Discriminative Topic Modeling Based on Manifold Learning

SEUNGIL HUH and STEPHEN E. FIENBERG, Carnegie Mellon University

Topic modeling has become a popular method used for data analysis in various domains including text documents. Previous topic model approaches, such as probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA), have shown impressive success in discovering low-rank hidden structures for modeling text documents. These approaches, however do not take into account the manifold structure of the data, which is generally informative for nonlinear dimensionality reduction mapping. More recent topic model approaches, Laplacian PLSI (LapPLSI) and Locally-consistent Topic Model (LTM), have incorporated the local manifold structure into topic models and have shown resulting benefits. But they fall short of achieving full discriminating power of manifold learning as they only enhance the proximity between the low-rank representations of neighboring pairs without any consideration for non-neighboring pairs. In this article, we propose a new approach, Discriminative Topic Model (DTM), which separates non-neighboring pairs from each other in addition to bringing neighboring pairs closer together, thereby preserving the global manifold structure as well as improving local consistency. We also present a novel model-fitting algorithm based on the generalized EM algorithm and the concept of Pareto improvement. We empirically demonstrate the success of DTM in terms of unsupervised clustering and semisupervised classification accuracies on text corpora and robustness to parameters compared to state-of-the-art techniques.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Application—*Data mining*; I.7.0 [**Document and Text Processing**]: General

General Terms: Algorithms

Additional Key Words and Phrases: Topic modeling, dimensionality reduction, document clustering and classification, semisupervised learning

ACM Reference Format:
Huh, S. and Fienberg, S. E. 2012. Discriminative topic modeling based on manifold learning. ACM Trans. Knowl. Discov. Data 5, 4, Article 20 (February 2012), 25 pages.
DOI = 10.1145/2086737.2086740 http://doi.acm.org/10.1145/2086737.2086740

1. INTRODUCTION

Topic models are based on the notion that each data component (e.g., a document) can be represented by a mixture of basic components (or *topics*). In text analysis, topic models typically adopt the bag-of-words assumption, which ignores information regarding the ordering of words. Each document in a given corpus thus has a representation in the form of a histogram containing the occurrence of words. The form of this histogram comes from a distribution over a certain number of topics, each of which is a distribution over words in the vocabulary. By learning the distributions, we can create a corresponding low-rank representation of the high-dimensional histogram for each document. Topic models, such as probabilistic Latent Semantic Analysis (pLSA) [Hofmann 1999] and Latent Dirichlet Allocation (LDA) [Blei et al. 2003] have shown

Seungil Huh is partially supported by a Samsung Fellowship.

Authors' address: S. Huh and S. E. Fienberg, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15232; email: seungilh@cs.cmu.edu, fienberg@stat.cmu.edu.

DOI 10.1145/2086737.2086740 http://doi.acm.org/10.1145/2086737.2086740

impressive empirical success by improving classification accuracy through the discovery of low-rank hidden structures. In addition, these models provide probabilistic interpretations of the generative process of data.

Recently, several topic models, namely, Laplacian Probabilistic Latent Semantic Indexing (LapPLSI) [Cai et al. 2008] and Locally-consistent Topic Modeling (LTM) [Cai et al. 2009], were developed by additionally considering the manifold structure of data. Since data from texts or images are often found to be placed on a low-rank nonlinear manifold within the high-dimensional space of the original data, learning the manifold structure can provide better dimensionality reduction mapping and visualization [Belkin and Niyogi 2001; Roweis and Saul 2000; Tenenbaum et al. 2000]. Both the topic models increase the proximity between the probability distributions of the data pairs with favorable relationships (i.e., within-class pairs or neighbors in manifolds) by adding proximity as a regularization term to the log-likelihood function of pLSA. As a result, these models obtain probabilistic distributions concentrated around the manifold and show higher accuracy than pLSA and LDA for text clustering and classification tasks. However, LapPLSI and LTM fall short of achieving the full discriminating power of manifold learning because the global manifold structure is often not well preserved by only enhancing the proximity between favorable pairs. To achieve the full benefit, they would also need to consider unfavorable relationships (i.e., between-class pairs or non-neighbors in manifolds) between data pairs.

In this work, we propose a new topic model to focus more on discriminating power, which we refer to as Discriminative Topic Model (DTM). In order to address clustering or classification problems in an unsupervised or semisupervised setting, i.e., using no, or a small amount, of labeled data with a large amount of unlabeled data, DTM maintains the local consistency of data by increasing the proximity between the probability distributions of the data pairs with favorable relationships, as do LapPLSI and LTM. In addition, in contrast to the previous models, DTM explicitly aims to increase the separability between those of the data pairs with unfavorable relationships. Due to the effectiveness of this more complete manifold learning formulation, DTM also preserves the global manifold structure, showing better performance in document clustering and classification tasks than the previous approaches. We also present an efficient algorithm to solve the proposed regularized log-likelihood maximization problem based on a generalized Expectation-Maximization algorithm [Dempster et al. 1977] and the concept of Pareto improvement [Barr 2004]. Our model-fitting algorithm does not require a regularization parameter to which the clustering and classification performance can be sensitive. We offer empirical evidence on three real-world text corpora (20 newsgroups, Yahoo! News K-series, and Reuters-21578) and demonstrate the superiority of DTM to state-of-the-art techniques.

It is worth mentioning that this work is an enlarged version of the paper with the same title, which appeared in the proceeding of ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) [Huh and Fienberg 2010]. In this extended version: (1) the formulation of DTM is generalized to adopt any user-defined unfavorable relationships; (2) the reestimation equations for the parameter set representing topic distributions are presented in complete matrix form; (3) an effective binary search replaces a typical line search when a Pareto optimum is explored; (4) a dynamic weighting scheme on unfavorable relationships is newly proposed; and (5) the clustering performance is reported in addition to the classification performance.

The remainder of this article is organized as follows. Section 2 provides the background and notation. Section 3 gives an overview of the previous work. In Section 4, we formulate DTM and describe how to fit the proposed model. We present the experimental setup and discuss the experimental results in Sections 5 and 6, respectively, followed by conclusions in Section 7.

2. BACKGROUND AND NOTATION

We begin by describing the two basic components of our method: probabilistic Latent Semantic Analysis (pLSA) [Hofmann 1999] as a topic model and Laplacian Eigenmaps [Belkin and Niyogi 2001] or graph embedding [Yan et al. 2007] as a manifold learning algorithm.

2.1 Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (pLSA) [Hofmann 1999] evolved from Latent Semantic Indexing (LSA) [Deerwester et al. 1990], and defines a proper generative model based on a solid statistical foundation.

Suppose that we have a corpus that consists of N documents $\{d_1, d_2, \cdots, d_N\}$ with words from a vocabulary containing M words $\{w_1, w_2, \cdots, w_M\}$. In pLSA, we associate the occurrence of a word w in a particular document d with one of K unobserved topic variables $\{z_1, z_2, \cdots, z_K\}$. More formally, we can define pLSA by the following generative process:

— select a document d with probability $P(d)$;
— pick a latent class z with probability $P(z|d)$;
— generate a word w with probability $P(w|z)$.

