lee, "Probabilistic Graphical Model for Auto-Annotation, Content-based Retrieval and Classification of TV clips containing Audio, Video, and Text", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

D. Putthividhya, H. T. Attias, S. Nagarajan, "Independent Factor Topic Models", *International Conference on Machine Learning (ICML)*, 2009.

D. Putthividhya, H. T. Attias, S. Nagarajan, "Supervised Topic Model for Automatic Image Annotation", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010.

D. Putthividhya, H. T. Attias, S. Nagarajan, "Topic-Regression Multi-Modal Latent Dirichlet Allocation for Auto-Annotation", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

D. Putthividhya, H. T. Attias, S. Nagarajan. "Statistical Topic Models for Image, Video, and Multimedia Annotation and Text-based Retrieval", *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2010 (in submission).

D. Putthividhya, H. T. Attias, S. Nagarajan, "On Topic Models for Image, Video, and Multimedia Annotation", 2010 (in preparation).

D. Putthividhya, H. T. Attias, S. Nagarajan, "Independent Factor Topic Models for Learning Topic Correlations", 2010 (in preparation).

ABSTRACT OF THE DISSERTATION

A Family of Statistical Topic Models for Text and Multimedia Documents

by

Duangmanee (Pew) Putthividhya

Doctor of Philosophy in Electrical Engineering (Signal and Image Processing)

University of California, San Diego, 2010

Kenneth Kreutz-Delgado, Chair

In this thesis, we investigate several extensions of the basic Latent Dirichlet Allocation model for text and multimedia documents containing images and texts, video and texts, or audio-video and texts.

For exploratory analysis of large-scale text document collections, we present Independent Factor Topic Models (IFTM) which captures topic correlations using linear latent variable models to directly uncover the hidden sources of correlations. Such a framework offers great flexibility in exploring different forms of source prior, and in this work we investigate 2 source distributions: Gaussian and Laplacian. When the sparse source prior is used, we can indeed visualize and give interpretation to the sources of correlations and construct a simple topic graph which can be used to navigate large-scale archives.

In extending IFTM to learn correlations between latent topics of different data modalities in multimedia documents, we present a topic-regression multi-modal Latent Dirichlet Allocation (tr-mmLDA) which uses a linear regression module to learn the precise relationships between latent variables in different modalites. We employ tr-mmLDA in an image and video annotation task, where the goal is to learn statistical association between images and their corresponding captions, so that the caption data can be accurately inferred in the test set. When dealing with annotation data that act more similar to class labels, the assumption in tr-mmLDA which allows caption words in the same document to be generated from multiple hidden topics might be overly complex. For such annotation data, we propose a novel statistical topic model called sLDA-bin, which extends supervised Latent Dirichlet Allocation (sLDA) [BM07] model to handle a multivariate binary response variable of the annotation data. We show superior image annotation and retrieval results comparing sLDA-bin with correspondence LDA [BJ03] on standard image datasets.

We also extend the association model for the case of image-text and video-text to perform automatic annotation of multimedia documents containing audio and video, we find that unlike cLDA, tr-mmLDA and sLDA-bin can be straight-forwardly extended to include influence from additional data modalities in predicting annotation by incorporating the latent topics from the additional modality as another set of covariates into the linear and logistic regression module respectively.

Chapter 1

Introduction

Large-scale text document collections have become increasingly available online in the era of pervasive internet. With new forms of documents, e.g. twitter messages, blogs, emails, and traditional print documents that are going digital with e-books being published online, and printed newspapers becoming 24-hour cycle of e-news, new documents are being produced at an extremely fast rate that one can easily gather large-scale document archives comprising several millions of entries over a short span of time. The scale and magnitude of modern day document collections therefore makes a compelling case for the developments of automated tools in exploring and organizing such collections. The main focus of this thesis is to propose novel algorithms based on statistical modeling of document archives for the purpose of exploratory analysis and information extraction for efficient indexing, retrieval, organization, and navigation through documents in large-scale archives.

In recent years, statistical topic models such as Latent Dirichlet Allocation (LDA) [BNJ02, BNJ03] and Correlated Topic Models (CTM) [BL06a, BL07] have emerged as powerful tools in analyzing the content and extracting key information contained in large-scale text document archives. By treating a document as a collection of words, and modeling words in the same document as deriving characteristics from multiple latent components, topic models efficiently summarize a large-scale document archive using a small number of topics. The popularity of such methods indeed stems from their ability to uncover underlying patterns of word co-occurrences that form interpretable topics. More sophisticated models, e.g. [BGJT04], even allow complicated structures such as topic hierarchies to be learned from the data. In managing large-scale unstructured document repositories, such topical information proves to be an invaluable cue for indexing, organizing, and cross-referencing documents for efficient navigation through the archives.

Recent years have also seen a great number of extensions of topic models beyond their traditional usage in analyzing collections of text documents. These extension models use the basic Latent Dirichlet Allocation (LDA) model as the core building block and adapt/append the basic model to be more appropriate for a variety of different types of data. Examples of these models include extensions of LDA to model image data [FfP05, CFf07, VT07, WG07], image-caption data [BJ03], multimedia data [PANL07], user profile data [GK04], network data [ABFX07], genotype data [SX08], and recommendation systems [Mar04]. Inspired by the success and applicability of topic models to data of different types, in this work, we present extensions of LDA to model multi-modal data. Indeed, with the explosion of the amount of multimedia documents made available online with the advent of youtube, facebook, myspace, twitter, and other social networking tools, we focus on extensions of LDA to represent multimedia documents containing images and texts, video and texts, or audio-video and texts in the same documents. In many scenarios, one may be interested in inferring the textual information given the other data types, the challenge in this application is how one learns the joint correlations between data of different modalities to allow for accurate prediction when the textual information is not available.

We present a family of statistical topic models for multi-modal data, with immediate applications in modeling multimedia documents. First, we present Independent Factor Topic Models (IFTM) [PAN09] which uses linear latent variable models to uncover the hidden sources of correlation between topics and learn topic correlations. In extending IFTM to learn correlations between latent topics of different data modalities, we arrive at a second topic model called Topic-Regression Multi-Modal Latent Dirichlet Allocation (tr-mmLDA) [PAN10b] which uses a linear regression module to learn the precise relationships between latent variables in different modalites. We employ tr-mmLDA in an image and video annotation task, where the goal is to learn statistical association between images (or videos) and their corresponding captions, so that we can make accurate predictions when the caption data is not available. In many image annotation datasets used in performance evaluation, we encounter caption data that are binary and scarce, we therefore modify tr-mmLDA to work with such binary annotation data and present a third topic model called supervised LDA-binary (sLDA-bin) [PAN10a]. By eliminating the hidden topics in the caption modality and model the binary data as a multi-variate Bernoulli random variable, sLDA-bin is more suitable for the binary annotation data and is able to obtain comparable annotation performance on standard image datasets with previously proposed models that are many times more computationally intensive [LMJ03, FML04].

For a multimedia annotation task, we extend tr-mmLDA and sLDA-bin from the 2-modality case for modeling image-caption or video-caption data, to a 3-modality version which allows the models to use additional input modalities to make more accurate predictions of caption data. We show experimental results using both the audio and video modalities to infer the missing captions. Our 3-modality models are able to reap benefits of the independence and correlation structures in all the input modalities to further improve the prediction performance over using each individual input modality alone.

Lastly, we present a novel video representation based on statistics of natural videos. We learn a set of spatio-temporal ICA basis for video blocks, where each ICA filter, resembling localized Gabor wavelet moving in time, captures "independent motion" in each block. Since ICA features are known to be sparse, we fit a mixture of Laplacian model on a large collection of features and a pattern of co-activation of basis with the same speed or velocity emerges in each mixture. We adopt this video feature representation in all experiments performed on video in this thesis.

1.1 Thesis Outline

The rest of the thesis is organized as follows:

- Chapter 2 presents Independent Factor Topic Models (IFTM)—a novel statistical topic model for learning correlations between topics.

- Chapter 3 discusses related work on image and video annotation and gives an overview of the use of statistical topic models for the problem of annotation.

- Chapter 4 presents a novel statistical topic model called Topic-Regression Multi-modal Latent Dirichlet Allocation (tr-mmLDA) for capturing statistical association between data of different types. In particular, we focus on image-caption and video-caption data for the task of auto-annotation.

- Chapter 5 presents a novel statistical topic model called supervised Latent Dirichlet Allocation-binary (sLDA-bin) which extends the basic supervised LDA model to handle multi-variate binary response variable. We employ this model for modeling image-caption, video-caption data where the caption data is restricted to binary.

- Chapter 6 presents extensions of tr-mmLDA in Chapter 4 and sLDA-bin in Chapter 5 to the problem of multimedia annotation.

- Chapter 7 presents a novel motion representation learned by using clustering of spatio-temporal ICA features.

- Chapter 8 summarizes the work and presents final remarks.

Chapter 2

Independent Factor Topic Models

Statistical topic models such as Latent Dirichlet Allocation (LDA) and Correlated Topic Models (CTM) have recently emerged as powerful statistical tools for text document analysis. With the ability to learn clusterings of words that are semantically related (these clusters are loosely termed topics), i.e. words that form a common theme or describe a common concept, topic models have been used more extensively in recent years as automated tools in exploring and organizing large-scale document collections. One of the main goals is to use semantic structures uncovered by topic models to index and organize documents in the archive so that navigation through the large-scale collections can be done efficiently based on such topical information. In considering a choice of statistical topic models for the task of exploratory analysis, LDA is found to be lacking in its inability to learn relationships between topics, which can be indeed useful in building a navigational tool to explore documents in the archive. For such a task, we pay special attention to more sophisticated topic models, such as Correlated Topic Models (CTM), hierarchical LDA (hLDA) that can capture correlation structures between topics.

In this chapter, we describe Independent Factor Topic Models (IFTM)—a novel a statistical topic model that presents an alternative to previously proposed models such as Correlated Topic Models (CTM) and Pachinko Allocation Model (PAM) in learning correlations between topics. More specifically, IFTM proposes the use of a linear latent variable framework to model the sources of topic correlation directly. Such a framework in capturing correlation offers great flexibility in exploring different forms of source prior model, and in this work we investigate 2 source distributions: Gaussian and Laplacian distribution. When the sparse source prior is used, we can indeed visualize and give interpretation to the sources of topic correlations and construct a topic graph which can be used to navigate documents in the archive. The introduction of the latent sources in our formulation leads to a fast variational inference algorithm for IFTM. Our experiments show that IFTM is on average 3-5 times less computationally demanding, while still performs very competitively with CTM.

2.1 Introduction

Exploratory analysis of document repositories poses an interesting challenge in the era of pervasive internet. Now more than ever, large-scale document collections have become increasingly available online in the forms ranging from twitter messages and blogs to emails and news articles. Gathering these documents within a short span of time, one can create a large repository comprising several millions of entries. With such an unprecedented magnitude of modern days document archives, there is a compelling need to develop automated tools for exploring and organizing large-scale document collections. In recent years, statistical topic models [BNJ02, BNJ03, GS04, RZGSS04, BL06a, BL06b] have emerged as powerful tools in analyzing the content and extracting key information contained in document archives. The popularity of such methods stems from their ability to uncover underlying patterns of word co-occurrences that form interpretable *topics*. More sophisticated models, e.g. [BGJT04], even allow complicated structures such as topic hierarchies to be learned from the data. In managing large-scale unstructured document repositories, such topical information proves to be an invaluable cue for indexing, organizing, and cross-referencing documents for efficient navigation through the archives.

Latent Dirichlet Allocation (LDA) [BNJ03] is the most basic and widely-used model in the family of statistical topic models. Under LDA, words in the same document are allowed to exhibit characteristics from multiple components (topics). A document, which is treated as a collection of words, can be summarized in terms of the components' overall relative influences on the collection. More specifically, LDA models a document as a random proportion of topics, and assumes the form of a Dirichlet distribution, while each topic, in turn, is modeled as a multinomial distribution over words in the vocabulary. Despite computational intractability in performing exact inference, the choice in modeling the topic proportion as a Dirichlet greatly simplifies the computation in approximate inference for LDA. In particular, a computationally efficient variational inference algorithm [BNJ03] and an efficient Rao-blackwellized Gibbs sampling for LDA [GS04] can be derived due to the Dirichlet-Multinomial conjugacy.

Nonetheless, Dirichlet distribution has a serious restriction. Under a Dirichlet, topic proportions are modeled as nearly independent, thus hampering the ability of LDA to model topic co-occurrences that are common-place in real-world documents. Correlated Topic Models (CTM) were consequently proposed in [BL06a] to address such a limitation. By replacing Dirichlet with a powerful logistic normal distribution, the correlation between topics is now captured in the full covariance structure of the normal distribution. This choice of prior, however, poses significant challenges in inference and parameter estimation. In particular, closed-form analytic solutions in LDA inference are now replaced by a conjugate gradient update in CTM, causing a significant slowdown in the inference step. Moreover, parameter estimation for the full-rank covariance matrix can be very inaccurate in high dimensional space.

In this work, we seek an alternative approach to characterize topic correlations. Consider an archive of news articles on 4 topics: Wall Street collapse, Iraq war, Subprime mortgage crisis, and September 11 attack. These 4 topics are found to co-occur often in the news archive of 2008, i.e. if an article contains a discussion about the Wall Street collapse, then we can predict, with high probability, the presence of discussions related to the Iraq war or September 11. An interesting question one might ask is whether such a relationship can be explained by a hidden factor, e.g. the presidential election 2008, that accounts for the co-occurrence of these topics.

The above example motivates the idea of employing latent variable models to uncover the source of correlations between topics. Indeed, the use of latent variable framework offers great flexibility in specifying the form of correlation being captured (linear vs non-linear), as well as in the choice of the latent source prior used (continuous vs discrete, Gaussian vs Non-Gaussian). In this work, we focus on the use of well-studied linear models where the latent source variables are continuous and modeled as independent, and the correlated topic vectors are formed by linearly combining these sources. We, therefore, adopt the name *Independent Factor Topic Models* (IFTM) to reflect the generative process of how correlation between topics is modeled.

Two choices of latent source prior model are investigated in this work. In the first scenario, we present IFTM with Gaussian source prior, where the independent sources are assumed to be drawn from an Isotropic Gaussian distribution. Using such a prior implies that the topic proportion for each document is drawn from a logistic normal distribution. From this viewpoint, IFTM with Gaussian prior can be seen as a generalized version of CTM with adjustable covariance structures. Assume that we have L sources and $L \ll K$ (K is the number of topics), IFTM is indeed CTM with a constrained structure of the covariance matrix. When $L = K$, IFTM becomes equivalent to CTM with no restriction in the form of the covariance matrix. By choosing $L \ll K$, however, we indeed reduce the number of covariance parameters from $\mathcal{O}(K^2)$ to $\mathcal{O}(KL)$ while still allowing the most significant correlations to be captured. With fewer parameters, IFTM is therefore more robust to over-fitting. In addition, by eliminating the full covariance structure of CTM, the objective function factorizes and each component of the variational parameters can be optimized independently, leading to an efficient Newton-Raphson based variational inference algorithm, which is found to be 5 times faster than that of CTM.

Motivated by the desire to visualize and interpret the individual sources, we go beyond the linear-Gaussian assumption and explore non-Gaussian source distribution. In particular, a sparse source prior in the form of Laplacian distribution is used. Such a prior favors a configuration where only a handful number of sources are “active” for each document, giving rise to more interpretable results. We adopt the convex dual representation of the Laplacian prior [Gir01] and derive a variational inference algorithm for the Laplacian source prior as a straightforward modification of the Gaussian case. Due to additional variational parameters, inference in this case is more computationally demanding but still runs 3 times faster than CTM.

This chapter is organized as follows. In section 2.2, we describe the model definition, the generative process of IFTM, and derive an approximate inference algorithm for IFTM with Gaussian and non-Gaussian sources using the variational framework. In section 3, we give a visualization and interpretation of what the hidden sources represent on 2 text datasets: (a) 4000-document archive of NSF abstracts and (b) 15000 articles from NYTimes collection. We show that IFTM performs competitively with CTM, as measured using perplexity and the log-likelihood over held-out data. IFTM displays a superior performance over LDA on a document retrieval task, demonstrating the power of topic models that can capture correlations between topics.

2.2 Independent Factor Topic Models

2.2.1 Model Definition

We establish the notation used throughout the paper. A word is denoted as a unit-basis vector w of size T with exactly one non-zero entry representing the membership to only one word in a dictionary of T words. A document is a collection of N word occurrences denoted by $\mathbf{w} = \{w_1, \ldots, w_N\}$. A set of D training documents is denoted by $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_D\}$.

Similar to LDA, IFTM assumes that there are k underlying latent topics corresponding to patterns of word co-occurrences in the document archive. With each topic modeled as a multinomial distribution over words in the vocabulary, we represent a document as a mixture weight of these K basis patterns (topics) and denote it by θ. To generate a document with N words, we first specify the proportion of the K topics that the document contains; the topics, in turn, govern the probability of generating each word in the document.

The key distinction between LDA, CTM, and IFTM lies in the modeling assumption for θ. In LDA, θ is drawn from a Dirichlet distribution, which models the components θ_i and θ_j as nearly independent . To allow for correlation among topics, CTM assumes θ is a draw from a logistic normal distribution. First, a continuous-valued random variable $\mathbf{x}$ is drawn from a Gaussian distribution with full-covariance structure: $\mathbf{x} \sim \mathcal{N}(\mathbf{x}; \mu, \Sigma)$, where $\mathcal{N}(\mathbf{x}; \mu, \Sigma)$ denotes a multi-variate Gaussian distribution with mean μ and covariance Σ. The topic proportion θ is then obtained by a deterministic mapping from $\mathbf{x} \in \mathbb{R}^K$ to a point on a $K-1$ dimensional simplex $\mathbb{S}^{K-1}$ through the softmax operation: $\theta_k = \frac{e^{x_k}}{\sum_l e^{x_l}}$. Correlations between pairs of topics are encoded in the entries of Σ. If the presence of topic i in a document boosts the chance of observing topic j, then x_i and x_j are positively correlated and is reflected in the covariance entry Σ_{ij}.

Inspired by the use of linear latent variable models, e.g. Factor Analysis, Independent Component Analysis, to uncover hidden factors that explain correlations in the data, IFTM assumes the existence of independent sources and model the correlation structure between topics by linearly mixing these sources to form correlated topic vectors. Specifically, we introduce, for each document, an L-dimensional latent variable $\mathbf{s}$ to represent the sources of topic correlations. The correlated topic proportion $\mathbf{x}$ is then obtained as a linear transformation of $\mathbf{s}$ with additive Gaussian noise: $\mathbf{x} = \mathbf{A}\mathbf{s} + \mu + \epsilon$, where $\mathbf{A}$ is a $K \times L$ mixing matrix (factor loading matrix), μ is a K-dimensional mean vector, and ϵ is a zero-mean Gaussian noise with *diagonal* inverse covariance $\mathbf{\Lambda}$: $\epsilon \sim \mathcal{N}(\epsilon; 0, \mathbf{\Lambda}^{-1})$. We explore 2 latent source distributions: (1) $p(\mathbf{s}) \sim \mathcal{N}(\mathbf{s}; 0, \mathbf{I}_L)$ in Section 2.2.2 and (2) $p(\mathbf{s})$ distributed as a Laplacian distribution in Section 2.2.3.

The diagonality assumption of $\mathbf{\Lambda}$ is crucial for 2 reasons. First, it implies conditional independence, which states that given the sources, the proportion of topics i and j: x_i and x_j are independent. Therefore, the correlations between topics are entirely due to their dependency on the latent sources (via the linear mixing encoded in matrix $\mathbf{A}$). The second reason for modeling the noise variance separately is so that we can model the variance unique to each topic in $\mathbf{\Lambda}$, while putting all the correlation structure into $\mathbf{A}$.

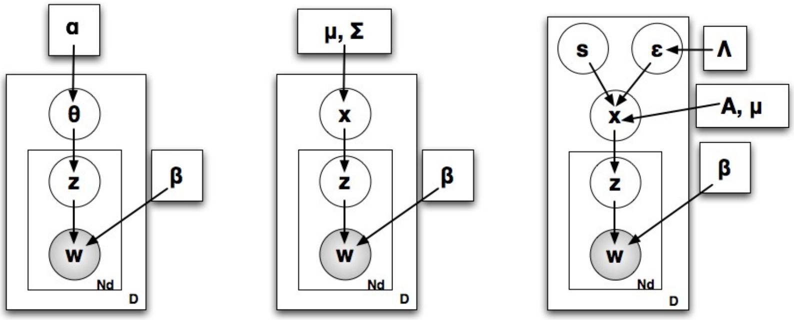
Given the number of topics K and the model parameters $\Psi = \{\beta, \mathbf{A}, \mathbf{\Lambda}, \mu\}$, to generate a document with N word occurrences: $\{w_1, \ldots, w_N\}$, we follow the generative process of IFTM as depicted in Figure 2.1(c):

- Draw contribution of independent sources: $\mathbf{s} \sim p(\mathbf{s})$.
- Draw continuous-valued correlated topic proportion vector $\mathbf{x}$ from the conditional distribution $p(\mathbf{x}|\mathbf{s})$: $\mathbf{x} \sim \mathcal{N}(\mathbf{x}; \mathbf{A}\mathbf{s} + \mu, \mathbf{\Lambda}^{-1})$.
- Given $\mathbf{x}$, the topic proportion vector θ is defined as: $\theta_k = \frac{e^{x_k}}{\sum_l e^{x_l}}$.
- For each word $n \in \{1, 2, \ldots, N\}$,
  1. Draw a topic assignment $z_n|\theta \sim \text{Mult}(\theta)$: $p(z_n = k|\theta) = \theta_k$.
  2. Under the topic z_n, draw a word $w_n|z_n \sim \text{Mult}(\beta_{z_n}) : p(w_n = t|z_n = k) = \beta_{kt}$.

2.2.2 IFTM with Gaussian source prior

When the latent source distribution is an Isotropic Gaussian, the generative process of $\mathbf{x}$ is indeed identical to the well-known factor analysis model [Eve84]. IFTM with $p(\mathbf{s}) = \mathcal{N}(\mathbf{s}; 0, \mathbf{I})$ can thus be thought as using the factor analysis model to explain the correlation structure in the topic proportions. Since $p(\mathbf{s})$ is Gaussian and the conditional distribution of $p(\mathbf{x}|\mathbf{s})$ is Gaussian, we can integrate out the latent factors $\mathbf{s}$ and obtain the marginal distribution of $\mathbf{x}$, which is also Gaussian as follows: $p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{s})p(\mathbf{s})d\mathbf{s} = \mathcal{N}(\mathbf{x}; \mu, \mathbf{C})$, where $\mathbf{C} = \mathbf{A}\mathbf{A}^\top + \mathbf{\Lambda}^{-1}$. Assuming $\mathbf{s} \in \mathbb{R}^L$, where $L \ll K$, IFTM with Gaussian source prior can thus be seen as a special case of CTM with a constrained covariance matrix parameterized by $\mathbf{A}, \mathbf{\Lambda}$, where the number of free parameters is $K + KL - \frac{L(L-1)}{2}$ instead of the $\frac{K(K+1)}{2}$ in the full covariance case. As we increase L to be the same as K, however, our model becomes equivalent to CTM in that we are no longer restricted to the factorized structure and can model any form of covariance matrix.

To learn the model parameters $\Psi = \{\mathbf{A}, \mathbf{\Lambda}, \mu, \beta\}$, we use Expectation Maximization



(a) **LDA** (b) **CTM** (c) **IFTM**

Figure 2.1: Graphical Model Representation comparing **(a)** Latent Dirichlet Allocation (LDA) **(b)** Correlated Topic Models (CTM) **(c)** our proposed model Independent Factor Topic Model (IFTM). The shaded nodes represent the observed variables. The key difference between the 3 models lies in how the topic proportion for each document, θ, is generated . In LDA, θ is a draw from a Dirichlet distribution. In CTM, θ is a softmax output of $\mathbf{x}$, where $\mathbf{x} \sim \mathcal{N}(\mathbf{x}; \mu, \Sigma)$. In IFTM, we first sample the independent sources $\mathbf{s}$ that govern how the topics are correlated within a document; then θ is defined as a softmax output of a linear combination of the independent sources plus additive Gaussian noise contribution.

(EM) framework to find Ψ that maximizes the data log-likelihood. EM iterates between the E-step where we infer the posterior distribution over the hidden variables, and the M-step where the model parameters are updated from the relevant sufficient statistics of the inferred posterior. Note that since IFTM with Gaussian source prior is a special case of CTM, we could use the inference algorithm of CTM, by replacing Σ with $\mathbf{A}\mathbf{A}^\top + \mathbf{\Lambda}^{-1}$. In the M step, however, the closed-form update of Σ must be replaced by a quasi-newton optimization for $\mathbf{A}$ and $\mathbf{\Lambda}$, see [Jor67]. Nonetheless, such an approach cannot be applied to the non-Gaussian source prior case in Section 2.2.3, as the marginal distribution $p(\mathbf{x})$ is no longer Gaussian. As we shall see, the formulation of variational inference for IFTM that explicitly incorporate the latent sources $\mathbf{s}$ simplifies the computation by taking advantage of the diagonality of $\mathbf{\Lambda}$, while in the M-step closed-form updates similar to those derived for EM for factor analysis can be obtained.

Variational Inference

We begin with the expression of the log-likelihood for 1 document:

$$\log p(\mathbf{W}|\Psi) \geq \int q(\mathbf{Z}, \mathbf{x}, \mathbf{s})(\log p(\mathbf{W}, \mathbf{Z}, \mathbf{x}, \mathbf{s}|\Psi) - \log q(\mathbf{Z}, \mathbf{x}, \mathbf{s}))d\mathbf{Z}d\mathbf{x}d\mathbf{s}, \tag{2.1}$$

where equality in (2.1) holds when the posterior over the hidden variables $q(\mathbf{Z}, \mathbf{x}, \mathbf{s})$ equals the true posterior $p(\mathbf{Z}, \mathbf{x}, \mathbf{s}|\mathbf{W})$. While the graphical model of IFTM in Figure 2.1 shows several missing arrows representing the conditional independence properties that exist between the hidden nodes $\{\mathbf{Z}, \mathbf{x}, \mathbf{s}\}$, when conditioned on the observed words $\mathbf{W}$, these hidden variables are no longer independent. Since IFTM has a mix of discrete and continuous-valued hidden variables, the true joint posterior $p(\mathbf{Z}, \mathbf{x}, \mathbf{s}|\mathbf{W})$ takes a really complicated form and computing the posterior exactly thus proves to be computationally intractable. We employ a mean-field approximation to approximate the joint posterior distribution with a variational posterior in a factorized form [Att00]: $p(\mathbf{Z}, \mathbf{x}, \mathbf{s}|\mathbf{W}) \approx \prod_n q(z_n)q(\mathbf{x})q(\mathbf{s})$. The problem now becomes one of finding, within such family of factorized distributions, the variational posterior that maximizes the lower bound of the data log-likelihood in (2.1). With $q(\mathbf{Z}, \mathbf{x}, \mathbf{s})$ now in a factorized form, the RHS of (2.1) becomes a strict lowerbound of the data log-likelihood and can be expressed as:

$$\begin{aligned}\log p(\mathbf{W}|\Psi) \geq \sum_n E_q[\log p(w_n|z_n, \beta)] + E_q[\log p(z_n|\mathbf{x})] + E_q[\log p(\mathbf{x}|\mathbf{s}, \mathbf{A}, \mathbf{\Lambda}, \mu)] + E_q[\log p(\mathbf{s})] \\ + \mathcal{H}[q(\mathbf{Z})] + \mathcal{H}[q(\mathbf{x})] + \mathcal{H}[q(\mathbf{s})] \quad = \quad \mathcal{F}. \qquad (2.2)\end{aligned}$$

Due to the normalization term in the softmax operation, the expectation term $E_q[\log p(z_n|\mathbf{x})]$ will be difficult to compute, regardless of the form of $q(\mathbf{x})$. We make use of convex duality and represents a convex function, i.e. $-\log(\cdot)$ function, as a point-wise supremum of linear functions as shown in Figure 2.2(a). In particular, the log normalization term is replaced with adjustable lower bounds parameterized by variational parameters ξ:

$$\log p(z_n = k|x) \geq x_k - \log \xi - \frac{1}{\xi}(\sum_l e^{x_l} - \xi). \qquad (2.3)$$

Since the logistic normal distribution is not a conjugate prior to the Multinomial, the form of the variational distributions $q(z_n)$, $q(\mathbf{x})$, and $q(\mathbf{s})$ need to be examined as follows.

1. $q(z_n)$ is a discrete probability distribution, whose parameters $q(z_n = k)$ are denoted as ϕ_{nk}.
2. $q(\mathbf{x})$: Under the diagonality assumption of $\mathbf{\Lambda}$ and the use of convex variational bound of the log-normalizer term in (4.4), the free-form maximization of $\mathcal{F}$ w.r.t $q(\mathbf{x})$ shows the variational posterior taking on a factorized form $q(\mathbf{x}) = \prod_k q(x_k)$. However, $q(x_k)$ obtained from the free-form maximization is not in the form of a distribution that we recognize. We thus approximate $q(x_k)$ as a Gaussian distribution: $q(x_k) \sim \mathcal{N}(x_k; \bar{x}_k, \gamma_k^{-1})$, with the mean parameter $\bar{x}_k$ and precision parameter γ_k.
3. $q(\mathbf{s})$: The free-form maximization of $\mathcal{F}$ w.r.t $q(\mathbf{s})$ results in $q(\mathbf{s}) \sim \mathcal{N}(\mathbf{s}; \bar{\mathbf{s}}, \mathbf{B}^{-1})$, where $\mathbf{B}$ is an $L \times L$ non-diagonal precision matrix.

Given the model parameters $\Psi = \{\mathbf{A}, \mu, \mathbf{\Lambda}, \beta\}$, the variational inference maximizes the lower bound of the data log-likelihood given in (2.2) w.r.t the variational parameters $\{\xi, \phi_{nk}, \bar{x}_k, \gamma_k, \bar{\mathbf{s}}, \mathbf{B}\}$. This culminates in a coordinate ascent algorithm, where we optimize one parameter while holding the rest of the parameter fixed. First, we maximize $\mathcal{F}$ w.r.t. ξ, ϕ_{nk}, and γ_k^{-1}, which attain the maximum at:

$$\xi = \sum_k e^{\bar{x}_k + \frac{0.5}{\gamma_k}}, \tag{2.4}$$

$$\phi_{nk} \propto \prod_t \beta_{kt}^{1(w_n=t)} \cdot \exp(\bar{x}_k), \tag{2.5}$$

$$\frac{\partial \mathcal{F}}{\partial \gamma_k^{-1}} = -\frac{N}{\xi} e^{\bar{x}_k} \cdot e^{\frac{\gamma_k^{-1}}{2}} - \lambda_k + \frac{1}{\gamma_k^{-1}} = 0. \tag{2.6}$$

Since there's no analytical solution available for (2.6), we use a Newton-Raphson algorithm to update γ_k.

Secondly, we maximize $\mathcal{F}$ w.r.t $\bar{x}_k$. This is where we improve significantly upon the variational inference of CTM. The choice of diagonal precision matrix $\mathbf{\Lambda}$, in effect, converts a K-dimensional optimization problem into K one-dimensional optimization problems which are easy to solve and extremely fast. To simplify the notation, we denote $n_k = \sum_n \phi_{nk}$ and $\mathbf{c} = \mathbf{A}\bar{\mathbf{s}} + \mu$. The derivative of $\mathcal{F}$ w.r.t. $\bar{x}_k$ can be written as:

$$\frac{\partial \mathcal{F}}{\partial \bar{x}_k} = \frac{n_k}{\lambda_k} - \frac{N}{\xi \lambda_k} e^{\bar{x}_k} \cdot e^{\frac{0.5}{\gamma_k}} - \bar{x}_k + c_k = 0. \tag{2.7}$$

As we can see, $\bar{x}_k$ that makes the gradient vanish cannot be obtained analytically. We thus employ a Newton-Raphson algorithm to find such $\bar{x}_k$. First, we rewrite (2.7) in the form:

$$t_k e^{t_k} = u_k, \tag{2.8}$$

where we now substitute $u_k = \frac{N}{\xi \lambda_k} e^{\frac{n_k}{\lambda_k} + c_k + \frac{0.5}{\gamma_k}}$ and $t_k = \frac{n_k}{\lambda_k} + c_k - \bar{x}_k$. We find that the newton algorithm derived from (2.8) (with proper initialization e.g. $t_k \approx \log u_k$) converges in just a few iterations.

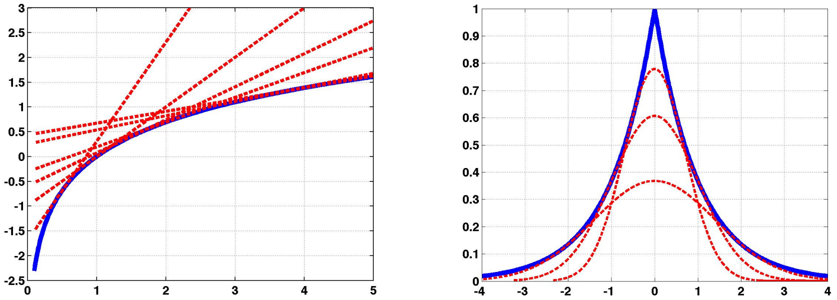
Lastly, we maximize w.r.t $\bar{\mathbf{s}}$ and obtain an analytic solution for the maximum at

$$\bar{\mathbf{s}} = \mathbf{B}^{-1} \mathbf{A}^\top \mathbf{\Lambda} (\bar{\mathbf{x}} - \mu), \tag{2.9}$$

where $\mathbf{B} = (\mathbf{A}^\top \mathbf{\Lambda} \mathbf{A} + \mathbf{I}_L)$ is the precision of the posterior $q(\mathbf{s})$. Note here that the update rule for $\mathbf{B}$ depends only on the model parameters $\mathbf{A}, \mathbf{\Lambda}$. $\mathbf{B}$ thus only needs updating in the M-step, avoiding the expensive matrix inversion in each variational inference step. Variational Inference for IFTM constitutes iteratively updating the variational parameters according to eqn (2.4)-(2.9) until convergence.

2.2.3 IFTM with Sparse source prior

In an attempt to give an interpretation to the individual sources $\mathbf{s}$, two main problems arise with IFTM that uses Gaussian source prior. First, the mixing matrix $\mathbf{A}$ learned from Gaussian source prior is often uninterpretable. The reason is that, under a Gaussian assumption, the sources $\mathbf{s}$, which control how the columns of $\mathbf{A}$ are combined together to form the correlated topic vector $\mathbf{x}$, are non-sparse. Therefore, to generate $\mathbf{x}$ for each document, all the columns of



(a)

(b)

Figure 2.2: Convex variational bounds used to simplify the computation in variational inference of IFTM. **(a)** $\log(\cdot)$ function (in blue) represented as pointwise infemum of linear functions (red lines). **(b)** Laplacian distribution (in blue) represented as pointwise supremum of Gaussian distributions with zero means and different variances.

the mixing matrix will be used. As a result, the individual columns of $\mathbf{A}$ do not carry meaningful patterns of topic correlations, while linear combinations of all the columns do. Second, as with the Factor Analysis model, the mixing matrix $\mathbf{A}$ when the source prior is Gaussian is identifiable only up to a rotation, since the log-likelihood remains unchanged when multiplying $\mathbf{A}$ with any arbitrary rotation matrix $\mathbf{Q} \in \mathbb{R}^{L \times L}$.

By assuming the independent sources are drawn from a sparse distribution, we remove nonidentifiability associated with rotations. In addition, a sparse distribution favors a representation of $\mathbf{x}$ that uses a small number of "active" sources for each document, allowing more interpretable correlation structures to emerge in the columns of the mixing matrix $\mathbf{A}$. In this work, we propose to model the independent sources as a Laplacian distribution given in the form:

$$p(\mathbf{s}) = \prod_{l}^{L} \frac{1}{2} e^{-|s_l|} \tag{2.10}$$

$$\log p(\mathbf{s}) = -\sum_{l} |s_l| - L \log 2. \tag{2.11}$$

Indeed, the choice of Laplacian distribution implies an L1-norm constraint on the solutions of $\mathbf{s}$. Unlike the L2-norm constraint of the Gaussian source case, L1 regularization penalizes the configurations that distribute activities among many sources, while encourages the configurations which a few sources are actively used to explain correlations between topics.

