Independent Factor Topic Models

Duangmanee (Pew) Putthividhya PUTTHI@UCSD.EDU
University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92037

Hagai T. Attias HTATTIAS@GOLDENMETALLIC.COM
Golden Metallic Inc., P.O. Box 475608, San Francisco, CA 94147

Srikantan Nagarajan SRI@RADIOLOGY.UCSF.EDU
University of California, San Francisco, 513 Parnassus Ave., San Francisco, CA 94143

Abstract

Topic models such as Latent Dirichlet Allocation (LDA) and Correlated Topic Model (CTM) have recently emerged as powerful statistical tools for text document modeling. In this paper, we improve upon CTM and propose Independent Factor Topic Models (IFTM) which use linear latent variable models to uncover the hidden sources of correlation between topics. There are 2 main contributions of this work. First, by using a sparse source prior model, we can directly visualize sparse patterns of topic correlations. Secondly, the conditional independence assumption implied in the use of latent source variables allows the objective function to factorize, leading to a fast Newton-Raphson based variational inference algorithm. Experimental results on synthetic and real data show that IFTM runs on average 3-5 times faster than CTM, while giving competitive performance as measured by perplexity and log-likelihood of held-out data.

1. Introduction

Large-scale document collections have become increasingly available online in the era of pervasive internet. Popular online message boards, blogs, emails, news articles gathered over a short span of time comprise several millions of entries. The magnitude of such document archives alone makes a compelling case for the need for automated tools in exploring and organizing such collections. In recent years, statistical topic models (Blei et al., 2003; Blei & Lafferty, 2006) 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 discover underlying patterns of word co-occurrences that form interpretable *topics*. More sophisticated models, e.g. (Blei et al., 2004), even allow 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) (Blei et al., 2003) is the most basic and widely-used model in the family of statistical topic models. Given a collection of documents, LDA decomposes the distribution of word counts from each document into contributions from K topics. Under LDA, a document is modeled as a draw from a Dirichlet distribution, while each topic, in turn, is modeled as a multinomial distribution over words in the vocabulary. Despite 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, an efficient variational inference algorithm (Blei et al., 2003) and an efficient Rao-blackwellized Gibbs sampling for LDA (Griffiths & Steyvers, 2004) 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 (Blei & Lafferty, 2006) to address such a limitation. By replacing Dirichlet with a powerful logistic normal

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009. Copyright 2009 by the author(s)/owner(s).

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 paper, 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 paper. In the first scenario, we present IFTM with Gaussian source prior, where the independent sources are 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 special case of CTM with a constrained structure of the covariance matrix. Indeed, assuming we have L sources, where generally $L \ll K$ (the number of topics), we 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 overfitting. 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 (Girolami, 2001) 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.

The paper is organized as follows. In section 2, we describe 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 an archive of NSF abstracts. We show that IFTM performs competitively with CTM, as measured using perplexity and the log-likelihood over held-out dataset. 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. Independent Factor Topic Models

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 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 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, 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 and (2) $p(\mathbf{s})$ distributed as a Laplacian pdf in Section 2.3.

To generate a document with N word occurrences: $\{w_1, \ldots, w_N\}$, we follow the generative process of IFTM as depicted in Figure 1(c):

- Draw contribution of independent sources: $\mathbf{s} \sim p(\mathbf{s})$.
- Draw correlated topic proportion $\mathbf{x}$ from the conditional distribution $p(\mathbf{x}|\mathbf{s})$:

$$\mathbf{x} \sim \mathcal{N}(\mathbf{x}; \mathbf{A}\mathbf{s} + \mu, \mathbf{\Lambda}^{-1}).$$

- For each word $n \in \{1, 2, \ldots, N\}$,
  1. Draw a topic $z_n \sim \mathcal{M}(\theta)$, where $\theta_k = \frac{e^{x_k}}{\sum_l e^{x_l}}$.
  2. Draw a word $w_n \sim \mathcal{M}(\beta_{z_n})$.