By summing out the latent variable z, we can compute the joint probability of an observed pair (d, w) as

$$P(d, w) = P(d)P(w|d) = P(d) \sum_{k=1}^{K} P(w|z_k)P(z_k|d).$$

Based on this joint probability, we can calculate the log-likelihood as

$$\tilde{\mathcal{L}} = \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, w_j) \log \Big(P(d_i) \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \Big), \tag{1}$$

where $n(d, w)$ denotes the number of times word w occurred in document d. Following the likelihood principle, we can determine $P(w|z)$ and $P(z|d)$ by maximizing the relevant part of Eq. (1)

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, w_j) \log \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i). \tag{2}$$

2.2 Laplacian Eigenmaps and Graph Embedding

Traditional manifold learning algorithms [Belkin and Niyogi 2001; Hinton and Roweis 2002; Roweis and Saul 2000; Tenenbaum et al. 2000] have given way to graph-based semisupervised learning algorithms [Belkin et al. 2006; Zhou et al. 2003; Zhu et al. 2003]. The goal of manifold learning is to recover the structure of a given dataset by a non-linear mapping into a low-dimensional space. One such manifold learning algorithm, Laplacian Eigenmaps [Belkin and Niyogi 2001], utilizes spectral graph theory [Chung 1997].

Suppose that we have N data points $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_N\}$, each of which is an $M \times 1$ vector. From the nearest neighbor graph of these data points, a local similarity matrix W is defined to contain favorable pair-wise relationships among them, e.g.,

$$W_{ij} = \begin{cases} 1, & \text{if } \mathbf{u}_i \in \mathcal{N}_r(\mathbf{u}_j) \text{ or } \mathbf{u}_j \in \mathcal{N}_r(\mathbf{u}_i) \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

where $\mathcal{N}_r(\mathbf{u})$ is the set of the r nearest neighbors of $\mathbf{u}$.

Let $\mathbf{x}_i$, which is a $K \times 1$ vector, be a low-rank representation of $\mathbf{u}_i$ on the manifold ($K \ll M$). Intuitively, if two data points $\mathbf{u}_i$ and $\mathbf{u}_j$ are close to each other in the original space, the corresponding low-rank representations $\mathbf{x}_i$ and $\mathbf{x}_j$ should also lie near each other. From this intuition, Laplacian Eigenmaps solve the following optimization problem.

$$\min \sum_{i,j=1}^{N} W_{ij} ||\mathbf{x}_i - \mathbf{x}_j||^2 \qquad s.t. \qquad XDX^T = I, \tag{4}$$

where X is the matrix the ith column of which is $\mathbf{x}_i$, and D is a diagonal matrix such that $D_{ii} = \sum_{j=1}^{N} W_{ij}$.

The constraint in Eq. (4) plays a role in avoiding all $\mathbf{x}$ to converge to 0 as well as leading to a generalized eigenvalue problem. (Note that without the constraint, the objective function decreases only by scale reduction.) To avoid a trivial solution, which is $\mathbf{x}_1 = \mathbf{x}_2 = \cdots = \mathbf{x}_N = \mathbf{c}$, where $\mathbf{c}$ is a nonzero constant vector, the eigenvectors corresponding to a zero eigenvalue are ignored and the ones corresponding to the smallest nonzero eigenvalues are selected among the eigenvalue solutions [Belkin and Niyogi 2001].

The idea of Laplacian Eigenmaps can be generalized with the idea of graph embedding [Yan et al. 2007], a general framework for dimensionality reduction, as follows.

$$\min \sum_{i,j=1}^{N} W_{ij} ||\mathbf{x}_i - \mathbf{x}_j||^2 \qquad s.t. \qquad XBX^T = I, \tag{5}$$

where B is either a diagonal matrix for scale normalization (as in Laplacian Eigenmaps) or the Laplacian matrix of a graph representing unfavorable relationships between data points.

From these notations, we can conclude that to control favorable relationships is not sufficient for learning manifold structure.

3. PREVIOUS WORK

Cai et al. [2008] recently proposed two topic models, Laplacian pLSI (LapPLSI) and Locally-consistent Topic Modeling (LTM) [Cai et al. 2009], which use manifold structure information based on pLSA. These models are formalized by regularizing the original log-likelihood of pLSA with the proximity between low-dimensional probability distributions of data pairs that are likely to be closely located on the manifold.

LapPLSI adopts the objective of Laplacian Eigenmaps for the measure

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, w_j) \log \sum_{k=1}^{K} P(w_j|z_k) P(z_k|d_i) - \frac{\lambda}{2} \sum_{k=1}^{K} \sum_{i,j=1}^{N} W_{ij} \big(P(z_k|d_i) - P(z_k|d_j)\big)^2, \tag{6}$$

where λ is the regularization parameter and W is an $N \times N$ matrix measuring the local similarity of document pairs based on word occurrences.

LTM uses the Kullback-Leibler divergence (KL divergence) instead of the squared Euclidean distance

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j)\log\sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) - \frac{\lambda}{2}\sum_{i,j=1}^{N} W_{ij}\Big(D_{KL}\big(P(z|d_i)||P(z|d_j)\big) + D_{KL}\big(P(z|d_j)||P(z|d_i)\big)\Big), \tag{7}$$

where the KL divergence of probability distributions $P(z|d_i)$ and $P(z|d_j)$ is

$$D_{KL}\big(P(z|d_i)||P(z|d_j)\big) = \sum_k P(z_k|d_i) log\frac{P(z_k|d_i)}{P(z_k|d_j)}. \tag{8}$$

By discovering the local neighborhood structure, these two models show higher discriminating power than pLSA and LDA on document clustering and classification tasks. Both models, however, fall short of the full discriminating power of manifold learning because the global manifold structure is often not well preserved by only enhancing the proximity between favorable pairs without maintaining or increasing the separability between unfavorable pairs. In addition, these models are limited in that their performance depends on the choice of the regularization parameter λ. It is unclear how to appropriately determine the value of λ, particularly when no labeled data is available.

4. DISCRIMINATIVE TOPIC MODEL

Here, we formalize our proposed model, named Discriminative Topic Model (DTM). We also present an algorithm to solve the proposed regularized log-likelihood maximization problem based on the generalized Expectation Maximization (EM) algorithm [Dempster et al. 1977] and the concept of Pareto improvement [Barr 2004].

4.1 Regularized Model

When increasing the local consistency in manifold learning, we also need to maintain or increase the separability of the low-rank representations of the data that are not likely to be placed close to each other. To do this, we define two matrices, W and $\overline{W}$: W measures the local similarity of document pairs based on word occurrences; on the other hand, $\overline{W}$ measures the local or global dissimilarity of document pairs. We will introduce the definitions of W and $\overline{W}$ adopted for our experiments in the following section. We constrain these two matrices to have only nonnegative elements.