Variational Inference

Variational inference for IFTM with Laplacian source prior proceeds similarly to the Gaussian case. Again, we adopt a factorized variational posterior $p(\mathbf{Z}, \mathbf{x}, \mathbf{s}|\mathbf{W}) \approx \prod_n q(z_n) q(\mathbf{x}) q(\mathbf{s})$. Unlike the Gaussian case, however, the form of Laplacian in (2.11) renders the posterior distribution $q(\mathbf{s})$ to be in the form that is not easy to recognize and thus further approximation of

$q(\mathbf{s})$ is required. To this end, we adopt the convex variational approximation that replaces the Laplacian source distribution with an adjustable lower bound as used in [Gir01]. By proving that $\log p(s) = -\sqrt{s^2}$ is convex in s^2 (square-convex), we can express the dual representation of $\log p(s)$ in the form of a pointwise supremum of functions of s^2, and in the process introduce a variational parameter η, which will be optimized out. Dropping the supremum, we obtain the following lower bound, as a function of the adjustable parameter η. For more details, refer to [Jaa97, Gir01].

$$-\sum_l |s_l| \geq -\sum_l (\frac{|\eta_l|}{2} + \frac{s_l^2}{2|\eta_l|}) \tag{2.12}$$

The dual form representation of the log prior of Laplacian source distribution in (2.12) takes a quadratic form, which, in essence, expresses the Laplacian distribution in terms of adjustable lower-bounds in the Gaussian form as seen in Figure 2.2(b). This dual representation allows the variational inference algorithm for the Laplacian source prior to be derived as a slight modification of the Gaussian case. In particular, the variational updates for $\{\xi, \phi_{nk}, \bar{x}_k, \gamma_k\}$ remain the same as in (2.4)-(2.7). The variational posterior over sources $\mathbf{s}$ also conveniently takes a Gausisan form: $q(\mathbf{s}) \sim \mathcal{N}(\mathbf{s}; \bar{\mathbf{s}}, \mathbf{B})$, but now with the update for the precision matrix $\mathbf{B}$ that depends on the variational parameter η.

$$\bar{\mathbf{s}} = \mathbf{B}^{-1}\mathbf{A}^\top \mathbf{\Lambda}(\bar{\mathbf{x}} - \mu) \tag{2.13}$$

$$\mathbf{B} = \mathbf{A}^\top \mathbf{\Lambda}\mathbf{A} + \text{diag}(\frac{1}{|\eta|}) \tag{2.14}$$

$$|\eta_l| = \sqrt{E[s_l^2]} \tag{2.15}$$

where $E[\mathbf{s}\mathbf{s}^\top] = \bar{\mathbf{s}}\bar{\mathbf{s}}^\top + \mathbf{B}^{-1}$ and $\text{diag}(\frac{1}{|\eta|})$ denotes a diagonal matrix with $|\eta_l|$ in entry (l, l).

Parameter Estimation

To update the model parameters $\Psi = \{\mathbf{A}, \mathbf{\Lambda}, \mu, \beta\}$, we maximize the lower bound of the log-likelihood in (2.2) w.r.t. Ψ and obtain the following closed-form updates:

$$\mathbf{A} = \mathbf{R}_{xs} \cdot \mathbf{R}_{ss}^{-1} \tag{2.16}$$

$$\mu = \frac{1}{D}(\sum_d \bar{\mathbf{x}}_d - \mathbf{A}\sum_d \bar{\mathbf{s}}_d) \tag{2.17}$$

$$\mathbf{\Lambda}^{-1} = \text{diag}(\frac{1}{D}(\mathbf{R}_{xx} - \mathbf{A}\mathbf{R}_{xs}^\top) + \frac{1}{D}\sum_d \mathbf{\Gamma}_d^{-1}) \tag{2.18}$$

$$\beta_{kt} = \frac{\sum_{d,n} \phi_{nk}^d 1(w_n^d = t)}{\sum_{t,d,n} \phi_{nk}^d 1(w_n^d = t)} \tag{2.19}$$

where the sufficient statistics of the variational posteriors are defined as follow: $\mathbf{R}_{xx} = \sum_d (\bar{\mathbf{x}}_d - \mu)(\bar{\mathbf{x}}_d - \mu)^\top$; $\mathbf{R}_{xs} = \sum_d (\bar{\mathbf{x}}_d - \mu)\bar{\mathbf{s}}_d^\top$; $\mathbf{R}_{ss} = \sum_d (\bar{\mathbf{s}}_d\bar{\mathbf{s}}_d^\top + \mathbf{B}_d^{-1})$.

2.3 Experimental Results

In this section, we show several benefits of IFTM over previously proposed topics models like CTM and LDA. First, we show how we can directly visualize meaningful patterns of topic correlations and build a topic graph, using a sparse source prior model of IFTM. Second, we compare the performance of IFTM and CTM using 2 metrics: held-out data likelihood and perplexity over held-out words. With similar performance between the 2 models, we show significant differences between computational intensity required by the 2 algorithms. We also present results on real-world applications, comparing IFTM and LDA on a document retrieval task, to showcase the benefits of topic models that can capture correlations between topics.

2.3.1 Correlated Topics Visualization

One of the main motivations in developing statistical topic models is the use of such models in exploratory analysis of text document collections. Given a large archive of documents, the goal is to use topic models to automatically explore and discover hierarchies and hidden topical structures in the archive, with the ultimate goal of using such information to build a browsing interface or navigational tool that allows users to explore documents in the collection based on their topics. In using LDA to build such a navigational tool, despite being able to discover a set of dominant topics contained in a document collection, LDA lacks a mechanism to learn the relationships between these topics, which is important in building a recommendation system to users. By replacing a Dirichlet prior in LDA with a logistic normal distribution with full covariance structure, CTM is able to learn statistics of topic co-occurrences in a document archive with a more powerful prior model for θ. While such an approach is indeed straightforward, the covariance parameter Σ only describes pairwise relationships between topics, which is often difficult to visualize and interpret. In [BL07], a separate sparse regression step is therefore needed to generate a topic graph which shows relations between groups of topics that are activated together in the archive.

Unlike the approach of CTM, IFTM postulates the existence of another layer of hidden variables corresponding to the independent sources that are linearly mixed together (using the mixing weights given in $\mathbf{A}$) to form a correlated topic proportion vector for each document. With the explicit assumption that the sources $\mathbf{s}$ are independent, patterns of topic co-occurrences emerge in the mixing matrix $\mathbf{A}$. Assuming that only one source is allowed to be active at a time (the other sources provide no contribution), we can indeed visualize the pattern of topic correlation each source represents by directly examining the corresponding column of the mixing matrix $\mathbf{A}$. More specifically, for each source, we rank all the topics according the weight values in the corresponding column of the mixing matrix $\mathbf{A}$ in ascending order to discover a pattern of topic co-occurrences. By representing each source using the most likely topics and repeating this step for all columns of $\mathbf{A}$, we can construct a simple topic graph that shows the most dominant

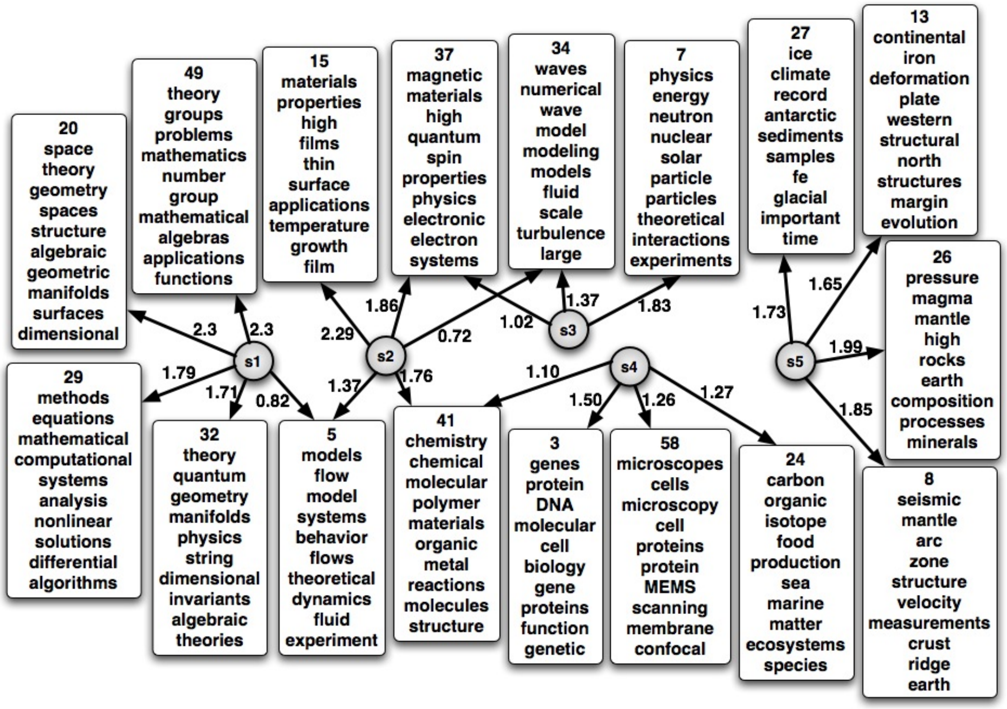


Figure 2.3: Visualization and Interpretation of sources of topic correlations on 4000 NSF abstracts data. For each source, we rank the corresponding column of **A** and show the most likely topics to co-occur when the source is "active". A source is shown as a circle, with arrows pointing to the most likely topics. The number on each arrow corresponds to the weight given to the topic.

patterns of topic co-occurrences in the archive.

To give a concrete example, we learn IFTM with Laplacian sources, using K=60 and L=10, on a corpus of 3,946 NSF abstracts[1] submitted in 2003. After removing function words and common/rare words, this corpus has vocabulary size 5261 words, with an average of 120 words per document. Figure 2.3 shows 5 independent sources—s_1, s_2, s_3, s_4, and s_5 denoted by the 5 shaded circles at the center of the figure. For each source, we show 3-5 most likely topics to co-occur when the source is active. Each topic is represented by choosing the 10 most likely words from the topic-specific Multinomial parameters over words. As seen in Figure 2.3, s_1 represents a grouping of topics corresponding to different branches of mathematics, e.g. Topic 20 is geometry, Topic 49 is algebra. s_2 groups the topics related to material science together. s_3 represents a group of topics from various areas of physics. s_4 represents a grouping of topics under the subjects biology and chemistry (biochemistry), while s_5 shows a grouping of topics related to earth, atmospheric, and planetary sciences. Some of the topics are shared by many independent sources. For example, Topic 37 discusses the physical and magnetic properties of the material, thus Topic 37 is likely under the subjects of material science (s_1) and physics (s_2).

[1] UCI KDD archive: http://kdd.ics.uci.edu/

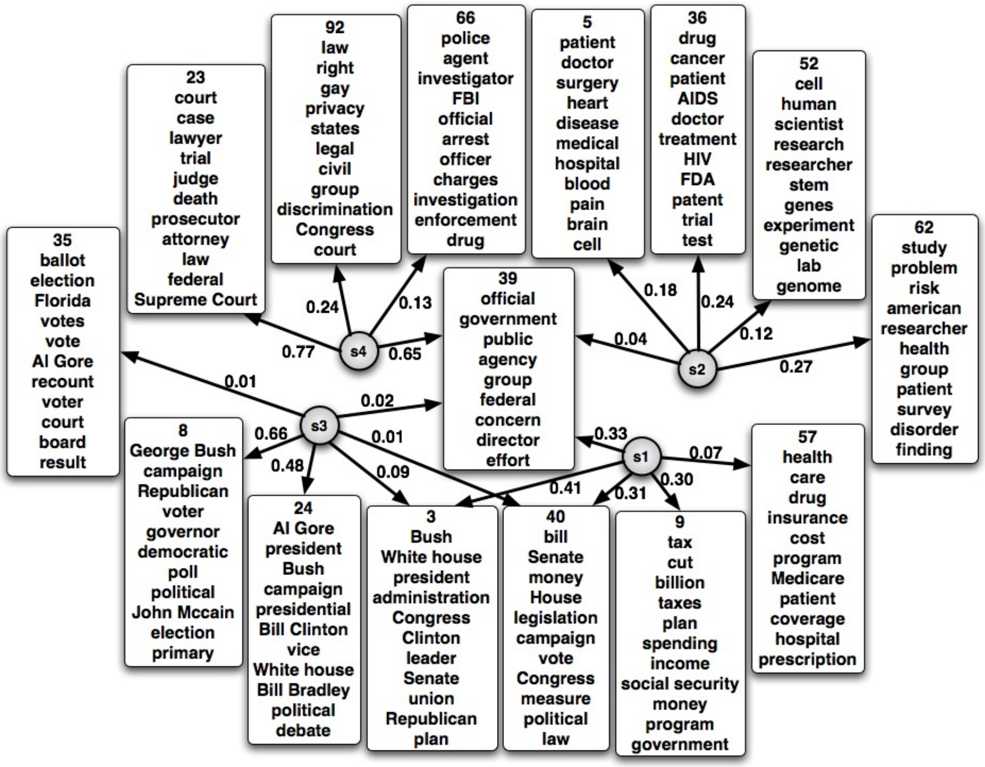


Figure 2.4: Visualization of 4 sources of topic correlations for 15000 NYTimes articles.

We show another example of topic co-occurrences that IFTM can capture in Figure 2.4. We run IFTM with Laplacian sources with $K = 150$ and $L = 30$ on 15000 articles randomly selected from 300000 articles in the UCI NYTimes[2] archive. After removing common and rare words, this dataset has a vocabulary size of 47996 words, with an average of 336 words per document. Shown in Figure 2.4 are 4 sources—s_1, s_2, s_3, and s_4 that share Topic 39 (government) in common. s_1 corresponds to grouping of topics related to legislative side of government, i.e. tax cut bill, health care bill. s_2 represents a group of topics most likely from the health section of the NYTimes archive. s_3 represents a grouping of topics related to the 1992 election, most likely from the politics section of the newspaper archive. s_4 groups together topics related to different aspects of law enforcements, e.g. police, FBI agents, court trials. Since these 4 sources share Topic 39 (government) in common, these sources thus represents different aspects or branches of government: s_1 shows the legislative branch of government, s_2 represents the health related role of government (FDA, Medicare), s_3 shows the politics side of government, and s_4 corresponds to the law enforcement aspect of government.

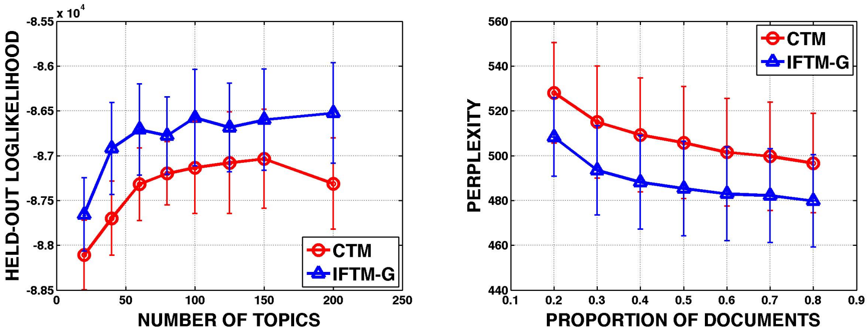
[2] UCI KDD archive: http://kdd.ics.uci.edu/

2.3.2 Model Comparison

Synthetic Data

Since IFTM with Gaussian source prior model (denoted by IFTM-G) can be seen as a special case of CTM with a constrained covariance structure when $L < K$, we first compare the performance of IFTM-G and CTM using simulated data generated according to the generative model of CTM (with full covariance structure). In particular, we sample 1,000 documents with an average of 80 words per document. The CTM model parameters $\{\mu, \mathbf{\Sigma}, \beta\}$ used to generate this dataset are drawn randomly from some distribution, with $K = 300$ and the vocabulary size $T = 625$. Unless specified otherwise, the number of sources L used for IFTM-G and IFTM-L is set to $\frac{K}{4}$. 800 documents are used for training, 200 for testing, with 5-fold cross validation. We run variational inference and EM until the relative change in the likelihood bound falls below 10^{-5}.

We evaluate the generalization ability of the model to explain unseen documents by using log-likelihood over held-out documents as a performance measure. Following the work in [BL06a], the log-likelihood of test documents is computed by employing importance sampling that uses the variational posterior as the proposal distributions. We are interested in observing how the 2 models perform as we increase the number of hidden topics K. As seen from the left panel of Figure 2.5, IFTM-G gives higher likelihood than CTM on all values of K. When K is small, the difference between the 2 models seems smaller, but as K increases their difference gets magnified. CTM clearly overfits the data as the likelihood drops after K=150. Since IFTM-G has much fewer parameters to fit, it is able to support higher model complexity (larger numbers of topics).



(a)

(b)

Figure 2.5: Results on Synthetic Data comparing CTM and IFTM with Gaussian source prior. The data is synthesized according to the CTM generative process. We compare the 2 models on 2 metrics **(a)** Held-out log-likelihood as a function of the number of topics, and **(b)** Predictive perplexity as a function of the proportion of test documents observed. CTM overfits more easily with more parameters to be learned, and as a result yields worse performance than IFTM-G.

Another important metric to compare the 2 models is how well they can predict unseen words in a test document, given that some portions of the words are observed. We conduct the following experiment where each test document is divided into 2 parts—$a\%$ of the words will be observed, while the rest will be unobserved. We employ a predictive perplexity score given below which measures how well the models can use the observed words to predict the unseen words. To compute $\log p(\mathbf{W}^{\text{mis}}|\mathbf{W}^{\text{obs}})$ for IFTM-G and CTM, we run variational inference on the observed portions of the test documents until convergence, and use the inferred variational posterior to predict the unseen words. As we increase the proportion of test documents that are observed, we expect the perplexity score to decrease, since the inferred variational posterior of the topic proportion should become more accurate as more words are observed.

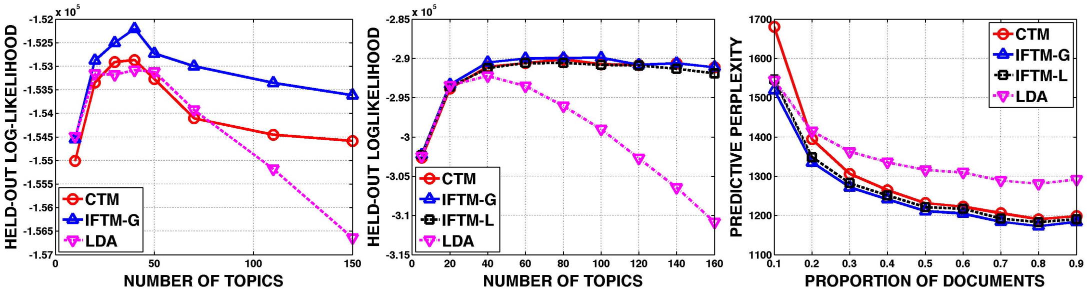
$$\text{Perp}(\mathbf{W}^{\text{mis}}|\mathbf{W}^{\text{obs}}) = \exp\left(-\frac{\sum_d \log p(\mathbf{W}_d^{\text{mis}}|\mathbf{W}_d^{\text{obs}})}{\sum_d N_d}\right).$$

Using the above described synthetic data, we obtain the plots shown on the right panel of Figure 2.5. We compare perplexity of LDA, CTM, and IFTM-G using $K = 100$. The perplexity score obtained from IFTM-G is lower than CTM by 20 words, since CTM overfits the data due to a much larger number of parameters that needs to be learned. Nonetheless, both CTM and IFTM-G outperform LDA (not shown in graph). This is to be expected because the data is generated according to the CTM generative model.

NSF Abstract Data

In this section, we show how IFTM with Gaussian and Laplacian source prior models perform comparatively to CTM and LDA, on real data. To this end, we train the 4 models on the corpus of 3,946 NSF abstracts (used in previous section), using 90% of the data for training and 10% as a held-out test set. 4-fold cross-validation is used and the results are averaged. Performance is measured by log-likelihood over held-out dataset. To avoid comparing the likelihood bounds of the 4 models, we again employ importance sampling that uses the variational posterior as the proposal distributions. In particular, for IFTM with Laplacian source prior, since $\log p(\mathbf{x})$ cannot be computed analytically, we sample $\mathbf{s} \sim q(\mathbf{s})$ first to compute $\log p(\mathbf{x})$, which is then used in estimating the true log-likelihood $\log p(\mathbf{W})$.

First we show the held-out likelihood comparing CTM, LDA, and IFTM-G on a 1500-document subset of the 4000-document archive in Figure 2.6 (a). LDA performance peaks at $K = 40$ but as we increase the number of topics, the performance drops dramatically. This is due to the nearly independent assumption of the topic proportion generated from a Dirichlet. As K increases, more topics will likely become correlated, and the Dirichlet distribution will no longer be a good fit for such topic proportions. On the contrary, topic models that capture correlations between topics, i.e. IFTM and CTM, are able to support larger numbers of topics. While IFTM and CTM also peak at 40 topics, compared to LDA, these 2 models give higher likelihood when K is large. When comparing CTM and IFTM on this dataset, we find that the



(a) (b) (c)

Figure 2.6: Results on NSF Abstracts Data. **(a)** Held-out averaged log-likelihood score as a function of the number of topics (K) using a smaller NSF dataset with 1500 documents. **(b)** Held-out averaged log-likelihood as a function of K on a 4000-document NSF dataset. **(c)** Predictive perplexity as a function of the percentage of observed words in test documents, comparing the 4 models when $K = 60$.

2 models yield similar log-likelihood when K is small, but as K increases IFTM generalizes to the unseen data better as CTM is found to overfit. In Figure 2.6 (b), we increase the training set size to the full 4000 NSF documents. With more than twice the data than in Figure 2.6 (a) CTM performs competitively with IFTM (for both Gaussian and Laplacian source priors) as seen in the comparable held-out likelihood scores.

Figure 2.6 (c) shows the predictive perplexity as a function of percentage of observed words in test documents, for $K = 60$. Since perplexity is the inverse of the data likelihood, the lower number means the model can generalize better to unseen words in the documents. When only a small portion of the test documents is observed, using statistics of topic correlation, IFTM and CTM can use more topics than the ones that actually occur in the observed portion of the test documents to explain the unseen test words. As seen in Figure 2.6 (c), IFTM and CTM both give lower perplexities than LDA which lacks a mechanism to learn correlations between topics. IFTM-G and IFTM-L give lower perplexity than CTM by almost 200 words, when only 10% of the words are observed, since CTM overfits the data and the parameter estimates have large variances. The difference between the 3 models, however, becomes smaller as more words are observed and the inferred posterior become more accurate.

Indeed, with comparable performance, CTM is found to be much more computationally demanding than IFTM. Table 2.1 shows the averaged training time required to train the 4 models in Figure 2.6. All of our simulations run on an Intel™Q6600 quad-core computer (one core is used for each algorithm). We use fairly optimized and comparable C implementation of the 4 models. The CTM implementation used in our experiment is obtained directly from the CTM's authors' website. On average, if IFTM-G requires 1 day to train, without the availability of distributed computing, CTM will take 5 days to train, which is quite prohibitive for practical applications. Note that IFTM-L is computationally more expensive than IFTM-G. This is due to

Table 2.1: Training time (in hours) as K increases, comparing LDA, IFTM with Gaussian prior (IFTM-G), IFTM with Laplacian prior (IFTM-L), and CTM.

K	LDA	IFTM-G	IFTM-L	CTM
20	0.546	0.878	1.177	3.648
40	1.108	1.833	4.033	13.973
60	2.795	3.861	7.733	22.551
80	3.651	8.705	13.030	43.156
100	4.147	9.296	18.599	53.376
120	4.840	13.836	19.963	65.446
140	6.521	17.340	22.946	70.833
160	10.173	20.287	25.523	90.900
	×**0.45**	×**1**	×**1.5**	×**5**

the new dependency between $\mathbf{B}$ and the convex variational parameter η in (2.14), which requires $\mathbf{B}$ to be inverted every time η and $\bar{\mathbf{s}}$ are updated.

2.3.3 Document Retrieval

In this section, we demonstrate the performance of IFTM in real world applications. One application that can directly benefit from statistics of topic correlations is the process of query expansion in document retrieval. Given a set of query words, the goal is to retrieve documents in the database that are most relevant to the query. There are indeed cases of relevant documents that might not contain the specific query words but instead contain similar words or synonyms of the query words. In such a scenario, it is beneficial to perform query expansion to retrieve those additional relevant documents. One form of query expansion is to retrieve documents using the inferred topics of the query words instead of directly matching the query words to documents in the dataset. Now with statistics of topic correlations, we can further expand the query words by using additional topics that are known to be highly correlated with the actual topics inferred from the query documents. With such expanded queries, additional relevant documents can be retrieved.

To this end, we are interested in comparing IFTM to LDA in a document retrieval task on 20 Newsgroup dataset. This version of the 20 Newsgroup[3] contains 19,949 messages from 20 Usenet newsgroups (each class contains $\approx$ 1000 documents). With function words, and rare and common words removed, we are left with 43,586 words in the vocabulary. Using 70 topics, we train 3 models: (*i*) IFTM with Gaussian source prior, (*ii*) IFTM with Laplacian sources, and (*iii*) LDA on 16,000 training documents, with 3,949 documents used for testing. 5-fold cross validation is used.

Given the query, the task of a retrieval system is to return items in the database that is closest in some distance measure to the query. To use LDA and IFTM for a retrieval task, we

[3] http://shi-zhong.com/research

use the following method. First, we train the models on a training set with 16000 documents. For each query document, we run variational inference until convergence, and use the variational posterior—$q(\mathbf{x})$ for IFTM and $q(\theta)$ for LDA—to compute the "distance": $\log p(\mathbf{W}^{\text{train}}|\mathbf{W}^{\text{query}})$ between a document in the training set to the query. Indeed, this distance metric is directly related to the perplexity score, which measures how well the words in the query can predict the words in the training set. Using such a distance measure, IFTM has a built-in advantage over LDA when presented with short queries, since IFTM can draw upon other topics, known to be correlated with the topics inferred from the query words to explain the training documents.

The performance of our retrieval system is measured by a precision-recall curve. Precision measures the percentage of relevant items in the returned set, while recall is the percentage of all relevant documents that gets returned. In this case, if a query belongs to class 1, then all the other documents under the same class are considered relevant in the returned set. As expected, Figure 2.7 shows that IFTM with both Gaussian source prior and Laplacian source prior give higher precision values at the same recall rate, as compared to LDA.

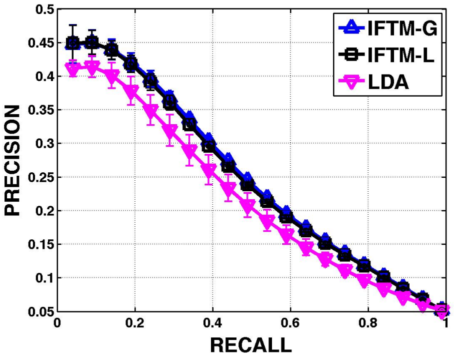


Figure 2.7: Precision-recall curve comparing LDA, IFTM-G, and IFTM-L on 20 NewsGroup dataset. The distance measure used in comparing the query document to a document in the training set is $p(\mathbf{W}^{\text{train}}|\mathbf{W}^{\text{query}})$

2.4 Appendix A

We briefly derive the convex variational bound used in (2.12). We refer interested readers to [JJ97, JJ00, Jaa97, Roc70] for more rigorous treatments of convex duality.

A convex function $f(s)$ can be represented as pointwise supremum of linear functions (dual representation). That is:

$$f(s) = \sup_{\xi}\{\xi s - f^*(\xi)\}$$
$$f^*(\xi) = \sup_{s}\{\xi s - f(s)\}.$$

A function $f(s)$ is a convex function in s^2, thus its dual representation has the form:

$$f(s) = g(s^2) = \sup_{\xi}\{\xi s^2 - g^*(\xi)\} \tag{2.20}$$

$$g^*(\xi) = \sup_{s}\{\xi s^2 - g(s^2)\} \tag{2.21}$$

For a given ξ, the point $s = \tilde{s}$ that attains the maximum of $g^*(\xi)$ in eqn (2.21) must satisfy: $\xi = \frac{\partial f}{\partial s^2}|_{s=\tilde{s}}$. We can rewrite $f(s)$ in eqn (2.20) as a function of the new variational parameter $\tilde{s}$ as follows:

$$f(s) = \sup_{\tilde{s}}\{f(\tilde{s}) + \left.\frac{\partial f}{\partial s^2}\right|_{s=\tilde{s}}(s^2 - \tilde{s}^2)\}.$$

Dropping the supremum, we obtain the following bound on $f(s)$:

$$f(s) \geq f(\tilde{s}) + \left.\frac{\partial f}{\partial s^2}\right|_{s=\tilde{s}}(s^2 - \tilde{s}^2). \tag{2.22}$$

Now consider a function $f(s) = \log\cosh^{-\frac{1}{\beta}}(\beta s)$, which is convex in s^2. We can apply the above convex bound and, in the process, introduce an adjustable variational parameter ξ (which will be optimized out) as follows:

$$\log\cosh^{-\frac{1}{\beta}}(\beta s) \geq \log\cosh^{-\frac{1}{\beta}}(\beta\xi) - \frac{\tanh(\beta\xi)}{2\xi}(s^2 - \xi^2)$$

By taking the limit of $\beta \to \infty$ of the above bound and make use of the following identities: $\lim_{\beta\to\infty} \log\cosh^{-\frac{1}{\beta}}(\beta s) = -|s|$, and $\lim_{\beta\to\infty} \tanh(\beta\xi) = \text{sign}(\xi)$ we obtain the following bounds:

$$\begin{aligned} -|s| &\geq -|\xi| - \frac{\text{sign}(\xi)}{2\xi}(s^2 - \xi^2) \\ -|s| &\geq -\frac{|\xi|}{2} - \frac{s^2}{2|\xi|}. \end{aligned}$$

Chapter 2, in part, is a reprint of the material as it appears in: D. Putthividhya, H. T. Attias, S. Nagarajan, “Independent Factor Topic Models,” in International Conference on Machine Learning (ICML) 2009. I was the primary researcher of the cited materials and the co-author listed in these publications supervised the work which forms the basis of this chapter.

Chapter 3

Statistical Topic Models for Image and Video Annotation

This chapter presents an overview of 2 novel statistical topic models for image and video annotation which will be examined in detail in Chapter 4 and 5. We first describe the problem of image and video annotation and give a brief literature review of related work. We then briefly introduce the 2 novel probabilistic topic models proposed in this work. First, we present topic-regression multi-modal Latent Dirichlet Allocation (tr-mmLDA), a powerful statistical topic model that uses a latent variable regression approach to capture the conditional relationship between annotation texts and the corresponding image or video features. Second, we present supervised Latent Dirichlet Allocation with multi-variate binary response variable (sLDA-bin), which extends the basic supervised LDA model to handle binary annotation data modeled as a multi-variate Bernoulli variable. We describe the details of 2 standard image annotation datasets used in performance evaluation: COREL and LabelMe datasets. For video annotation experiments, we use a dataset of television program recordings where each video clip contains video, audio, and the corresponding closed captions (we discard the audio modality for the moment). We discuss the representations for images, video, and caption texts, as well as the choices of image and video features used in the 2 proposed models. In evaluating annotation system performance, we adopt 2 metrics: caption perplexity in [BJ03] which measures how well the held-out caption words can be predicted from the test image and a precision-recall metric which measures an auto-annotation performance when treating annotation as class labels.

3.1 Introduction

Today's advanced digital media technology has led to explosive growth of multimedia data in a scale that has never occurred before. The availability of such large-scale quantities of multimedia documents prompts the need for effective and efficient algorithms to search and index multimedia files. Traditional approaches in multimedia retrieval have mainly focused on query-by-example systems, also known as content-based retrieval, where users submit an image or video query, and the system retrieves an item in the database closest in some distance measure to the query. While many algorithms have been developed along this line of work (see [SWS$^+$00] for an extensive review), users often find providing example queries a cumbersome way of interfacing with retrieval systems. As today's multimedia content becomes increasingly multi-modal with texts accompanying images and videos in the form of content description, transcribed text, or captions, the new trend in multimedia search technology advocates the use of collateral annotation texts to identify and retrieve images and video. Besides the fact that users prefer textual queries over examples, an important benefit of a text-based approach is the high-level semantic retrieval, e.g. retrieval of abstract concepts, that could not be achieved with low-level visual cues used in most query-by-example systems.

With annotation texts playing an increasingly vital role in modern multimedia retrieval systems, a relevant question one might ask is how to deal with numerous fast-growing user-generated content that often lacks descriptive annotation texts which would enable accurate semantic retrieval to be performed. Indeed, the traditional solution is to employ manual labeling—a tedious process that is costly, error-prone, and unscalable to large-scale repositories. While there has been interesting research in developing labeling games to motivate users to have fun while challenging them to correctly annotate image and video content [vAD04, vA06, CMS07], ultimately the real challenge is how to sidestep human intervention altogether and develop automated tools that can automatically generate semantic descriptors of multimedia content—automatic annotation systems. Given a training set comprising images or videos together with their corresponding annotation words, the goal of an automatic annotation algorithm is to learn the underlying patterns of image-text, video-text association so that when presented with an image or video without its accompanying annotation texts, the system can use the fitted model to accurately predict the missing annotation.

Previous work on multimedia annotation can be broadly classified into 2 groups. The first line of work treats annotation words as target variables to be predicted and formulate the problem of automatic annotation as a supervised learning problem. In such a framework, one directly models the relevant conditional distribution of target (annotation texts) given the input variables (image or video features). With annotation words modeled as discrete valued categorical variables, the multi-class classification framework is adopted [LW03, CV05, CCMV07]. For each word in the vocabulary, the class-conditional density is learned from all the images (or videos) tagged with that word. During annotation, the posterior over class labels is computed and ranked, and the top concepts with the highest probabilities are used as predicted annotation. Despite obtaining a good predictive performance, one serious drawback of the classification approach to annotation is the lack of a mechanism to capture word co-occurrences that can be useful in prediction. As we attempt to model annotation words beyond a small set of keywords or concepts, i.e. free-form texts, the ability to learn word semantics based on statistics of word co-occurrences in the collections will prove to be essential in a successful auto-annotation system.

Another trend in multimedia annotation considers probabilistic models with latent variables to explicitly capture the joint statistical associations between image/video and their corresponding annotation texts. In such a framework, by postulating the existence of a small set of shared hidden factors that are the underlying causes of cross-correlations between data types, the goal is to infer latent variable representations of images or videos that are predictive of annotation texts. While some successes have been reported in modeling image-caption data for automatic image annotation, see [BDdF$^+$03, BJ03, LMJ03], empirical results in [BJ03] show that a latent variable model that maximizes the joint distributions in the data does not always translate to a good model for predicting one data type from another. The problem indeed lies in the unsupervised nature of these models. Unlike in the supervised setting, this line of approach models the joint correlations in the multi-modal data and do not allocate enough modeling power to capture the precise conditional relationships between data types that are important for prediction. While models that learn the joint correlations serve as good representations in important tasks such as multimedia classification, when the goal is prediction, a representation that better models the precise relations or mapping from the input variables to target variables will indeed be a better predictive model.

3.2 Overview of Proposed Models

Drawing from the strengths and weaknesses of the 2 paradigms of previous work, we explore a new annotation model that attempts to combine the best of both worlds. From the latent variable framework, we maintain the use of latent variables to explicitly capture correlation structures in the data, while from the supervised learning framework, we adopt the approach of directly modeling the conditional relationship between the input (image or video modality) and target variables (annotation words). Combining the benefits of both approaches, we present a latent variable regression approach that incorporates a regression framework into latent variable models. Instead of having a set of latent variables shared between the two data types, we propose the use of 2 latent variable models (one for each data modality) and introduce a regression module to learn the precise relation between the 2 sets of latent variables. In line with the goal of prediction, the regression module allows latent variables of the target modality to be predicted from latent variables of the input modality. Our approach therefore strikes a delicate balance between learning a good predictive model and a good representation for multi-modal data.

Indeed, by using a regression module to capture precise relationships and allowing regression coefficients to be learned from the data, our framework can more accurately adjust to the varying degrees of correlations that exist between modalities. When the 2 data types are independent, all the regression parameters will be driven close to 0 and our model reduces to learning 2 separate latent variable models, one for each modality. As the dependencies between the 2 data types grow, the regression parameters of the latent variables with large influence on the other data type will take values further away from 0. Indeed, the naive sharing of latent variables between 2 data modalities can be seen as a special case of our latent variable regression model with non-zero coefficients restricted to only the diagonal entries. We believe that a more powerful framework that can more accurately capture the true complex relations between data in different modalities will indeed lead to a better performance in a prediction task.

While our latent variable regression framework is not specific to a particular choice of latent variable models used, in order to develop a concrete annotation model we need to specify the choices of latent models in each modality. For text modeling, we consider a family of statistical topic models such as Latent Dirichlet Allocation (LDA), Correlated Topic Models (CTM), and Independent Factor Topic Models (IFTM) which have been shown to learn meaningful latent structures corresponding to topics and groups of topics from text document collections [BNJ03, BL07, PAN09]. For image or video data, while there exists a great variety of specialized models that can be tailored to suit the unique statistical structures of each data type, we make a case for a single unified representation to allow for a simple integration into our latent variable regression framework. In this work, we propose to delegate the task of capturing important and relevant statistics of image and video modalities to the appropriate choices of features. In doing so, we free the models from learning such structures and allow for an exploration of simpler density models, e.g. mixture of Gaussians, mixture of Laplacians [PL04, PL06]. Indeed, inspired by the recent successes of topic models in image modeling and scene classification [FfP05, SRE$^+$05, QMO$^+$05, CFf07, VT07], we adopt statistical topics models for image and video representations. Our framework thus unifies the representation of all data modalities considered in this work.