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 (Everitt, 1984). 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 $\mathbf{C} = \mathbf{A}\mathbf{A}^\top + \mathbf{\Lambda}^{-1}$. IFTM with Gaussian source prior can thus be seen as a special case of CTM with the constrained covariance matrix parameterized by $\mathbf{A}, \mathbf{\Lambda}$. Assuming $\mathbf{s} \in \mathbb{R}^L$, where $L \ll K$, we are modeling the covariance structure with $K + KL - \frac{L(L-1)}{2}$ free parameters instead of the $\frac{K(K+1)}{2}$ free parameters in the full covariance case.

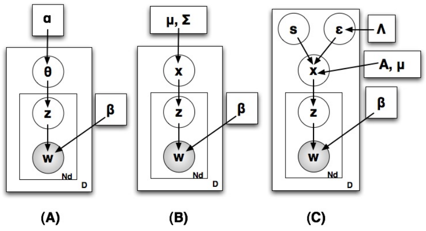

Figure 1. Graphical Model Representation comparing (a) LDA (b) CTM (c) our proposed model IFTM. The shaded nodes represent the observed variables.

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 (Joreskog, 1967). Nonetheless, such an approach cannot be applied to the non-Gaussian source prior case in Section 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 from EM for factor analysis can be obtained.

2.2.1. Variational Inference

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

$$\log p(\mathbf{W}|\Psi) \geq \int q(\mathbf{Z}, \mathbf{x}, \mathbf{s})(\log p(\mathbf{W}, \mathbf{Z}, \mathbf{x}, \mathbf{s}|\Psi) - \log q(\mathbf{Z}, \mathbf{x}, \mathbf{s}))d\mathbf{Z}d\mathbf{x}d\mathbf{s}, \quad (1)$$

where equality in (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 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. Computing the exact joint posterior $p(\mathbf{Z}, \mathbf{x}, \mathbf{s}|W)$ 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 (Attias, 2000): $p(\mathbf{Z}, \mathbf{x}, \mathbf{s}|\mathbf{W}) \approx \prod_n q(z_n)q(\mathbf{x})q(\mathbf{s})$. 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 (1). With $q(\mathbf{Z}, \mathbf{x}, \mathbf{s})$ now in a factorized form, the RHS of (1) becomes a strict lowerbound of the data log-likelihood and can be expressed as:

$$\begin{aligned} \log p(\mathbf{W}|\Psi) \geq \sum_n E_q[\log p(w_n|z_n, \beta)] + E_q[\log p(z_n|\mathbf{x})] \\ + E_q[\log p(\mathbf{x}|\mathbf{s}, \mathbf{A}, \mathbf{\Lambda}, \mu)] + E_q[\log p(\mathbf{s})] \\ + \mathcal{H}[q(\mathbf{Z})] + \mathcal{H}[q(\mathbf{x})] + \mathcal{H}[q(\mathbf{s})] \quad = \quad \mathcal{F}. \qquad (2) \end{aligned}$$

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. In particular, the log normalization term is replaced with adjustable lower bounds parameterized by variational parameters ξ.

$$\log p(z_n = k|x) \geq x_k - \log \xi - \frac{1}{\xi}(\sum_l e^{x_l} - \xi). \qquad (3)$$

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 (3), 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: $q(x_k) \sim \mathcal{N}(x_k; \bar{x}_k, \gamma_k^{-1})$.
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) 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:

$$\xi = \sum_k e^{\bar{x}_k + \frac{0.5}{\gamma_k}}, \qquad (4)$$

$$\phi_{nk} \propto \prod_t \beta_{kt}^{1(w_n=t)} \cdot \exp(\bar{x}_k), \qquad (5)$$

$$\frac{\partial \mathcal{F}}{\partial \gamma_k^{-1}} = -\frac{N}{\xi} e^{\bar{x}_k} \cdot e^{\frac{\gamma_k^{-1}}{2}} - \lambda_k + \frac{1}{\gamma_k^{-1}} = 0. \qquad (6)$$

Since there's no analytical solution available for (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:

$$\frac{\partial \mathcal{F}}{\partial \bar{x}_k} = \frac{n_k}{\lambda_k} - \frac{N}{\xi \lambda_k} e^{\bar{x}_k} \cdot e^{\frac{0.5}{\gamma_k}} - \bar{x}_k + c_k = 0. \qquad (7)$$

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 (7) in the form:

$$t_k e^{t_k} = u_k, \qquad (8)$$

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 (8) (with proper initialization e.g. $t_k \approx \log u_k$) converges in a few iterations.