The proximity of favorable pairs can then be expressed by the weighted sum of the squared Euclidean distances between the low-rank probability distributions

$$\sum_{i,j=1}^{N}\sum_{k=1}^{K} W_{ij}\big(P(z_k|d_i) - P(z_k|d_j)\big)^2. \tag{9}$$

Similarly, the separability of unfavorable pairs can be expressed as

$$\sum_{i,j=1}^{N}\sum_{k=1}^{K} \overline{W}_{ij}\big(P(z_k|d_i) - P(z_k|d_j)\big)^2. \tag{10}$$

To minimize the proximity while maintaining or maximizing the separability, we combine these two objectives as a fraction form and maximize the following objective function:

$$\frac{\sum_{i,j=1}^{N}\sum_{k=1}^{K}\overline{W}_{ij}\big(P(z_k|d_i)-P(z_k|d_j)\big)^2}{\sum_{i,j=1}^{N}\sum_{k=1}^{K}W_{ij}\big(P(z_k|d_i)-P(z_k|d_j)\big)^2}. \tag{11}$$

Our model is regularized with this term to learn the manifold structure of data, in addition to adopting the generative process of pLSA. The log-likelihood of our model is thus as follows:

$$\begin{aligned}\mathcal{L} = &\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j)\log\sum_{k=1}^{K}P(w_j|z_k)P(z_k|d_i)\\ &+\lambda\frac{\sum_{i,j=1}^{N}\sum_{k=1}^{K}\overline{W}_{ij}\big(P(z_k|d_i)-P(z_k|d_j)\big)^2}{\sum_{i,j=1}^{N}\sum_{k=1}^{K}W_{ij}\big(P(z_k|d_i)-P(z_k|d_j)\big)^2},\end{aligned} \tag{12}$$

where λ is a regularization parameter. Although our model includes the regularization parameter, we do not need to consider it directly; it is handled implicitly by the nature of our model-fitting algorithm, as we elaborate in the following subsection.

4.2 Model-Fitting

When a probabilistic model involves unobserved latent variables, the EM algorithm offers a general approach for the maximum likelihood estimation of the model. Here we use the generalized EM algorithm, which finds parameters that improve the expected value of the log-likelihood function in the M-step rather than maximizing it. For further details, see Dempster et al. [1977].

Let $\phi = [P(w_j|z_k)]$ and $\theta = [P(z_k|d_i)]$, which are parameters of DTM. Thus, we need to estimate $MK + KN$ parameters, the same as for pLSA.

E-step. The E-step of DTM is exactly the same as that of pLSA [Hofmann 1999]. By applying Bayes' formula, we compute posterior probabilities

$$P(z_k|d_i, w_j) = \frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{k'=1}^{K}P(w_j|z_{k'})P(z_{k'}|d_i)}. \tag{13}$$

M-step. In the M-step of DTM, we improve the expected value of the log-likelihood function which is

$$\begin{aligned}\mathcal{Q}(\phi,\theta) &= \mathcal{Q}_1(\phi,\theta)+\lambda\mathcal{Q}_2(\theta) \qquad (14)\\ &= \sum_{i=1}^{N}\sum_{j=1}^{M}n(d_i, w_j)\sum_{k=1}^{K}P(z_k|d_i, w_j)\log[P(w_j|z_k)P(z_k|d_i)]\\ &\qquad+\lambda\frac{\sum_{i,j=1}^{N}\sum_{k=1}^{K}\overline{W}_{ij}\big(P(z_k|d_i)-P(z_k|d_j)\big)^2}{\sum_{i,j=1}^{N}\sum_{k=1}^{K}W_{ij}\big(P(z_k|d_i)-P(z_k|d_j)\big)^2}.\end{aligned}$$

The M-step reestimation equation for ϕ is exactly the same as those for pLSA because the regularization term of DTM does not include $P(w_j|z_k)$; the M-step reestimation equation is as follows:

$$P(w_j|z_k) = \frac{\sum_{i=1}^{N}n(d_i, w_j)P(z_k|d_i, w_j)}{\sum_{j'=1}^{M}\sum_{i=1}^{N}n(d_i, w_{j'})P(z_k|d_i, w_{j'})} \tag{15}$$

Before describing the M-step reestimation algorithm for θ, we introduce the concept of Pareto improvement [Barr 2004], based on which, we propose our algorithm. Pareto improvement is the change from one status to another that can improve at least one objective without worsening any other objectives. More formally, in our problem, an update $\theta^{(t)} \rightarrow \theta^{(t+1)}$ is a Pareto improvement if either of the following two conditions holds.

(1) $\mathcal{Q}_1(\phi, \theta^{(t+1)}) > \mathcal{Q}_1(\phi, \theta^{(t)})$ and $\mathcal{Q}_2(\theta^{(t+1)}) \geq \mathcal{Q}_2(\theta^{(t)})$;
(2) $\mathcal{Q}_1(\phi, \theta^{(t+1)}) \geq \mathcal{Q}_1(\phi, \theta^{(t)})$ and $\mathcal{Q}_2(\theta^{(t+1)}) > \mathcal{Q}_2(\theta^{(t)})$.

Based on the concepts of generalized EM and Pareto improvement, we re-estimate θ by (1) increasing $\mathcal{Q}(\phi, \theta)$ rather than maximizing it, and (2) increasing at least one of $\mathcal{Q}_1(\phi, \theta)$ and $\mathcal{Q}_2(\theta)$ without decreasing the other.

Among many possible Pareto improvements, we choose the one that has the greatest improvement of $\mathcal{Q}_2(\theta)$ at each iteration, because $\mathcal{Q}_2(\theta)$, which is optimized to reveal manifold structure, is more critical to the discriminative power of the model. One advantage of this strategy is that $\mathcal{Q}(\phi, \theta)$ is improved regardless of the regularization parameter λ whose value affects the performance of previous models, and yet is hard to determine appropriately.

4.2.1 Reestimation of θ for each of $\mathcal{Q}_1(\phi, \theta)$ and $\mathcal{Q}_2(\theta)$. In order to present a reestimating algorithm for θ to increase $\mathcal{Q}(\phi, \theta)$ based on Pareto improvement, we first propose reestimation equations to increase each of $\mathcal{Q}_1(\phi, \theta)$ and $\mathcal{Q}_2(\theta)$ in the following theorems.

THEOREM 4.1. *If $\theta^{(t+1)}$ is computed from $\theta^{(t)}$ by applying the following reestimation equation*

$$P(z_k|d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j) P(z_k|d_i, w_j)}{\sum_{j=1}^{M} n(d_i, w_j)}, \tag{16}$$

then $\mathcal{Q}_1(\phi, \theta)$ monotonically increases when θ moves from $\theta^{(t)}$ to $\theta^{(t+1)}$ along the line with fixed ϕ.