In chapter 4, we present topic-regression multi-modal Latent Dirichlet Allocation (tr-mmLDA), a novel statistical topic model for image and video annotation. At the core of tr-mmLDA lies the latent variable regression approach that learns the precise mapping from latent variables of the input modality (image or video features) to those of the target modality (annotation texts). For each data modality, we learn a separate topic model, and in order to capture the conditional relationship between different data types, we introduce a regression module which allows one set of latent topics to be predicted from the others. More specifically, we adopt a linear regression model where the proportion of topic variable for annotation texts is the target variable and is modeled as a (Gaussian) noise corrupted version of a linear combination of the topic proportion variable of the image modality. Inspired by the good predictive performance of supervised Latent Dirichlet Allocation (sLDA) model [BM07], we adopt the empirical image topic frequency as the covariates to ensure that only the topics that actually occur in the image modality are used in predicting caption texts. We derive an efficient variational inference algorithm using a mean-field approximation and a convex variational bound to handle intractable posterior computations.

In many annotation datasets that we encounter, caption data act more similarly to class labels in that the words either appear once in the captions or not at all. When combined with the fact that only a handful number of caption words are available per image, statistical topic models, which allow words in the same document to be allocated from multiple hidden topics, might indeed be overly complex for modeling such documents. In chapter 5, we present a novel annotation model to handle such binary annotation data. Our proposed model which we call supervised Latent Dirichlet Allocation model for multi-variate binary response variable (sLDA-bin) extends the basic supervised Latent Dirichlet Allocation (sLDA) model proposed in [BM07] to account for binary response variables. By modeling annotation words as a multivariate Bernoulli variable, we introduce, for each word, a logistic regression module that maps the empirical image topic frequency to the target binary annotation data. The use of logistic link function makes the computation for inference and parameter learning intractable. We therefore derive an approximate inference algorithm based on mean-field approximation and adopt a tight convex variational bound for the logistic function, similar to the derivation of the logistic PCA model in [SSU03].

3.3 Related Work

In modeling statistical association between image features and caption texts for the task of image annotation, two main sets of techniques have been proposed. In the first line of work, the problem of image annotation is cast as a supervised learning problem where annotation words are treated as concept classes [LW03, CV05]. For each word in the vocabulary, the class conditional density—modeled using mixture of Gaussians in [CV05] and 2-dimensional multi-scale Hidden Markov Models in [LW03]—is learned from *all* the images annotated with that word. To assign a word to an un-annotated image, the posterior over class labels for each concept is computed. The top concepts that yield highest posterior probabilities are chosen. This line of approaches suffers a scalability issue, since the class-conditional density must be learned for each word. In practical situations, these models are capable of handling only a small vocabulary, e.g. a set of generic keywords of size 300-600 words as available in the COREL database (and not 50,000 words in the English language vocabulary).

Another set of techniques proposed to model images and text treats text and visual features on a more equal footing. These models learn the joint probabilistic distribution of text and image features by assuming that for each document there is a hidden factor that governs the association between the image features and the corresponding caption words, see [TSK01, BF01, BDdF$^+$03, JLM03, LMJ03, FML04, VH06, ZZL$^+$06]. Several variations of the same word-image association model are presented with different forms of probability distribution assumed for the word distribution (multinomial vs. bernoulli distribution) and different image feature modeling (non-parametric kernel density estimation vs. mixture of Gaussians).

Indeed, strong annotation performance on the COREL dataset was reported using the relevance models proposed in [LMJ03, FML04]. These models are non-parametric and each image-caption pair in the training set is treated as a prototype to be compared (there are as many hidden factors as there are training samples). For each image-caption pair in the training set, image features are extracted and fitted using a non-parameteric kernel density model, while caption words are modeled as either a Multinomial [LMJ03] or a multi-variate Bernoulli distribution [FML04]. Given a test image, annotation words are chosen by averaging over all the caption models of the training samples weighted by how well the density models of the training images fit the test image. Despite excellent annotation performance on the COREL dataset, this type of non-parametric association models are not amenable to large-scale datasets often found in real-world applications.

The work of [DBdFF02, DW03, BJ03] addresses the problem of image *region* annotation, where each annotation word is assumed to be associated with a specific region in the image. In [DBdFF02], the problem of region annotation is formulated as a classic machine translation problem. Images are first segmented into regions, where each region is assigned to one of the region vocabulary. These image regions and the annotation words are then aligned by postulating a hidden variable that assigns each word to one of the image regions. The work of [BJ03] relaxes the strict association between all the image regions and annotation words in the same document due to the sharing of the same hidden variable (1 latent variable per document), as proposed in [TSK01, BF01]. Now the hidden factors are allowed to be repeatedly allocated within a given document, i.e. each region is given its own hidden variable. This formulation more flexibly enables the model to associate one image region with multiple annotation words, or each region can be associated with only one distinctive annotation word. The flexibility of such a model has been shown to pay off and [BJ03] obtains a good region annotation performance on a 7500-image subset of COREL dataset.

Despite wide-spread interests in the problem of automatic image annotation, relatively little work has been done in modeling audio and words. In [WR02, WE04, Whi05, TBL06], the joint distribution of music and the associated text (from song reviews or from web documents associated with the artists) are learned. Music annotation is cast as a problem of class-conditional density estimation for each concept class, similar to the modeling of image and text in [CV05]. These methods thus suffer the same scalability issue when the vocabulary size becomes large. In [Sla02a, Sla02b], animal soundtracks and their associated description words are modeled by learning the probability density of each soundtrack in the training set fitting a mixture of Gaussians model to acoustic features extracted from each soundtrack. Annotation of a test audio is done by choosing description words from the soundtrack whose density model best explains the test audio. The computation in these type of models indeed scale linearly with the size of the training set; these methods therefore do not lend themselves applicable to large audio databases used in most real-world scenarios.

3.4 Multimedia Data and their Representation

In this work, we consider annotation models for multimedia documents containing 1 non-text modality (e.g. image, video, or audio), together with their corresponding text description (captions, closed-captions). Examples of this type of documents include images and their captions, speech and their transcriptions, music and lyrics, video and their closed-captions. Since there has been a great deal of interests in the problem of image annotation with several standard data sets that can be used in performance evaluation, we focus our attention on image-caption data. The statistical association models presented in this work, however, make no assumptions that are restrictive to the statistics of image features. In fact, without modification, the exact same models can be readily applied to modeling association between, for example, acoustic or video features and their corresponding text descriptors.

3.4.1 Image-Caption Data

We focus on 2 standard datasets for image annotation: COREL and LabelMe datasets [RTMF08]. The COREL database with 600 image categories annotated with keywords from a vocabulary of 374 words is widely used in training and performance evaluation of image annotation models. In particular, we focus on the 5,000-image subset of COREL in our experiments as used in [FML04, CV05]. This subset contains 50 classes of images, with 100 images per class. Each image in the collection is reduced to size 117×181 (or 181×117). 4,500 images are used in training (90 per class), and 500 for testing (10 per class). Each image in the dataset is annotated with 1-5 captions with a collection average of 3.5 words per image. Note that in this dataset, annotation words either appear once in an image or not at all.

For the LabelMe dataset, following the work in [WBFf09], we use the 8-category subset which contains 2,687 images from the classes: ‘coast’, ‘forest’, ‘highway’, ‘inside city’, ‘mountain’, ‘open country’, ‘street’, and ‘tall building’. 80% of the data in each class (2,147 images total) are used for training and 20% for testing (540 total). Each image is of size 256×256 and has caption words ranging from 2 to 71 words, with a collection average of 13.5 words per image. Note that unlike in the COREL dataset, each caption word can appear more than once in an image. If multiple instances of the same object appear in a scene, e.g. multiple cars in a highway scene, multiple bikes in a city scene, the words ’car’ and ’bike’ will be repeated multiple times in the corresponding caption data.

3.4.2 Video-Closed captions Data

The multimedia database used in this work is obtained by recording TV programs from cable television broadcast using a PC that runs Windows Media Center Edition™ personal video recorder, as used in the experiments in [PANL07]. Each recorded TV show contains video and closed-caption text provided from the television station (we discard the audio modality for the moment). Because of the large amount of video data corresponding to each hour-long TV show, a document is defined as a 20-sec segment (clip). 2 datasets of recorded TV clips are used in our experiment. In the first dataset, a total of 2,502 clips have been gathered from 24 episodes of Modern Marvels program (each episode is 1 hour long) from the History channel. 2100 documents are used in training while 402 documents is withheld for testing. In the second recordings, we gather 2,267 clips from 6 TV shows shown in Figure 3.1: $40-a-day, Crime Scene Investigation (CSI), Good Eats, Law and Order, Modern Marvels, and The West Wing. 6 episodes of the 30-min shows ($40-a-day and Good Eats) and 3 episodes of hour-long shows (CSI, Law and Order, Modern Marvels, and the West Wing) have been recorded to create the 2,267-document dataset. 2000 documents are used for training while 267 documents for used in testing.

We preprocess the data by eliminating the portion of the video and closed caption texts corresponding to TV commercials. Each frame in the video is first converted from RGB to gray scale and down-sampled from 720×480 to 80×64. By eliminating very common and rare words in closed-caption texts, we are left with 4,171-term vocabulary in the 2,502 clip datataset. Each video contains 5-51 annotation words with an average of 25.63 words per clip. The same preprocessing step is done on the second video dataset, and we are left with 4,511-term vocabulary. Each video contains 3-46 caption words with a collection average of 22.68 words per video.

Each episode of Modern Marvels show delves into the history and technology behind the subject of the episode, which varies drastically from episode to episode. Examples of the episodes used in our experiments feature the subjects of ‘demolition’, ‘dams’, ‘gold mines’, and ‘cheese’. $40-a-day is a travel show where in each episode the host travels to different destination cities and try local restaurants within a $40 budget. Good Eats is a cooking show, featuring the scientific aspect of the cooking process. CSI and Law and Order are criminal dramas, while The West Wing is a political drama show.

$40-a-Day CSI Good Eats Law & Order Modern Marvels The West Wing

Figure 3.1: Examples of Recorded TV shows used in Multimedia Annotation.

3.4.3 Data Representation

We borrow a tool from statistical text document analysis and adopt a bag-of-word representation for all the data modalities considered in this work. In such a representation, the ordering of words in sentences and paragraph structures are ignored and a document is treated as a collection of words. In a similar fashion, an image is considered as a collection of patches and a video clip is treated as a collection of spatio-temporal blocks. It can be argued that unlike text modality, image and video data are characterized by strong spatial and temporal correlation; using a bag-of-word representation (histogram representation) could throw away valuable information encoded in these dependencies. Indeed, in the case of images, the importance of spatial correlations is repeatedly emphasized in numerous previous work in computer vision that makes extensive use of geometric constraints and spatial continuity in various object detection and recognition tasks, e.g. [VJ01]. However, recent state-of-the-art results on scene modeling and classification [FfP05, QMO$^+$05, SRE$^+$05] have been obtained using a histogram of visual words representation. These surprising favorable outcomes were explained in [QMO$^+$05] with the observation that a scene is characterized more by a co-occurrence of a set of objects, not by their relative positions in the scene. In a city scene, for example, a car and a building can be placed at various different positions while still describing the same scene. The histogram representation, while simple and crude, does indeed capture precisely the statistics of co-ocurrences of patches in an image, which as reported in [FfP05, QMO$^+$05, SRE$^+$05] prove to be sufficient in discriminating one type of scenes from the others. The success of these recent work with the histogram model thus inspires adopting such a representation in our model.

In our unified representation, a multimedia document consisting of an image (or video) and the corresponding caption text is summarized in a pair of vectors of word counts. For text, we simply count how many of word t from a dictionary of T_w words appears in each document. In the case of image and video, such a concept of *word* needs to first be identified and extracted via clustering of visual features. An image is thus modeled as a collection of patches where each patch is assigned to a codeword from a dictionary of T_r visual words. Similarly for the case of video, a video clip is modeled as a collection of spatio-temporal blocks where each block is assigned to a codeword from a dictionary of T_r words. By tallying the number of word occurrences in a multimedia document, we obtain a corresponding pair of vectors of word counts.

3.5 Feature Extraction

While the focus of this thesis is on novel annotation models for capturing statistical associations between visual features and their corresponding caption texts, the particular choice of image or video features used does have a great impact on the overall performance of annotation systems. In general, the goal of feature extraction is to transform the original input pixels to features in a new space that contains more discriminatory power pertinent to the task at hand. In addition, for visual features, since small changes in viewpoints, lighting conditions, scale, translation, and rotation can drastically affect the original input pixels, we are interested in a feature space that remains approximately insensitive to these variations so long as the semantics of the scenes stay unchanged. Generally speaking, however, invariance and discriminatory power do not go hand in hand, and as invariance increases, the features start loosing their discriminatory power. In designing an effective visual feature for object matching and classification, we therefore need to strike a delicate balance between making the feature more robust by increasing invariance and retaining important discriminatory information.

3.5.1 Image words

To accompany the histogram representation we focus on image feature descriptors which have been shown to perform well on various object matching and recognition tasks. Following the work in [FfP05], we use the 128-dimensional Scale Invariant Feature Transform (SIFT) descriptors [Low04] extracted from 20×20 gray-scale patches. The SIFT descriptor computes a set of histograms of magnitudes and orientations on the 4×4 sub-regions around a given patch. To ensure invariance to scale, location, and rotation, these magnitudes and orientation features are computed with respect to the scale and orientation of the patch itself. By tallying an 8-bin histogram on each of the 4×4 sub-regions, we obtain a 128-dimensional vector of SIFT descriptor to represent each 20×20 image patch. The SIFT descriptor has been shown to be highly distinctive and partially invariant to other variations such as illumination and viewpoint, hence good performance have been reported in various object and feature matching tasks.

Although highly distinctive, SIFT features computed on gray-scaled patches only describe shape and texture information while the color information of the patches has been completely ignored. Following the work in [vdWS06, VT07] we adopt 36-dimensional robust color descriptors which have been designed to complement the SIFT descriptors extracted from gray-scale patches. To ensure the color descriptor is invariant to different illuminating conditions, e.g. shading and specularities, normalization of each color channel is proposed. A hue descriptor defined as a histogram of the angles of opponent colors [vdWS06] is computed. To obtain a dictionary of visual words, we run a k-means clustering algorithm on a large collection of 164-dimensional features to learn a set of T_r words. To account for the different statistics in different image datasets, separate visual word dictionaries are learned for each image dataset.

3.5.2 Video words

In video processing, motion has long been useful in a variety of tasks ranging from high-level object segmentation, activity recognition, to low-level video compression. In this work, we adopt the motion representation which seamlessly combines the spatial and temporal dynamics in the video as proposed in [PL06] and described in detail in Chapter 7. Each input video is first divided into spatio-temporal blocks of size 8×8×8. By postulating that the desired representation be as sparse and efficient as possible, a set of spatio-temporal filters that transform input pixels into ICA coefficients are learned from the data. Such spatio-temporal ICA filters are found to resemble spatially-localized oriented gabor wavelet filters (edge detectors) that are moving in time. The coefficients of the video block associated with a set of such filters are then used as our motion features. To learn a dictionary of video words, we fit a mixture of Laplacians model, with as many clusters as the desired number of video words, to a collection of ICA coefficients derived from the training data. For more details, see Chapter 7.

Chapter 4

Topic-regression Multi-modal Latent Dirichlet Allocation Model (tr-mmLDA)

In this chapter, we examine several topic models for learning statistical associations between image (or video) and text for the task of image and video annotation. The goal is to draw from the strengths and compensate for the weaknesses of the 2 main approaches to auto-annotation: (a) supervised learning approach where the relevant conditional distribution of annotation data given the corresponding images is directly modeled in a classification framework, with annotation words treated as class labels, and (b) unsupervised learning approach where probabilistic latent variable models are used to learn the joint correlations between images and the corresponding texts. In this work, we explore an alternative approach that combines the best of both worlds by incorporating the discriminative power of the supervised learning formulation into latent variable structures of the generative methods. With the use of latent variables, our approach is able to capture important correlations in annotation data which will prove useful in prediction, while the introduction of the discriminative module allows our approach to learn the precise conditional relations between caption words given the corresponding images, which will lead to a better predictive performance.

We first review Latent Dirichlet Allocation (LDA) model and examine 2 extensions of LDA—correspondence LDA (cLDA) [BJ03] and multi-modal LDA (mmLDA) [BDdF$^+$03, BJ03]—which modify the basic LDA model to capture the joint correlations/statistical association between images and texts. We identify several limitations of the 2 association models and present a novel statistical topic model called topic-regression Multi-modal LDA (tr-mmLDA) to address these limitations. Instead of sharing a set of latent variables between the 2 data modalities as in the formulation of cLDA and mm-LDA, our approach learns 2 separate sets of latent topics and introduces a regression module that allows the topic proportion of annotation texts to be linearly predicted from the hidden topics of the image (video) modality. Our proposed formulation is more general and can capture varying degrees of correlations between the two data modalities. We derive an efficient variational inference algorithm using a mean-field approximation to handle intractable posterior computations. To demonstrate the predictive power of the new association model, we compare image annotation performance on 2 standard datasets: a 5000-image subset of COREL and 2687-image 8-category subset of the LabelMe dataset.

4.1 Notations

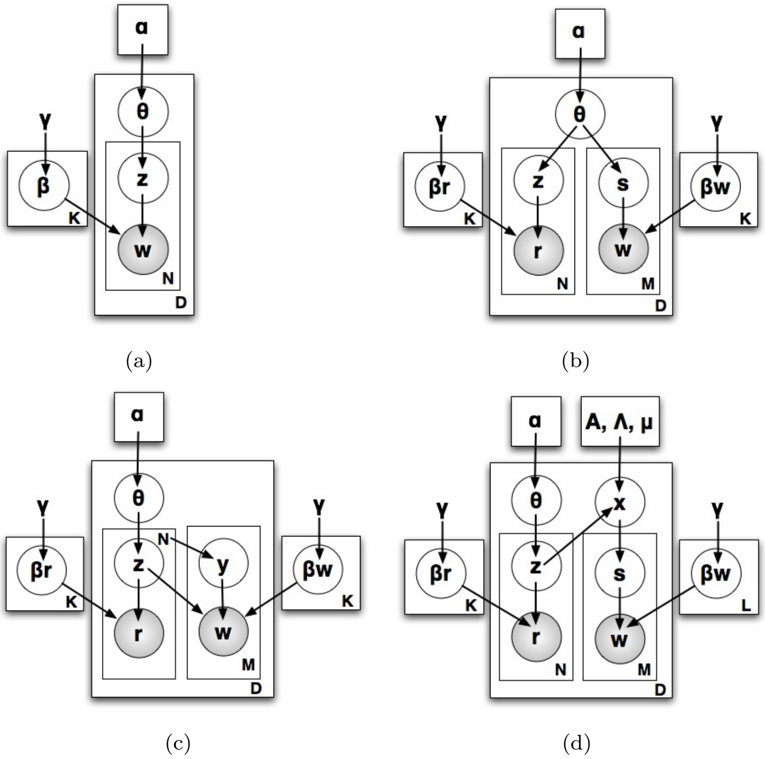
As mentioned in Chapter 3, we adopt a bag-of-word representation for both image (video) and text. In such a representation, word ordering is ignored and a document is simply reduced to a vector of word count. A multimedia document consisting of an image and the corresponding caption text is thus summarized in our representation as a pair of vectors of word counts.

An image word is denoted as a unit-basis vector r of size T_r with exactly one nonzero entry representing the membership to only one word in a dictionary of T_r visual words. A caption word w_n is similarly defined for a dictionary of size T_w. An image is a collection of N word occurrences denoted by $\mathbf{R} = \{r_1, r_2, \ldots, r_N\}$; the caption text is a collection of M word occurrences denoted by $\mathbf{W} = \{w_1, w_2, \ldots, w_M\}$. A training set of D image-caption pairs is denoted as $\{\mathbf{R}_d, \mathbf{W}_d\}$, $d \in \{1, 2, \ldots, D\}$.

4.2 Probabilistic Models

All the models discussed in this work builds on Latent Dirichlet Allocation (LDA) [BNJ03] which is a powerful generative model for modeling words in documents. Unlike the mixture of unigrams model [NMTM00] which assumes that all the words in the same document are generated from the same hidden factor, under LDA words in a document are allowed to exhibit characteristics from multiple factors (topics). A document, which is a collection of words, is then summarized in terms of the factors' overall relative influences on the collection. To this end, LDA employs 2 sets of latent variables for each document as seen in Fig 4.1(a): (i) a discrete-valued hidden variable z_n which assigns a word w_n to one of the K topics (factors) and (ii) a latent variable θ representing the random proportion of the topics' influence in the document. In more specific terms, LDA decomposes the distribution of word counts for each document into contributions from K topics and model the proportion of topics θ as a Dirichlet distribution, while each topic, in turn, is a multinomial distribution over words. Given the Multinomial parameters for each topic β and Dirichlet parameter α, we can generate a document with N word occurrences by taking the following steps:

- Draw a proportion of topic θ from $p(\theta|\alpha) \sim \mathrm{Dir}(\alpha)$.



(a) (b)

(c) (d)

Figure 4.1: Graphical model representations for **(a)** Latent Dirichlet Allocation (LDA) and 3 extensions of LDA for the task of image annotation: **(b)** Multi-modal LDA (mmLDA) **(c)** correspondence LDA (cLDA) **(d)** Topic-regression Multi-modal LDA (tr-mmLDA).

- Given θ, for each word in the document w_n, $n \in \{1, 2, \ldots, N\}$,
  1. Draw a topic assignment $z_n = k|\theta \sim \text{Mult}(\theta_k)$.
  2. Given the topic z_n, draw a word $w_n = t|z_n = k \sim \text{Mult}(\beta_{kt})$.

4.2.1 Multi-modal LDA (mmLDA) and Correspondence LDA (cLDA)

In order to extend the basic LDA model to learn the joint correlations between data of different types, a traditional solution under a probabilistic framework is to assume the existence of a small set of shared latent variables that are the common causes of correlations between the 2 modalities. This is precisely the design philosophy behind multi-modal LDA (mmLDA) and correspondence LDA (cLDA) [BJ03], which extend LDA to describe the joint distributions of image and caption words in multimedia documents. While adopting the same core assumption of LDA in allowing words in a document to be generated from multiple topics, the two extension models differ in their choices of latent variables being shared between images and texts.

Multi-modal LDA (mmLDA)

Originally proposed in [BDdF$^+$03], mmLDA postulates that the mean topic proportion variable θ is the common factor shared by the 2 data modalities. By forcing the topic proportion to be the same in image and caption modality, the 2 sets of Multinomial topic parameters therefore are assumed to correspond. For each document, we sample a proportion of topics θ shared by the 2 data modalities; conditioned on θ, image and caption words are then generated independently. To generate an image word, we first sample a hidden topic k with probability θ_k; then using a Multinomial parameters of topic k, we sample an image word. Conditioned on the same θ, caption words can be generated in a similar manner. Given the number of topics K, the Dirichlet parameter α, Multinomial parameters over image words β^r, and Multinomial parameters over caption words β^w, we can generate an image-caption pair with N visual words and M caption words as follows:

- Draw a shared topic proportion $\theta|\alpha \sim \text{Dir}(\alpha)$
- For each image word r_n, $n \in \{1, 2, \ldots, N\}$
  1. Draw topic assignment $z_n = k|\theta \sim \text{Mult}(\theta_k)$.
  2. Given the topic z_n, draw visual word $r_n = t|z_n = k \sim \text{Mult}(\beta^r_{kt})$.
- For each caption word w_m, $m \in \{1, 2, \ldots, M\}$,
  1. Draw topic assignment $s_m = k|\theta \sim \text{Mult}(\theta_k)$.
  2. Given the topic s_m, draw caption word $w_m = t|s_m = k \sim \text{Mult}(\beta^w_{kt})$.

The decision to share θ between the two data modalities indeed implies that image and caption words become independent conditioned on θ. Without the plate notation as depicted in the graphical model of Fig 4.1(b), it is not hard to see that the association of mmLDA assumes that image and caption words are exchangeable—a key assumption which allows the 2 types of words in the same document to potentially be generated from non-overlapping sets of hidden topics. As K becomes large, annotation experiments on the COREL dataset in [BJ03] show that more than 50% of the caption words are assigned to topics that do not occur in the corresponding images. In such a scenario, instead of learning the 2 sets of Multinomial topic parameters β^r, β^w that correspond, the knowledge about the image modality renders essentially half useless at predicting the missing caption words. The flexibility of mmLDA provides a good fit for the joint distribution of the data but is a bad fit for a prediction task, hence a poor annotation performance.

Correspondence LDA (cLDA)

To ensure that only the set of topics that actually generate the image words are those used in generating the caption words, correspondence LDA (cLDA), with the graphical representation shown in Fig 4.1 (c), was designed so that image is the primary modality and is generated first; conditioned on the topics used in the image, caption words are then generated. More specifically, cLDA introduces for each caption word a uniform random variable y which selects one topic, from a set of topics used in the image modality, to be associated with the caption word. Indeed more strictly speaking, since cLDA has been designed for the task of image region annotation, a random variable y associated with each caption word selects an image word to share a topic with. Note that while each caption word is restricted to be associated with one particular image word (region), the association of cLDA does allow the same image word (region) to be associated with multiple caption words, accounting for the scenario where more than one caption words are used to describe a single object in the image.

Given the model parameters $\{\alpha, \beta^r, \beta^w\}$, the generative process of cLDA for an image-caption pair with N image words and M caption words is given as follows:

- Draw an image topic proportion $\theta|\alpha \sim \text{Dir}(\alpha)$
- For each image word r_n, $n \in \{1, 2, \ldots, N\}$:
    1. Draw topic assignment $z_n = k|\theta \sim \text{Mult}(\theta_k)$.
    2. Draw visual word $r_n = t|z_n = k \sim \text{Mult}(\beta^r_{kt})$.
- For each caption word w_m, $m \in \{1, 2, \ldots, M\}$:
    1. Select one of the image words by drawing a uniform random variable $p(y_m = n|N) = \frac{1}{N}$.
    2. Draw caption word $w_m = t|y_m = n, z_n = k \sim \text{Mult}(\beta^w_{kt})$.

By forcing each caption word to directly share a hidden topic with a randomly selected image word, cLDA guarantees that the topics in caption texts are indeed a subset of the topics that occur in the corresponding image. With the same set of hidden topics employed in modeling images and the corresponding caption texts, the hidden topics in cLDA capture patterns of co-occurences of words of different types. Each topic now describes how certain caption words co-occur more often with certain image or video words. The knowledge about the presence of a group of image words will therefore be highly predictive of which caption words to occur in its corresponding caption data. cLDA therefore is a better predictive model than mmLDA.

Despite a good annotation performance as reported in [BJ03], the constrained association of cLDA proves to be too restrictive in practice. When dealing with annotation words that globally describe the scene as a whole, the type of association that restricts each caption word to one image region can be very inaccurate. A more powerful association model should indeed allow the captions words to be influenced by topics from all image regions as well as those from a particular subset of regions. In addition, the assumption imposed by cLDA which require the set of topics occurring in caption words be a subset of those used in the corresponding image could indeed be easily violated in practice. An example of such a scenario is when we have a small number of image regions (hence a small number of topics used in the image modality) paired with a large number of caption words (hence a large number of hidden topics required). Depending on the datasets, this type of documents can be very common and cLDA will indeed be a poor fit for such data. In the next section, we will describe a more flexible association model that address these limitations of cLDA while still maintaining a good predictive power.

4.2.2 Topic-regression Multi-modal LDA (tr-mmLDA) [PAN10b]

To get past the issues associated with sharing latent variables between data modalities, in this work we explore a novel approach in modeling correlations between data of different types. Instead of using a set of shared latent variables to explain correlations in the data, we propose a latent variable regression approach to correlate latent variables of the two modalities. By incorporating a regression framework into the latent variable models, our framework learns the precise relations between 2 sets of latent variables and allows latent variables of one type to be predicted from latent variables of another type. In the specific case of extending LDA to learn correlations between image and caption words, we propose a formulation that uses 2 separate topic models one for each data modality and introduce a regression module to correlate the 2 sets of hidden topics. To this end, we draw insights from several recent topic models [BL07, PAN09] and adopt a linear regression module which takes in the image topic proportion as its input and target the hidden topic proportion for annotation texts as the response variable (in line with the task of predicting captions given an image). Our approach is similar in spirit to the way topic correlations are captured in Independent Factor Topic Models in [PAN09] by explicitly modeling the independent sources and linearly combining them to obtain the correlated topic vectors. In our case, the hidden sources of topic correlations in caption data correspond to the hidden topics of the image modality.

Our model which we call a topic-regression multi-modal Latent Dirichlet Allocation (tr-mmLDA) has the graphical representation as shown in Fig 4.1(d). Given K image topics and L text topics, from an image side we have an LDA model with hidden topics $\mathbf{Z} = \{z_1, z_2, \ldots, z_N\}$ and topic proportion θ. A real-valued topic proportion variable for caption text $\mathbf{x} \in \mathcal{R}^L$ is given by: $\mathbf{x} = \mathbf{A}\bar{\mathbf{z}} + \mu + \mathbf{n}$, where $\mathbf{A}$ is an $L \times K$ regression coefficients matrix, μ is a vector of the mean parameters, $\mathbf{n} \sim \mathcal{N}(\mathbf{n}; 0, \mathbf{\Lambda}^{-1})$ is a zero-mean uncorrelated Gaussian noise with a diagonal precision matrix $\mathbf{\Lambda}$. Instead of regressing over the mean topic proportion variable θ as done in mmLDA, we follow the formulation in supervised LDA in [BM07, WBFf09] and adopt the empirical topic frequency covariates $\bar{\mathbf{z}} = \frac{1}{N}\sum_n z_n$ as an input into our regression module so that the topic proportion of annotation data depends directly on the actual topics that do occur in the image. Given $\mathbf{x}$, the topic proportion of caption text η is deterministically obtained via a softmax transformation of $\mathbf{x}$, i.e. the probability of observing topic l is given by $\eta_l = \frac{\exp(x_l)}{\sum_{k=1}^{L}\exp(x_k)}$. The generative process of tr-mmLDA for an image-caption pair with N visual words and M caption words is given as follows:

- Draw an image topic proportion $\theta|\alpha \sim \text{Dir}(\alpha)$
- For each image word r_n, $n \in \{1, 2, \ldots, N\}$
    1. Draw topic assignment $z_n = k|\theta \sim \text{Mult}(\theta_k)$
    2. Draw visual word $r_n = t|z_n = k \sim \text{Mult}(\beta^r_{kt})$
- Given the empirical image topic proportion $\bar{\mathbf{z}} = \frac{1}{N}\sum_{n=1}^{N} z_n$, we sample a real-valued topic proportion variable for caption text: $\mathbf{x}|\bar{\mathbf{z}}, \mathbf{A}, \mu, \mathbf{\Lambda} \sim \mathcal{N}(\mathbf{x}; \mathbf{A}\bar{\mathbf{z}} + \mu, \mathbf{\Lambda})$.

- Compute topic proportion $\eta_l = \frac{\exp(x_l)}{\sum_{k=1}^{L}\exp(x_k)}$.
- For each caption word w_m, $m \in \{1, 2, \ldots, M\}$
  1. Draw topic assignment $s_m = l|\eta \sim \text{Mult}(\eta_l)$
  2. Draw caption word $w_m = t|s_m = l \sim \text{Mult}(\beta^w_{lt})$

The formulation of tr-mmLDA can be seen as linking the LDA model for images and the IFTM model [PAN09] for caption texts using a linear regression module, which is a flexible way of capturing correlations in the data. Under this framework, varying degrees of correlations can be captured by adjusting the regression coefficients in the matrix $\mathbf{A}$ accordingly. When the 2 data modalities are independent, the coefficients in $\mathbf{A}$ are driven close to 0 and tr-mmLDA is reduced to 2 independent topic models one for each data modality. As correlations between the 2 data types strengthen, more regression coefficients in the matrix $\mathbf{A}$ will take values further away from 0. In fact, the association of correspondence LDA (cLDA) can be derived as a special case of tr-mmLDA by restricting the non-zero coefficients to be in the diagonal entries of $\mathbf{A}$ (assuming $K = L$) and setting precision $\mathbf{\Lambda}$ to ∞, which has the effect of forcing the empirical topic proportions in the 2 data modalities to be identical. As the regression coefficient matrix $\mathbf{A}$ moves away from being diagonal (with more non-zero coefficients in off-diagonal entries), tr-mmLDA allows the hidden topics from more than one image region to collectively exert influence on each caption word, which depicts a more accurate relation for annotation words that globally describe the scene as a whole. Note that since tr-mmLDA employs 2 sets of hidden topics, it allows the number of topics in images and captions to be different, i.e. we can use as many topics as needed to explain correlations between caption words without being restricted to those that occur in the corresponding image. Our framework in capturing correlations is thus more flexible than cLDA and allow more general forms of correlation to be modeled.

4.3 Variational EM

To learn parameters of tr-mmLDA that maximizes the likelihood of the training data, we employ the Expectation Maximization (EM) framework that iteratively estimates the model parameters of latent variable models. Using Jensen's inequality, the E step of the EM algorithm derives an auxiliary function which tightly lower-bounds the data likelihood function to allow for a more simple optimization to be performed in the M step. Indeed, for most probabilistic models involving a large number of latent variables of different types, computing the exact posterior distribution over latent variables in the E step is a computationally intractable task. In variational EM [Att00, JJ00], we replace the exact inference in the E step with an approximate inference algorithm. Variational EM framework thus alternates between computing a strict likelihood lower bound in the variational E step, and maximizing the bound to obtain a new parameter estimate in the M step. In this section, we derive an approximate inference algorithm using mean-field variational approximation [Att00]. As in the inference algorithm for IFTM in [PAN09], most parameter updates are obtained in closed-form and those parameters without closed-form updates are learned very efficiently using Newton-Raphson algorithms.