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

$$\bar{\mathbf{s}} = \mathbf{B}^{-1} \mathbf{A}^\top \mathbf{\Lambda} (\bar{\mathbf{x}} - \mu), \qquad (9)$$

where $\mathbf{B} = (\mathbf{A}^\top \mathbf{\Lambda} \mathbf{A} + \mathbf{I}_L)$ is the precision of $q(\mathbf{s})$. 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 using (4)-(9) until convergence.

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 the mixing matrix will be needed. As a result, the individual columns of $\mathbf{A}$ do not carry meaningful patterns of correlation, 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 $\mathbf{A}$. In this work, we propose to model the independent sources as a Laplacian distribution:

$$p(\mathbf{s}) = \prod_{l}^{L} \frac{1}{2} e^{-|s_l|}, \log p(\mathbf{s}) = -\sum_l |s_l| - L \log 2. \quad (10)$$

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 use many sources to explain correlations between topics, while encourages those which use only a few sources.

2.3.1. Variational Inference

Variational inference for IFTM with Laplacian source prior proceeds similarly to the Gaussian case. We propose 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 (10) will require further approximating $q(\mathbf{s})$. To this end, we adopt the convex variational approximation that replaces the Laplacian source distribution with an adjustable lower bound as used in (Girolami, 2001). 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 (Jaakkola, 1997; Girolami, 2001).

$$-\sum_l |s_l| \geq -\sum_l (\frac{|\eta_l|}{2} + \frac{s_l^2}{2|\eta_l|}) \quad (11)$$

The dual form representation of the log prior of Laplacian source distribution in (11) takes a quadratic form, which, in essence, expresses the Laplacian distribution in terms of adjustable lower-bounds in the Gaussian form. This dual representation allows the variational inference algorithm 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 (4)-(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.

$$\bar{\mathbf{s}} = \mathbf{B}^{-1} \mathbf{A}^\top \mathbf{\Lambda} (\bar{\mathbf{x}} - \mu) \quad (12)$$

$$\mathbf{B} = \mathbf{A}^\top \mathbf{\Lambda} \mathbf{A} + \mathrm{diag}(\frac{1}{|\eta|}) \quad (13)$$

$$|\eta_l| = \sqrt{E[s_l^2]} \quad (14)$$

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

2.3.2. 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) w.r.t. Ψ and obtain the following closed-form updates:

$$\mathbf{A} = \mathbf{R}_{xs} \cdot \mathbf{R}_{ss}^{-1} \quad (15)$$

$$\mu = \frac{1}{D} (\sum_d \bar{\mathbf{x}}_d - \mathbf{A} \sum_d \bar{\mathbf{s}}_d) \quad (16)$$

$$\mathbf{\Lambda}^{-1} = \mathrm{diag}(\frac{1}{D}(\mathbf{R}_{xx} - \mathbf{A}\mathbf{R}_{xs}^\top) + \frac{1}{D} \sum_d \mathbf{\Gamma}_d^{-1}) \quad (17)$$

$$\beta_{kt} = \frac{\sum_{d,n} \phi_{nk}^d 1(w_n^d = t)}{\sum_{t,d,n} \phi_{nk}^d 1(w_n^d = t)} \quad (18)$$

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

3. Experimental Results

3.1. Correlated Topics Visualization

To understand what each independent source represents, we look at the most likely topics to co-occur when the source is "active". In particular, for each source, we sort the corresponding column of the mixing matrix $\mathbf{A}$ first in ascending order to discover the most likely patterns of topic cooccurrence when $s > 0$. Another set of pattern emerges by sorting the same column of $\mathbf{A}$ in reverse order (for the case of $s < 0$). As 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 shows 3 independent sources—s_1, s_2, and s_3 learned from the data, denoted by the 3 circles in the figure. For each source, we show 3-4 most likely topics to co-occur when the source is active. Each topic is represented by 10 most likely words. As seen in Figure 2, s_1 groups the topics related to material science together. s_2 represents a group of topics from various areas of physics, while s_3 represents a grouping of topics under the subject of biology and chemistry. 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).