PROOF. $\mathcal{Q}_1(\phi, \theta)$ is the expected value of the log-likelihood function of pLSA and Eq. (16) is the reestimation equation for $P(z_k|d_i)$ of pLSA; thus, $\theta^{(t+1)}$ maximizes $\mathcal{Q}_1(\phi, \theta)$ when ϕ is fixed. Since $\mathcal{Q}_1(\phi, \theta)$ is a concave function of θ and $\theta^{(t+1)}$ is the maximum solution of $\mathcal{Q}_1(\phi, \theta)$, $\mathcal{Q}_1(\phi, \theta)$ monotonically increases when θ moves from $\theta^{(t)}$ to $\theta^{(t+1)}$ along the line. □

THEOREM 4.2. *Let α be the estimated value of the regularization term under the current estimates of the parameters with nonnegative matrices W and $\overline{W}$*

$$\alpha = \frac{\sum_{i,j=1}^{N} \sum_{k=1}^{K} \overline{W}_{ij} \big(P(z_k|d_i) - P(z_k|d_j)\big)^2}{\sum_{i,j=1}^{N} \sum_{k=1}^{K} W_{ij} \big(P(z_k|d_i) - P(z_k|d_j)\big)^2}, \tag{17}$$

and we define β_{pi} for topic id p *and document id* i *as*

$$\beta_{pi} = \min\left(\frac{\overline{D}_{ii} P(z_p|d_i) + \alpha \sum_{j=1}^{N} W_{ij} P(z_p|d_j)}{\sum_{j=1}^{N} \overline{W}_{ij} P(z_p|d_j) + \alpha D_{ii} P(z_p|d_i)}, \frac{1}{P(z_p|d_i)} \right), \tag{18}$$

where D and $\overline{D}$ are diagonal matrices such that

$$D_{ii} = \sum_{j=1}^{N} W_{ij} \quad \textit{and} \quad \overline{D}_{ii} = \sum_{j=1}^{N} \overline{W}_{ij}. \tag{19}$$

Then, $\mathcal{Q}_2(\theta)$ is nondecreasing by the following reestimation equation for $[P(z_1|d_i), P(z_2|d_i), \cdots, P(z_K|d_i)]$:

$$P(z_k|d_i) = \begin{cases} \beta_{pi} P(z_p|d_i), & \textit{if } k = p \\ \frac{1-\beta_{pi} P(z_p|d_i)}{1-P(z_p|d_i)} P(z_k|d_i), & \textit{otherwise} \end{cases}. \tag{20}$$

PROOF. See Appendix A. □

In Eq. (18), the minimum operator is used to ensure that $[P(z_1|d_i), \cdots, P(z_K|d_i)]$ is a probability distribution after reestimation. It can be easily verified that $\sum_{k=1}^{K} P(z_k|d_i) = 1$ and $\forall k$, $P(z_k|d_i) \geq 0$ after the reestimation, when $P(z_p|d_i) \neq 0$. If $P(z_p|d_i) = 0$, we replace $P(z_p|d_i)$ with a tiny value to avoid a division by zero.

4.2.2 Reestimation Equation of θ for $\mathcal{Q}_2(\theta)$ in Matrix Form. The reestimation equation in Theorem 4.2 can be converted into a matrix form and thus the computation of the equations can be parallelized as follows. Let P be a matrix such that $P_{ki} = P(z_k|d_i)$.

First, we compute the numerator of α in Eq. (17) as

$$\begin{aligned} \sum_{i,j=1}^{N} \sum_{k=1}^{K} W_{ij} \big(P(z_k|d_i) - P(z_k|d_j)\big)^2 &= \sum_{i,j=1}^{N} \sum_{k=1}^{K} W_{ij} (P_{ki} - P_{kj})^2 \\ &= \sum_{i,j=1}^{N} \sum_{k=1}^{K} W_{ij} {P_{ki}}^2 - 2 \sum_{i,j=1}^{N} \sum_{k=1}^{K} W_{ij} P_{ki} P_{kj} + \sum_{i,j=1}^{N} \sum_{k=1}^{K} W_{ij} {P_{kj}}^2 \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} P_{ki} \big(\sum_{j=1}^{N} W_{ij}\big) P_{ki} - 2 \sum_{i,j=1}^{N} \sum_{k=1}^{K} P_{ki} W_{ij} P_{kj} + \sum_{j=1}^{N} \sum_{k=1}^{K} P_{kj} \big(\sum_{i=1}^{N} W_{ij}\big) P_{kj} \\ &= Tr(PDP^T) - 2Tr(PWP^T) + Tr(PDP^T) \\ &= 2Tr(P(D-W)P^T) = 2Tr(PLP^T), \end{aligned} \tag{21}$$

where $L = D - W$, which is the graph Laplacian of the similarity graph. In the same way, we compute the denominator of α in Eq. (17) as

$$\sum_{i,j=1}^{N} \sum_{k=1}^{K} \overline{W}_{ij} \big(P(z_k|d_i) - P(z_k|d_j)\big)^2 = \sum_{i,j=1}^{N} \sum_{k=1}^{K} \overline{W}_{ij} \big(P_{ki} - P_{kj}\big)^2 = 2Tr(P\overline{L}P^T), \tag{22}$$

where $\overline{L} = \overline{D} - \overline{W}$, which is the graph Laplacian of the dissimilarity graph.

Therefore, from Equations (21) and (22), we compute α in Eq. (17) as

$$\alpha = \frac{Tr(P\overline{L}P^T)}{Tr(PLP^T)}. \tag{23}$$

Second, we re-express β_{pi} for topic id p and document id i in Eq. (18) in matrix form

$$\beta_{pi} = \min\left(\frac{\big(P(\overline{D} + \alpha W)\big)_{pi}}{\big(P(\overline{W} + \alpha D)\big)_{pi}}, \frac{1}{P_{pi}}\right). \tag{24}$$

Considering all documents and all topic ids together, we define a matrix B whose (p, i) element is β_{pi}

$$B = \min\left(\frac{\left(P(\overline{D} + \alpha W)\right)}{\left(P(\overline{W} + \alpha D)\right)}, \frac{1_{K\times N}}{P}\right), \tag{25}$$

where $1_{K\times N}$ is the $K \times N$ matrix with all ones and the min operation and divisions are element-wise.

To convert the reestimation equation in Eq. (20) to a matrix form, we reformalize the equation. The reestimation equation is equivalent to the following two-step update.

Step 1. Update $P(z_p|d_i)$ with the following equation with β_{pi}.

$$P(z_p|d_i) = \beta_{pi}\frac{1 - P(z_p|d_i)}{1 - \beta_{pi}P(z_p|d_i)}P(z_p|d_i) \tag{26}$$

Step 2. Normalize $P(z_1|d_i), P(z_2|d_i), \cdots, P(z_K|d_i)$ to be summed to 1.

$$P_{ki} = \frac{P_{ki}}{\sum_{k'=1}^{K} P_{k'i}} \qquad for\ \forall k \tag{27}$$

Finally, applying this transformed update to all topics and documents yields a two-step matrix form reestimation for $\{P(z_k|d_i)\}$:

Step 1.

$$P = B \otimes \frac{1_{K\times N} - P}{1_{K\times N} - B \otimes P} \otimes P \tag{28}$$

Step 2.

$$P = \frac{P}{1_{K\times K}P}, \tag{29}$$

where $\otimes$ denotes element-wise multiplication and the divisions are also element-wise.