4.3.1 Variational Inference

To infer the posterior over hidden variables, we begin with the expression of the true log-likelihood for an image-caption pair $\{\mathbf{W}, \mathbf{R}\}$:

$$\log p(\mathbf{W}, \mathbf{R}|\Psi) \geq \int q(\mathbf{Z}, \theta, \mathbf{x}, \mathbf{S}) \left\{\log p(\mathbf{W}, \mathbf{R}, \mathbf{Z}, \theta, \mathbf{x}, \mathbf{S}|\Psi) - \log q(\mathbf{Z}, \theta, \mathbf{x}, \mathbf{S})\right\} d\mathbf{Z} d\theta d\mathbf{x} d\mathbf{S} = \mathcal{F}, \tag{4.1}$$

where Ψ denotes the model parameters for tr-mmLDA $\{\beta^r, \beta^w, \gamma, \mathbf{A}, \mu, \mathbf{\Lambda}\}$. Using the concavity of the log function, we apply Jensen's inequality and derive a lower bound of the log-likelihood as seen in (4.1). Indeed, equality holds when the posterior over the hidden variables $q(\mathbf{Z}, \theta, \mathbf{x}, \mathbf{S})$ equals the true posterior $p(\mathbf{Z}, \theta, \mathbf{x}, \mathbf{S}|\mathbf{W}, \mathbf{R})$. Like in LDA, computing the exact joint posterior is computationally intractable; we employ a mean-field variational approximation to approximate the joint posterior distribution with a variational posterior in a factorized form: $p(\mathbf{Z}, \theta, \mathbf{x}, \mathbf{S}|\mathbf{w}, \mathbf{R}) \approx \prod_n q(z_n) \prod_m q(s_m) q(\theta) q(\mathbf{x})$. With such a posterior, the RHS of (4.1) becomes a strict lower bound of the data likelihood. The goal of the variational E step now is to find within a family of factorized distributions the variational posterior that maximizes the lower bound. Writing out the likelihood lower bound $\mathcal{F}$ on the right hand side of (4.1), we obtain the following expression:

$$\begin{aligned}\mathcal{F} = &\sum_n (E[\log p(r_n|z_n, \beta^r)] + E[p(z_n|\theta)]) + E[\log p(\theta|\alpha)] \\ &+ \sum_m (E[\log p(w_m|s_m, \beta^w)] + E[\log p(s_m|\mathbf{x})]) + E[\log p(\mathbf{x}|\bar{\mathbf{z}}, \mathbf{A}, \mathbf{\Lambda}, \mu)] \\ &+ \mathcal{H}(q(\mathbf{Z})) + \mathcal{H}(q(\theta)) + \mathcal{H}(q(\mathbf{S})) + \mathcal{H}(q(\mathbf{x})),\end{aligned} \tag{4.2}$$

where the expectations are taken with respect to the factorized posteriors. $\mathcal{H}(p(x))$ denotes the entropy of $p(x)$. The fifth expectation term in (4.2)

$$E_{q(\mathbf{x})}[\log p(s_m = l|\mathbf{x})] = E_{q(\mathbf{x})}[x_l - \log(\sum_j e^{x_j})] \tag{4.3}$$

contains a normalization term from the softmax operation that will be difficult to evaluate in closed-form, regardless of the form of the variational posterior $q(\mathbf{x})$. We make use of convex duality and represents a convex function ($-\log(\cdot)$ function) as a point-wise supremum of linear functions. More specifically, the log normalization term is replaced with adjustable lower bounds parameterized by a convex variational parameter ξ:

$$x_l - \log \sum_j e^{x_j} \geq x_l - \log \xi - \frac{1}{\xi} \sum_j e^{x_j} + 1. \tag{4.4}$$

Under the diagonality assumption of $\mathbf{\Lambda}$ and the use of convex variational bound of the log-normalizer term in (4.4), the free-form maximization of $\mathcal{F}$ w.r.t $q(\mathbf{x})$ results in the variational posterior $q(\mathbf{x})$ automatically taking on a factorized form $q(\mathbf{x}) = \prod_l q(x_l)$. However, $q(x_l)$ obtained by the free-form maximization is not in the form of a distribution that we recognize. We thus approximate $q(x_l)$ as a Gaussian distribution: $q(x_l) \sim \mathcal{N}(x_l; \bar{x}_l, \gamma_l^{-1})$ with mean parameter $\bar{x}_l$ and precision γ_l. To simplify the notation, we denote $q(z_n = k)$ as ϕ_{nk} and $q(s_m = l)$ as η_{ml}. Since the prior $p(\theta|\alpha)$ is a Dirichlet distribution, which is a conjugate prior to the multinomial distribution, we can conclude that the posterior $q(\theta)$ is also a Dirichlet distribution. More specifically, we denote the posterior Dirichlet parameters as $\tilde{\alpha}$: $q(\theta) \sim \mathrm{Dir}(\tilde{\alpha})$. By taking the expectation with respect to the variational posterior, we can write out the terms in the lower bound $\mathcal{F}$ explicitly as a function of the variational parameters. The first two terms of $\mathcal{F}$ in (4.2) are given by:

$$\sum_{n,k} \phi_{nk} \sum_t 1(r_n = t) \log \beta^r_{kt} + \sum_{n,k} \phi_{nk} E_{q(\theta)}[\log \theta_k] \tag{4.5}$$

where $E_{q(\theta)}[\log \theta_k] = \Psi(\tilde{\alpha}_k) - \Psi(\sum_j \tilde{\alpha}_j)$, with $\Psi(x)$ denoting the first derivative of the log-gamma function $\frac{\partial \log \Gamma(x)}{\partial x}$. The third term in (4.2) can be written as:

$$\log \Gamma(\sum_j \alpha_j) - \sum_j \log \Gamma(\alpha_j) + \sum_j (\alpha_j - 1) E_{q(\theta)}[\log \theta_j].$$

By evaluating the expectation with respect to a Gaussian posterior $q(x_j) \sim \mathcal{N}(x_j; \bar{x}_j, \gamma_j)$, we have that $E_{q(x_j)}[e^{x_j}] = e^{\bar{x}_j + \frac{0.5}{\gamma_j}}$ and the fourth and fifth expectation terms in (4.2) can be written as:

$$\sum_{m,l} \eta_{ml} \sum_t 1(w_m = t) \log \beta^w_{lt} + \sum_{m,l} \eta_{ml} \bar{x}_l - M \log \xi - \frac{M}{\xi} \sum_j e^{\bar{x}_j + \frac{0.5}{\gamma_j}} + M. \tag{4.6}$$

Making use of the following expectation $E[\mathbf{x}^\top \mathbf{\Lambda} \mathbf{x}] = \bar{\mathbf{x}}^\top \mathbf{\Lambda} \bar{\mathbf{x}} + \mathrm{tr}(\mathbf{\Lambda} \mathbf{\Gamma}^{-1})$, the sixth term in (4.2) is given by:

$$-\frac{1}{2}\left((\bar{\mathbf{x}} - \mu)^\top \mathbf{\Lambda} (\bar{\mathbf{x}} - \mu) + \mathrm{tr}(\mathbf{\Lambda} \mathbf{\Gamma}^{-1}) - 2(\bar{\mathbf{x}} - \mu)^\top \mathbf{\Lambda} \mathbf{A} E[\bar{\mathbf{z}}] + E[\bar{\mathbf{z}}^\top \mathbf{A}^\top \mathbf{\Lambda} \mathbf{A} \bar{\mathbf{z}}]\right), \tag{4.7}$$

where $E[\bar{\mathbf{z}}] = \frac{1}{N} \sum_n \phi_n$ and $E[\bar{\mathbf{z}}^\top \mathbf{A}^\top \mathbf{\Lambda} \mathbf{A} \bar{\mathbf{z}}]$ is evaluated to be $\mathrm{tr}(\mathbf{A}^\top \mathbf{\Lambda} \mathbf{A} \frac{1}{N^2}(\sum_n \mathrm{diag}(\phi_n) + \sum_n \phi_n \sum_{m \neq n} \phi_m^\top))$. The entropy terms in (4.2) can now be expressed as functions of the variational parameters and are given by:

$$\begin{aligned}
\mathcal{H}(q(\mathbf{Z})) &= -\sum_{n,k} \phi_{nk} \log \phi_{nk}, \\
\mathcal{H}(q(\mathbf{S})) &= -\sum_{m,l} \eta_{ml} \log \eta_{ml}, \\
\mathcal{H}(q(\theta)) &= -\log \Gamma(\sum_j \tilde{\alpha}_j) + \sum_j \log \Gamma(\tilde{\alpha}_j) - \sum_j (\tilde{\alpha}_j - 1) E[\log \theta_j], \\
\mathcal{H}(q(\mathbf{x})) &= -\sum_l \frac{1}{2} \log \gamma_l + \frac{L}{2} \log 2\pi + \frac{L}{2}.
\end{aligned}$$

To update these variational parameters, we employ a coordinate ascent algorithm where we update one set of parameters while holding the rest fixed. By computing the gradient of $\mathcal{F}$ w.r.t. ϕ_n and set the derivative to 0, we obtain the following update rule for ϕ_n:

$$\begin{aligned}\log \phi_n = &\sum_t 1(r_n = t) \log \beta^r_{\cdot t} + E[\log \theta] \\ &+ \frac{1}{N}\mathbf{A}^\top \mathbf{\Lambda}(\bar{\mathbf{x}} - \mu) - \frac{1}{2N^2}\text{diag}(\mathbf{A}^\top \mathbf{\Lambda} \mathbf{A}) - \frac{1}{N^2}\mathbf{A}^\top \mathbf{\Lambda} \mathbf{A} \sum_{m \neq n} \phi_m. \end{aligned} \tag{4.8}$$

The variational parameters $\eta_{ml}, \xi, \tilde{\alpha}_k$ can be similarly re-estimated, resulting in the following closed-form updates:

$$\log \eta_{ml} = \sum_t 1(w_m = t) \log \beta^w_{lt} + \bar{x}_l, \tag{4.9}$$

$$\xi = \sum_j e^{\bar{x}_j + \frac{0.5}{\gamma_j}}, \tag{4.10}$$

$$\tilde{\alpha}_k = \sum_n \phi_{nk} + \alpha_k. \tag{4.11}$$

To update the parameters of the variational posterior $q(x_l) \sim \mathcal{N}(x_l; \bar{x}_l, \gamma_l)$, we differentiate $\mathcal{F}$ w.r.t. $\bar{x}_l$ and obtain the following expression for the gradient:

$$\frac{\partial \mathcal{F}}{\partial \bar{x}_l} = \sum_m \eta_{ml} - \frac{M}{\xi} e^{\frac{0.5}{\gamma_l} + \bar{x}_l} - \lambda_l (x_l - \mu_l - \mathbf{a}_l^\top E[\bar{\mathbf{z}}]). \tag{4.12}$$

However, the value of $\bar{x}_l$ that makes the gradient vanish cannot be obtained in closed-form. We employ a Newton algorithm that finds a zero-crossing solution for (4.12) efficiently. First, by substituting $\sum_m \frac{\eta_{ml}}{\lambda_l} - \bar{x}_l + a_l^\top E[\bar{\mathbf{z}}] + \mu_l$ with t_l, we can re-write the expression in (4.12) as follows:

$$t_l e^{t_l} = \frac{M}{\xi \lambda_l} e^{\frac{0.5}{\gamma_l}} \cdot e^{\frac{\sum_m \eta_{ml}}{\lambda_l} + a_l^\top E[\bar{\mathbf{z}}] + \mu_l} = u_l. \tag{4.13}$$

The Newton update rule for t_l is thus given by:

$$t_l^n = t_l^o + \frac{u_l e^{-t_l^o} - t_l^o}{t_l^o + 1}. \tag{4.14}$$

Starting from a good initial solution, the Newton algorithm converges in just a few iterations. The precision parameter γ_l can be similarly updated using a fast Newton algorithm, with the gradient given as:

$$\frac{\partial \mathcal{F}}{\partial \gamma_l^{-1}} = -\frac{\lambda_l}{2} - \frac{M}{2\xi} e^{\bar{x}_l} \cdot e^{\frac{\gamma_l^{-1}}{2}} + \frac{1}{2\gamma_l^{-1}} = 0. \tag{4.15}$$

4.3.2 Parameter Estimation

Closed-form parameter updates can be obtained for all our model parameters. Again, by taking a derivative of the objective function $\mathcal{F}$ w.r.t the regression parameters and set the derivative to 0, the re-estimation equations can be written as follow:

$$\mathbf{A} = \left(\sum_d (\bar{\mathbf{x}}_d - \mu) E[\bar{\mathbf{z}}_d]^\top\right)\left(\sum_d E[\bar{\mathbf{z}}_d \bar{\mathbf{z}}_d^\top]\right)^{-1}, \tag{4.16}$$

$$\mu = \frac{1}{D}\sum_d \left(\bar{\mathbf{x}}_d - \mathbf{A}E[\bar{\mathbf{z}}_d]\right), \tag{4.17}$$

$$\mathbf{\Lambda}^{-1} = \frac{1}{D}\sum_d \left((\bar{\mathbf{x}}_d - \mu)(\bar{\mathbf{x}}_d - \mu)^\top + \mathbf{\Gamma}_d^{-1} - \mathbf{A}E[\bar{\mathbf{z}}_d]\bar{\mathbf{x}}_d^\top\right). \tag{4.18}$$

where $E[\bar{\mathbf{z}}_d] = \frac{1}{N_d}\sum_n \phi_n^d$ and $E[\bar{\mathbf{z}}_d\bar{\mathbf{z}}_d^\top] = \frac{1}{N_d^2}(\sum_n \text{diag}(\phi_n^d) + \sum_n \phi_n^d \sum_{m\neq n}\phi_m^{d\top})$. The Multinomial parameters for each topic can be similarly re-estimated with the following update rules:

$$\tilde{\gamma}_{lt}^w = \sum_{d=1}^D \sum_m \eta_{ml}^d 1(w_m^d = t) + \gamma, \tag{4.19}$$

$$\tilde{\gamma}_{kt}^r = \sum_{d=1}^D \sum_n \phi_{nk}^d 1(r_n^d = t) + \gamma. \tag{4.20}$$

4.3.3 Annotation

During annotation, we are given a test image without its annotation words. The goal is to infer the most likely captions for the test image using the fitted model. The key step lies in how to compute the conditional probability $p(w|\mathbf{R})$ given by:

$$p(w|\mathbf{R}) = \int p(w, \mathbf{x}, \mathbf{Z}|\mathbf{R}) d\mathbf{x} d\mathbf{Z} \tag{4.21}$$

$$= \int p(w|\mathbf{x})p(\mathbf{x}|\mathbf{Z})p(\mathbf{Z}|\mathbf{R}) d\mathbf{x} d\mathbf{Z}. \tag{4.22}$$

By approximating $p(\mathbf{Z}|\mathbf{R})$ with a variational posterior $q(\mathbf{Z}|\mathbf{R})$ inferred from the image modality alone and using a point estimate of $p(\mathbf{x}|\mathbf{Z}) \approx \delta(x - \hat{x})$, where $\hat{x} = \mathbf{A}\bar{\mathbf{z}} + \mu$, and $\bar{\mathbf{z}} = \frac{1}{N}\sum_n \phi_n$, we obtain the following approximation of the conditional probability:

$$p(w = t|\mathbf{R}) \approx \sum_l \beta_{lt}^w \cdot \frac{\exp(\hat{x}_l)}{\sum_j \exp(\hat{x}_j)}. \tag{4.23}$$

4.4 Experimental Results

In this work, we employ 2 standard image annotation datasets: COREL and LabelMe datasets, which have been described in some detail in Chapter 3. The COREL dataset used in our experiment has 5,000 documents, with 4,500 used in training and 500 as a test set. Each image in the COREL collection is reduced to size 117 × 181 (or 181 × 117) and is treated as a collection of 20 × 20 patches obtained by sliding a window with a 20-pixel interval, resulting in 45 patches per image. The caption word dictionary for COREL has $T_w = 374$ words, with the number of captions in each image ranging from 1-5 words, with a collection average of 3.5 words. For the LabelMe dataset, we use the 8-category subset which contains 2,687 images from 8 classes. 2,147 images are used in training and 540 as a test set. Again we represent an image as a collection of 20 × 20 patches; with a 20-pixel interval, we obtain 144 patches per image. The caption word dictionary of this dataset is $T_w = 454$, with the number of caption words per image ranging from 2 to 71 words, with a collection average of 13.5 words. As described in Chapter 3, we use 128-dimensional SIFT descriptors computed on 20 × 20 gray-scale patches to describe shape and texture information. In addition, we follow the work in [vdWS06] and add additional 36-dimensional robust color descriptors which have been designed to complement the SIFT descriptors extracted from the gray-scale patches. We run k-means on a collection of 164-dim features to learn a dictionary of $T_r = 256$ visual words.

Caption words in the COREL dataset indeed act more similar to class labels in that they are binary (0/1) and as a result they do not distinguish between an image with 1 instance of an object appearing in the scene and another image containing multiple instances of the same object. For the LabelMe dataset, on the other hand, the manual annotation process to obtain the ground-truth data more accurately captures this frequency statistics. Hence, in the scenario where multiple cars appear on a highway scene, the word 'car' will be repeated multiple times in the corresponding caption data. With the choice of a Multinomial distribution for caption data, our model indeed is more suitable for handling free-form annotation texts. Experiments on image annotation in this chapter, therefore, are focused on the LabelMe dataset.

In addition to modeling image-caption data, we show some results using tr-mmLDA to model video-caption data. As described in detail in Chapter 3, the multimedia dataset used in our experiment is obtained by recording television program broadcasts. By dividing each long recording into a set of 20-sec clips, we gather, from 24 episodes of a TV show called Modern Marvels, a total of 2,502 multimedia documents (each clip is a document). 2,100 clips are used in training 402 documents for testing. Each video is treated as a collection of 8 × 8 × 8 blocks, resulting in approximately 500 blocks per video. By extracting the spatio-temporal ICA features, described in detail in Chapter 7, from each video block, we run a mixture of Laplacians model on a collection of ICA features to learn a dictionary of $T_r = 256$ video words. Caption data associated with each video is obtained from the closed-caption texts provided by the broadcasting station. By removing common and rare words, we are left with a 4,171-term vocabulary.

4.4.1 Caption Perplexity

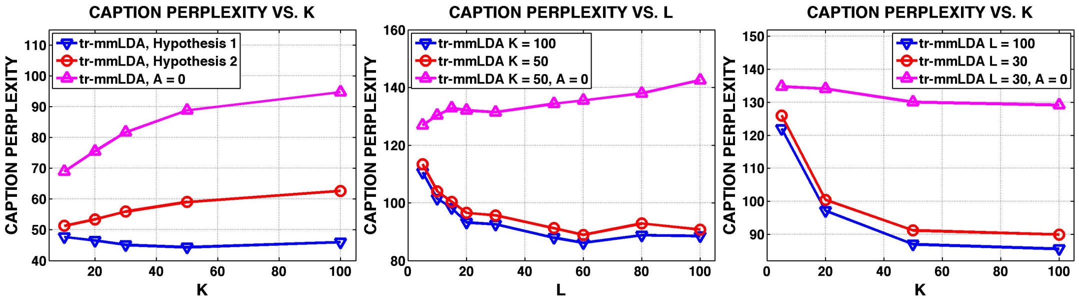
To measure the quality of annotations predicted by annotation models, we follow [BJ03] and adopt caption perplexity as a performance measure. The essential quantity that we need to compute is the conditional probability of caption words given a test image $p(w|\mathbf{R})$, which is computed with respect to the variational posterior, as given in (4.23). In comparing the different association models for image-caption data, a model that better captures the conditional relationships between images features (input variables) and their captions (target variables) will therefore lead to a higher conditional probability on unseen data. As seen in the definition in

(4.24), perplexity is indeed the inverse of the geometric mean of the likelihood, which implies that the model with higher conditional likelihood will give a lower perplexity. Hence, counter-intuitvely the lower perplexity value means a better model.

$$\text{Perp} = \exp\left(-\frac{\sum_{d=1}^{D}\sum_{m=1}^{M_d}\log p(w_m|\mathbf{R}_d)}{\sum_d M_d}\right) \tag{4.24}$$

We start our analysis with an experiment that explores properties of tr-mmLDA. For all the experiments done on this model, unless specified otherwise, we will assume for convenience that the number of image and caption topics are the same, i.e. $K = L$. In the first experiment, we show the benefit of simultaneously learning the latent topic representations and the regression module that captures their relationships. Since tr-mmLDA can be seen as linking 2 topic models using a regression module, one could argue that these 2 stages need not be learned simultaneously. In such a simplistic approach, we first learn 2 topic models which give low-dimensional representations of image and their caption data. Subsequently, we then fit a regression module which learns the mapping between the low-dimensional representations of the 2 data types. We denote Hypothesis 1 as the hypothesis that supports concurrent updating and Hypothesis 2 as subsequent learning of topic models and the regression parameters. As shown on the plot in Fig 4.2 (a) for the LabelMe dataset, Hypothesis 1, shown in blue, gives lower perplexity than Hypothesis 2, in red, for all values of K. We also compare Hypothesis 1 and 2 to the benchmark (magenta curve) which is a plot of perplexity as a function of K when the regression coefficient matrix $\mathbf{A}$ is set to 0. The models that learn regression coefficients $\mathbf{A}$ significantly reduce the caption perplexity by over 30-40 words compared to when setting $\mathbf{A} = 0$, and the model that updates $\mathbf{A}$ together with the topic parameters yields the lowest perplexity. We argue that the concurrent updating of the 2 sets of parameters allows the topics of the caption modality to influence how image topics are chosen; hence we obtain a low-dimensional projection of images that are more predictive of caption words, which leads to a better prediction performance.

Note that in Fig 4.2 (a) as the number of topics K increases, the perplexity gap between Hypothesis 1 and 2 also grows. We attribute this phenomenon to the problem of over-fitting when topic model parameters of the 2 modalities are learned separately. As seen in the update for the variational posterior $\log \phi_n$ in (4.8), the regression module connecting the two topic models introduces the last 3 terms of RHS of (4.8), which encourages more topics to be used to explain each image word (image words in each document to be softly assigned to more clusters (topics)). Consequently, the posterior $q(z_n)$ under Hypothesis 1 will have higher entropy compared to Hypothesis 2. As K increases, the separate updating of the 2 sets of topic parameters (Hypothesis 2) will require more data to obtain good estimate of the topic parameters, and as a result will over-fit more easily when data is scarce. Indeed, when learning the latent variable representations of the 2 data modalities together along with the regression coefficients, caption data exert influence on how image topics are selected and can be thought of as an extra set of virtual image words, which helps with over-fitting. Image topics, on the other hand, can be regarded as sources



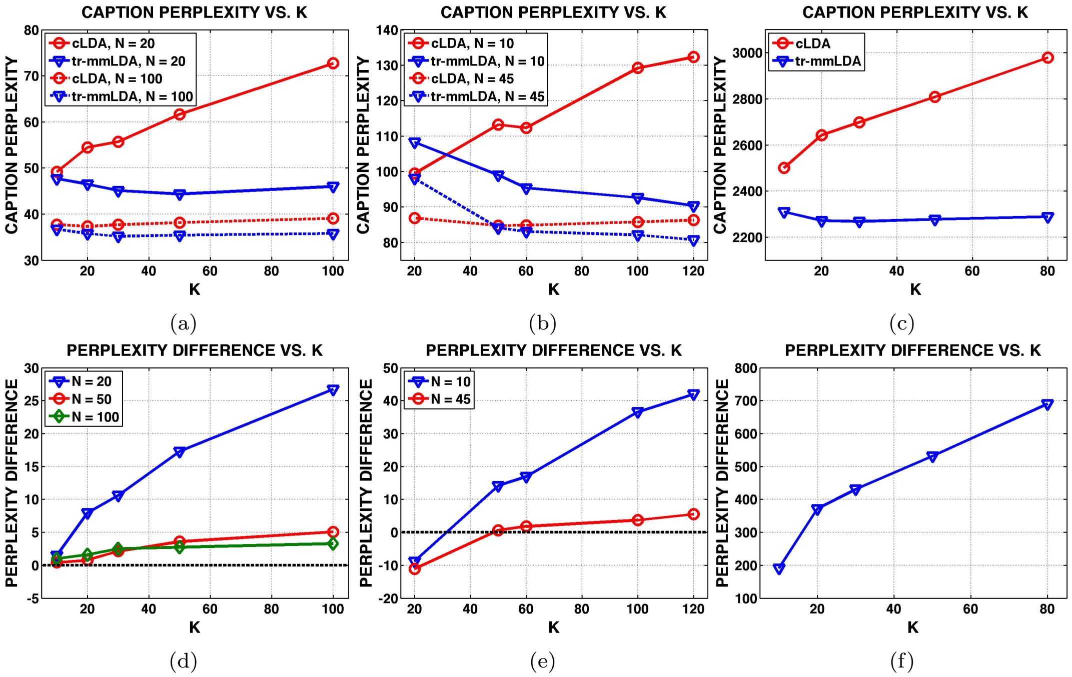
(a) (b) (c)

Figure 4.2: **(a)**: Caption perplexity for LabelMe dataset as a function of K for Hypothesis 1(blue) and Hypothesis 2 (red). **(b)**: Caption perplexity for COREL dataset as a function of L, while holding K fixed. **(c)**: Caption perplexity for COREL dataset as a function of K, while holding L fixed.

or factors that control the activation of certain groups of caption topics, similar to the Factor analysis prior of IFTM in [PAN09] that replaces the Dirichlet prior for θ in LDA. Using the same analogy, caption words in each document will be softly assigned to more clusters (topics) under Hypothesis 1, which allows parameter estimates of the topic parameters to be more accurate when data is limited. When the 2 sets of parameters are learned together, the 2 data modalties indeed help each other out and over-fitting becomes less of an issue.

We also observe a similar pattern in the COREL dataset as shown in the plot of caption perplexity as a function of number of caption topics L while holding the number of image topics K fixed in Fig 4.2 (b). Generally, caption perplexity decreases as L increases for $K = 50$ (in red) and $K = 100$ (in blue), and we do not observe problems with over-fitting. For all values of L, we obtain a lower perplexity with $K = 100$ than with $K = 50$. We also compare these 2 plots against the benchmark (magenta curve) which shows caption perplexity as a function of L for $K = 50$ but with the regression coefficient matrix $\mathbf{A}$ set to 0. Indeed, the improvement over the benchmark can reach over 50 words. A similar pattern is again observed in the plot of caption perplexity as a function of K while holding L fixed in Fig 4.2 (c). We conclude that tr-mmLDA is quite robust to over-fitting as model complexity increases (K or L increases).

In the second experiment, we compare the predictive performance of cLDA and tr-mmLDA. We use $K = L$ for tr-mmLDA throughout this experiment and plot caption perplexity as a function of the number of topics K comparing the 2 association models. To see if over-fitting is an issue for cLDA, we observe how perplexity (as a function of K) changes as the number of patches in each image—N—changes. To obtain images with varying values of N, we sub-sample N patches from all the patches to represent each image. As shown in Fig 4.3 (a), for a small value of $N = 20$, cLDA seriously overfits the data as caption perplexity increases dramatically as K increases. With larger N, over-fitting becomes less of an issue, but tr-mmLDA still outperforms cLDA as seen in the plots in Fig 4.3 (d) showing the perplexity difference between cLDA and



(a) (b) (c)

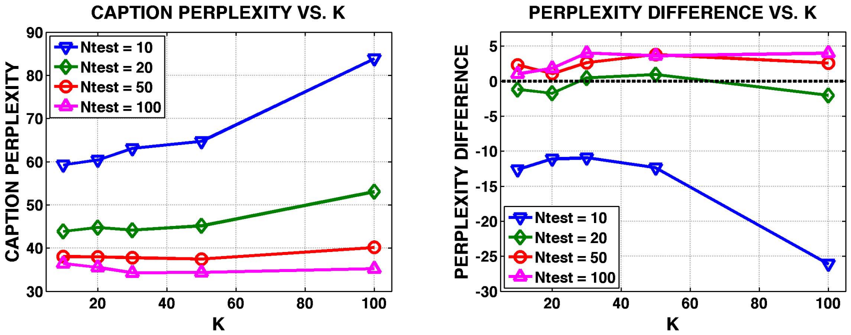
(d) (e) (f)

Figure 4.3: **(a)** Perplexity as a function of K as N increases (LabelMe). **(b)** Perplexity as a function of K as N increases (COREL). **(c)** Perplexity as a function of K (video-caption). **(d)** Perplexity difference between cLDA and tr-mmLDA as a function of K for varying values of N (LabelMe). **(e)** Perplexity difference between cLDA and tr-mmLDA as a function of K for varying values of N (COREL). **(f)** Perplexity difference between cLDA and tr-mmLDA as a function of K (video-caption).

tr-mmLDA rising above the x-axis for all values of K and N.

Indeed, the severe over-fitting in cLDA can be directly attributed to the restrictive association between the 2 modalities. When N is small, since cLDA enforces that the topics used in the caption modality must be a subset of the topics occurring in the image modality, a small value of N implies that the words in each document will be assigned to only a small number of topics (clusters). For a large value of K, in order not to have empty clusters (topics), a large number of documents will be required. With less than 2200 training documents in the LabelMe dataset, the topic parameters for caption texts will therefore be estimated poorly. For the COREL dataset with a larger training set (4500 documents), the problem of over-fitting does not become severe until we reduce N down to 10, as shown in Fig 4.3 (b) and (e). Since tr-mmLDA imposes no such restriction with regards to the number of topics used for the caption modality, over-fitting becomes less of an issue using the same dataset size. We observe a similar trend of over-fitting for cLDA in the video-caption dataset with 2,100 training documents, even though we use $N = 200$ in the plot in Figure 4.3 (c). The corresponding perplexity difference between cLDA and tr-mmLDA is shown in Figure 4.3 (f).

To better understand the differences between the 2 association models, we conduct another experiment where cLDA and tr-mmLDA models are fitted to the LabelMe dataset for $N = 100$ (over-fitting is no longer an issue for cLDA at this value of N for the LabelMe dataset).



(a)

(b)

Figure 4.4: **(a)**: Caption perplexity as a function of K as N increases (LabelMe dataset). **(b)**: Perplexity difference between cLDA and tr-mmLDA as a function of K for varying values of N (LabelMe dataset).

During prediction, however, we present a set of test images with varying values of N. Fig 4.4 (a) shows caption perplexity obtained from tr-mmLDA and (b) perplexity difference between cLDA and tr-mmLDA when the number of patches in the test set takes values $N = \{10, 20, 50, 100\}$. When N is very small, cLDA indeed gives lower caption perplexity due to the restrictive association, but as N increases the advantage over tr-mmLDA disappears and tr-mmLDA gives a superior prediction performance. We attribute the poor performance of tr-mmLDA when the number of patches N is small to a poor estimate of the empirical image topic frequency $\bar{\mathbf{z}}$. Since $\bar{\mathbf{z}}$ serves to activate groups of correlated caption topics via the mapping specified in the regression matrix $\mathbf{A}$, when the estimate for $\bar{\mathbf{z}}$ is poor, the activation of caption topics is more random and hence a poor perplexity score (high perplexity). As N increases, the estimate of $\bar{\mathbf{z}}$ is more accurate, and tr-mmLDA is able to draw upon a group of correlated topics to explain caption words, while cLDA is restricted to a small set of topics that occur in the image modality. Just as CTM and IFTM outperform LDA on a similar task, as discussed in Chapter 2, tr-mmLDA outperforms cLDA by making use of the information on topic correlations. Since N can indeed be controlled by our choice of image representation, when an image is modeled as a collection of 6-10 regions obtained via image segmentation as in [BJ03], the association model of cLDA might indeed be better at modeling such a restrictive association. In representing an image as a bag of patches, the value of N tends to be quite large. In such a scenario, tr-mmLDA is therefore a more suitable association model.

4.4.2 Example Topics

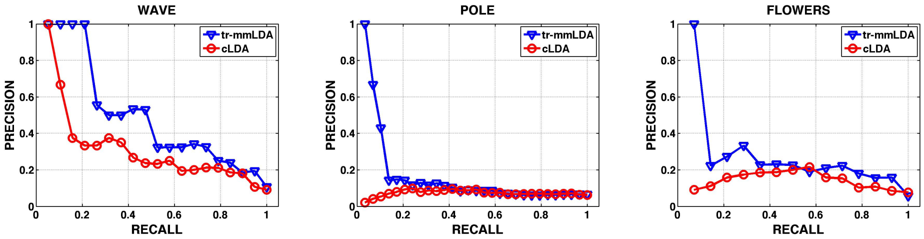
We compare examples of caption topics from the LabelMe dataset learned using cLDA and tr-mmLDA. As in other previously proposed topic models, we examine the topic Multinomial parameters over words and employ 10 most probable caption words to represent each topic. We learn $K = 50$ topics using cLDA and found around 50% of those topics learned has the word *car* in its top 10 caption words, while only 4 out of 50 topics learned using tr-mmLDA contains the word *car*, see Table 4.1. The *car* example used here illustrates that topics learned using cLDA are found to contain more general terms, while tr-mmLDA correctly uncover the 4 contexts *car* appears in the dataset: class highway, street, inside city, tall building. In using these topic parameters to predict caption words, it is therefore more likely that caption words predicted by cLDA will contain more general and vague terms, while predicted captions from tr-mmLDA, with more semantically meaningful topic parameters, will be more relevant and accurate.

Table 4.1: Example topics from 50 topics learned using cLDA (top panel) and tr-mmLDA (bottom panel) on the LabeMe dataset. Each topic is represented by 10 most likely words.

Topic 1	car, sky, road, tree, building, mountain, trees, window, buildings, stone
Topic 2	building, window, car, person, buildings, skyscraper, walking, sky, sidewalk
Topic 5	sky, road, car, fence, mountain, trees, sign, street light, tree, cabin
Topic 6	car, bulding, buildings, person, sidewalk, cars, walking, road, sky, van
Topic 14	window, building, car, door, person, pane, road, sidewalk, column, sky
Topic 15	sky, tree, building, mountain, car, road, trees, buildings, aerial, snowy
Topic 16	buildings, car, sky, tree, cars, building, road, van, person, trees
Topic 17	car, road, sign, trees, street light, highway, sky, freeway, central, reservation
Topic 20	car, road, highway, freeway, sign, trees, streetlight, sky, central, reservation
Topic 24	building, tree, sky, person, car, buildings, skyscraper, mountain, walking, road
Topic 26	building, car, buildings, sky, mountain, sidewalk, road, person, cars, tree
Topic 27	car, building, buildings, person, sidewalk, road, tree, sky, walking, window
Topic 33	window, building, door, car, pane, plant, road, sky, side walk, tree
Topic 39	person, walking, car, building, sidewalk, buildings, window, road, standing
Topic 40	window, balcony, door, building, car, shop, person, sidewalk, road, awning
Topic 41	building, skyscraper, buildings, sky, car, road, window, sidewalk, person, tree
Topic 42	window, car, building, road, sidewalk, sky, door, sign, plant, person
Topic 45	car, building, buildings, road, sidewalk, person, cars, bus, sky, van
Topic 46	mountain, snowy, sky, rocky, car, snow, tree, trees, car, glacier, person
Topic 50	sky, trees, road, car, sign, freeway, highway, tree, building, mountain
Topic 15	car, road, sign, trees, highway, freeway, sky, fence, central, reservation
Topic 32	car, buildings, building, sidewalk, cars, road, sky, van, poster, crosswalk
Topic 39	car, road, car back, car top back, van, car right, car left, car top front, car front
Topic 48	balcony, shop, building, door, car, terrace, light, road, attic, metal

4.4.3 Example Annotation

We show examples of predicted annotation on the LabelMe dataset, comparing tr-mmLDA and cLDA using $K = 50$. As seen in Table 4.2, captions generated from cLDA contain slightly more irrelevant terms. In 3 examples in Table 4.2 (row 3 left, row 4 and 5 right) showing scenic pictures of trees, lake, and mountain tops belonging to class ‘mountain’ and ‘open country’, cLDA predicts ‘car’ as the first or second annotation word. Indeed, this poor performance should not come as a surprise due to the poor topic parameter estimate obtained by cLDA as shown in Table 4.1 which assigns high probabilities to the word car in many topics. In the examples in row 6 left-right which belong to class ‘street’, cLDA also predict irrelevant words such as



(a) (b) (c)

Figure 4.5: Precision-recall curve for 3 single word queries: 'wave', 'pole', 'flower', comparing cLDA (red) and tr-mmLDA (blue).

'mountain' and 'snowy' for the content of the 2 images describing streets in a city scene. In all examples shown in Table 4.2, the predicted words generated from tr-mmLDA are more related to the ground-truth captions and to the image category than those predicted by cLDA.

4.4.4 Text-based Retrieval

We can also use the annotation model of tr-mmLDA and cLDA to perform image retrieval on a database of un-annotated images using word queries. This retrieval method is called text-based retrieval, to contrast with content-based retrieval where queries are given in the form of examples. Given a single word query, we perform retrieval by ranking the test images according the probability that each image will be annotated with the query word. More specifically, the score used in ranking is $p(w|\mathbf{R}_{test})$ which can be computed using variational posterior inferred for each test document as in (4.23).

Table 4.3 shows examples of images retrieved using tr-mmLDA with query words 'hill', 'buildings', 'road', 'brushes', 'snowy', and 'sidewalk'. To evaluate the performance of a retrieval system, we need a performance measure that evaluate the rank order in which relevant images are retrieved. In this work, we use the mean average precision measure, which computes for each query word an average over the precision values computed at every point where a relevant item is retrieved from the database. We define an image to be relevant if the true caption contains the query word of interest. At $K = 50$, on the LabelMe dataset, using tr-mmLDA, the mean average precision is 16%, while for cLDA the mean average precision is 13%. More extensive experiments need to be performed to obtain optimized results. Figure 4.5 shows precision-recall curves for 3 single word queries, comparing the rank order of the retrieved images using cLDA and tr-mmLDA. Our model generally yields higher precisions at the same recall values for all the 3 queries and give a better overall retrieval performance.

Chapter 4, in part, is a reprint of the material as it appears in: D. Putthividhya, H. T. Attias, S. Nagarajan. "Topic-Regression Multi-Modal Latent Dirichlet Allocation for Auto-Annotation," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010,

Table 4.2: Examples of predicted annotation generated from tr-mmLDA and cLDA.