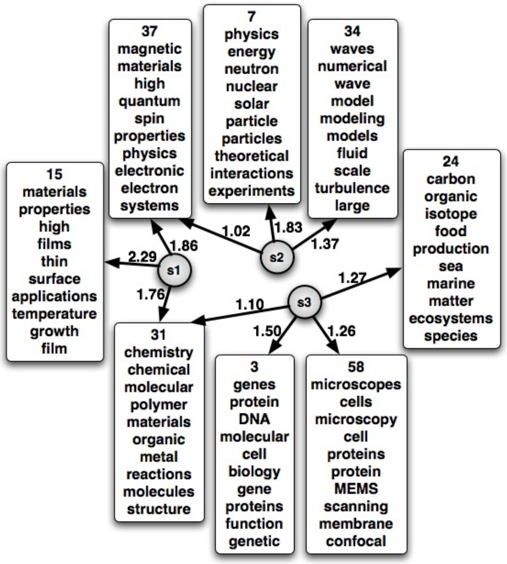

Figure 2. Visualization of 3 sources of topic correlations.

3.2. Model Comparison

3.2.1. 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, 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 used for IFTM-G and IFTM-L are 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 log-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 lead of (Blei & Lafferty, 2006), 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 3, 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, it is able to support larger numbers of topics.

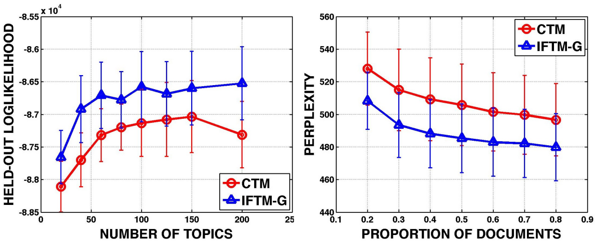

Figure 3. Results on Synthetic Data. Left: Held-out log-likelihood vs. the number of topics. Right: Per-word perplexity vs. the proportion of test document observed.

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 use a perplexity score given below to measure the performance of the 2 models. 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 should become more accurate as more words are observed.

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

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

Using the above described synthetic data, we obtain the results shown on the right panel of Figure 3. 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.

3.2.2. 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. We avoid comparing the different bounds used in the 4 models, and 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 $\log p(\mathbf{W})$.

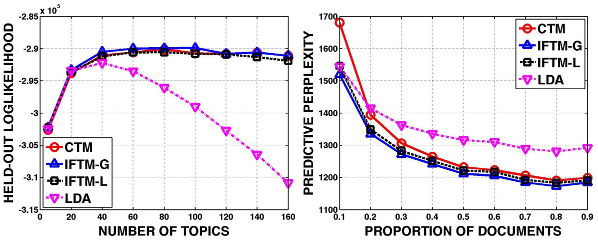

Figure 4. Results on NSF Abstracts Data. Left: Held-out log-likelihood score as a function of the number of topics, averaged over 4 folds with average standard deviation $= 1.13 \times 10^3$. Right: Predictive perplexity as a function of the percentage of observed words in test documents, comparing the 3 models when $K = 60$.