4.2.3 Reestimation Algorithm for θ. Based on Theorems 4.1 and 4.2, we propose a reestimating algorithm for θ. Let the current parameter set θ be θ_0. We first compute θ_1 by applying Eqs. (23) through (29) to θ_0. Theorem 4.2 guarantees that $\mathcal{Q}_2(\theta_1) \geq \mathcal{Q}_2(\theta_0)$. θ_2 is then computed from θ_1 by applying the pLSA M-step in Eq. (16). Theorem 4.1 ensures that $\mathcal{Q}_1(\phi, \theta)$ monotonically increases when θ moves from θ_1 to θ_2 along the line.

On the line segment between θ_1 and θ_2, we find a Pareto optimum between $\mathcal{Q}_1$ and $\mathcal{Q}_2$. Among the possible Pareto optima, we are interested in the Pareto optimum that maximizes $\mathcal{Q}_2$ since $\mathcal{Q}_2$ is decisive for the discriminating power of the model, and $\mathcal{Q}_1$ for the generative process. To find the Pareto optimum, we perform a binary search between θ_1 and θ_2 as follows. Let ρ be the center point between θ_1 and θ_2. If $\mathcal{Q}_1(\phi, \rho) < \mathcal{Q}_1(\phi, \theta_0)$, we exclude the interval between θ_1 and ρ from further consideration because Theorem 4.1 ensures that $\mathcal{Q}_1$ is not improved and thus there exists no Pareto optimum in the interval. Otherwise, we compute the directional derivative of $\mathcal{Q}_2$ at ρ along the direction $\theta_2 - \theta_1$

$$\nabla_{\theta_2-\theta_1}\mathcal{Q}_2(\theta)|_{\theta=\rho} = vec\left(\frac{Tr(PLP^T)P\overline{L} - Tr(P\overline{L}P^T)PL}{Tr(PLP^T)^2}\right)\cdot(\theta_2 - \theta_1)|_{P=mat(\rho)}, \tag{30}$$

where *vec* denotes the operation of vectorization with the same order of elements as θ_1 (or θ_2), and *mat* is its inverse operation.

If this derivative is positive, $\mathcal{Q}_2(\theta)$ increases as θ moves towards θ_2; thus, we exclude the interval between θ_1 and ρ from further consideration. In the same way, if the derivative is negative, we exclude the interval between ρ and θ_2. Then, by regarding the remaining interval as $[\theta_1, \theta_2]$, we perform the next iteration of the binary search until the derivative is equal to zero or the length of the remaining interval is sufficiently short. As a result, we obtain a local optimum of $\mathcal{Q}_2$ that also improves $\mathcal{Q}_1$ on the line segment between θ_1 and θ_2.

After obtaining the result, denoted by $\rho*$, we examine whether it yields a Pareto optimum between $\mathcal{Q}_1$ and $\mathcal{Q}_2$, i.e., $\big(\mathcal{Q}_1(\phi, \rho*) \geq \mathcal{Q}_1(\phi, \theta_0)$ and $\mathcal{Q}_2(\rho*) > \mathcal{Q}_2(\theta_0)\big)$ or $\big(\mathcal{Q}_1(\phi, \rho*) > \mathcal{Q}_1(\phi, \theta_0)$ and $\mathcal{Q}_2(\rho*) \geq \mathcal{Q}_2(\theta_0)\big)$. If this is true,, we reestimate θ with $\theta = \rho*$. Otherwise, we keep θ as θ_0 without updating in the M-step and continue to the next E-step.

In our model, we iteratively repeat the E-step and M-step until both the parameters ϕ and θ converge. This convergence is typically evaluated by examining whether the change of the parameters is less than a small threshold. In our experiments, our model required approximately 100 to 300 iterations to achieve convergence.

Algorithm 1 summarizes our model-fitting algorithm. As described in lines 6 through 8, we selectively give weights on the elements of $\overline{W}$ for every, or a certain number, of interations, which often improves the discriminative power of the model. The weights can be given based on the low-dimensional representation in order to prioritize the pairs not sufficiently separated from each other. The weighting scheme used in our experiments will be presented in the following section.

5. EXPERIMENTS

We tested DTM using three widely used text corpora (20 newsgroups, Yahoo! News K-series, and Reuters-21578) and compared it with other topic models and dimensionality reduction methods. In this section, we describe our experimental setup and implementation details.

5.1 Datasets and Experimental Setup

The 20 newsgroups corpus is a collection of approximately 20,000 newsgroup documents, partitioned almost evenly across 20 different newsgroups.[1] We downloaded the preprocessed version from R. F. Corrêa's Web page,[2] which includes 18821 documents with 8156 distinct words. Among the documents, we randomly selected 100 documents from each category for each test run; as a result, in total 2000 documents were used for test. Yahoo! News K-series is a collection of 2340 news articles belonging to one of 20 different categories, which includes documents of varying sizes ranging from 494 to 9 [Boley 1998]. We downloaded the preprocessed version including 8104 distinct words from D. L. Boley's Web page.[3] For every test run, we used all the 2340 documents. Reuters-21578 is a corpus of newswire stories made available by Reuters, Ltd., and corrected by D. D. Lewis.[4] The entire Reuters-21578 corpora consists of documents in 135 categories. Among the documents, we extracted the 10 largest categories with unique category labels, the sizes of which are unbalanced, ranging from 3923 to 112. We downloaded a preprocessed version of Reuters-21578 from R. F. Corrêa's Web page, which includes 5180 distinct words. For each test run, we randomly selected