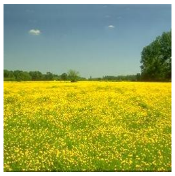
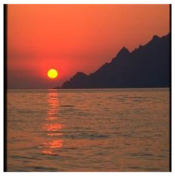
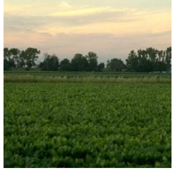
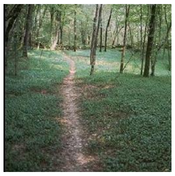
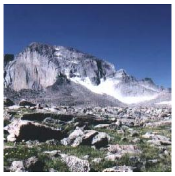
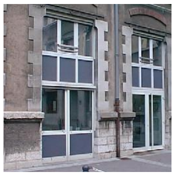
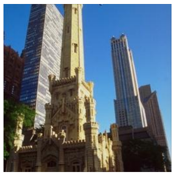
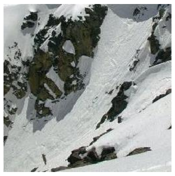
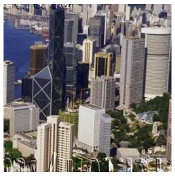
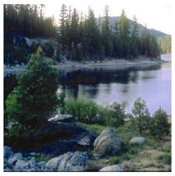
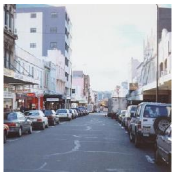
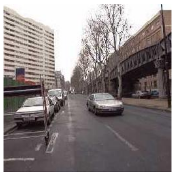
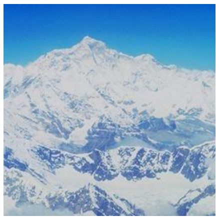
	**Groundtruth** sky, trees, flowers, field meadow **tr-mmLDA** sky, flowers, trees, field brushes **cLDA** mountain, trees, sky, tree field		**Groundtruth** water, sea, sky, mountain sunset, sun **tr-mmLDA** sky, clouds, water, sea sunset, mountain **cLDA** sea, water, sky, pane sand, rock
	**Groundtruth** sky,trees,field **tr-mmLDA** field, sky, mountain, grass trees **cLDA** sky, trees, mountain, tree water		**Groundtruth** tree, trunk, trees, ground grass, path **tr-mmLDA** tree, trees, trunk, sky ground, path, grass, forest **cLDA** tree, trees, trunk, sky mountain, grass, ground
	**Groundtruth** sky, mountain, rock, rocky plain **tr-mmLDA** sky, trees, mountain, building **cLDA** car, sky, trees, building buildings		**Groundtruth** building, road, pole, door pane **tr-mmLDA** window, building, door, car road, sidewalk **cLDA** window, building, skyscraper, door
	**Groundtruth** sky,building,tree,skycraper **tr-mmLDA** building, skyscraper, buildings, sky **cLDA** building, skyscraper, sky, mountain		**Groundtruth** mountain, rocks, snow, snowy cap **tr-mmLDA** snowy, mountain, sky, tree snow **cLDA** car, mountain, building, buildings sky
	**Groundtruth** water, building, river, buildings skyscraper **tr-mmLDA** building, skyscraper, buildings, sky **cLDA** window, building, car, sky door		**Groundtruth** water, sky, mountain, tree rocks, shrubs, river **tr-mmLDA** mountain, trees, sky, tree rpclu. ground, river **cLDA** trees, car, river, tree water, sky, field
	**Groundtruth** car, sky, building, bridge tree, road **tr-mmLDA** car, building, building, sidewalk road, cars **cLDA** window, car, sky, building mountain, trees		**Groundtruth** car, sky, building, bridge tree **tr-mmLDA** car, buildings, building, sidewalk road **cLDA** car, building, snowy, mountain person

Table 4.3: Examples of images (with no captions) retrieved using single word queries.

hill				
buildings				
road				
brushes				
snowy				
sidewalk				

and D. Putthividhya, H. T. Attias, S. Nagarajan, T.-W. Lee, “Probabilistic Graphical Model for Auto-Annotation, Content-based Retrieval and Classification of TV clips containing Audio, Video, and Text,” in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2007. I was the primary researcher of the cited materials and the co-author listed in these publications supervised the work which forms the basis of this chapter.

Chapter 5

Supervised Latent Dirichlet Allocation with Multi-variate Binary Response Variable (sLDA-bin)

So far we have discussed several topic models for learning statistical association between image or video features and their corresponding caption texts. All the topic models discussed in Chapter 4—multi-modal LDA (mmLDA), correspondence LDA (cLDA), and topic-regression multi-modal LDA (tr-mmLDA)—assume free-form caption texts and maintain the core assumption of LDA in allowing caption words in the same document to be generated from multiple hidden topics. We find that such an assumption might not be a good fit for many datasets used in performance evaluation, as captions in such datasets are often manually obtained and only a few words are used to annotate each image. Combined with the fact that no words are used more than once in labeling each image, each entry in a vector of word count representing the caption data reduces to a binary value (0/1). For such annotation data, topic models which learn patterns of word co-occurrences from statistics of word frequencies (word counts) might indeed not be the most appropriate choice.

In this chapter we focus on annotation models more suitable for such binary annotation data. Previous work in the related literature has mainly boiled down to the 2 extremes. The first line of approach treats binary annotation as class labels [LW03, CV05, CCMV07, YCH06, JCL07, EXCS06] and employ a classification framework for predicting annotation. In [LW03, CV05, CCMV07], a generative classifier is built for each annotation word by learning class-conditional density from all the images tagged with that word. Annotation on a test image is done by selecting the top few words whose class-conditional density models provide the best fit on the test image. For video annotation in [YCH06, JCL07], classifiers of different semantic concepts are learned together with the statistics of co-occurrences of concepts using a unified framework of conditional random field models. Temporal label correlations between adjacent video clips are captured in [EXCS06] using hidden markov models.

The second line of approach models the joint correlation of image features and the corresponding captions using a set of shared latent variables to represent common causes of correlations between the 2 data modalities. In such an approach, traditionally Multinomial distribution [BDdF^{+}03, JLM03, LMJ03] is assumed in modeling caption data. When considering binary annotation data as class labels, the multinomial distribution might indeed be a poor choice as it splits the probability mass between multiple caption words appearing in the same image [FML04]. Under a multinomial model, an image labeled with the words 'horse' and 'grass' will be less likely to be retrieved using the query word 'horse' than another image labeled with the word 'horse' alone. In [FML04], a multi-variate Bernoulli model is adopted to avoid such a problem. Under a multi-variate Bernoulli model, the 2 images tagged with the word 'horse' in the above example will be equally likely since the decision to label an image with a particular label is made independently of other labels.

In this work, we modify the topic-regression multi-modal LDA (tr-mmLDA) in Chapter 4 to work with binary annotation data. Following the footstep of [FML04], we replace a Multinomial distribution which is more suitable for capturing statistics of word frequencies with a multi-variate Bernoulli model, which is more suitable for binary annotation data. Instead of a linear regression module that predicts the latent topics of the caption modality, we introduce a logistic regression module that maps from the latent topics of the image modality directly to the binary response variable corresponding to the presence/absence of each annotation word. By adopting the image topic frequency covariate as in tr-mmLDA, our model indeed can be seen more as a direct extension of supervised LDA (sLDA) to handle a multi-variate binary response variable, with the use of a logistic sigmoid function. We therefore adopt the name supervised LDA-binary (sLDA-bin) for our model. The use of a logistic sigmoid function makes the computation for inference and parameter learning intractable. We derive an efficient variational inference algorithm based on mean-field approximation and adopt a tight convex variational bound for the logistic function [SSU03, JJ97]. Using a subset of the COREL dataset with 5000 images, we demonstrate the power of our model on an image annotation task. Experimental results show that sLDA-bin performs more favorably to cLDA as measured by caption prediction probability.

5.1 Notations

As mentioned in Chapter 3, we borrow a tool from statistical text document analysis and represent an image as a bag of words. In such a representation, word ordering is ignored and an image is simply reduced to a vector of word count. We adopt the following notation. Each image is a collection of N patches and is denoted as $\mathbf{R} = \{r_1, r_2, \ldots, r_N\}$ where r_n is a unit-basis vector of length T_r with exactly one non-zero entry representing the membership of the patch n to only 1 codeword in a dictionary of T_r visual words. For caption data, we simply record the presence/absence of each caption word in an image. We denote this as a $T_w \times 1$ binary vector $\mathbf{w}$, where the entry w_i takes a value 1 if word i is present in an image and 0 otherwise. A collection of D image-caption pairs is denoted as $\{\mathbf{R}_d, \mathbf{w}_d\}$, $d \in \{1, 2, \ldots, D\}$.

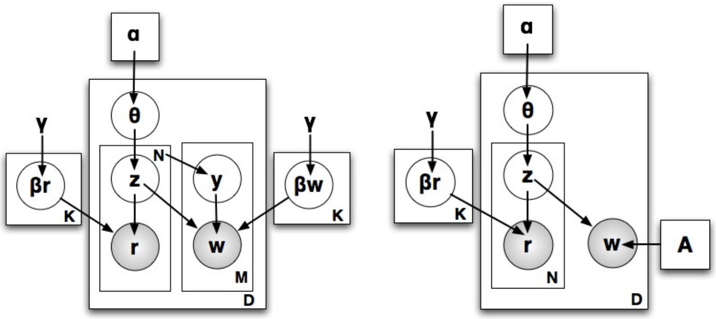
5.2 Proposed Model

We first briefly review the basic Latent Dirichlet Allocation model and describe how supervised LDA model (sLDA) in [BM07] extends LDA to model documents paired with one-dimension real-valued response variables. Then we describe our proposed model—sLDA-bin—which extends the basic sLDA model in [BM07] to handle multi-variate binary response variable.

5.2.1 Supervised Latent Dirichlet Allocation (sLDA)

Supervised LDA builds on the basic LDA model which uses hidden variables, loosely termed topics, to cluster words and model word co-occurrences. Given a collection of documents in a bag-of-word form, LDA decomposes the distribution of word counts from each document into contributions from K topics. A document under LDA is modeled as a proportion of topics which assumes the form of a Dirichlet distribution, while each topic, in turn, is a multinomial distribution over terms. When the number of topics K is much smaller than the size of vocabulary, LDA can be seen as performing dimensionality reduction by learning a small set of projection directions (topics) that account for most of the correlations in the data. The low-dimensional subspaces (topics) uncovered by LDA often reveal semantic structures useful for visualization, browsing, and navigation through large repositories of text collections.

When working with labeled documents or documents paired with their associated response values, where the goal is to predict the response values given the document, it might be useful to incorporate these response variables into learning the low-dimensional mapping of documents. This is precisely the motivation behind the supervised LDA (sLDA) model proposed in [BM07]. Instead of inferring hidden topics that best explain correlations between words in documents, sLDA finds latent topics that are best predictive of the response variables. More specifically, sLDA incorporates a linear regression module to LDA which allows a real-valued one-dimensional response variable to be linearly predicted from the latent topics of the corresponding document. In order for the response variable to directly influence the selection of topics, instead of regressing over the mean topic proportion θ as in mmLDA [BJ03], the formulation in [BM07] makes the response variable directly dependent on the actual topics that occur in



(a) cLDA

(b) sLDA-bin

Figure 5.1: Graphical model representation comparing (a) Correspondence LDA (cLDA) and (b) Supervised Latent Dirichlet Allocation with a multi-variate binary response variable (sLDA-bin).

the document, by using the empirical topic proportion $\bar{\mathbf{z}} = \frac{1}{N}\sum_{n=1}^{N} 1(z_n)$ as an input into the regression model. Such a setup treats the response variable as non-exchangeable with words in the document and indeed prevents the learning of topics that are used entirely to explain either the response variables or the words in the documents, as these topics will not be useful in predicting the response variables. In [WBFf09], sLDA was extended to handle response variables that are of categorical type (e.g. class labels) with the use of a softmax function for a multi-class classification problem.

5.2.2 Supervised Latent Dirichlet Allocation with Multi-variate Binary Response Variable (sLDA-bin) [PAN10a]

Motivated by such a success of sLDA in a prediction task, here we adopt the framework of sLDA in learning statistical association between image features and the corresponding captions. In such a setting, documents are images while caption data is the response variable that we want to predict. The goal now is to infer low-dimensional representation of images (image topics) that are predictive of caption words. In order to handle the multi-variate binary response variables of the annotation data, instead of a linear regression module used in the basic sLDA model, we introduce a logistic regression module which maps from the empirical image topic frequency to the target binary variable corresponding to the presence/absence of each caption word. We model the conditional distribution of response variables given image topics as a multi-variate Bernoulli and use the logistic function of linear combinations of $\bar{\mathbf{z}}$ to define its probability. Given the empirical image topic frequency $\bar{\mathbf{z}}$, the probability of labeling the current image with caption word i is given by

$$p(w_i|\mathbf{Z},\mathbf{A}) = \sigma(\mathbf{a}_i^\top \bar{\mathbf{z}})^{w_i}\sigma(-\mathbf{a}_i^\top \bar{\mathbf{z}})^{1-w_i}, \tag{5.1}$$

where $w_i \in \{0, 1\}$ with $w_i = 1$ denoting the presence of word i in an image. $\sigma(x) = \frac{1}{1+e^{-x}}$ denotes a logistic sigmoid function. $\mathbf{a}_i$ are the corresponding regression coefficients for word i which will be learned from the training data. Each regression coefficient describes the relative influence that each topic exerts on the probability that the outcome will be 1, corresponding to word w_i being used in labeling a given image. For example, if the occurrence of topic k in an image highly increases the probability of the outcome, the corresponding regression coefficient a_{ik} will take a positive value further away from 0.

Given the number of topics K and the model parameters $\{\beta^r, \alpha, \mathbf{A}\}$, to generate an image-caption pair $\{\mathbf{R}, \mathbf{w}\}$ with N image patches, we follow the generative process of sLDA-bin, which is illustrated in the graphical model representation in Figure 5.1(b):

- Draw a topic proportion $\theta|\alpha \sim \mathrm{Dir}(\alpha)$
- For each image patch r_n, $n \in \{1, 2, \ldots, N\}$
    1. Draw topic assignment $z_n|\theta \sim \mathrm{Mult}(\theta)$
    2. Draw word $r_n = t|z_n = k \sim \mathrm{Mult}(\beta^r_{kt})$
- Given the empirical topic proportion $\bar{\mathbf{z}} = \frac{1}{N}\sum_{n=1}^{N} 1(z_n)$, for each caption word w_i, $i \in \{1, 2, \ldots, T_w\}$:
    1. Draw a Bernoulli r.v. w_i with probability given as follows:

$$p(w_i = 1|\mathbf{Z}, \mathbf{A}) = \sigma(\mathbf{a}_i^\top \bar{\mathbf{z}}).$$

Indeed, a special case of sLDA-bin with univariate Bernoulli response variables can be straight-forwardly obtained for modeling documents paired with univariate binary responses. When considering each annotation word independently and discarding the presence/absence information of the other words, we can train a specialized sLDA-bin model from a training set of images paired with the binary response values of each annotation word. In such a scenario, instead of inferring one unified set of image topics influenced by the presence/absence of all annotation words in the vocabulary, we learn a total of T_w sets of specialized image topics, each set is highly predictive of the presence/absence of a particular annotation word. Annotation on a test image is done by fitting all the models on the test image and make a 0/1 prediction on each annotation word independently. We select the top few words with the highest probability of predicting an outcome 1, corresponding to the event that the words will be used to label the test image. Since a binary classifier is trained for each annotation word to reflect the decision of whether or not to label a given image with each word, the annotation system described here is equivalent to learning a set of T_w one-vs-all (OVA) probabilistic discriminative classifiers (since each classifier is trained on both positive and negative samples). This is in contrast with the use of generative classifiers for the problem of auto-annotation in [LW03, CV05, CCMV07] where only positive samples are used in learning each classifier.

Our model is related to correspondence LDA (cLDA) [BJ03] described in detail in Chapter 4, which also extends the basic LDA model for an image/video annotation task. Beside the obvious difference in the choice of distributions assumed for caption words (Multinomial vs.

multi-variate Bernoulli), another main difference between the 2 models lies in how image features are associated with their captions. With the goal of image region annotation, each caption word under cLDA is restricted to be associated with 1 particular image region. As seen from the graphical model of cLDA in Figure 5.1(a), each caption word is generated by first selecting an image region to associate with; now using the hidden topic of that region, we generate a word according to its topic-specific Multinomial distribution over terms. In practice, however, some annotation words do globally describe the scene as a whole, using such a restrictive association model could prove to be very inaccurate. By regressing over the empirical topic proportion, the formulation of sLDA-bin allows each caption word to be influenced by the topics from all image regions as well as by a particular image region depending on the corresponding regression coefficients, which are learned entirely from the data. Our association model is thus more general and accurate in capturing the true nature of the relationship between hidden structures in the image modality and the corresponding caption words.

5.2.3 Supervised LDA with other covariates

Besides the empirical topic frequency $\bar{\mathbf{z}} = \frac{1}{N}\sum_{n=1}^{N} 1(z_n)$ as proposed in the basic sLDA model, we are also interested in extending sLDA to include other forms of input variables (covariates) into the regression module. Indeed, since we are adopting a statistical topic model for representing images, it might be helpful to partially retain some spatial information which has long been known to be useful in various image classification and recognition tasks. To this end, instead of averaging over the indicator variables of all the topics occurring in an image, we propose to concatenate these variables. More specifically, we replace the covariates $\bar{\mathbf{z}} = \frac{1}{N}\sum_{i=1}^{N} 1(z_n)$ with $\mathbf{Z} = [1(z_1), \ldots, 1(z_N)]$. The generative process of an image-caption pair remains the same, but we now plug in a different input variable $\mathbf{Z}$ to the logistic regression module for predict binary annotation. The real difference between the 2 choices of covariates, however, lies in a much larger number of regression coefficients that needs to be learned. More specifically, in the case of the topic frequency input variables $\bar{\mathbf{z}}$ we have $T_w \times K$ parameters where T_w is the number of words in the caption dictionary and K is the number of hidden topics. In the case of concatenated topic variables $\mathbf{Z}$, we need to learn $T_w \times K \times N$ regression parameters, where N is the number of patches in each image. One caveat to this choice of covariates is that the size of images used in training and testing must be the same. To be more precise, it is the number of patches in each image that needs to be the same, not the actual dimensions of the images. Such a constraint, however, could further restrict the applicability of this extension.

5.3 Variational EM

To learn the model parameters for sLDA-bin, we use the maximum likelihood estimation where the goal is to find a model parameter setting that maximizes the likelihood of the data. The Expectation Maximization (EM) algorithm is a general framework for iteratively estimating parameters of statistical models with latent variables. Using Jensen's inequality, the E step of the EM algorithm derives an auxiliary function which lower-bounds a complex likelihood function to allow for a more simple optimization to be performed in the M step. This auxiliary lower bound is indeed a function of the posterior distribution over the hidden variables. When the posterior can be computed exactly, the lower bound obtained in the E step will be tight and touches the likelihood function. It can be shown that the new parameter updates obtained by maximizing such a lower bound is guaranteed not to decrease the data likelihood. Starting from an initial parameter estimate, EM lower-bounds the likelihood function at the current parameter estimate in the E step and optimizes such a lower bound to obtain a new estimate of the model parameters in the M step. By iteratively alternating between these 2 steps, the algorithm converges quickly to a local maximum nearby the starting point.

5.3.1 Varational Inference

To infer the posterior over hidden variables, we begin with the expression of the log-likelihood for an image-caption pair:

$$\log p(\mathbf{w}, \mathbf{R}|\Psi) \geq \int q(\mathbf{Z}, \theta) \left(\log p(\mathbf{w}, \mathbf{R}, \mathbf{Z}, \theta|\Psi) - \log q(\mathbf{Z}, \theta) \right) d\mathbf{Z} d\theta. \tag{5.2}$$

where Ψ denotes the model parameters $\{\beta^r, \mathbf{A}\}$. Equality in (5.2) holds when the posterior over the hidden variables $q(\mathbf{Z}, \theta)$ equals the true posterior $p(\mathbf{Z}, \theta|\mathbf{w}, \mathbf{R})$. Similarly to LDA, computing the exact joint posterior is computationally intractable as the hidden variables $\{\mathbf{Z}, \theta\}$ become highly dependent when conditioned on the observed document $\{\mathbf{w}, \mathbf{R}\}$. We employ an effiicient mean-field approximation which approximates the joint posterior distribution with a variational posterior in a factorized form: $p(\mathbf{Z}, \theta|\mathbf{w}, \mathbf{R}) \approx \prod_n q(z_n) q(\theta)$. With the use of a factorized posterior, the RHS in (5.2) is a strict lower bound of the data log-likelihood. The problem now becomes one of finding, within such family of factorized distributions, the variational posterior that maximizes such a lower bound. Evaluating the expectation w.r.t the factorized posterior $\prod_n q(z_n) q(\theta)$, the lower bound in RHS of (5.2) can be expressed as:

$$\begin{aligned} \mathcal{F} = &\sum_n E[\log p(r_n|z_n, \beta_r)] + E[\log \theta] + E[\log p(\theta|\alpha)] + \\ &\sum_{i=1}^{T_w} E[\log p(w_i|\mathbf{Z}, \mathbf{A})] - \sum_n E[\log q(z_n)] - E[\log q(\theta)]. \end{aligned} \tag{5.3}$$

The logistic function complicates the evaluation of the expectation term $E[\log p(w_i|\mathbf{Z}, \mathbf{A})]$ (no matter which posterior distribution we use). We make use of convex duality and represents the logistic function as a point-wise supremum of a square function of $\mathbf{a}_i^\top \bar{\mathbf{z}}$ [JJ97, SSU03]. Dropping the supremum, we obtain the following lower-bound of $\log p(w_i|\mathbf{Z}, \mathbf{A})$

$$\begin{aligned}
\log p(w_i|\mathbf{Z}, \mathbf{A}) &= w_i \log \sigma(\mathbf{a}_i^\top \bar{\mathbf{z}}) + (1 - w_i) \log \sigma(-\mathbf{a}_i^\top \bar{\mathbf{z}}) && (5.4)\\
&= \frac{(2w_i - 1)}{2} \mathbf{a}_i^\top \bar{\mathbf{z}} - \log(e^{\frac{\mathbf{a}_i^\top \bar{\mathbf{z}}}{2}} + e^{-\frac{\mathbf{a}_i^\top \bar{\mathbf{z}}}{2}}) && (5.5)\\
&\geq \frac{2w_i - 1}{2} \mathbf{a}_i^\top \bar{\mathbf{z}} - \log(2\cosh(\frac{\xi_i}{2})) - \lambda(\xi_i)(\mathbf{a}_i^\top \bar{\mathbf{z}}\bar{\mathbf{z}}^\top \mathbf{a}_i - \xi_i^2), && (5.6)
\end{aligned}$$

where going from (5.5) to (5.6), we introduce variational parameters ξ_i which correspond to the point of contact where the lower bound touches the logistic function. $\lambda(\xi_i)$ in (5.6) is a shorthand for $\frac{\tanh(0.5\xi_i)}{4\xi_i}$.

Due to the Dirichlet-multinomial conjugacy, the posterior $q(\theta)$ takes the form of a Dirichlet and we use $\tilde{\alpha}$ to denote its parameters. To simplify the notation, we write $q(z_n = k)$ as ϕ_{nk}. By making use of the following expectations $E[\bar{\mathbf{z}}] = \frac{1}{N} \sum_n \phi_n$ and $E[\bar{\mathbf{z}}\bar{\mathbf{z}}^\top] = \frac{1}{N^2}(\sum_n \operatorname{diag}(\phi_n) + \sum_n \phi_n \sum_{m \neq n} \phi_m^\top)$, we can now express the likelihood lower-bound $\mathcal{F}$ as a function of the variational posterior parameters

$$\begin{aligned}
\mathcal{F} = &\sum_{n,k} \phi_{nk} \sum_t 1(r_n = t) \log \beta^r_{kt} + \sum_{n,k} \phi_{nk} E[\log \theta_k] + E[\log p(\theta|\alpha)] - E[\log q(\theta)]\\
&+ \frac{2w_i - 1}{2} \mathbf{a}_i^\top E[\bar{\mathbf{z}}] - \log(2\cosh(\frac{\xi_i}{2})) - \lambda(\xi_i)(\mathbf{a}_i^\top E[\bar{\mathbf{z}}\bar{\mathbf{z}}^\top]\mathbf{a}_i - \xi_i^2) - \sum_{n,k} \phi_{nk} \log \phi_{nk}. \qquad (5.7)
\end{aligned}$$

We employ a coordinate ascent algorithm and update one set of parameters at a time, while holding the rest of the parameters fixed. The process keeps on repeating until changes in the likelihood lower-bound falls below 1e-5. By differentiating $\mathcal{F}$ w.r.t to the variational parameters and set the derivatives to 0, we obtain the following closed-form updates:

$$\begin{aligned}
\log \phi_n &= \log \beta^r_{\cdot t} + E[\log \theta] + \sum_{i=1}^{T_w} \left[\frac{2w_i - 1}{2N} \mathbf{a}_i - \frac{\lambda(\xi_i)}{N^2} \left(\operatorname{diag}(\mathbf{a}_i \mathbf{a}_i^\top) + 2\mathbf{a}_i \mathbf{a}_i^\top \sum_{m \neq n} \phi_m \right) \right], && (5.8)\\
\tilde{\alpha}_k &= \sum_n \phi_{nk} + \alpha_k, && (5.9)\\
\xi_i^2 &= \mathbf{a}_i^\top E[\bar{\mathbf{z}}\bar{\mathbf{z}}^\top]\mathbf{a}_i. && (5.10)
\end{aligned}$$

When using the concatenated topic indicator variables $\mathbf{Z} = [1(z_1), \ldots, 1(z_N)]$ as an input variable into logistic regression, the variational inference algorithm can be obtained in a similar manner as in the case of $\bar{\mathbf{z}}$. The expectations $E[\mathbf{Z}]$ and $E[\mathbf{Z}\mathbf{Z}^\top]$ have the following expressions:

$$E[\mathbf{Z}] = \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_N \end{pmatrix}, \qquad E[\mathbf{Z}\mathbf{Z}^\top] = \begin{pmatrix} \operatorname{diag}(\phi_1) & \phi_1\phi_2^\top & \ldots & \phi_1\phi_N^\top \\ \phi_2\phi_1^\top & \operatorname{diag}(\phi_2) & \ldots & \phi_2\phi_N^\top \\ \vdots & \vdots & \ldots & \vdots \\ \phi_N\phi_1^\top & \phi_N\phi_2^\top & \ldots & \operatorname{diag}(\phi_N) \end{pmatrix}.$$

The terms $\mathbf{a}_i^\top E[\mathbf{Z}]$ and $\mathbf{a}_i^\top E[\mathbf{Z}\mathbf{Z}^\top]\mathbf{a}_i$ in the likelihood lower-bound in (5.7) can therefore be written out as:

$$\mathbf{a}_i^\top E[\mathbf{Z}] = \sum_{n=1}^{N} \mathbf{a}_i^{n\top} \phi_n, \tag{5.11}$$

$$\mathbf{a}_i^\top E[\mathbf{Z}\mathbf{Z}^\top]\mathbf{a}_i = \sum_{i=1}^{N} \mathbf{a}_i^{n\top} \mathrm{diag}(\phi_n)\mathbf{a}_i^n + \sum_{n=1}^{N} \mathbf{a}_i^{n\top}\phi_n \sum_{m\neq n} \phi_m^\top \mathbf{a}_i^m, \tag{5.12}$$

where $\mathbf{a}_i = [\mathbf{a}_i^1, \dots, \mathbf{a}_i^N]$ and $\mathbf{a}_i^n \in \mathcal{R}^K$ are the regression coefficients of patch n for predicting the outcome of word i. By plugging in the above expressions for $\mathbf{a}_i^\top E[\mathbf{Z}]$ and $\mathbf{a}_i^\top E[\mathbf{Z}\mathbf{Z}^\top]\mathbf{a}_i$ into (5.3), we can write the likelihood lower bound $\mathcal{F}$ as a function of the variational parameters. Differentiating $\mathcal{F}$ w.r.t. the variational parameters and setting the derivatives to 0, we obtain the following closed-form updates:

$$\log\phi_n = \log\beta^r_{\cdot t} + E[\log\theta] + \sum_{i=1}^{T_w}\left(\frac{2w_i-1}{2}\mathbf{a}_i^n - \lambda(\xi_i)[\mathrm{diag}(\mathbf{a}_i^n\mathbf{a}_i^{n\top}) + 2\mathbf{a}_i^n \sum_{m\neq n}\mathbf{a}_i^{m\top}\phi_m]\right) \tag{5.13}$$

$$\tilde{\alpha}_k = \sum_n \phi_{nk} + \alpha_k, \tag{5.14}$$

$$\xi_i^2 = \mathbf{a}_i^\top E[\mathbf{Z}\mathbf{Z}^\top]\mathbf{a}_i. \tag{5.15}$$

5.3.2 Parameter Estimation

To update the model parameters $\Psi = \{\beta^r, \mathbf{A}, \alpha\}$, we maximize the lower bound of the log-likelihood in (5.3) w.r.t. Ψ and obtain the following closed-form updates for the multinomial parameters β^r and regression parameters $\mathbf{a}_i$. Note that the Dirichlet parameter α is not learned from the data and we fix its value to 0.01.

$$\beta^r_{kt} = \frac{\sum_{d,n}\phi^d_{nk} 1(r^d_n = t)}{\sum_{t,d,n}\phi^d_{nk} 1(w^d_n = t)} \tag{5.16}$$

$$\mathbf{a}_i = \left(2\sum_d \lambda(\xi_i^d) E[\bar{\mathbf{z}}_d \bar{\mathbf{z}}_d^\top]\right)^{-1} \left(\sum_d \frac{2w_i^d - 1}{2} E[\bar{\mathbf{z}}_d]\right) \tag{5.17}$$

5.3.3 Caption Prediction

In annotation, given a test image without caption $\mathbf{R}$, the task is to infer the most likely caption words. For this task, we run variational inference on $\mathbf{R}$ until convergence and use the inferred posterior parameters ϕ_n to approximate the conditional probability $p(\mathbf{w}|\mathbf{R})$ as follows:

$$\begin{aligned} p(\mathbf{w}|\mathbf{R}) &= \int p(\mathbf{w}|\mathbf{Z},\mathbf{A})p(\mathbf{Z}|\mathbf{R})d\mathbf{Z} \approx \int p(\mathbf{w}|\mathbf{Z},\mathbf{A})q(\mathbf{Z}|\mathbf{R})d\mathbf{Z} \\ &\approx \prod_{i=1}^{T_w} \sigma(\mathbf{a}_i^\top E[\bar{\mathbf{z}}])^{w_i}\sigma(-\mathbf{a}_i^\top E[\bar{\mathbf{z}}])^{1-w_i} \end{aligned} \tag{5.18}$$

where $E[\bar{\mathbf{z}}] = \frac{1}{N}\sum_n \phi_n$, with ϕ_n inferred from each test image by withholding the caption.

5.4 Experimental Results

We demonstrate the performance our annotation model on the 5,000 image subset of the COREL dataset, which is described in some detail in Chapter 3. The COREL subset used in our experiment contains 50 classes of images, with 100 images per class. Each image in the collection is reduced to size 117×181 (or 181×117). 4,500 images are used in training (90 images from each class), and 500 for testing (10 images per class). Each image is treated as a collection of 20×20 patches obtained by sliding a window with 20-pixel interval, resulting in 45 patches per image. To represent an image as a bag of words, the concept of visual words needs to first be identified and extracted using clustering of features. While there exists a great variety of image features in the related literature to choose from, we adopt the SIFT feature descriptors, which have been shown to be discriminative in numerous classification and recognition tasks. In addition, we incorporate the 36-dimensional robust color descriptors proposed in [vdWS06] which have been designed to complement the SIFT-descriptors extracted from the gray-scale patches. To learn a dictionary of visual words, we run a k-means algorithm on a collection of 164-dimensional features and obtain a set of T_r visual words. In some experiments in this section, we use the SIFT and color descriptors independently to represent an image, in order to understand the effect and influence that each feature type has on the overall annotation performance. The results indeed shed some lights on how best to combine the different feature types, which will be the main topic of discussion in the next chapter.

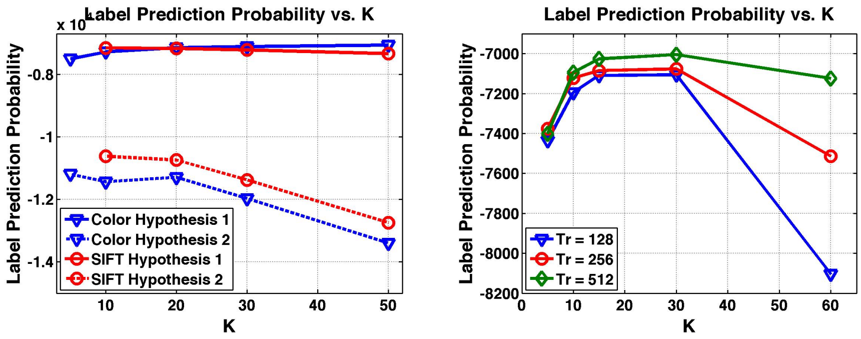
5.4.1 Label Prediction Probability

To compare the quality of annotation predicted by various parameter settings of sLDA-bin, we compute the label prediction probability defined as:

$$\text{score} = \sum_{d} \log p(\mathbf{w}_d | \mathbf{R}_d), \tag{5.19}$$

where the quantity $p(\mathbf{w}|\mathbf{R}) \approx \prod_{i=1}^{T_w} \sigma(\mathbf{a}_i^\top E[\bar{\mathbf{z}}])^{w_i} \sigma(-\mathbf{a}_i^\top E[\bar{\mathbf{z}}])^{1-w_i}$ is computed with respect to the ground-truth captions $\mathbf{w}$, and the expectation$E[\bar{\mathbf{z}}]$ is computed using the variational posterior inferred from the image portion of each test document. The summation in (5.19) is performed over all the words in the vocabulary and all 500 test images. This performance measure can indeed be viewed as computing an inverse-distance (affinity) measure between the held-out ground-truth data and the soft prediction inferred by the model.

In the first experiment, we demonstrate that a better prediction performance can be obtained when learning the latent low-dimensional topic representation of documents together with the regression parameters that map from the topics to the response variable. We denote Hypothesis 1 as the hypothesis where the topic parameters β are learned together with the logistic regression coefficients $\mathbf{A}$ using sLDA-bin. Hypothesis 2 denotes the experiment where we fit an



(a)

(b)

Figure 5.2: (a) Label prediction probability as a function of K, comparing Hypothesis 1 where the model parameters β and $\mathbf{A}$ are learned together with Hypothesis 2 where β and $\mathbf{A}$ are learned separately. (b) Label prediction probability as a function of K, using varying different number of terms in the vocabulary $T_r = \{128, 256, 512\}$. In general, a better label prediction performance is obtained with more words.

LDA model to a training set of images, and use the inferred latent topic representation $E[\bar{\mathbf{z}}]$ as an input into a separate logistic regression model, hence, under Hypothesis 2, β and $\mathbf{A}$ are learned separately. Figure 5.2 compares the label prediction probability under the two hypotheses when representing an image using SIFT features alone (red curves) or color features alone (blue curves). The findings in Figure 5.2 indeed confirms our conjecture that the latent topics uncovered by LDA are the ones that best explain correlations between words in the training documents; they are not necessarily discriminative in a prediction or classification task. By learning the topic parameters together with the associated response variables (caption words), sLDA-bin allows the selection of latent topics to be influenced by the values in the response variables, which lead to the topic parameters that perform better in a prediction task. Figure 5.2 (b) shows a plot of label prediction probability as a function of the number of topics K, while we vary the size of visual word vocabulary T_r. As we increase the number of visual words in the dictionary, we also obtain better predictive probabilities, with best performance obtained when T_r is 512.

Another important performance evaluation metric for auto-annotation systems is a precision-recall metric, which can be seen as another form of distance measure computed between the ground-truth binary annotation and the predicted annotation. The important quantity to compute for this metric is still $p(w|\mathbf{R})$, instead of the soft prediction in the case of label prediction probability, the fitted sLDA-bin model is used to predict actual annotation words for each test image. We follow the experiments in [FML04] and pick the top 5 most likely words as our prediction. For a given annotation word of interest, we define A to be the number of images with that word in the predicted annotation, B is defined as the number of images correctly annotated with that word, and C be the number of images that contain the word of interest in the ground-truth annotation. Precision is then given as $\frac{B}{A}$ and recall $\frac{B}{C}$. We pick the top 49 words with best performance and compute an average of precision and recall values in that set to obtain the mean-per-word precision and recall values presented in Table 5.1- 5.2. The column with header # recall ≥ 0 in the 2 tables reports the number of words with non-zero recall values, i.e. $B \neq 0$. If less than 49 words have been used to annotate the 'correct' images (less than 49 words with non-zero recalls), we leave the corresponding entries in the table blank, as in the first 2 rows of Table 5.1.

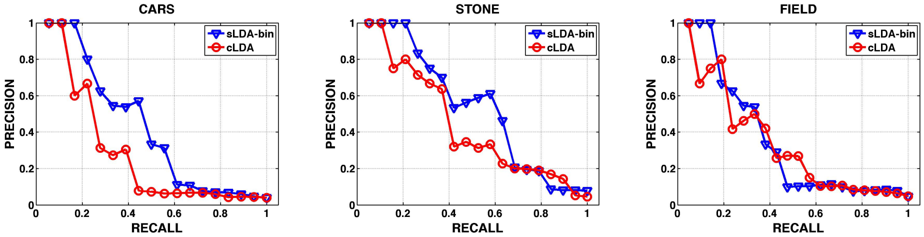
Table 5.1: Mean-per-word precision and mean-per-word recall for 49 best annotation words, using (a) Color features with $T_r = 256$, (b) SIFT features with $T_r = 512$.