Figure 4 (left) shows the averaged log-likelihood of the held-out data as a function of the number of topics. 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. IFTM peaks at somewhere between 60-80 topics, and compared to LDA, IFTM always gives higher likelihood for different number of topics. The performance of CTM and IFTM on this dataset turns out to be very similar so far up to $K = 160$, since this dataset is much larger than the synthetic dataset. However, we do expect IFTM to outperform CTM, as K increases, again due to overfitting. Figure 4 (right) shows the predictive perplexity as a function of percentage of observed words in test documents, for $K = 60$. IFTM-G and IFTM-L give lower perplexity than CTM by almost 200 words, when only 10% of the words are observed. The difference between the 3 models 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 1 shows the averaged training time required to train the 4 models in Figure 4. All of our simulations run on an Intel$^{\text{TM}}$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 the new dependency between $\mathbf{B}$ and the convex variational parameter η in (13), which requires $\mathbf{B}$ to be inverted every time η and $\bar{\mathbf{s}}$ are updated.

3.3. Document Retrieval

In this section, we demonstrate that the ability of IFTM to capture topic correlations can be beneficial in a real-world application. To this end, we are interested in comparing IFTM to LDA in a document retrieval task. We use 20 Newsgroup dataset[2], containing 19,949 messages from 20 Usenet newsgroups (each class contains ≈ 1000 documents). This version of the dataset has been pre-processed: function words and rare/common words have been removed, and we are left with 43,586 words in the vocabulary. Using 70 topics, we train 3 models: IFTM with Gaussian source prior, IFTM with Laplacian sources, and LDA with 16,000 training documents, while 3,949 documents are 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 most similar to

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

the query. To use LDA and IFTM for a retrieval task, we use the following method. First, we train the models on 16,000-document training set. 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 any given 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 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 words in the query documents, to explain the training data.

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 5 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.

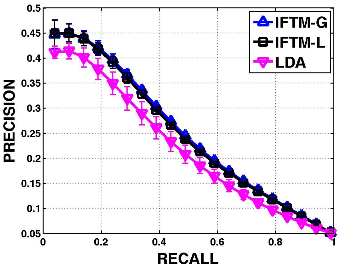

Figure 5. Precision-recall curve on 20 NewsGroup dataset.

4. Conclusions

In this paper, we describe Independent Factor Topic Models (IFTM), which present an alternative to CTM in modeling correlations between topics. IFTM proposes the use of a latent variable framework to model the sources of topic correlation directly. Such a framework for capturing correlation offers great flexibility in exploring different source prior models. In this work, we show the results from using Gaussian sources and Laplacian sources. When the sources are modeled as Gaussian distribution, we found that IFTM can be thought of as CTM with a constrained covariance structure. When the sparse source prior, e.g. Laplacian, is used, we can visualize and give interpretation to the sources of topic correlation by examining the corresponding columns of the mixing matrix **A**. The introduction of the latent sources in our formulation leads to a fast variational inference algorithm for IFTM. Our results show that IFTM is, on average, 3-5 times less computationally demanding, while still performing very competitively with CTM. In the future work, we are interested in finding out a way to determine the number of hidden sources automatically. One direction we are pursuing is the introduction of a prior over the mixing matrix **A**.

Table 1. Training Time (in hours) as K increases.

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**

Acknowledgement

The authors thank Shinko Cheng, Jiucang Hao, and Prof. Charles Elkan for many helpful discussions.

References

Attias, H. (2000). A variational bayesian framework for graphical models. *Advances in Neural Information Processing Systems (NIPS)* (pp. 209–215).

Blei, D. M., Griffiths, T., Jordan, M. I., & Tenenbaum, J. (2004). Hierarchical topic models and the nested chinese restaurant process. *Advances in Neural Information Processing Systems (NIPS)* (pp. 17–24).

Blei, D. M., & Lafferty, J. D. (2006). Correlated topic models. *Advances in Neural Information Processing Systems (NIPS)* (pp. 147–154).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Everitt, B. S. (1984). *An introduction to latent variable models*. London: Chapman and Hall.

Girolami, M. (2001). A variational method for learning sparse and overcomplete representations. *Neural Computation*, *13*, 2517–2532.

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences* (pp. 5228–5235).

Jaakkola, T. S. (1997). *Variational methods for inference and estimation in graphical models*. Doctoral dissertation, MIT.

Joreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, *32*, 443–482.