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

[2] http://sites.google.com/site/renatocorrea02/textcategorizationdatasets/

[3] http://www-users.cs.umn.edu/~boley/ftp/PDDPdata/

[4] http://www.daviddlewis.com/resources/testcollections/reuters21578/

Algorithm 1 Model fitting for DTM

```
Input:
$\{n(d_i, w_j)\}$: a set of (weighted and normalized) word occurrences,
$N$: number of documents, $M$: size of vocabulary, $K$: number of topics,
$MaxIter$: the maximum number of iteration allowed,
$W$: similarity matrix, $\overline{W}$: dissimilarity matrix
Output:
$\phi = \{P(w_j|z_k)\}$ and $\theta = \{P(z_k|d_i)\}$.

Compute $D$ and $\overline{D}$ as in Eq. (19).
$L \leftarrow D - W$, $\overline{L} \leftarrow \overline{D} - \overline{W}$.
Randomly initialize $\phi$ and $\theta$.
$iter \leftarrow 0$.
repeat
    Optional step:
    Give weights on $\overline{W}$ based on $\theta$.
    Recompute $\overline{D}$ and $\overline{L}$.
    E-step:
    Compute $P(z_k|d_i, w_j)$ using $\phi$ and $\theta$ as in Eq. (13).
    M-step:
    Reestimate $\phi$ as in Eq. (15).
    Compute $\theta_1$ from $\theta$ by applying Eqs. (23), (25), (28), and (29).
    Compute $\theta_2$ from $\theta_1$ by applying Eq. (16).
    $low \leftarrow 0$, $high \leftarrow 1$, $\delta \leftarrow \theta_2 - \theta_1$
    while $high - low > \epsilon$ where $\epsilon$ is a tiny value greater than 0 do
        $mid \leftarrow (low + high)/2$
        $\rho \leftarrow \theta_1 + mid \times \delta$
        if $\mathcal{Q}_1(\phi, \rho) < \mathcal{Q}_1(\phi, \theta)$ then
            $low \leftarrow mid$
        else
            Compute the directional derivative $\nabla_{\theta_2 - \theta_1} \mathcal{Q}_2(\theta)|_{\theta=\rho}$ as in Eq. (30).
            if $\nabla_{\theta_2 - \theta_1} \mathcal{Q}_2(\theta)|_{\theta=\rho} > 0$ then
                $low \leftarrow mid$
            else
                if $\nabla_{\theta_2 - \theta_1} \mathcal{Q}_2(\theta)|_{\theta=\rho} < 0$ then
                    $high \leftarrow mid$
                else
                    $high \leftarrow mid$, break
                end if
            end if
        end if
    end while
    $\rho* \leftarrow \theta_1 + high \times \delta$
    if $\big(\mathcal{Q}_1(\phi, \rho*) \geq \mathcal{Q}_1(\phi, \theta)$ and $\mathcal{Q}_2(\rho*) > \mathcal{Q}_2(\theta)\big)$
      or $\big(\mathcal{Q}_1(\phi, \rho*) > \mathcal{Q}_1(\phi, \theta)$ and $\mathcal{Q}_2(\rho*) \geq \mathcal{Q}_2(\theta)\big)$ then
        $\theta \leftarrow \rho*$
    end if
    $iter \leftarrow iter + 1$.
until ($\phi$ and $\theta$ converge) or ($iter \geq MaxIter$)
```

100 documents from each category; as a result, in total 1000 documents were used for testing.

We evaluated the performance of DTM and provide comparison with previous topic models, including LapPLSI and LTM, and other traditional dimension reduction algorithms:

- Locally-Consistent Topic Modeling (LTM) [Cai et al. 2009];
- Laplacian Probabilistic Latent Semantic Indexing (LapPLSI) [Cai et al. 2008];
- Latent Dirichlet Allocation (LDA) [Blei et al. 2003];
- Probabilistic Latent Semantic Analysis (pLSA) [Hofmann 1999];
- Principal Component Analysis (PCA) [Jolliffe 2002];
- Non-Negative Matrix Factorization (NMF) [Lee and Seung 2000].

Additionally, we tested the approach using word histograms with the tf-idf weight scheme [Salton and Buckley 1988] and L1-normalization but without any dimension reduction.

5.2 Implementation Details

Given the word occurrences of each document in a text corpus, we applied the tf-idf weight scheme [Salton and Buckley 1988] and subsequently L1-normalization. This preprocessing is optional, but we found that it generally improves overall clustering and classification performance. We then computed the histogram intersection to measure the similarity of two documents. More formally, we calculated the histogram intersection of two documents d_i and d_j as

$$HI(d_i, d_j) = \sum_{k=1}^{M} \min(n(d_i, w_k), n(d_j, w_k)), \tag{31}$$

where $n(d, w)$ is the occurrence of word w in document d, which is obtained by applying the tf-idf weight scheme and L1-normalization to the original word occurrence. We found that this histogram intersection is more effective than the Euclidean distance or cosine distance in discovering the nearest neighbors in documents.

Based on this similarity measure, we define a local similarity matrix W as

$$W_{ij} = \begin{cases} HI(d_i, d_j), & \text{if } d_i \in \mathcal{N}_r(d_j) \text{ or } d_j \in \mathcal{N}_r(d_i), \\ 0, & \text{otherwise} \end{cases} \tag{32}$$

where $\mathcal{N}_r(d)$ is the set of the r nearest neighbors of document d that have the r highest HI values. In addition, we define a local dissimilarity matrix $\overline{W}^{original}$ as

$$\overline{W}_{ij}^{original} = \begin{cases} 1, & \text{if } \exists k\ W_{ik} > 0, W_{kj} > 0, \text{ and } W_{ij} = 0 \\ 0, & \text{otherwise} \end{cases}. \tag{33}$$

In other words, if two documents that are not directly neighboring each other are connected through another document, the two documents are linked in the local dissimilarity graph. We can reexpress Eq. (33) in matrix form:

$$\overline{W}_{ij}^{original} = \begin{cases} 1, & \text{if } (W^2)_{ij} > 0 \text{ and } W_{ij} = 0 \\ 0, & \text{otherwise} \end{cases} \tag{34}$$

We give weights on the elements of this dissimilarity matrix during model-fitting, based on the current topic distributions of documents, which are the low-dimensional representation of the documents, in order to more focus on relatively less separated pairs as follows:

$$\overline{W}_{ij} = \overline{W}_{ij}^{original} / \Big(\big(P(z_k|d_i) - P(z_k|d_j) \big)^2 + \sigma \Big), \tag{35}$$

where σ is a small positive value for avoiding a division by zero. We set σ to be 0.1 in our experiment.

The locally-defined dissimilarity matrix and weighting scheme are effective in separating unfavorable data pairs in a global manner, as we will empirically show in the following result section. Although we used the same definitions of W and $\overline{W}$ and weight scheme for all three text corpora in our experiments, we could find more effective definitions for each of the data based on their intrinsic properties or experimental results.

It is worth mentioning that class label information can be additionally used for construction of similarity matrix W, as described in the previous work [Cai et al. 2009]. More specifically, after an r-nearest neighbor graph is generated in an unsupervised manner, edges can be added between documents belonging to the same category and removed between documents belonging to different categories. This scheme can be extended to the construction of dissimilarity matrix $\overline{W}$ in the opposite way: after constructing a dissimilarity graph, edges can be removed between documents belonging to the same category and added between documents belonging to different categories. We did not apply this scheme in our implementation because the performance gain is marginal due to the lack of labeled documents in our setting.

For performance comparison, we implemented the other approaches as follows. For pLSA, we downloaded the source codes from Peter Gehler's code and dataset page.[5] For LDA, we used Matlab Topic Modeling Toolbox 1.3.2.[6] For LapPLSI and LTM, we downloaded the source codes from the author's Web page.[7] We directly implemented the other two methods: PCA and NMF. For pLSA and LDA, we did not apply the tf-idf scheme and L1-normalization; higher performances were achieved without these schemes in our experiments.

In LTM and LapPLSI, the regularization parameter needs to be determined. Instead of tuning the parameter, we tested four values (1, 10, 100, and 1000) and selected the best one based on the average performance. Although we found the best parameter by referring to the results, this parameter can be tuned through a typical validation scheme if more than one labeled document is given per category. However, if no, or only one, labeled document is available, it is unclear how to tune the parameter. We will discuss the performance variation of LTM and LapPLSI due to this choice of regularization parameter in the following section. For LTM and LapPLSI, the same W is used as DTM. To determine r, which is the number of the nearest neighbors, we tested two values, 10 and 20, and selected the better one based on the performance.