K	Color			SIFT		
	Recall	Precision	# recall ≥ 0	Recall	Precision	# recall ≥ 0
5	-	-	18	-	-	21
10	-	-	34	-	-	36
20	-	-	46	0.3472	0.2728	49
30	0.4313	0.3453	53	0.3394	0.2465	50
50	0.5007	0.4184	58	0.3648	0.3185	63

Table 5.1 shows annotation results for different numbers of topics K in the sLDA-bin model, for the 2 cases where we represent each image using (a) SIFT descriptors alone or (b) color descriptors alone. For SIFT descriptors, we use $T_r = 512$, while for color descriptor $T_r = 256$. In general, as we increase the number of hidden topics K, we obtain higher precision-recall values. The mean-per-word precision and recalls obtained using color features are significantly higher than the results using SIFT features, for the same number of hidden topics K used. This result indeed should not come as a surprise since many classes in the COREL dataset do employ unique and distinct color schemes, which indicates that color features should be more discriminant than texture features, e.g. SIFT, in predicting annotation. We compare the mean-per-word precision and recall values in Table 5.1 when the 2 sets of features are used separately with the results shown in Table 5.2 where each image patch is represented as a concatenated vector of SIFT and color features. For this experiment, we learn $T_r = 512$ visual words. For the same number of hidden topics $K = 30$ and $K = 50$, we find that the results obtained using 2 sets of features are indeed significantly worse than the results obtained using color features alone. In fact, using 2 sets of features we obtain similar results to using SIFT features alone. We attribute this poor showing to the dominance of SIFT features in learning clusters of features representing visual words and the assignment of each image patch to a word. With 128-dimensional descriptors dominating the distance computation, a green patch and a red patch with similar SIFT descriptors might be indeed assigned to the same word. This result raises an interesting question of how to best combine features of different types in order to benefit from the independence and correlation structures of different feature types and improve the performance of a prediction task. Indeed, we will explore this topic in more detail in the next chapter.

Table 5.2: Mean-per-word precision and mean-per-word recall for 49 best annotation words, using concatenated SIFT and Color features with $T_r = 512$.

K	Recall	Precision	# word with recall ≥ 0
30	0.3741	0.3384	51
50	0.4123	0.3297	60
60	0.4251	0.3960	63
80	0.4188	0.3864	68
100	0.4203	0.3861	71



(a)

(b)

(c)

Figure 5.3: Precision-recall curve for 3 single word queries: ‘wave’, ‘pole’, ‘flower’, comparing cLDA (red) and tr-mmLDA (blue).

5.4.2 Annotation Examples

Examples of predicted captions generated by sLDA-bin and cLDA are shown comparatively in Table 5.3. In all of the examples shown here, captions predicted by sLDA-bin contain specific words that are semantically related to the visual content and the true captions, while cLDA has preference for more general words (e.g. sky, water, tree) in its choice of captions. We attribute the superior prediction performance over cLDA to the use of multi-variate Bernoulli variable which is a more suitable choice for binary annotation data of the COREL dataset, compared to the use of Multinomial in cLDA. By learning the topic parameters together with the regression coefficients that map from the latent topics to to the caption word, our model infers a low-dimensional latent variable representation of an image predictive of the caption words, which indeed leads to a more accurate prediction performance.

5.4.3 Text-based Image Retrieval

Similar to the case of tr-mmLDA in Chapter 4, we use sLDA-bin to perform image retrieval using single word queries. Table 5.4 shows examples of images ranked order and returned using the queries ‘sculptures’, ‘beach’, ‘plane’, ‘snow’ and ‘cars’. To evaluate the performance of a retrieval system, we adopt the mean average precision measure, described in detail in section 4.4.4. Using sLDA-bin with $K = 30$ we obtain a mean average precision of 0.19. For cLDA, the best value is 0.13. Figure 5.3 shows precision-recall curves for 3 single-word queries comparing sLDA-bin and cLDA. Our model in general can obtain higher precisions at the same recall values and attains a better ranked order retrieval performance than cLDA.

Table 5.3: Examples of predicted annotation inferred using sLDA-bin and correspondence LDA.

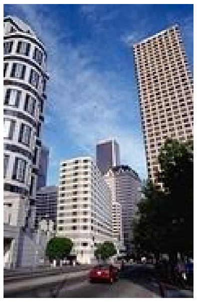
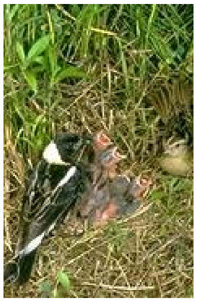
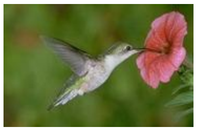
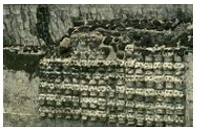
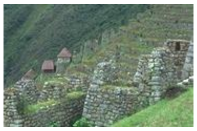
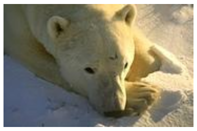
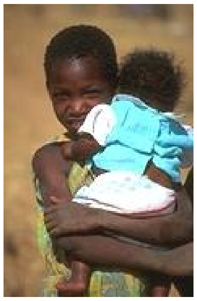
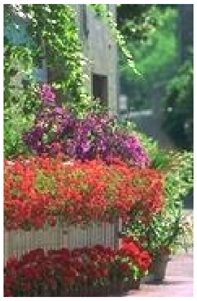
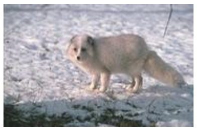
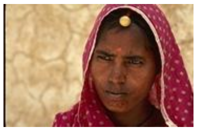
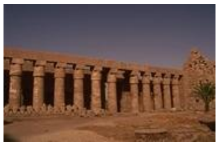
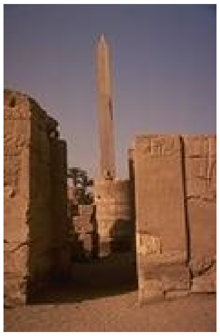
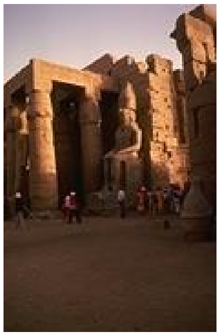
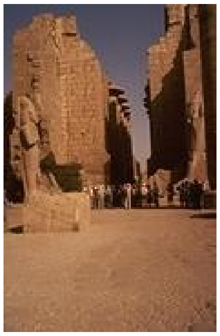
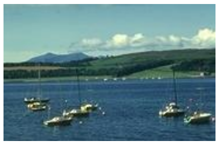
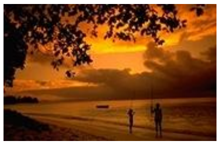
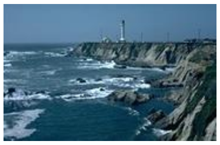
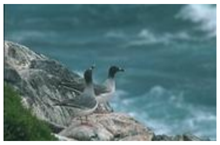
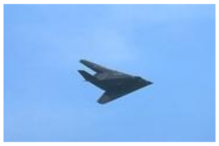
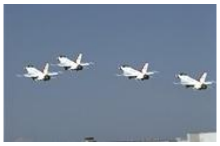
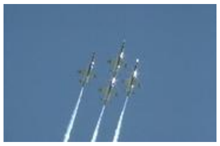
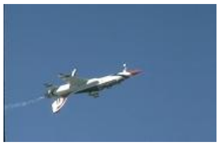
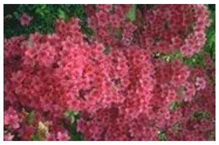
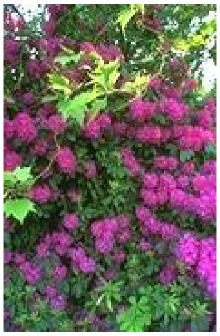
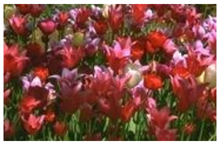
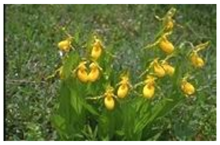
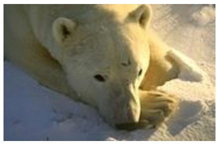
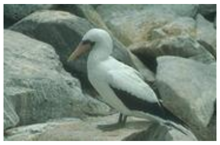
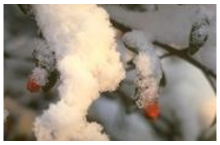
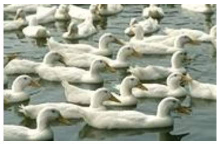
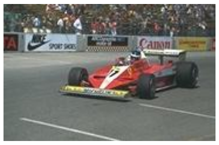
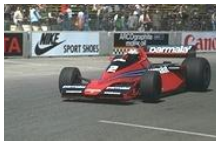
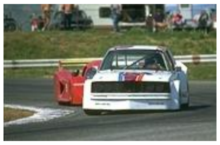
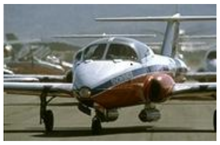
	**Groundtruth**: buildings, street, car, skyline **sLDA-bin**: sky, street, buildings, car **cLDA**: sky, tree, water, buildings		**Groundtruth**: grass, birds, nest **sLDA-bin**: grass, nest, birds **cLDA**: tree, grass, water
	**Groundtruth**: flowers, birds, fly **sLDA-bin**: leaf, birds, close-up, flowers **cLDA**: tree, flowers, grass, plants		**Groundtruth**: ruins, wall, sculpture **sLDA-bin**: ruins, field, stone **cLDA**: tree, grass, water
	**Groundtruth**: mountain, rocks, valley, terrace **sLDA-bin**: valley, mountain, rocks, field **cLDA**: water, tree, grass, train		**Groundtruth**: snow, bear, polar, head **sLDA-bin**: snow, bear, polar, face **cLDA**: sky, water, grass, tree
	**Groundtruth**: people, close-up, baby, shirt **sLDA-bin**: head, close-up, snow, sky **cLDA**: people, water, tree, sky		**Groundtruth**: flowers, street, plants, bloom **sLDA-bin**: flowers, garden, tree, grass **cLDA**: flowers, tree, people, ocean
	**Groundtruth**: grass, snow, fox, arctic **sLDA-bin**: snow, rocks, fox, arctic **cLDA**: snow, water, sky, tree		**Groundtruth**: people, head, woman, indian **sLDA-bin**: people, indian, close-up, man **cLDA**: people, buildings, flowers, sky

Table 5.4: Examples of images (with no captions) retrieved using single word queries.

sculpture				
beach				
plane				
flowers				
snow				
cars				

Chapter 5, in part, is a reprint of the material as it appears in: D. Putthividhya, H. T. Attias, S. Nagarajan. “Supervised Topic Model for Automatic Image Annotation”, in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2010. I was the primary researcher of the cited materials and the co-author listed in these publications supervised the work which forms the basis of this chapter.

Chapter 6

Statistical Topic Models for Multimedia Annotation

In Chapter 4 and Chapter 5, we explore annotation models for multimedia documents comprising 2 data streams, i.e. image-text (image and captions) and video-text (video and closed captions), with the goal of predicting the text modality from the image or video modality. While most recent research efforts in multimedia annotation have been tailored to documents with 2 modalities, modern multimedia content is often characterized by having multiple varied forms, i.e. movies consisting of video-audio streams with closed caption texts; web pages containing pictures, text, songs, and video clip animation. For such documents, since it is often difficult, if not possible, to isolate the textual descriptors that are used to describe each individual stream, in this chapter we investigate how one can use all the data modalities to make predictions about the associated captions. In particular, we are interested in exploring how statistical topic models discussed in Chapter 4 and 5 can be modified to deal with multiple sources of input variables.

Generally speaking, features extracted from different data modalities (multimodal features) describe different characteristics of the documents and when used in combination should increase the prediction accuracy of an auto-annotation system. Since most previous work on multimedia annotation has formulated the problem of auto-annotation as a semantic concept detection (classification) task [AAD$^+$02, ABC$^+$03, NNLS04, NS04], the main question for such systems is how one combines features from different modalities to optimally make predictions about the semantic class of each multimedia clip. While separate classifiers can be built for each modality and the results of each classifier later combined, for the case of generative classifiers, the work in [YLXH07, PANL07] show that representations that capture correlations between data of different modalities yield better classification results than using the single modalities in isolation. For the case of SVM [NNLS04, NS04], one fuses data from different modalities by simply concatenating multimodal features. Care must be taken, however, as problems can arise when the concatenated features become very high-dimensional. The problem of how to optimally combine features of different types is indeed not limited to the case of multimodal features but also occurs when employing a variety of features from a single modality (unimodal features) in concert to infer the missing annotation.

In this chapter, we investigate how different statistical topic models for modeling association between image-text or video-text discussed in Chapter 4—cLDA, mmLDA, tr-mmLDA—can be extended to include additional data modalities in making predictions. We find that the 2-modality tr-mmLDA model can be straight-forwardly modified to allow features of different types to influence the prediction of annotation texts, by introducing the latent topics of the new data modality as another set of covariates to the regression module. We focus on modeling multimedia documents containing audio-video streams and their associated captions obtained from speech transcribed data. We use this type of multimedia documents in our experiments because the ground-truth captions are readily available without resorting to manual labeling, and the same type of multimedia documents has been employed in previous studies on auto-annotation, e.g. [VH06]. Given an un-annotated multimedia clip, the goal is now to infer the missing caption texts using both the audio and video modalities. In a typical scenario where caption words serve to describe objects that manifest themselves in both the audio and video content, using all the available observations to predict closed-captions data is expected to yield improved performance over the use of audio or video modality alone.

An extension to incorporate additional input modalities can be similarly derived for the sLDA-bin model from Chapter 5. For this model, we show experimental results using 2 types of image features (unimodal features) to predict annotation on the COREL dataset. Superior annotation quality, as measured by label prediction probability and precision-recall metrics, is obtained when multiple feature types are used.

6.1 Notations

As in Chapter 4 and 5, we adopt a bag-of-word representation for all the data modalities of our multimedia document. In such a representation, a video clip is represented as a bag of video blocks, and an audio clip is represented in the same fashion as a collection of audio frames (note that frame length and block length need not be the same and we will elaborate more on this point in the experiment section). We adopt the same terminology as in Chapter 4 and 5 and where applicable introduce a superscript a to denote variables associated with the audio modality and superscript v for variables for the video modality. A caption word is denoted as a unit-basis vector w of size T_w with exactly one non-zero entry representing the membership to only one word in the dictionary of T_w words. An audio word is similarly denoted as a unit-basis vector r^a of size T_r^a, and a video word is denoted as a unit-basis vector r^v of size T_r^v respectively.

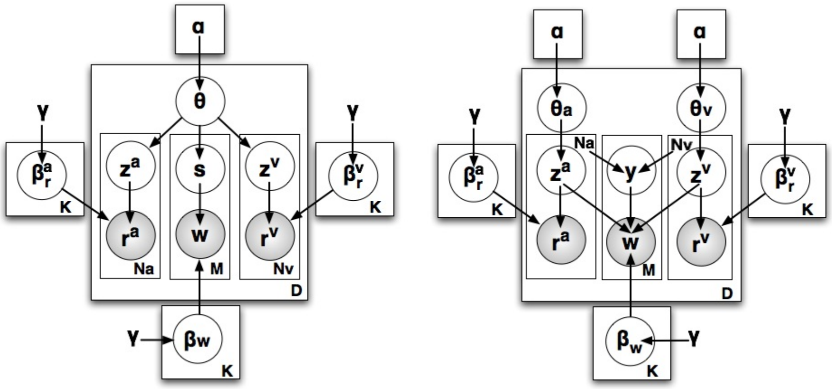
A multi-modal document containing 3 data modalities is represented as a triplet of vectors of word counts denoted as $\{\mathbf{w}, \mathbf{r^a}, \mathbf{r^v}\}$, where $\mathbf{w} = \{w_1, \ldots, w_M\}$ denotes a collection of M text words in the caption modality; $\mathbf{r}^a = \{r_1^a, ..., r_{N_a}^a\}$ denotes a collection of N_a audio frames in the audio modality; $\mathbf{r}^v = \{r_1^v, ..., r_{N_v}^v\}$ denotes N_v video blocks in the video modality.

6.2 3-modality Extensions

To extend mmLDA and cLDA to model the association between audio-video streams and their closed-captions, we postulate the existence of a set of shared latent variables that are the common causes of correlations in the 3 data types. In the case of mmLDA, this can be easily done by sharing the mean topic proportion variable θ as shown in Figure 6.1(a). As in the case of the 2-modality model for image-caption data, such an association will suffer a similar drawback in that the set of topics that are associated with caption words might indeed be non-overlapping with the set of topics used in the audio and video modalities, due to the underlying exchangeability assumption of words of different types. In such a scenario, the knowledge about the audio-video content will be not be particularly useful in predicting the missing captions.

For cLDA, however, an extension from the model for image-caption data to accommodate another input source is not quite as straight-forward and clear-cut. In the 2-modality model, in order to ensure that the set of topics used to generate caption words are a subset of those used in the corresponding image, each caption word is generated conditioned on a latent topic of a randomly chosen image region. To this end, a uniform random variable $y \sim \text{Unif}(1, N)$ is employed for each caption word to select a particular image region (one of N regions) whose topic is used to generate the caption word. To extend the association model of cLDA to include another input modality, potentially we could associate each caption word with a video or an audio word that has been selected at random. More specifically, as seen in the graphical model in Figure 6.1(b), a uniform random variable $y \sim \text{Unif}(1, N_a + N_v)$ is introduced to model a random selection of one of the audio or video words whose topic is then used to generate the caption word. Indeed, such a selection mechanism favors the input modality with more words. For a multimedia clip with 1,000 video blocks and 100 audio frames, when selecting y according to $y \sim \text{Unif}(1, N_a + N_v)$, the caption words will be 10 times more likely to serve as descriptions of the video content over its audio counterpart. A potential alternative to such a scheme is to first select one of the input modalities at random, and choose among the words of the selected modality.

The extensions of cLDA described above indeed have serious limitations, as the mechanism introduced to ensure that caption topics are a subset of the topics used in the input modalities, also restricts each caption word to be associated with either one of the input modalities, but not both. Such an association model might therefore be a good fit for modeling caption words describing the background music, or words describing objects in the scene that do not emit sound, but the model lacks a mechanism to account for the majority of caption words that are



(a) **mmLDA**

(b) **cLDA**

Figure 6.1: Potential extensions of **(a)** Multi-modal LDA (mmLDA) and **(b)** correspondence LDA (cLDA) to the task of multimedia annotation using both audio and video data to predict caption texts.

used to describe objects that manifest themselves visually and audibly in the scene.

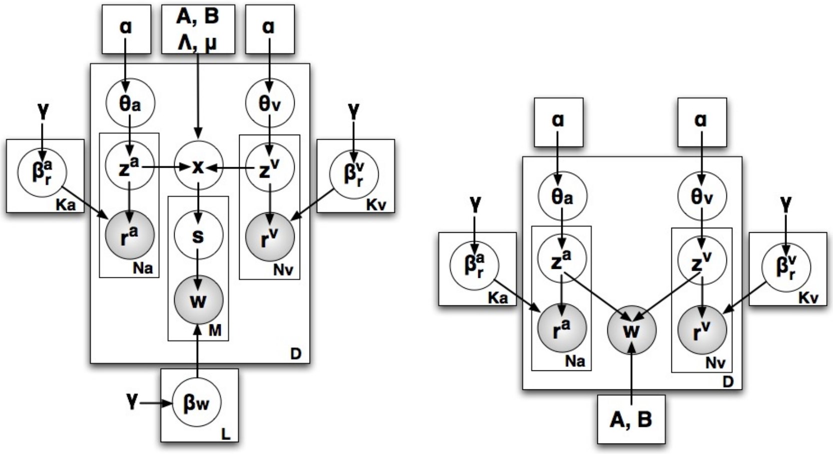
6.2.1 3-modality Topic-regression Multi-modal LDA (3-mod tr-mmLDA)

With the latent variable regression approach of tr-mmLDA, we can easily extend our 2-modality annotation model to learn statistical associations between audio-video streams and their corresponding captions. By treating audio and video topic proportion variables as covariates into the regression modules, and allowing the 2 sets of regression coefficients to be learned from the data, we can capture the precise mapping and quantify the strength of the relationship between latent topics of the target modality (closed-captions) and those of each individual input modality (audio and video streams). More specifically, we model the conditional distribution of the real-valued topic proportion variable for caption modality $\mathbf{x}$, given the empirical topic proportion of audio $\bar{\mathbf{z}}^a$ and empirical topic frequency of video $\bar{\mathbf{z}}^v$ as

$$p(\mathbf{x}|\bar{\mathbf{z}}^a, \bar{\mathbf{z}}^v, \mathbf{A}, \mathbf{B}, \Lambda, \mu) \sim \mathcal{N}(\mathbf{x}; \mathbf{A}\bar{\mathbf{z}}^a + \mathbf{B}\bar{\mathbf{z}}^v + \mu, \mathbf{\Lambda}^{-1}), \tag{6.1}$$

where $\mathbf{A}$ and $\mathbf{B}$ are the respective regression coefficient matrices for the audio and video modality, and Λ and μ are the same noise precision and mean parameter as defined in the 2-modality case. The graphical model representation of the 3-modality tr-mmLDA is illustrated in Figure 6.2(a). Given K_a audio topics, K_v video topics, L caption topics, and the Multinomial and regression parameters of the 3-modality tr-mmLDA model $\{\beta_r^a, \beta_r^v, \beta_w, \alpha, \mathbf{A}, \mathbf{B}, \mathbf{\Lambda}, \mu\}$, we can generate a multimedia document $\{\mathbf{w}, \mathbf{r}^a, \mathbf{r}^v\}$ with N_a audio words, N_v video words, and M closed caption words by taking the following steps:

- Draw an audio topic proportion $\theta_a|\alpha \sim \text{Dir}(\alpha)$



(a) **tr-mmLDA** (b) **sLDA-bin**

Figure 6.2: Extensions of **(a)** Topic-regression Multi-modal LDA (tr-mmLDA) and **(b)** Supervised LDA-binary (sLDA-bin) to the task of multimedia annotation using both audio and video data to predict caption texts.

- For each audio word r_n^a, $n \in \{1, 2, \dots, N_a\}$
  1. Draw topic assignment $z_n^a = k|\theta_a \sim \text{Mult}(\theta_k^a)$
  2. Draw audio word $r_n^a = t|z_n^a = k \sim \text{Mult}(\beta_r^a)$
- Draw a video topic proportion $\theta_v|\alpha \sim \text{Dir}(\alpha)$
- For each video word r_n^v, $n \in \{1, 2, \dots, N_v\}$
  1. Draw topic assignment $z_n^v = k|\theta_v \sim \text{Mult}(\theta_k^v)$
  2. Draw video word $r_n^v = t|z_n^v = k \sim \text{Mult}(\beta_r^v)$
- Given the empirical topic proportions for audio and video: $\bar{\mathbf{z}}^a = \frac{1}{N}\sum_{n=1}^{N_a} z_n^a$, $\bar{\mathbf{z}}^v = \frac{1}{N_v}\sum_{n=1}^{N_v} z_n^v$, we sample a real-valued topic proportion variable for caption text according to the following conditional distribution:

$$\mathbf{x}|\bar{\mathbf{z}}^a, \bar{\mathbf{z}}^v, \mathbf{A}, \mathbf{B}, \mu, \mathbf{\Lambda} \sim \mathcal{N}(\mathbf{x}; \mathbf{A}\bar{\mathbf{z}}^a + \mathbf{B}\bar{\mathbf{z}}^v + \mu, \mathbf{\Lambda}).$$

- Compute topic proportion for the caption modality $\eta_l = \frac{\exp(x_l)}{\sum_{k=1}^{L}\exp(x_k)}$.
- For each caption word w_m, $m \in \{1, 2, \dots, M\}$
  1. Draw topic assignment $s_m = l|\eta \sim \text{Mult}(\eta_l)$
  2. Draw caption word $w_m = t|s_m = l \sim \text{Mult}(\beta_{lt}^w)$

The 3-modality tr-mmLDA model enjoys many of the same benefits as its 2-modality counterpart over the 3-modality extensions of mmLDA and cLDA. By using 2 linear regression modules to capture the relationships between modalities and allowing the associated regression coefficients to be learned from the data, tr-mmLDA can model caption words that serve as descriptions of either the audio or video content, by setting the corresponding regression coefficients of the other modality to 0, or words that are influenced by both input modalities. The latter scenario indeed cannot be captured in the 3 modality extension of cLDA as explained earlier. The flexibility of tr-mmLDA allows for a more precise learning of the true relationship between the audio-visual content of the scene and the associated caption texts and hence we expect an improved performance in a prediction task.

Another worthwhile property to note is that unlike mmLDA and cLDA, tr-mmLDA allows data in the 3 modalities to be modeled with different numbers of hidden topics: K_a, K_v, and L. Indeed, this flexibility can be beneficial as the vocabulary sizes for video, audio, and text modality are often quite different. To give a concrete example, in the specific case of modeling the TV show data in our experiments, the closed-caption vocabulary reaches over 4000 words, while we only have 256 video and 100 audio words. In such a scenario, models that allow for a larger numbers of hidden topics for the closed-caption data than the other 2 modalities will be better at modeling the joint distribution of multi-modal data.

6.2.2 3-modality Supervised LDA-binary (3-mod sLDA-bin)

In a similar fashion, we can extend the sLDA-bin model proposed for image-caption data in Chapter 5 to accommodate multiple input sources. More specifically, we include the influence from the audio modality in predicting caption words by introducing the latent topic representation of audio as another set of input variables into the logistic regression module. Given the empirical topic frequency for audio and video $\bar{\mathbf{z}}^a$, $\bar{\mathbf{z}}^v$, the probability of labeling the current image with caption word i is given by:

$$p(w_i|\bar{\mathbf{z}}^a, \bar{\mathbf{z}}^v, \mathbf{A}, \mathbf{B}) = \sigma(\mathbf{a}_i^\top \bar{\mathbf{z}}^a + \mathbf{b}_i^\top \bar{\mathbf{z}}^v)^{w_i} \sigma(-\mathbf{a}_i^\top \bar{\mathbf{z}}^a - \mathbf{b}_i^\top \bar{\mathbf{z}}^v)^{1-w_i},$$

where $w_i \in \{0,1\}$, $\sigma(x) = \frac{1}{1+\exp(-x)}$ denotes the logistic sigmoid function, and $\mathbf{a}_i$ and $\mathbf{b}_i$ are the corresponding regression coefficients which specify the relative influences that audio and video topics have over the presence/absence of word i in the current image. The graphical model representation of the 3-modality sLDA-bin model is shown in Figure 6.2(b). Given K_a audio topics, K_v video topics, and their corresponding Multinomial and regression parameters $\{\beta_r^a, \beta_r^v, \alpha, \mathbf{A}, \mathbf{B}\}$, to generate a multimedia document $\{\mathbf{w}, \mathbf{r}^a, \mathbf{r}^v\}$ with N_a audio frames, N_v video frames, and binary caption word vector $\mathbf{w}$, we follow the generative process of sLDA-bin as given by:

- Draw an audio topic proportion $\theta_a|\alpha \sim \text{Dir}(\alpha)$
- For each audio word r_n^a, $n \in \{1, 2, \ldots, N_a\}$
    1. Draw topic assignment $z_n^a = k|\theta_a \sim \text{Mult}(\theta_k^a)$
    2. Draw audio word $r_n^a = t|z_n^a = k \sim \text{Mult}(\beta_r^a)$
- Draw a video topic proportion $\theta_v|\alpha \sim \text{Dir}(\alpha)$
- For each video word r_n^v, $n \in \{1, 2, \ldots, N_v\}$
    1. Draw topic assignment $z_n^v = k|\theta_v \sim \text{Mult}(\theta_k^v)$
    2. Draw video word $r_n^v = t|z_n^v = k \sim \text{Mult}(\beta_r^v)$

- Given the empirical topic proportions for audio and video: $\bar{\mathbf{z}}^a = \frac{1}{N}\sum_{n=1}^{N_a} z_n^a$, $\bar{\mathbf{z}}^v = \frac{1}{N_v}\sum_{n=1}^{N_v} z_n^v$, the probability of labeling the current multimedia document with word i is given by:

$$p(w_i|\bar{\mathbf{z}}^a, \bar{\mathbf{z}}^v, \mathbf{A}, \mathbf{B}) = \sigma(\mathbf{a}_i^\top \bar{\mathbf{z}}^a + \mathbf{b}_i^\top \bar{\mathbf{z}}^v)^{w_i} \sigma(-\mathbf{a}_i^\top \bar{\mathbf{z}}^a - \mathbf{b}_i^\top \bar{\mathbf{z}}^v)^{1-w_i}$$

- For each caption word w_i, $i \in \{1, 2, \ldots, T_w\}$,
  1. Draw a Bernoulli sample for word $w_i \sim \text{Bernoulli}(p_i)$ where $p_i = \sigma(\mathbf{a}_i^\top \bar{\mathbf{z}}^a + \mathbf{b}_i^\top \bar{\mathbf{z}}^v)$.

6.2.3 Variational Inference and Parameter Estimation for 3-modality tr-mmLDA

Variational inference algorithm for the case of 3-modality tr-mmLDA can be derived in the same manner as its 2-modality counterpart. Given the model parameters $\{\beta^w, \beta_r^a, \beta_r^v, \mathbf{A}, \mathbf{B}, \mathbf{\Lambda}, \mu\}$, for each document we infer the posterior probability over the hidden variables given the observed words $p(\mathbf{Z}^a, \mathbf{Z}^v, \theta^a, \theta^v, \mathbf{x}, \mathbf{S}|\mathbf{r}^a, \mathbf{r}^v, \mathbf{w})$. To make the computation tractable, the key idea is to approximate the exact posterior with variational posterior in a factorized form. More specifically, we have $p(\mathbf{Z}^a, \mathbf{Z}^v, \theta^a, \theta^v, \mathbf{x}, \mathbf{S}|\mathbf{r}^a, \mathbf{r}^v, \mathbf{w}) \approx \prod_n q(z_n^a) \prod_n q(z_n^v) \prod_m q(s_m) q(\theta^a) q(\theta^v) q(\mathbf{x})$. Using the factorized posterior, we can write down an expression for the lower-bound of the likelihood function in terms of parameters of the variational posteriors. As in the inference algorithm for the 2 modality case, we employ a coordinate-ascent algorithm to update these parameters, where one set of parameters are updated at a time, while holding the rest of the parameters fixed. The main difference between the algorithm in the 3-modality case and its 2-modality counterpart is the following updates for ϕ_n^a, ϕ_n^v, and $\bar{\mathbf{x}}$ which are the parameters of the variational posteriors $q(z_n^a)$, $q(z_n^v)$, $q(\mathbf{x})$:

$$\begin{aligned}
\log \phi_{n\cdot}^a &= \sum_t 1(r_n^a = t) \log \beta_{r\cdot t}^a + E[\log \theta^a] + \frac{1}{N_a} \mathbf{A}^\top \Lambda(\bar{\mathbf{x}} - \mu - \mathbf{B}E[\bar{\mathbf{z}}^v]) \\
&\quad - \frac{1}{2N_a^2} \mathrm{diag}(\mathbf{A}^\top \mathbf{\Lambda} \mathbf{A}) - \frac{1}{N_a^2} \mathbf{A}^\top \mathbf{\Lambda} \mathbf{A} \sum_{m \neq n} \phi_m^a, && (6.2) \\
\log \phi_{n\cdot}^v &= \sum_t 1(r_n^b = t) \log \beta_{r\cdot t}^v + E[\log \theta^v] + \frac{1}{N_v} \mathbf{B}^\top \Lambda(\bar{\mathbf{x}} - \mu - \mathbf{A}E[\bar{\mathbf{z}}^a]) \\
&\quad - \frac{1}{2N_v^2} \mathrm{diag}(\mathbf{B}^\top \mathbf{\Lambda} \mathbf{B}) - \frac{1}{N_v^2} \mathbf{B}^\top \mathbf{\Lambda} \mathbf{B} \sum_{m \neq n} \phi_m^v, && (6.3) \\
\frac{\partial \mathcal{F}}{\partial \bar{x}_l} &= \sum_m \eta_{ml} - \frac{M}{\xi} e^{\bar{x}_l + \frac{0.5}{\gamma_l}} - \lambda_l(\bar{x}_l - \mu_l - \mathbf{a}_l^\top E[\bar{\mathbf{z}}^a] - \mathbf{b}_l^\top E[\bar{\mathbf{z}}^v]) = 0 && (6.4)
\end{aligned}$$

where $E[\bar{\mathbf{z}}^v] = \frac{1}{N_v}\sum_n \phi_n^v$ and $E[\bar{\mathbf{z}}^a] = \frac{1}{N_a}\sum_n \phi_n^a$. The remainder of the variational posterior parameters—$\eta_{ml}, \gamma, \tilde{\alpha}^a, \tilde{\alpha}^v, \xi$—can be updated using the update rules of the 2-modality case given in (4.9)-(4.15).

Parameter Estimation:

By differentiating the likelihood lower-bound w.r.t. to the model parameters $\Psi = \{\beta_r^a, \beta_r^v, \beta^w, \mathbf{A}, \mathbf{B}, \Lambda, \mu\}$, we obtain simple closed-form updates for all the model parameters, as in the 2-modality case. The multinomial parameter updates are exactly identical to the updates in (4.19)-(4.20) for the 2-modality case. The difference, however, lies in the regression parameter updates which are given as follow:

$$\mathbf{A} = \left(\sum_d (\bar{\mathbf{x}}_d - \mu - \mathbf{B}E[\bar{\mathbf{z}}_d^v])E[\bar{\mathbf{z}}_d^a]^\top\right)\left(\sum_d E[\bar{\mathbf{z}}_d^a\bar{\mathbf{z}}_d^{a\top}]\right)^{-1}, \tag{6.5}$$

$$\mathbf{B} = \left(\sum_d (\bar{\mathbf{x}}_d - \mu - \mathbf{A}E[\bar{\mathbf{z}}_d^a])E[\bar{\mathbf{z}}_d^v]^\top\right)\left(\sum_d E[\bar{\mathbf{z}}_d^v\bar{\mathbf{z}}_d^{v\top}]\right)^{-1}, \tag{6.6}$$

$$\mu = \frac{1}{D}\sum_d \left(\bar{\mathbf{x}}_d - \mathbf{A}E[\bar{\mathbf{z}}_d^a] - \mathbf{B}E[\bar{\mathbf{z}}_d^v]\right), \tag{6.7}$$

$$\mathbf{\Lambda}^{-1} = \frac{1}{D}\sum_d \left((\bar{\mathbf{x}}_d - \mu)(\bar{\mathbf{x}}_d - \mu)^\top + \mathbf{\Gamma}_d^{-1} - (\mathbf{A}E[\bar{\mathbf{z}}_d^a] + \mathbf{B}E[\bar{\mathbf{z}}_d^v])(\bar{\mathbf{x}}_d - \mu)^\top\right). \tag{6.8}$$

6.2.4 Variational Inference and Parameter Estimation for 3-modality sLDA-bin

Similar to the case of the 3-modality tr-mmLDA model, the variational inference algorithm for the 3-modality sLDA-bin model can be derived in the exact same fashion as its 2-modality counterpart. First, we employ the following convex variational bound to approximate $\log p(w_i|\bar{\mathbf{z}}^a, \bar{\mathbf{z}}^v)$ and in the process introduce additional variational parameters ξ_i:

$$\begin{aligned} w_i \log\sigma(\mathbf{a}_i^\top\bar{\mathbf{z}}^a + \mathbf{b}_i^\top\bar{\mathbf{z}}^v) + (1-w_i)\log\sigma(-\mathbf{a}_i^\top\bar{\mathbf{z}}^a - \mathbf{b}_i^\top\bar{\mathbf{z}}^v) &\geq (2w_i - 1)\frac{\mathbf{a}_i^\top\bar{\mathbf{z}}^a + \mathbf{b}_i^\top\bar{\mathbf{z}}^v}{2} \\ - \log(e^{\frac{\xi_i}{2}} + e^{-\frac{\xi}{2}}) - \lambda(\xi_i)(\mathbf{a}_i^\top\bar{\mathbf{z}}^a\bar{\mathbf{z}}^{a\top}\mathbf{a}_i &+ 2\mathbf{b}_i^\top\bar{\mathbf{z}}^v\bar{\mathbf{z}}^{a\top}\mathbf{a}_i + \mathbf{b}_i^\top\bar{\mathbf{z}}^v\bar{\mathbf{z}}^{v\top}\mathbf{b}_i - \xi_i^2), \end{aligned} \tag{6.9}$$

where $\lambda(\xi_i) = \frac{\tanh(\frac{\xi_i}{2})}{4\xi_i}$. The bound is tight at the point where the variational parameter ξ_i satisfies $\xi^2 = (\mathbf{a}_i^\top\bar{\mathbf{z}}^a + \mathbf{b}_i^\top\bar{\mathbf{z}}^v)^2$. By approximating the true posterior with a variational posterior in a factorized form: $p(\mathbf{Z}^a, \mathbf{Z}^v, \theta^a, \theta^v|\mathbf{r}^a, \mathbf{r}^v, \mathbf{w}) \approx \prod_n q(z_n^a)\prod_n q(z_n^v)q(\theta^a)q(\theta^v)$, we can derive an expression for the likelihood lower-bound in terms of variational posterior parameters. By differentiating the lower-bound w.r.t. all variational parameters and set the derivatives to 0, we obtain closed-form updates for all the variational parameters used in this model. The main difference between the 3-modality case and the inference algorithm of its 2-modality counterpart lies in the following updates for ϕ_{nk}^a and ϕ_{nk}^v which are shorthands for the posterior parameters $q(z_n^a = k)$ and $q(z_n^v = k)$.

$$\begin{aligned} \log\phi_n^a = \sum_t 1(r_n^a = t)\log\beta_{kt}^a + E[\log\theta^a] + \sum_{i=1}^{T_w}\Bigg[\frac{2w_i - 1}{2N_a}\mathbf{a}_i - \lambda(\xi_i)\Bigg(\frac{1}{N_a^2}\mathrm{diag}(\mathbf{a}_i\mathbf{a}_i^\top) \\ + \frac{2}{N_a^2}\mathbf{a}_i\mathbf{a}_i^\top\sum_{m\neq n}\phi_m^a + \frac{2}{N_a}\mathbf{a}_i\mathbf{b}_i^\top E[\bar{\mathbf{z}}^v]\Bigg)\Bigg], \end{aligned} \tag{6.10}$$

$$\log \phi_n^v = \sum_t 1(r_n^v = t) \log \beta_{lt}^v + E[\log \theta^v] + \sum_{i=1}^{T_w} \left[\frac{2w_i - 1}{2N_v} \mathbf{b}_i - \lambda(\xi_i) \left(\frac{1}{N_v^2} \text{diag}(\mathbf{b}_i \mathbf{b}_i^\top) \right.\right.$$
$$\left.\left. + \frac{2}{N_v^2} \mathbf{b}_i \mathbf{b}_i^\top \sum_{m \neq n} \phi_m^v + \frac{2}{N_v} \mathbf{b}_i \mathbf{a}_i^\top E[\bar{\mathbf{z}}^a] \right) \right]. \quad (6.11)$$

Since $p(\theta^a|\alpha)$ and $p(\theta^v|\alpha)$ are modeled as a Dirichlet which is a conjugate prior to the Multinomial distribution, the variational posteriors $q(\theta^a)$ and $q(\theta^v)$ also take the form of a Dirichlet with parameters $\tilde{\alpha}_k^a$ and $\tilde{\alpha}_k^v$ given by $\tilde{\alpha}_k^a = \sum_n \phi_{nk}^a + \alpha$ and $\tilde{\alpha}_k^v = \sum_n \phi_{nk}^v + \alpha$. The update for the variational parameter ξ_i is:

$$\xi_i^2 = \mathbf{a}_i^\top E[\bar{\mathbf{z}}^a \bar{\mathbf{z}}^{a\top}] \mathbf{a}_i + 2\mathbf{b}_i^\top E[\bar{\mathbf{z}}^v \bar{\mathbf{z}}^{a\top}] \mathbf{a}_i + \mathbf{b}_i^\top E[\bar{\mathbf{z}}^v \bar{\mathbf{z}}^{v\top}] \mathbf{b}_i, \quad (6.12)$$

where $E[\bar{\mathbf{z}}^a \bar{\mathbf{z}}^{a\top}] = \frac{1}{N_a^2}(\sum_n \text{diag}(\phi_n^a) + \sum_n \phi_n^a \sum_{m \neq n} \phi_m^{a\top})$, $E[\bar{\mathbf{z}}^v \bar{\mathbf{z}}^{a\top}] = E[\bar{\mathbf{z}}^v] E[\bar{\mathbf{z}}^a]^\top$, and $E[\bar{\mathbf{z}}^v \bar{\mathbf{z}}^{v\top}] = \frac{1}{N_v^2}(\sum_n \text{diag}(\phi_n^v) + \sum_n \phi_n^v \sum_{m \neq n} \phi_m^{v\top})$.

Parameter Estimation

We differentiate the likelihood lower-bound w.r.t. the model parameters and obtain closed-form updates for all the parameters. The Multinomial parameter updates for the 3-modality sLDA-bin are identical to the 2-modality case as given in (5.16). The regression coefficients associated with word i can be updated as in the following:

$$\mathbf{a}_i = \left(\sum_d 2\lambda(\xi_i^d) E[\bar{\mathbf{z}}_d^a \bar{\mathbf{z}}_d^a]^\top \right)^{-1} \left(\sum_d (\frac{2w_i^d - 1}{2} - 2\lambda(\xi_i^d) E[\bar{\mathbf{z}}_d^b]^\top \mathbf{b}_i) E[\bar{\mathbf{z}}_d^a] \right), \quad (6.13)$$

$$\mathbf{b}_i = \left(\sum_d 2\lambda(\xi_i^d) E[\bar{\mathbf{z}}_d^b \bar{\mathbf{z}}_d^b]^\top \right)^{-1} \left(\sum_d (\frac{2w_i^d - 1}{2} - 2\lambda(\xi_i^d) E[\bar{\mathbf{z}}_d^a]^\top \mathbf{a}_i) E[\bar{\mathbf{z}}_d^b] \right). \quad (6.14)$$

6.3 Experimental Results

We demonstrate the capability of the 3-modality tr-mmLDA model in a multimedia annotation task. As described in detail in Chapter 3, we use a collection of multimedia documents obtained by recording television programs and divide each long recording into small 20-second clips. Each multimedia clip contains audio-video streams and the associated closed-caption texts. In order to represent all the 3 data modalities in a histogram representation, we need to first define the concept of audio and video words. To this end, we use short-term spectral features to represent audio signals of each multimedia document. 8-kHz input audio signals are first divided into frames of size 256 samples. 256-pt FFT is then performed on each frame and the log of magnitude spectrum are used as our spectral features. A dictionary of audio words is learned by fitting a mixture of Gaussian model with diagonal covariance to a collection log spectrum features of the training data. The number of Gaussians used corresponds to the desired number of audio words T_r^a in the dictionary. For video features, we emphasize the use of motion features and adopt the motion representation of [PL06]. Each input video is first segmented into blocks of size 8 × 8 × 8, and spatio-temporal ICA features which describe the motion content of each video block are extracted. Since ICA features are known to be sparse, we learn a dictionary of video words by fitting a mixture of Laplacian model, with the number of mixtures corresponding to the desired number of video words T_r^v in the dictionary.

The 2 multimedia datasets used in our experiment are obtained from 2 separate recordings. As described in detail in Chapter 3, the first dataset is a collection of 2,502 clips obtained from a total of 24 episodes of a television program called Modern Marvels. A 20-sec clip in this dataset contains an average of 771 audio frames, 501 video blocks, and 5-51 annotation words with an average of 25.63 words per clip. We denote this dataset as MM24. The second dataset is a collection of 2,267 documents obtained from 30 episodes of 6 different TV programs. A 20-sec clip in this dataset contains an average of 677 audio frames, 868 video blocks, and 3-46 annotation words with an average of 22.68 words per clip. We denote this dataset as MM6. In our experiment on both datasets, we use $T_r^v = 256$ and $T_r^a = 100$ and $T_w = 4,171$ words for the MM24 dataset and $T_w = 4,511$ for the MM6 dataset.

Table 6.1: Comparison of per-word caption perplexity on a test set of the MM24 dataset using (a) audio alone, (b) video alone, and (c) audio and video together.

K_v	K_a	L	Video alone	Audio alone	Video and Audio
10	10	10	2327.345	2312.088	**2270.406**
20	20	20	2284.468	2272.209	**2189.741**
30	30	30	2280.381	2273.523	**2206.605**
50	30	50	2289.372	2275.964	**2222.766**
80	30	80	2304.942	2283.715	**2243.270**

The goal is to demonstrate the ability of the 3-modality extension of tr-mmLDA in using the audio-visual content to predict the missing captions. We use the per-word caption perplexity metric as in Chapter 4, which describes how well the fitted model can be used to predict the closed-caption data given the corresponding audio and video portion of the document. We compare the prediction using the 3-modality model with the performance obtained from the 2-modality tr-mmLDA, where either audio or video alone are used in making predictions. Table 6.1 shows the predictive caption perplexity computed on a 402-document test set of the MM24 dataset, when using (a) audio modality alone, (b) video modality alone, and (c) audio and video together to make prediction. The numbers in boldface are the best performance for each setting of K_a, K_v, and L (the number of topics in audio, video, and text modalities). As seen in Table 6.1, the best performance for all values of K_a, K_v, and L is obtained when using audio and video together to infer the missing captions. The reduction in perplexity obtained in the 3-modality model over a 2-modality case is quite significant and can reach up to 100 words. In a typical multimedia document, most of the caption words are used to describe objects that manifest themselves in both the audio and video modalities. In such a scenario, using all the available observations to predict closed-captions data is expected to yield improved performance over using audio or video modality alone.

A similar conclusion can be drawn from an identical experiment carried out on the MM6 dataset. Table 6.2 shows the comparative results. Audio data seems to be more informative than video in predicting captions, as seen in the lower perplexity values obtained using the same number of hidden topics, even though we only use 100 audio words compared to 256 words in the video modality. This result, however, should come as no surprise since the caption data used in our experiments are obtained from audio transcriptions provided by the television broadcast companies. The reduction in perplexity can reach over 200 words when using the 2 observation modalities together to make prediction.

Table 6.2: Comparison of per-word perplexity on a test set of the MM6 dataset using (a) audio alone, (b) video alone, and (c) audio and video together.

K_v	K_a	L	Video alone	Audio alone	Video and Audio
10	10	10	2508.051	2338.159	**2249.515**
20	20	20	2516.615	2324.345	**2203.778**
30	30	30	2562.345	2435.008	**2358.814**
50	30	50	2644.595	2628.567	**2537.078**
80	30	80	2699.212	2653.566	**2604.788**

We demonstrate the prediction capability of the 3-modality extension of sLDA-bin on an image annotation task, employing the same COREL dataset as used in Chapter 4 and 5. Instead of concatenating different features extracted from each image patch, we treat each feature type as a separate observation modality, gaining different perspectives of the same document from the point of view of different feature types. This is indeed appropriate especially when combining features from multiple scales since feature concatenation is not an option, i.e. one set of features are obtained from small patch sizes (10×10 patches) to capture the fine details, while another set of features are extracted from larger image regions (40×40 regions) to provide a more global description of an image. In this experiment, we treat the SIFT and color features (described in Chapter 3) as 2 separate data modalities representing 2 manifestations of the same document. We show in Table 6.3 the comparison of label prediction probability, which can been as a form of an affinity measure (inverse distance measure) between the predicted binary annotation and the ground-truth data, for the 3 cases where we use (a) SIFT features alone, (b) color features alone, and (c) color and SIFT features together to make predictions. The results in Table 6.3 are obtained for different settings of numbers of hidden topics K_s (for SIFT) and K_c (for color), using 512 SIFT words and 256 color words. When $K_s = K_c = 10$, SIFT features give a higher prediction probability than color features. This can be seen as a direct result of our decision to use a larger number of SIFT words. K_s, however, seems to peak at $K_s = 20$ and as K_s and K_c continue to increase, the color modality turns out to be more predictive of annotation labels as seen in the higher label prediction probability. Again, a similar conclusion can be drawn, as in the case of multimodal features: using the 2 types of features in combination yield better performance over using each individual modality alone for all values of K_s and K_c.

Table 6.3: Comparison of label prediction probability when making inference using (a) SIFT features alone, (b) color features alone, and (c) SIFT and Color features together.

K_s	K_c	SIFT	Color	SIFT and Color
10	5	-7211.028	-7513.551	-6822.163
10	10	-7211.028	-7287.714	-6678.179
20	10	-7182.945	-7287.714	-6676.720
20	20	-7182.945	-7165.212	-6627.742
30	10	-7219.059	7287.714	-6699.018
30	20	-7219.059	-7165.212	-6694.553
30	30	-7219.059	-7126.554	-6580.760
30	50	-7219.059	-7146.669	-6591.940

Another popular metric for comparing performance of image annotation systems is the precision-recall metric, which was adopted in [BDdF$^+$03, JLM03, LMJ03, FML04]. The definition of the quantities involved are found in [FML04] and given as follow. Assume the annotation word of interest is 'tiger', we define A to be the number of test images automatically annotated with the word tiger; B is defined as the number of test images correctly annotated with the word tiger; C is the number of test images having the word tiger in ground-truth annotation. recall is therefore computed as $\frac{B}{C}$ while precision is $\frac{B}{A}$. As done in [BDdF$^+$03, FML04], for each test image we predict 5 annotation words that are the most likely under the fitted model, and compute the mean-per-word precision and recall by averaging over the best 49 annotation words. The results for different settings of K_s and K_c are presented in Table 6.4.

In general, as we increase the number of topics, we are able to obtain a better prediction performance as seen in the associated increase in the precision and recall values. Our best result is obtained when setting $K_s = K_c = 30$. While our results are slightly inferior (with slightly lower precision and recall values) to the previous results shown as CRM (continuous relevance model) [LMJ03] and MBRM (Multi-variate Bernoulli relevance model) [FML04] in the bottom 2 rows of Table 6.4, we take consolation in the fact that sLDA-bin, which is a parametric model, can produce very comparable results to those obtained from the doubly non-parameteric relevance models of CRM and MBRM. By treating each training document as a prototype in the association model and fit a non-parametric kernel density model to each training image, the computational requirements of CRM and MBRM make the models unscalable to the size and magnitude of today's multimedia collections. Our method, on the other hand, employs a much smaller number of parameters and is therefore more appropriate in modeling large-scale multimedia repositories.

Table 6.4: Comparison of mean-per-word precision and mean-per-word recall for 49 best annotation words comparing predictions made using (a) SIFT or Color alone, and (b) SIFT and Color features together.

K_s	K_c	Best of SIFT or Color			SIFT and Color		
		Recall	Precision	# recall ≥ 0	Recall	Precision	# recall ≥ 0
10	5	-	-	18	0.4303	0.3055	53
10	10	-	-	36	0.5028	0.4233	66
20	10	0.3472	0.2728	49	0.4915	0.4053	68
20	20	0.3472	0.2728	49	0.5687	0.4894	73
30	10	0.3394	0.2465	50	0.5137	0.4376	68
30	20	0.3394	0.2465	50	0.5529	0.4686	74
30	30	0.4313	0.3465	53	0.5565	0.5215	78
30	50	0.4604	0.3814	58	0.5983	0.5133	83
CRM					0.70	0.59	107
MBRM					0.78	0.74	122

Chapter 7

Video Features

This chapter proposes a new type of video features which capture patterns of motion and serve as a general measure of activities in the scene. Inspired by representations derived from statistics of natural images, in this work we investigate the use of nonlinear dependencies in the statistics of natural image sequence to learn higher-order structures in natural videos. We propose a two-layer model that learns variance correlation between linear ICA coefficients and present a novel nonlinear representation of natural videos. The first layer performs a linear mapping from pixel values to ICA coefficients. In doing so, the spatio-temporal dynamics in natural videos are decomposed into a set of independent basis each encoded with “independent motion”. By assuming that the nonlinear dependency of ICA coefficients takes the form of variance correlation, the second layer learns the joint distribution of ICA sources that captures how these independent bases co-activate. Experimental results show that the abstract representation correspond to various activation patterns of bases with similar motion, hence the term motion patterns. Our model offers a novel description of higher-order structures in natural videos. We illustrate the usefulness of the proposed representation in video segmentation and denoising tasks.

7.1 Introduction

It is generally assumed that sensory coding systems in mammals are functionally adapted to optimally process visual stimuli in natural environment. The past decade has witnessed several successful computational models that use the statistics of natural stimuli to learn low-level representations of natural scenes that have similar properties to the receptive fields of neurons in primary visual cortex [FO96, BS97]. One such seminal idea proposed by Bell and Sejnowski [BS97] suggests that biological systems perform a linear independent coding of visual stimuli. Independent Component Analysis (ICA) [SB95] performs a linear mapping that transforms input pixel values to ICA coefficients in such a way that the derived coefficients are as independent as possible. The bases learned under this independent principle constitute the building blocks of natural scenes and are found to resemble oriented edges/bars for natural images. For dynamic scenes, motion is known to play a critical role in how a human perceives and responds to the environment. The representation of natural image sequences learned by ICA indeed reflects this notion [vHR98]. ICA has been found to decompose the spatio-temporal dynamics of natural videos into a set of bases, each encoding an "independent" motion. The resulting independent components correspond to localized oriented edges/bars moving at constant velocities in the directions perpendicular to their orientations [vHR98].

The representation of natural scenes learned by a linear ICA model is nevertheless considered relatively primitive. Complex dynamic scene structures such as moving contours, shapes, temporal textures, objects, and other high-level scene properties cannot be well represented using a linear ICA model. The reason behind this limitation lies in that ICA captures linear dependency between pixel values and therefore has restricted capability of modeling nonlinear statistical regularities that exist in natural scenes. It is observed that even though ICA coefficients are linearly independent, their magnitudes and energies can exhibit strong dependencies [Sim97, HH01]. In the case of natural images, previous studies [KL03, PL04] have shown that higher-order image structures such as texture-related properties can be learned by capturing variance dependency in the ICA sources as a *post-processing* step after ICA. In [HH01, AHV03, HH00], a topographic organization in a manner similar to the topographic structure of simple cells in the visual cortex emerges as a result of incorporating variance dependency in the ICA source modeling.

Our work is related to the hierarchical model in [KL03] that learns higher-order image structures. By using the Bayesian formulation, Karklin and Lewicki [KL03] model the ICA source variances explicitly as products of variance bases $\mathbf{B}$ and a sparse hierarchical prior v. The prior v can be thought of as controlling the selection of variance bases (the different bases correspond to different types of variance dependencies). While various abstract image structures can successfully be learned, the Bayesian framework in [KL03] makes the learning complicated and several approximations are indeed required in estimating the model parameters. Park and Lee [PL04] proposed a simplification to the model in [KL03] by constraining the prior v to be either 0 or 1, thus allowing the ICA source distribution to have a simple analytic pdf and the model parameters can be easily learned via EM.

The focus of our work is in learning high-level representation of natural videos. Our work builds on the previous works in [PL04] by extending the mixture of Laplacian source model to learn higher-order structures in the spatio-temporal domain. In contrast to the works in [KL03] and [PL04] which fix the ICA bases before learning the variance correlated source prior, we propose a new learning algorithm that optimizes the ICA bases and ICA source distribution parameters simultaneously using the generalized EM framework. Experimental results show that our model captures variance dependencies that correspond to abstract structures in natural videos. We demonstrate the usefulness of our proposed video representation in 2 video processing applications: video segmentation and video denoising.

7.2 Modeling Natural Videos

We begin by explaining the linear ICA model when applied to natural image sequences as was done in [vHR98]. Then we introduce the nonlinear statistical dependency in the ICA sources in the form of variance correlated prior. Lastly, we present the learning algorithm for estimating the spatio-temporal ICA bases and ICA coefficient distribution simultaneously.

7.2.1 Linear ICA Model for Natural Image Sequences

The instantaneous ICA mixing model was used in [vHR98]. A video block $\mathbf{x}$, which consists of T consecutive frames of image patches of size n×n, is transformed by M linear filters $\mathbf{w}_m$ to coefficients u_m for $m = \{1, 2, ..., M\}$. The goal of ICA is to find such a linear mapping that makes the filter responses u_m's as independent as possible. Writing in the matrix form, we have:

$$\mathbf{u} = \mathbf{W}\mathbf{x} \tag{7.1}$$

$$\mathbf{x} = \mathbf{A}\mathbf{u} = \mathbf{W}^{-1}\mathbf{u} \tag{7.2}$$

The ICA coefficients u_m in the standard ICA model are assumed to be statistically independent. Thus the joint distribution factors into the product of individual component distributions: $P(\mathbf{u}) = \prod_{m=1}^{M} P(u_m)$. In natural image sequences, these coefficients are found to have a sparse, super-gaussian distribution and are often modeled using a Laplacian pdf. Therefore, the source distribution in the standard ICA linear model assumes the form:

$$P(\mathbf{u}) = \prod_{m=1}^{M} \frac{1}{2\lambda_m} \exp(-\frac{|u_m|}{\lambda_m}) \tag{7.3}$$

where variance λ_m of each source is fixed to 1.

7.2.2 Modeling Variance Correlation: Mixture of Laplacian Source Prior

Our model extends the basic ICA model above by introducing nonlinear dependency in the ICA source distribution that cannot be eliminated using a linear transformation. Previous studies have shown that after ICA the filter coefficients often show strong variance or magnitude correlation (though they are no longer linearly predictable from each other), the question remains in how to choose the specific form of variance dependency. Many classes of natural scenes tend to activate certain filters more frequently than others (the variance of the activated coefficients will thus be higher than the non-activated ones). One way to model variance correlation in ICA sources is by describing how the ICA source variances change under different types of scenes or different "contexts". Precisely, we denote "context" as a hidden variable that selects the variance correlation pattern which in turn controls the activation/deactivation of ICA bases. Given the "context" z of each video block, the ICA filter responses are assumed to be independent and have Laplacian distribution with variance λ_m for $\{m = 1, ..., M\}$ dependent on z. We assume there are a fixed number of "contexts", i.e. $z \in \{1, 2, ..., K\}$, the variance correlated source prior that we use thus has the form of a mixture of Laplacian distribution [PL04]:

$$P(\mathbf{u}|z = k) = \prod_{m=1}^{M} \frac{1}{2\lambda_{km}} \exp(-\frac{|u_m|}{\lambda_{km}}) \tag{7.4}$$

$$P(\mathbf{u}) = \sum_{k=1}^{K} \pi_k P(\mathbf{u}|z = k) \tag{7.5}$$

$$P(\mathbf{u}) = \sum_{k=1}^{K} \pi_k \prod_{m=1}^{M} \frac{1}{2\lambda_{km}} \exp(-\frac{|u_m|}{\lambda_{km}}) \tag{7.6}$$

Our model for nonlinear video representation is described in the diagram in Figure 7.1. The problem is to find a linear mapping $\mathbf{W}$ that maps the natural video data $\mathbf{x}$ into $\mathbf{u}$ such that the coefficients $\mathbf{u} = \mathbf{W}\mathbf{x}$ follow the mixture of Laplacian distribution as described in Eqn 7.6.

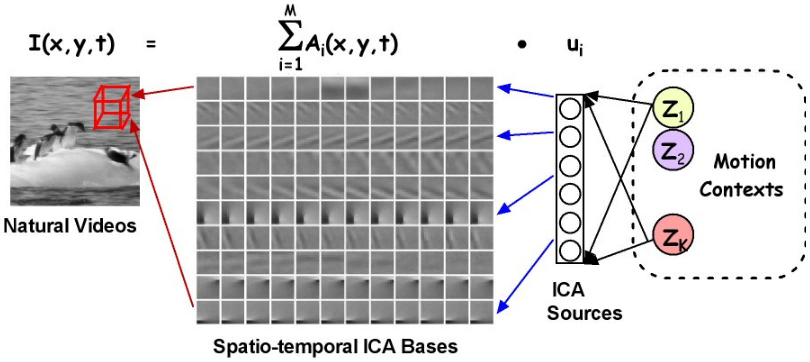


Figure 7.1: Diagram of our proposed model for nonlinear video block representation. A linear transformation is used in the first layer to decompose the spatio-temporal dynamics in natural videos using a set of bases that encode "independent motion". The second layer learns the variance correlation patterns that govern the activation of these bases. In our model, the variance correlated source prior has a mixture of Laplacian distribution. Each mixture thus describes the "motion context" that generates the video blocks.

7.2.3 Maximum Likelihood Estimation

The parameters of our model are the ICA separating matrix and the Laplacian mixture parameters: $\Psi = (\mathbf{W}, \pi_k, \Lambda_k)$ where $\mathbf{W}$ denotes the separating matrix, $\Lambda_k = (\frac{1}{\lambda_{k1}}, \frac{1}{\lambda_{k2}}, ..., \frac{1}{\lambda_{kM}})^T$ denotes the ICA source variances for mixture k, and π_k denotes $P(z = k)$. We formulate the learning problem using the maximum likelihood framework by looking for a set of parameters that maximizes the log-likelihood of the observed data. Because the ”context” z is hidden, we use the EM formulation for missing data problem. Let $D = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ be N video blocks gathered from a collection of natural scene. We write $P(\mathbf{x}_l)$ in terms of $P(\mathbf{u}_l)$ using the linear transformation in Eqn(7.1) as $P(\mathbf{x}_l) = |\det(\mathbf{W})|P(\mathbf{u}_l)$. The data log-likelihood then becomes:

$$\begin{aligned} \log P(D) &= \sum_{l=1}^{N} \log P(\mathbf{x}_l) \qquad (7.7) \\ &= N \log|\det(\mathbf{W})| + \sum_{l=1}^{N} \log P(\mathbf{u}_l) \qquad (7.8) \end{aligned}$$

Rewrite $P(\mathbf{u_l})$ in the form of mixture of Laplacian pdf with K mixtures as shown in Eqn(7.6), we then have:

$$\log P(D) = N \log|\det(\mathbf{W})| + \sum_{l=1}^{N} \log \sum_{k=1}^{K} P(\mathbf{u}_l|k)\pi_k \tag{7.9}$$

Using Jensen's inequality, the E step constitutes computing the expected value of the complete data log-likelihood F written as:

$$\begin{aligned} F &= N \log|\det(\mathbf{W})| + \sum_{l}\sum_{k} Q_l(k)\log(P(\mathbf{u}_l|k)\pi_k) \qquad (7.10) \\ &= \sum_{l,k} Q_l(k) \left[\log \pi_k - \sum_{m=1}^{M} (\log 2\lambda_{km} + \frac{|u_{lm}|}{\lambda_{km}})\right] + N \log|\det(\mathbf{W})| \end{aligned}$$

where $Q_l(k)$ denotes the soft assignment probability of sample $\mathbf{x}_l$ $P(z_l = k|\mathbf{x}_l)$. By differentiating F with respect to the parameters Ψ, we derive the updates for parameters in the M step. First, we differentiate F w.r.t. λ_{km} and set the derivative to 0 and apply the same procedure to π_k this time subject to the constraint that $\sum_k \pi_k = 1$, we have the following update rules for the mixture parameters:

$$\begin{aligned} \lambda_{km}^{(t+1)} &= \frac{\sum_{l=1}^{N} |u_{lm}| Q_l(k)}{\sum_{l=1}^{N} Q_l(k)} \qquad (7.11) \\ \pi_k^{(t+1)} &= \frac{\sum_{l=1}^{N} Q_l(k)}{\sum_{k=1}^{K}\sum_{l=1}^{N} Q_l(k)} \qquad (7.12) \end{aligned}$$

where $Q_l(k) = P(z_l = k|\mathbf{x}_l; \Psi^{(t)})$ is the soft assignment probability at iteration t and is computed as in the following (note $\log|\det(W)|$ cancels out and disappears):

$$Q_l(k) = \frac{P(\mathbf{u}_l|z_l = k; \Psi^{(t)})\pi_j^{(t)}}{\sum_{k=1}^{K} P(\mathbf{u}_l|z_l = k; \Psi^{(t)})\pi_j^{(t)}} \tag{7.13}$$

To estimate the ICA separating matrix, we differentiate F with respect to $\mathbf{W}$. Since there is no closed form solution, we derive the gradient update as follows:

$$\begin{aligned} \nabla F_{\mathbf{W}} &= \mathbf{W}^{-T} - E\left[\sum_{k=1}^{K} Q_l(k) \sum_{m=1}^{M} \frac{\text{sign}(u_{lm})}{\lambda_{km}} \frac{\partial u_{lm}}{\partial \mathbf{W}}\right] \\ &= \mathbf{W}^{-T} - E\left[\sum_{k=1}^{K} Q_l(k)\Lambda_k \circ \text{sign}(\mathbf{u}_l)\mathbf{x}_l^T\right] \qquad (7.14) \end{aligned}$$

$$\nabla F_{\mathbf{W}}^{nat} = \left[I - E \left(\sum_{k=1}^{K} Q_l(k) \Lambda_k \circ \text{sign}(\mathbf{u}_l)\mathbf{u}_l^T \right) \right] \mathbf{W} \tag{7.15}$$

where $\circ$ denotes element-wise multiplication. The expectation is the sample average taken over all training samples $\{l = 1, 2, ..., N\}$. The natural gradient in Eqn(7.15) is derived by right-multiplying the right hand term in Eqn(7.14) with $\mathbf{W}^T\mathbf{W}$. This step eliminates the need to invert the matrix $\mathbf{W}$ at every iteration. The update rule shown in Eqn(7.17) for $\mathbf{W}$ is thus a gradient descent(ascent) algorithm with the increment of $\mathbf{W}$ given by the sum of natural gradient and the momentum term as shown in Eqn(7.16):

$$\Delta\mathbf{W}^{(t)} = \eta \nabla F_{\mathbf{W}}^{nat} + \mu \Delta\mathbf{W}^{(t-1)} \tag{7.16}$$

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} + \Delta\mathbf{W}^{(t)} \tag{7.17}$$

7.3 Experiments

We collect 40,000 video blocks of size n×n×T (each video block consists of T consecutive frames of n×n gray-scale image patches centering at the same location; in our experiments we try various sizes of video blocks: 8×8×8, 10×10×10, 12×12×12, 16×16×16) by sampling randomly from the database of 216 natural image sequences (same database as used in [vHR98]; courtesy of J. H. van Hateren). Each video sequence is 192s long and contains 9600 frames of size 128×128. As a pre-processing step, the mean of each video block is subtracted off and the variance is normalized to 1. PCA is then used to reduce the dimensionality of the data. In our experiments with 12×12×12 and 16×16×16 video blocks which we show the results in the next section, we learn 160, 200, 250, 300, 400 independent components accordingly. For the mixture of Laplacian parameters, we try several numbers of mixtures K = 8, 16, 24, 32, 64 for the different video processing applications. Learning by EM is quite fast and after less than a few hundred iterations EM converges. Figure 7.2 shows a comparison of log-likelihood for learning 16 mixtures and 250 components from 20,000 samples when the 2 stages are learned separately (red) as in [PL04] and our learning algorithm that updates the 2 stages simultaneously (blue). As expected, our algorithm is able to achieve higher likelihood.

7.4 Learning results

7.4.1 Spatio-temporal Basis captures "Motion"

To show off how ICA encodes the spatio-temporal dynamics in each video block, the linear transformation described in Eqn(7.2) is re-written as shown in Eqn(7.18), where it is clear now that u_m is neither a function of space $(x,\ y)$ or time t. The spatio-temporal dynamics in $I(x, y, t)$ are encoded in the bases $A_m(x, y, t)$. We can think of u_m as controlling the activation

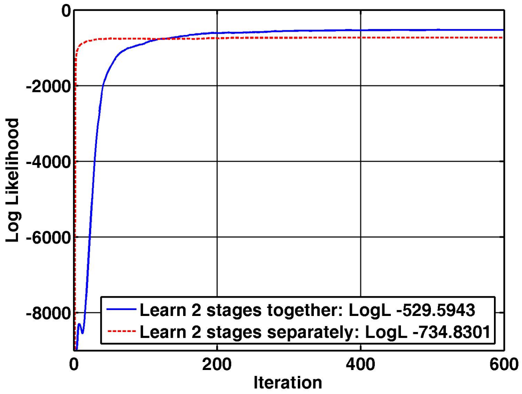
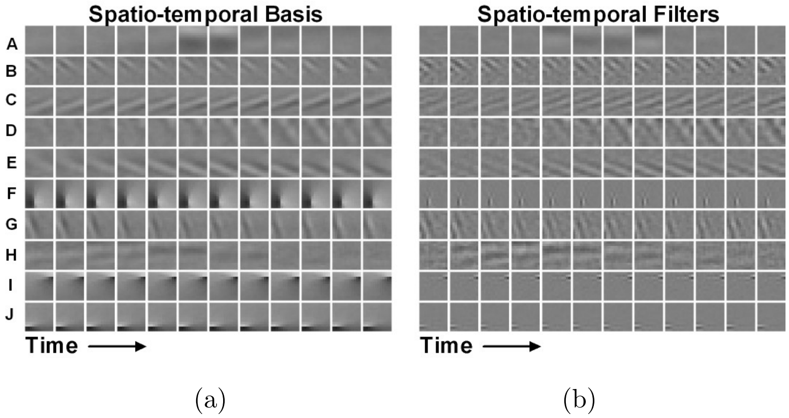


Figure 7.2: Log likelihood comparison when the 2 stages are learned separately (red) and our learning algorithm that updates the 2 stages simultaneously (blue).



(a)

(b)

Figure 7.3: (a)-(b) Examples of spatio-temporal ICA bases of 12×12×12 video blocks and their corresponding filters. Each row corresponds to the 12 frames of a 12×12×12 spatio-temporal filter.

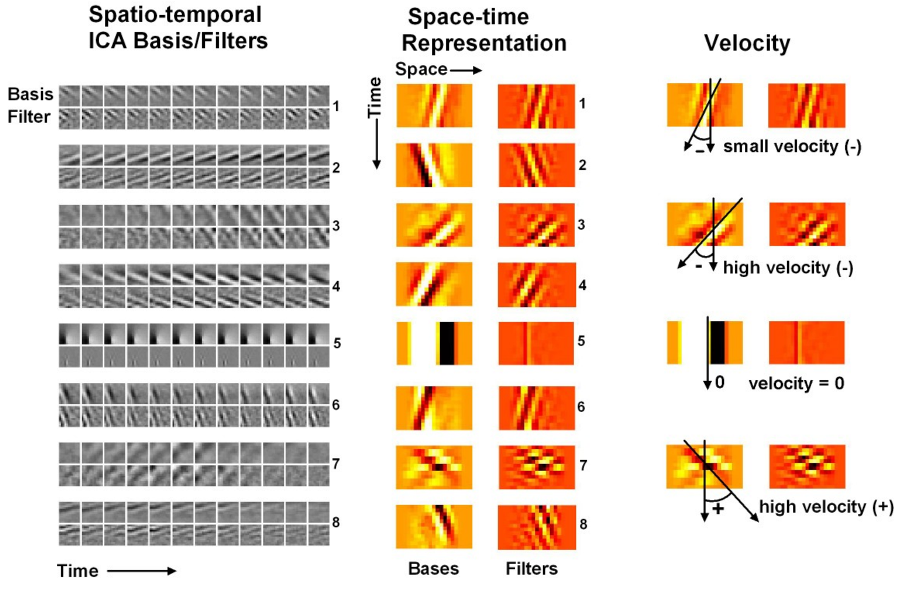
of the m^{th} basis. The goal of ICA is to find a set of bases that are activated as independently as possible (in a linear sense).

$$I(x, y, t) = \sum_{m=1}^{M} A_m(x, y, t)u_m \tag{7.18}$$

The spatio-temporal bases learned from natural video statistics using our algorithm are found to be qualitatively similar to the results from the model in [vHR98] in that they resemble Gabor filters moving at *constant* velocities in the direction orthogonal to their orientation. In essence, ICA decomposes the spatio-temporal dynamics in natural scenes using a set of spatio-temporal bases that encode ”independent motion”. Examples of the spatio-temporal bases and filters learned using our algorithm are shown in Figure 7.3(a)-(b).

7.4.2 Space-time diagram

To capture the temporal dynamics of the spatio-temporal ICA basis/filters, we adopt the space-time diagram as used in [vHR98, AB85]. The idea behind the use of the space-time representation is that the temporal behaviors can be summarized by integrating the bases/filters along their directions of propagation. Thus, each frame (n×n pixels) of a spatio-temporal basis/filter will be reduced to a one-dimensional signal; stacking T frames together creates a space-time representation of an n×n×T spatio-temporal basis/filter as a two-dimensional image as shown in Figure 7.4(b). It is observed that when displaying each spatio-temporal ICA basis/filter in the space-time representation, a Gabor-like pattern emerges. In [vHR98], the angle of this space-time Gabor pattern with respect to the time axis is used as an indicator of the velocity of the basis/filter. As seen in Figure 7.4(c), the tangent of the marked angles (the slope parameters) give the velocities of the bases/filters shown in (a). When the basis/filter moves fast, the slope will be large (and small for a slow moving basis). When the spatial characteristic of the basis/filter does not change with time, this quantity is 0. This measure of velocity also reflects the direction of movement in the sign of the slope (+/-). Since tan(θ) is a monotonically increasing function, we simply use the angle θ itself as a measure of velocity. The velocity for each basis/filter, therefore, ranges between $-\frac{\pi}{2}$ to $\frac{\pi}{2}$ radians/s or -90 to 90 degrees/s.



(a) (b) (c)

Figure 7.4: (a) Examples of spatio-temporal ICA bases/filters. (b) The corresponding space-time representations of bases/filters in (a). (c) The angle the orientation vector makes with respect to the time axis indicates the velocity of each basis/filter.

7.4.3 Visualization of "Motion Patterns"

To visualize the kind of higher-order structures in natural videos that our model captures, we look at the variance vectors Λ_k for mixture k which correspond to different patterns of co-activation of bases. In particular, since each spatio-temporal basis encodes "motion", we call the contexts that control the co-activation of these bases "motion contexts" and the variance correlation patterns "motion patterns".