Source codes of DTM are available online.[8]

[5] http://www.kyb.mpg.de/bs/people/pgehler/code/index.html

[6] http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

[7] http://www.zjucadcg.cn/dengcai/LapPLSA/index.html

[8] http://www.cs.cmu.edu/∼seungilh

Table I. Fifty Topics of 20 Newsgroups Discovered by DTM

T 1	T 2	T 3	T 4	T 5	T 6	T 7	T 8	T 9	T 10
year	drive	people	sale	software	card	computer	move	god	dog
night	disk	kill	ship	error	mac	system	clutch	christian	front
ohio	scsi	mormon	sell	fax	driver	phone	speed	jesu	ground
won	install	question	price	message	video	modem	trade	church	disclaim
job	hard	word	mail	read	model	ibm	condition	bible	back

T 11	T 12	T 13	T 14	T 15	T 16	T 17	T 18	T 19	T 20
list	bill	homosex	mhz	israel	money	change	good	car	current
mail	article	view	test	isra	gov	make	day	oil	circuit
post	expect	cramer	chip	arab	station	status	friend	insurance	audio
scott	apr	optilink	design	muslim	cost	handle	article	brake	electron
reply	rule	men	replace	jew	base	problem	hour	detector	sound

T 21	T 22	T 23	T 24	T 25	T 26	T 27	T 28	T 29	T 30
test	wing	code	mike	access	gun	state	buy	key	tax
ignore	hawk	process	georgia	copy	waco	back	mile	clipper	uiuc
article	blue	standard	gatech	product	fire	report	wave	encrypt	opinion
done	cup	routine	eng	toni	batf	number	engine	chip	talk
apr	leaf	keyboard	blah	protect	koresh	call	live	phone	new

T 31	T 32	T 33	T 34	T 35	T 36	T 37	T 38	T 39	T 40
sun	group	info	graphic	post	school	satan	read	window	program
time	lot	advance	bit	member	answer	make	mean	file	printer
run	honda	appreciate	color	freenet	colleague	man	reason	font	run
math	new	inform	image	cleveland	book	human	belief	manage	server
start	mirror	find	version	jewish	include	life	believe	ftp	print

T 41	T 42	T 43	T 44	T 45	T 46	T 47	T 48	T 49	T 50
drug	time	bike	bob	moral	law	religion	space	game	question
doctor	stop	dod	owner	atheist	david	islam	nasa	player	gif
disease	gui	ride	steve	keith	hand	exist	shuttle	stat	wonder
medicine	hole	bnr	fenwai	object	issue	act	orbit	yankee	surface
food	sort	motorcycle	people	mathew	agree	accept	gov	play	internet

6. RESULTS AND DISCUSSIONS

In this section, we provide qualitative and quantitative evaluations of DTM, showing that DTM produces discriminative topics and is thus superior to other topic models and dimensionality reduction methods in document clustering and classification tasks.

6.1 Topic-Modeling

First we study the topic-modeling capability of DTM on 20 newsgroups. Table I lists 50 topics produced by DTM, where the most frequent five words are reported. Table II shows the three major topics for each newsgroup with the largest proportions among the 50 topics. The proportion of topic z for newsgroup c is computed by averaging $P(z|d)$ over all documents d belonging to newsgroup c.

These results show that the discriminative learning idea of DTM does not damage the statistical structure of the generative topic modeling. For example, topics 4, 15, 19, 26, 29, 41, 43, and 48 obviously represent sale, the middle east, car, gun, cryptography, motorcycle, medical treatment, and space development, respectively. Therefore, it is consequent that these topics are the first major topics of newsgroups misc.forsale, talk.politics.mideast, rec.autos, talk.politics.guns, sci.crypt, rec.motorcycles, sci.med, and sci.space. Topics 3, 9, 38, and 47 are related to religion and the major topics of newsgroups alt.atheism, soc.religion.christian, and talk.religion.misc. Topics 2, 6, 7, 14, 34, 39, and 40 contain computer terminologies and are associated with computer related newsgroups.

Figure 1 demonstrates topic distributions of five documents from each of categories misc.forsale and sci.electronics generated by DTM, LTM, and pLSA. In the results of DTM, not only do documents in the same category show similar topic distributions, but also topic distributions in the different categories are distinguishable from one

Table II. Three Major Topics for each Category of 20 Newsgroups and their Proportions

Newsgroup	Three major topics and their proportions					
alt.atheism	T 9	0.1116	T 3	0.0948	T 47	0.0880
comp.graphics	T 39	0.0961	T 40	0.0636	T 34	0.0595
comp.os.ms-windows.misc	T 39	0.1393	T 40	0.0794	T 2	0.0765
comp.sys.ibm.pc.hardware	T 2	0.1305	T 6	0.1028	T 39	0.0696
comp.sys.mac.hardware	T 2	0.1112	T 6	0.1026	T 7	0.0592
comp.windows.x	T 39	0.1410	T 40	0.0972	T 34	0.0735
misc.forsale	T 4	0.0806	T 2	0.0793	T 6	0.0665
rec.autos	T 19	0.0841	T 28	0.0790	T 8	0.0692
rec.motorcycles	T 43	0.1147	T 28	0.1067	T 8	0.0728
rec.sport.baseball	T 49	0.1948	T 22	0.0883	T 44	0.0553
rec.sport.hockey	T 49	0.2317	T 22	0.1279	T 24	0.0682
sci.crypt	T 29	0.1944	T 25	0.0574	T 46	0.0516
sci.electronics	T 2	0.0679	T 6	0.0584	T 14	0.0532
sci.med	T 41	0.1479	T 13	0.0452	T 37	0.0449
sci.space	T 48	0.1116	T 16	0.0386	T 42	0.0361
soc.religion.christian	T 9	0.1595	T 3	0.1099	T 38	0.0904
talk.politics.guns	T 26	0.1590	T 3	0.0749	T 15	0.0535
talk.politics.mideast	T 15	0.1821	T 3	0.0759	T 47	0.0570
talk.politics.misc	T 13	0.1329	T 26	0.0703	T 3	0.0484
talk.religion.misc	T 9	0.1187	T 3	0.0975	T 38	0.0712

For each category and a given topic z, the topic proportion is computed by averaging $P(z|d)$ over all documents in the category.