We adapt the visualization scheme for viewing higher-order image structures as used in [KL03]. For each spatio-temporal ICA basis, we find a matching Gabor filter with the same centering position, Gaussian envelope size, orientation, and spatial frequency. These parameters determine the position of each basis in the spatial and frequency visualization (because ICA bases are localized in space/frequency, each basis corresponds to 1 dot in the spatial and frequency plot). Specifically, the location of each basis in the spatial plot is the centering position of its matched Gabor filter, while the location (x, y) in the frequency plot is chosen so that $\arctan(\frac{y}{x})$ corresponds to the basis' orientation and $\sqrt{x^2 + y^2}$ (the distance of point (x, y) to the origin) denotes its spatial frequency. The dot color indicates the variance value of the basis (i.e. red shades for high values and blue shades for values close to 0). The temporal characteristics of our spatio-temporal bases are visualized by adding the third axis Z to both the spatial and frequency plots. In the spatial domain, Z denotes the temporal center of activation (time index of the center of the Gabor pattern in space-time diagram). In the frequency plot, Z-axis denotes velocity to show off the movement of each basis in time. The resulting 3D visualization is, however, difficult to understand in a 2D plane. We therefore project each 3D plot of X-Y-Z onto three 2D planes–Y-X, Z-Y, Z-X, and different colors are used for different axes (X is green, Y is brown and Z is blue). Each variance correlation pattern is then shown in 8 plots–4 spatial plots (3D spatial plot, Y-X, Time-X, Time-Y) and 4 frequency plots (3D frequency plot, F_y-F_x, Velocity-F_x, Velocity-F_y) as seen in Figure 7.5.

Two main types of variance correlation patterns emerge. Type I (mixture A in Figure 7.5) corresponds to a mixture that activates the bases based on their spatial locations regardless of their orientations, spatial frequencies or velocity preferences. In the spatial plot of mixture A in Figure 7.5, the high variance bases (red dots) are concentrated in the first 8 frames of 16×16×16 video blocks, while in the frequency plot, these red dots seem to be at random locations. The second type of variance co-variation pattern (mixture B in Figure 7.5) shows co-activation of bases based on their frequency preferences. As seen in the frequency plot of mixture B, the fast-moving low-frequency bases (red shades) are activated together, while the slow-moving high-frequency bases have small variances and are considered inactive. Mixture B most likely corresponds to video blocks that contain large edges moving at a high speed. More examples of Type II variance correlation pattern are shown in Figure 7.6 (only their frequency plots are displayed). We see various co-activation patterns of bases based on similar speeds (B1 activates non-moving bases;

B2 activates slow to medium-speed bases; B3 activates slow-moving bases) and velocity (B4).

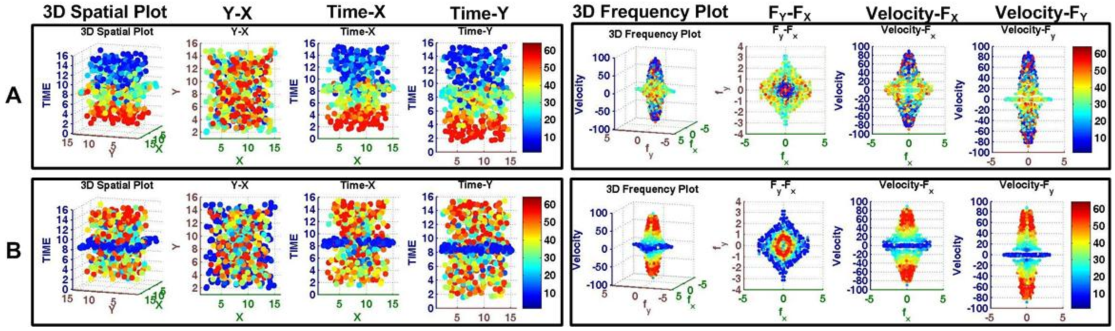


Figure 7.5: Visualization of the learned variance co-activation pattern. 2 main types of mixtures emerge. Mixture A illustrates type I pattern where bases are co-activated based on their spatial locations regardless of their frequency/orientation preferences. Mixture B illustrates type II pattern, where bases with the same frequency preferences are co-activated. The 3D visualization (X-Y-Z plot) is shown on the leftmost plot and the 3 2D projections onto Y-X, Z-X, and Z-Y planes are shown respectively.

To gain a better understanding of the kind of visual context each mixture represents, we use our model to learn the representation of dynamic textures that appear in natural scenes. For this purpose, we use the MIT temporal texture database which contains 15 classes of dynamic textures, e.g. boiling water, flags, fire, fountain, smoke, and etc. We collect 40,000 video blocks by drawing randomly from each class and learn K = 32 variance patterns. By using our algorithm to perform unsupervised clustering of video samples by assigning each video block $\mathbf{x}$ to the mixture that yields the highest posterior probability $\hat{z} = \arg\max P(z = k|\mathbf{x})$, we found that each mixture contains samples roughly from the same texture class. For each mixture in Figure 7.7, we display 5 texture sequences that have the highest probability of being generated from it. We thus give a definition of visual ”context” as the texture class each mixture corresponds to. Mixture A activates bases with high speed moving in all directions; the visual ”context” of this pattern is the texture created by water boiling. Mixture B activates the bases with horizontal orientation; the ”context” for this mixture is the stripe texture or texture of flowing water in the river viewed from far away. Mixture C activates the bases that have a particular spatial orientation; the corresponding visual ”context” is the rippling pattern of plastic sheet caused by the wind blowing. Mixture D activates high spatial frequency bases with small motion; the context is the texture from trees that are moved slightly by the wind. Mixture E activates bases with low spatial frequency and have vertical orientation; the context is the texture created by fire and clothes whirling in a dryer. Mixture F activates low spatial frequency bases moving with medium speed; the context is the texture created by water swirling in the toilet bowl or water in the shower.

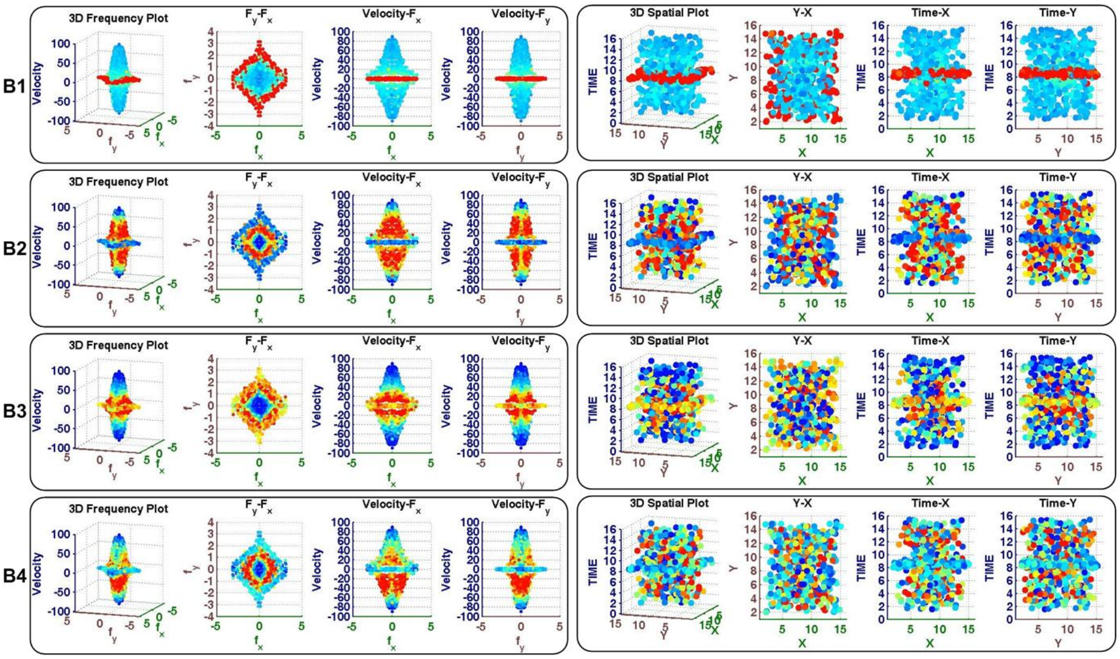


Figure 7.6: More examples of type II "Motion Patterns" learned from 16×16×16 video blocks. B1, B2, B3 activate the spatio-temporal bases based on their speeds, while B4 activates the bases with the same velocities (same direction and speed).

7.5 Applications

7.5.1 Video Segmentation

We illustrate the usefulness of the higher-order representation learned in an unsupervised video segmentation application. The idea behind this concept is to uncover the underlying higher-order "contexts" that control the visual appearance and motion in a video sequence. To do this, we label each video block with the "context" or mixture that gives the highest probability of generating it. That is, for each video block $\mathbf{x}_l$, we compute $\hat{z}_l = \text{argmax} P(z|\mathbf{x_l})$ (similar to the cluster assignment of temporal textures). We show the results of our simple segmentation algorithm on 2 video clip excerpts. Figure 7.8 shows the scene of a penguin jumping into the water. Our simple segmentation algorithm is able to successfully segment the moving penguin using its "motion pattern" as shown in the red blobs moving in time in Figure 7.8(b)-(c). Since each mixture corresponds to a co-activation pattern of ICA spatio-temporal bases, we compute the "energy" of each mixture by using the 2-norm of the variance vector $\Lambda_k = \sum_k \lambda_k^2$. The choice of the color map in Figure 7.8(b) reflects the *ordering* of mixtures by their "motion energy." The red shades correspond to high energy mixtures while the blue shades are used for low energy mixtures. In (c), only the pixels corresponding to the mixture with highest energy value are shown in red labels superimposed on the original video sequence.

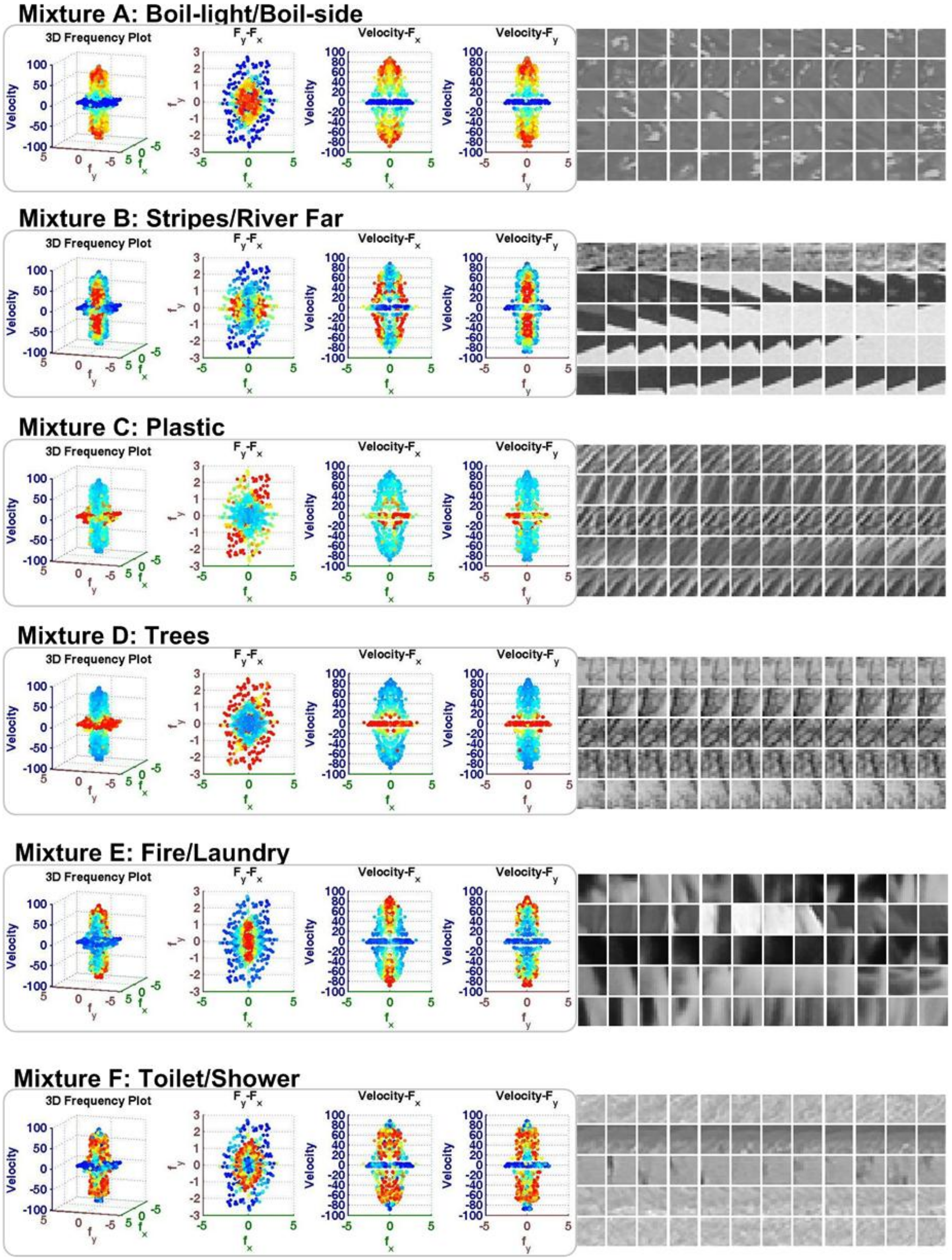


Figure 7.7: Understanding visual ”context” by learning a representation of dynamic textures drawn from 15 classes of the MIT temporal texture database. Roughly, the texture video blocks from the same texture category are assigned to the same mixture.

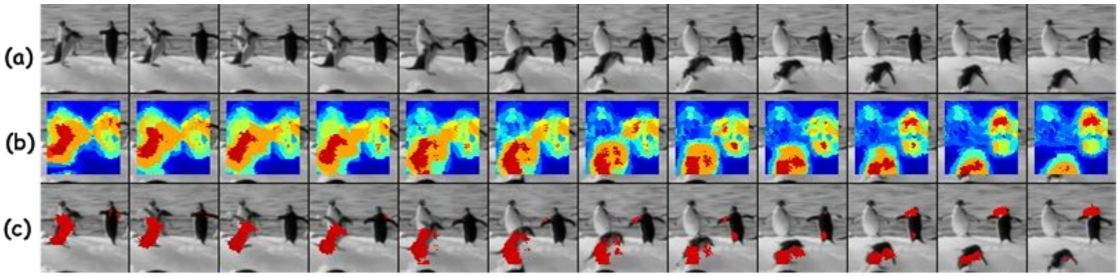


Figure 7.8: Examples of video segmentation by labeling each 12×12×12 video block with the label corresponding to the ”context” that gives the highest probability of generating the video block. (a) Original video sequence. (b) Segmentation labels for the whole scene. (c) Only pixels corresponding to the mixture with highest energy are shown in red.

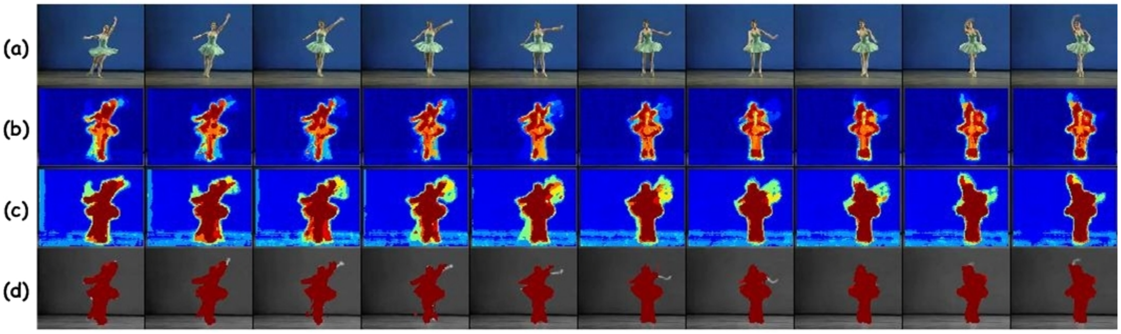


Figure 7.9: Segmentation results on a ballet dancer sequence by labeling 8×8×8 video blocks. (a) Original video. (b) Segmentation labels by assigning each pixel with a value from {1,2,.., K} corresponding to the mixture k that best explains the video block. (c) Segmentation labels by taking into account the *value* of the motion energy of each mixture k (d) Only pixels with motion energy exceeding a certain threshold are shown in red.

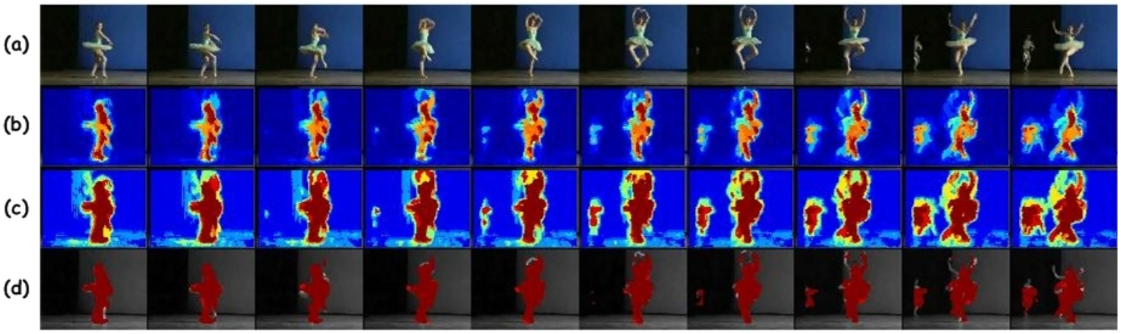


Figure 7.10: More segmentation example on the ballet dancer sequence. (a) Original video sequence. (b) Segmentation results labeled by mixture index (c) Segmentation labels using motion energy (d) Only pixels with motion energy exceeding a certain threshold are shown in red.

Figure 7.9 shows the segmentation results on a ballet dancer sequence. We first convert the color input into grayscale (color cue is not used in our algorithm). The original video sequence in (a) shows the articulated hand and leg movement of the ballet dancer. 7.9(b) shows the segmentation results by labeling each pixel with the mixture that best explains the motion pattern of the video block (each pixel thus takes a value between 1 and K). 7.9(c) shows a better segmentation result by assigning to each pixel the *value* of the motion energy of the mixture in 7.9(b). By taking into account the actual *value* of the motion energy in our segmentation algorithm, we are able to accentuate the contrast between the stationary region of the video and the region that contains moving objects (since the variance is less than 1 for non-activated coefficients and greater than 1 for activated coefficients, the squaring operation of the 2-norm enhances this difference). 7.9(d) shows only the pixels with motion energy above a certain threshold. The result confirms that our segmentation algorithm can track the articulated movement of the ballet dancer well.

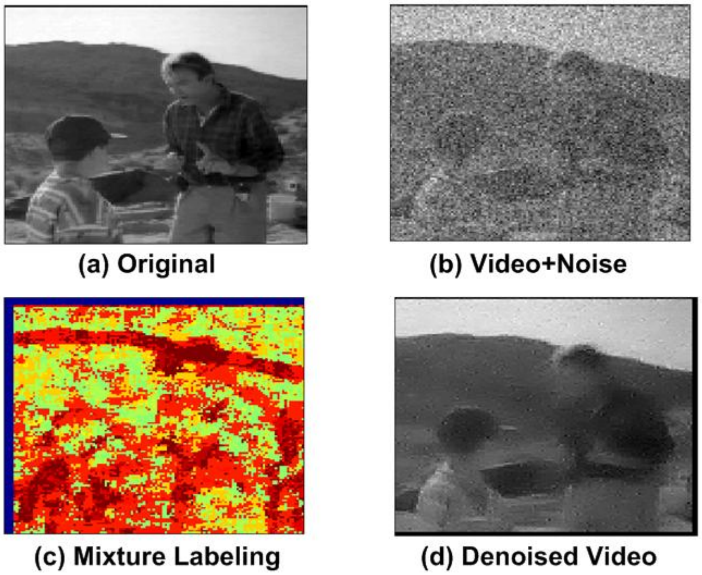


Figure 7.11: Adaptive video denoising based on the "context" of each region in the video. Noise variance is 0.81.

7.5.2 Adaptive video processing: Video Denoising

Another video processing application that shows the merit of our model is adaptive denoising where different regions in the input undergo different restoration procedures based on their local structures ("contexts"). Assuming an input video block $\mathbf{x}$ is corrupted with an additive Gaussian noise: $\mathbf{x} = \mathbf{A}\mathbf{u} + \mathbf{n}$. Given the noise variance, we can compute an MAP estimate of $\mathbf{u}$: $\hat{\mathbf{u}} = \arg\max P(\mathbf{u}|\mathbf{x};\mathbf{A})$. Using the ICA source prior given by the mixture that best explains the video block $\mathbf{x}$, i.e. $P(\mathbf{u}) \approx P(\mathbf{u}|z = k)$ where $k = \arg\max P(z = k|\mathbf{x})$ and assuming that $A^T = A^{-1}$, the solution to the MAP problem in Eqn 7.19 can be derived in closed form [Hyv99].

$$\begin{aligned}\hat{\mathbf{u}} &= \arg\max \log P(\mathbf{x}|\mathbf{u}) + \log P(\mathbf{u}) \\ &= \arg\min \frac{1}{2\sigma^2}|\mathbf{x} - \mathbf{A}\mathbf{u}|^2 + \sum_{m=1} \frac{|u_m|}{\lambda_{km}}\end{aligned} \tag{7.19}$$

The denoising result using 64 mixtures on 7×7×5 video block is shown in Figure 7.11 where the noise variance is 0.81. Due to the space limit, only one frame in the video sequence is shown. We compare our denoising performance with 3 other algorithms–ICA MAP using spatio-temporal bases, 2D Wiener Filter applied to a noisy video frame-by-frame and 3D Wiener Filter applied to video blocks. Figure 7.12 shows SNR comparison that our method performs better than ICA MAP or Wiener Filter. We expect even better results using K > 64.

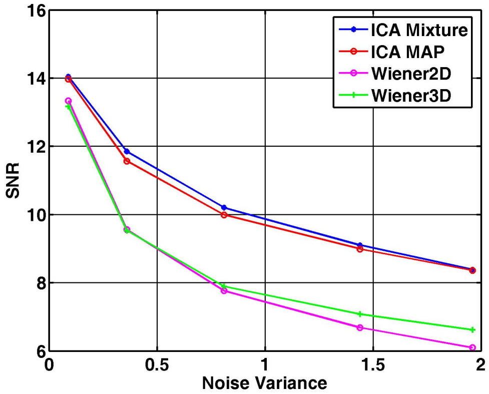


Figure 7.12: SNR comparison showing that our algorithm (K=64) performs better than ICA MAP, 2-Dimensional Wiener Filter and 3-Dimensional Wiener Filter.

7.6 Discussions

In this work, we have presented a mixture model for learning higher-order representation of natural videos. From the statistics of natural image sequences, our model learns in the first layer–the linear mapping that decomposes the spatio-temporal dynamics in natural image sequences into a set of bases that encode ”independent motion” and in the second layer–the patterns of co-activation of these bases which we term ”motion patterns.” These variance correlation patterns are found to correspond meaningfully to some high-level ”contexts” such as different types of textures. We show 2 applications of our representation: video segmentation and adaptive denoising, both give very good results. Because of the excellent motion segmentation results, in the future work we plan to explore the use of our representation in the task of behavior detection/classification. A video compression scheme based on our representation also seems promising.

Chapter 7, in part, is a reprint of the material as it appears in: D. Putthividhya, T.-W. Lee, “Motion Patterns: High-level Representations of Natural Video Sequences,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006. I was the primary researcher of the cited materials and the co-author listed in these publications supervised the work which forms the basis of this chapter.

Chapter 8

Final Remarks

In this thesis, we have investigated several extensions of the basic Latent Dirichlet Allocation model for text and multimedia documents. First, for the application of exploratory analysis, we have presented Independent Factor Topic Models (IFTM) which use linear latent variable models to directly model the hidden sources of correlation between topics and learn topic correlations. We have extended IFTM to learn correlations between latent topics of different data modalities in multimedia documents and presented a second topic model called Topic-regression multi-modal Latent Dirichlet Allocation (tr-mmLDA) which uses a linear regression module to learn the precise relationships between latent variables of different data types. We employed tr-mmLDA in an image and video annotation task, where the goal is to learn statistical association between images and their corresponding captions, so that we can accurately infer the missing caption data. In many image annotation datasets used in performance evaluation, we encounter annotation words that are binary and act more similarly to class labels. In such annotation data, the assumption of tr-mmLDA in modeling word frequency of annotation words might be overly complex. We therefore modified tr-mmLDA to work with binary annotation data and presented a third topic model called supervised LDA-binary (sLDA-bin). By eliminating the hidden topics in the caption modality and model the binary data as a multi-variate Bernoulli random variable, sLDA-bin is more suitable for the binary annotation data and is able to obtain reasonably comparable annotation performance on standard image datasets, where performance is measured using caption perplexity and a precision-recall metric.

For a multimedia annotation task, we presented an extension of tr-mmLDA and sLDA-bin from the 2-modality case for modeling image-caption or video-caption, to a 3-modality version which allows the models to use additional input modalities to make more accurate predictions of caption data. We showed experimental results using both the audio and video modalities to infer the missing captions. Our 3-modality models are able to reap benefits of the independence and correlation structures in all the input modalities to further improve the prediction performance over using each individual input modality alone.

Lastly, we have presented a novel video representation based on statistics of natural videos. We learn a set of spatio-temporal ICA basis with each video blocks, where each ICA filter, resembling localized Gabor wavelet moving in time, captures “independent motion” in each block. Since ICA features are known to be sparse, we fit a mixture of Laplacian model on a large collection of features and a pattern of co-activation of basis with the same speed or velocity emerges in each mixture. We adopted this video feature in all the experiments performed on video in this thesis.

For the future work, we would like to explore ways to use a sparse prior on the regression parameter matrix of IFTM and tr-mmLDA models. We also plan to incorporate more relevant temporal statistics of audio and video data in the models itself. Currently, we are investigating computationally tractable time-series extensions of tr-mmLDA and sLDA-bin.

Bibliography

[AAD+02] B. Adams, A. Amir, C. Dorai, S. Ghosal, G. Iyengar, A. Jaimes, C. Lang, C.-Y. Lin, A. P. Natsev, M. R. Naphade, C. Neti, H. J. Nock, H. H. Permuter, R. Singh, J. R. Smith, S. Srinivasan, B. L. Tseng, T. V. Ashwin, and D. Zhang. Ibm research trec-2002 video retrieval system. In *Proceedings of the 11th Text Retrieval Conference*, 2002.

[AB85] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Optical Society of America*, 2(2):284–299, 1985.

[ABC+03] A. Amir, M. Berg, S. Chang, W. Hsu, G. Iyengar, C.-Y. Lin, M. R. Naphade, A. P. Natsev, C. Neti, H. J. Nock, J. R. Smith, B. L. Tseng, Y. Wu, and D. Zhang. Ibm research trec-2003 video retrieval system. In *Proceedings of the TRECVID workshop*, 2003.

[ABFX07] E. Airoldi, D. M. Blei, S. Fienberg, and E. P. Xing. Combining stochastic block model and mixed membership for statistical network analysis. *Statistical Network Analysis: Models, Issues, and New Directions*, (4503):57–74, 2007.

[AHV03] J. Hurri A. Hyvarinen and J. Vayrynen. Bubbles: a unifying framework for low-level statistical properties of natural image sequences. *J. Optical Society of America*, pages 1237–1252, 2003.

[Att00] H. Attias. A variational bayesian framework for graphical models. In *Advances in Neural Information Processing Systems (NIPS)*, volume 12. Advances in Neural Information Processing Systems, 2000.

[AXCM09] A. Ahmed, E. P. Xing, W. W. Cohen, and R. Murphy. Structured correspondence topic models for mining captioned figures in biological literature. In *ACM SIG-KDD Conference Knowledge Discovery and Data Mining*, 2009.

[BDdF+03] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. M. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[BF01] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *International Conference on Computer Vision (ICCV)*, 2001.

[BGJT04] D. M. Blei, T. Griffiths, M. I. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

[BJ03] D. M. Blei and M. I. Jordan. Modeling annotated data. In *ACM SIGIR*, 2003.

[BL06a] D. M. Blei and J. D. Lafferty. Correlated topic models. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

[BL06b] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *International Conference on Machine Learning (ICML)*, 2006.

[BL07] D. M. Blei and J. D. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.

[BM07] D. M. Blei and J. D. McAuliffe. Supervised topic models. In *Neural Information Processing Systems (NIPS)*, 2007.

[BNJ02] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. In *Neural Information Processing Systems (NIPS)*, 2002.

[BNJ03] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[BS97] A. J. Bell and T. J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37:3327–3338, 1997.

[CCMV07] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29, 2007.

[CFf07] L. Cao and L. Fei-fei. Spatially coherent latent topic model for concurrent object segmentation and classification. In *International Conference on Computer Vision (ICCV)*, 2007.

[CMS07] S. Chang, W. Y. Ma, and A. W. M. Smeulders. Recent advances and challenges of semantic image/video search. In *ICASSP*, 2007.

[CV05] G. Carneiro and N. Vasconcelos. Formulating semantic image annotation as a supervised learning problem. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[DBdFF02] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon of a fixed image vocabulary. In *European Conference on Computer Vision*, 2002.

[DW03] P. Duygulu and H. D. Wactlar. Associating video frames with text. In *Multimedia Information Retrieval Workshop in ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, 2003.

[Eve84] B. S. Everitt. *An Introduction to Latent Variable Models*. Chapman and Hall, London, 1984.

[EXCS06] S. Ebadollahi, L. Xie, S.-F. Chang, and J. R. Smith. Visual event detection using multi-dimensional concept dynamics. In *IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2006.

[FfP05] L. Fei-fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[FML04] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[FO96] D. J. Field and B. A. Olshausen. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, (381):607–9, 1996.

[Gir01] M. Girolami. A variational method for learning sparse and overcomplete representations. *Neural Computation*, 2001.

[GK04] M. Girolami and A. Kaban. Simplicial mixtures of markov chains: Distributed modelling of dynamic user profiles. In *Neural Information Processing Systems (NIPS)*, 2004.

[GS04] T. Griffiths and M. Steyvers. Finding scientific topics. In *Proceedings of the National Academy of Sciences*, volume 101, pages 5228–5235, 2004.

[HH00] A. Hyvarinen and P. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12:1705–20, 2000.

[HH01] A. Hyvarinen and P. O. Hoyer. Topographic independent component analysis as a model of v1 receptive fields. *Neurocomputing*, pages 38–40, 2001.

[Hyv99] A. Hyvarinen. Sparse code shrinkage: denoising of nongaussian data by maximum likelihood estimation. *Neural Computation*, 11:1739–68, 1999.

[Jaa97] T. S. Jaakkola. *Variational Methods for Inference and Estimation in Graphical Models*. PhD thesis, MIT, 1997.

[JCL07] W. Jiang, S.-F. Chang, and A. C. Loui. Context-based concept fusion with boosted conditional random fields. In *IEEE ICASSP*, 2007.

[JJ97] T. S. Jaakkola and M. I. Jordan. Bayesian logistic regression: A variational approach. In *Proceedings of the 1997 Conference on Artificial Intelligence and Statistics*, pages 283–294, 1997.

[JJ00] T. S. Jaakkola and M. I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 2000.

[JLM03] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, 2003.

[Jor67] K. G. Joreskog. Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32(4), 1967.

[KL03] Y. Karklin and M. S. Lewicki. Learning higher-order structures in natural images. *Network: Computational Neural Systems*, 14:483–499, 2003.

[LMJ03] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Advances in Neural Information Processing Systems (NIPS)*, 2003.

[Low04] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[LW03] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10), 2003.

[Mar04] B. Marlin. Collaborative filtering: A machine learning perspective. Master's thesis, University of Toronto, 2004.

[NMTM00] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classificaton from labeled and unlabeled documents using em. *Machine Learning*, 39(2):103–134, 2000.

[NNLS04] M. R. Naphade, A. P. Natsev, C.-Y. Lin, and J. R. Smith. Multi-granular detection of regional semantic concepts. In *IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2004.

[NS04] M. R. Naphade and J. R. Smith. On the detection of semantic concepts at trecvid. In *ACM Int. Conf. on Multimedia*, 2004.

[PAN09] D. Putthividhya, H. Attias, and S. Nagarajan. Independent factor topic models. In *International Conference on Machine Learning (ICML)*, 2009.

[PAN10a] D. Putthividhya, H. Attias, and S. Nagarajan. Supervised topic model for automatic image annotation. In *ICASSP*, 2010.

[PAN10b] D. Putthividhya, H. T. Attias, and S. S. Nagarajan. Topic regression multi-modal latent dirichlet allocation for image and video annotation. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshop (CVPR)*, 2010.

[PANL07] D. Putthividhya, H. Attias, S. Nagarajan, and T.-W. Lee. Probabilistic graphical model for auto-annotation, content-based retrieval, and classification of tv clips containing audio, video, and text. In *ICASSP*, 2007.

[PL04] H. J. Park and T.-W. Lee. Modeling nonlinear dependencies in natural images using mixture of laplacian distribution. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

[PL06] D. Putthividhya and T.-W. Lee. Motion patterns: High-level representation of natural video sequences. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[QMO$^+$05] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool. Modeling scenes with local descriptors and latent aspects. In *International Conference on Computer Vision (ICCV)*, 2005.

[Roc70] R. Rockafellar. *Convex Analysis.* Princeton University Press, 1970.

[RTMF08] B. 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.

[RZGSS04] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *20th Conference on Uncertainly in Artificial Intelligence (UAI)*, 2004.

[SB95] T. J. Sejnowski and A. J. Bell. An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(1129-59), 1995.

[Sim97] E. P. Simoncelli. Statistical models for images: Compression, restoration and synthesis. *31st Asilomar Conf on Signals, Systems, and Computers*, pages 673–678, Nov. 1997.

[Sla02a] M. Slaney. Mixtures of probability experts for audio retrieval and indexing. In *IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2002.

[Sla02b] M. Slaney. Semantic audio retrieval. In *IEEE ICASSP*, 2002.

[SRE$^+$05] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering object categories in image collections. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[SSU03] A. I. Schein, L. K. Saul, and L. H. Ungar. Generalized linear model for principal component analysis of binary data. In *AISTATS*, 2003.

[SWS$^+$00] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[SX08] S. Shringarpure and E. P. Xing. mstruct: A new admixture model for inference of population structure in light of both genetic admixing and allele mutataions. In *International Conference on Machine Learning (ICML)*, 2008.

[TBL06] D. Turnbull, L. Barrington, and G. Lanckriet. Modeling music and words using a multi-class naive bayes approach. In *Int. Conf. on Music Information Retrieval (ISMIR)*, 2006.

[TSK01] B. Taskar, E. Segal, and D. Koller. Probabilistic classication and clustering in relational data. In *Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2001.

[vA06] L. von Ahn. Games with a purpose. *IEEE Computer*, 39:92–94, 2006.

[vAD04] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *ACM Int. Conf. on Human Factors in Computing Systems*, 2004.

[vdWS06] J. van de Weijer and C. Schmid. Coloring local feature extraction. In *ECCV*, 2006.

[VH06] A. Velivelli and T. S. Huang. Automatic video annotation by mining speech transcripts. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshop (CVPR)*, 2006.

[vHR98] J. H. van Hateren and D. L. Ruderman. Independent component analysis of natural image sequence yields spatio-temporal filters similar to simple cells in primary visual cortex. In *Proc. Royal Society of London*, pages 2315–2320, 1998.

[VJ01] P. Viola and M. Jones. Robust real-time objection detection. *International Journal of Computer Vision*, 2001.

[VT07] J. Verbeek and B. Triggs. Region classification with markov field aspect models. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[WBFf09] C. Wang, D. M. Blei, and L. Fei-fei. Simultaneous image classification and annotation. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[WE04] B. Whitman and D. Ellis. Automatic record reviews. In *Int. Conf. on Music Information Retrieval (ISMIR)*, 2004.

[WG07] X. Wang and E. Grimson. Spatial latent dirichlet allocation. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[Whi05] B. Whitman. *Learning the meaning of music.* PhD thesis, Massachusetts Institute of Technology, 2005.

[WR02] B. Whitman and R. Rifkin. Musical query-by-description as a multiclass learning problme. In *IEEE Multimedia Signal Processing Conference*, 2002.

[YCH06] R. Yan, M.-Y. Chen, and A. Hauptmann. Mining relationship between video concepts using probabilistic graphical models. In *IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2006.

[YLXH07] J. Yang, Y. Liu, E. P. Xing, and A. G. Hauptmann. Harmonium models for semantic video representation and classification. In *SIAM International Conference on Data Mining*, 2007.

[ZZL$^+$06] R. Zhang, Z. Zhang, M. Li, W.-F. Ma, and H.-J. Zhang. A probabilistic semantic model for image annotation and multi-modal image retrieval. *Journal of Multimedia Systems*, 12(1):27–33, 2006.