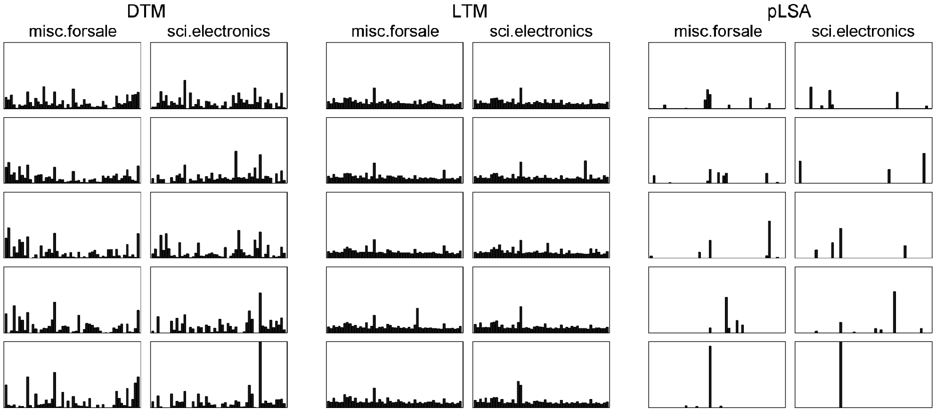

Fig. 1. Topic distributions of five documents from each of categories misc.forsale and sci.electronics generated by DTM (left), LTM (center), and pLSA (right), on 20 newsgroups.

another. In contrast, in the results of LTM, although topic distributions in the same category are similar as in DTM, topic distributions in the different categories are not quite separable. This phenomenon often happens when different categories share major topics because LTM does not explicitly take into account the separability of different categories. pLSA does not consider manifold structure at all so that even the documents in the same category often do not show similar topic distributions.

6.2 Clustering

For clustering, after performing topic modeling or dimensionality reduction, we applied K-means clustering to the low-dimensional representations ($P(z_k|d_i)$ for topic models). For each approach, we explored several numbers of topics or dimensionalities of the

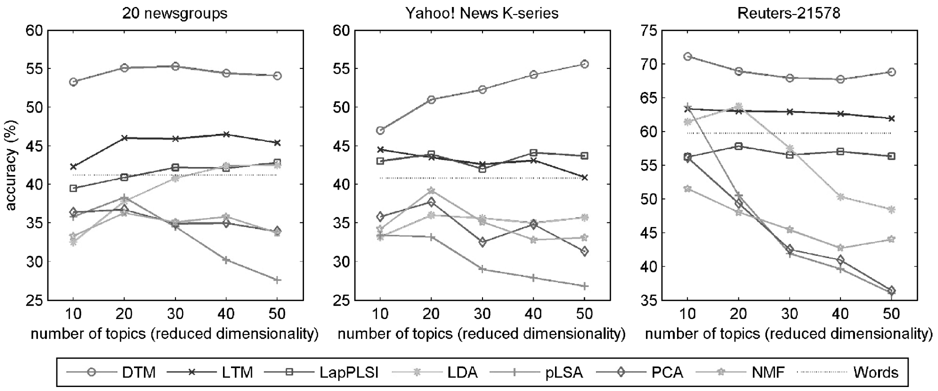

Fig. 2. Clustering performance of the K-means algorithm after various topic models and dimensionality reduction methods are applied on three text corpora: 20 newsgroups (left), Yahoo! News K-series (center), and Reuters-21578 (right) (best viewed in color). "Words" indicates that no dimensionality reduction is applied; word histograms weighted by the td-idf weight scheme and subsequently normalized by L1-normalization are used as features.

embedding space (10, 20, 30, 40, and 50). We evaluated the clustering results by referring to the labels in terms of clustering accuracy [Xu et al. 2003], defined as

$$\text{clustering accuracy} = \max_{P_\pi} \frac{\sum_{i=1}^{N} P_\pi(y_i, c_i)}{N}, \tag{36}$$

where P_π is a $C \times C$ permutation matrix (C is the number of categories), $P_\pi(j, k)$ is the (j, k) element of P_π, y_i is the original label of document i, and c_i is the cluster id of document i as a result of K-means clustering; both of y_i and c_i are one of $\{1, 2, \cdots, C\}$. We used the Hungarian method [Kuhn 1955] to find the best permutation matrix that maximizes $\sum_{i=1}^{N} P_\pi(y_i, c_i)$. We report the average of the clustering accuracies after 20 test runs.

Figure 2 demonstrates that DTM outperforms the other approaches, including LapPLSI and LTM, in the document clustering task. On all the three corpora, DTM achieves approximately 5 to 10% higher performance than the subsequent methods. From these results, we can conclude that DTM is more successful in exposing the manifold structures inherent in the text corpora. On the other hand, LapPLSI and LTM are not as effective as DTM because unfavorable pairs are not taken into account in discovering the manifold structures.

Figure 3 illustrates the distributions of low-dimensional representations of 20 newsgroup documents generated by DTM, LTM, and pLSA, where each dot represents a document and each marker indicates a category. To produce 2D embeddings, we applied t-Distributed Stochastic Neighbor Embedding (t-SNE) [van der Maaten and Hilton 2008] using Matlab Toolbox for Dimensionality Reduction.[9] When both favorable and unfavorable relationships are taken into account (DTM), documents belonging to the same category tend to be more separately grouped from those in the other categories. On the other hand, if only favorable relationships are considered (LTM), documents in different categories often overlap one another, resulting in less effective separation of categories. When neither of the relationships are considered (pLSA), documents in the same category are often divided into several groups and those in different categories are jumbled together, which leads to the worst clustering performance among the three models.

[9] http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html

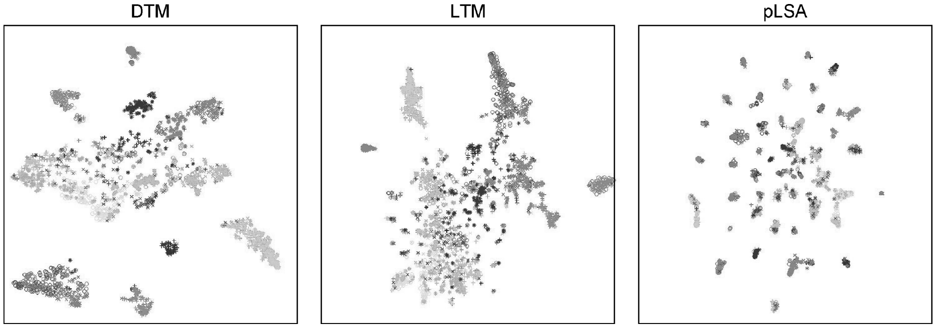

Fig. 3. 2D embeddings of the low-dimensional representations of 20 newsgroup documents generated by DTM (left), LTM (center), and pLSA (right). Different markers indicate different categories (best viewed in color).

6.3 Classification

For classification, in order to address real-world problems in a semisupervised setting, we randomly selected a small number of documents (one of 1, 3, and 5) from each category as labeled data and considered the remainder to be unlabeled. After we obtained low-dimensional representations through topic modeling or dimensionality reduction, we applied 1-nearest neighbor (1-NN) and linear-kernel Support Vector Machine (SVM) to the low-dimensional representations ($P(z_k|d_i)$ for topic models) for classification. We tuned the slack parameter of the SVM through leave-one-out-per-class cross validation when more than one document per category is labeled; when only one labeled document is available per category, we used the default parameter. We report the average of the classification accuracies after 20 test runs.

Figures 4 and 5 show that DTM outperforms the other approaches in terms of document classification accuracy. Among previous approaches, LapPLSI and LTM generally show higher performance than the other methods, as we expected. Although LapPLSI and LTM do not reach the full discriminating power of manifold learning, they can still find a low-rank nonlinear embedding space to which documents are mapped. On the other hand, pLSA and LDA, which do not adopt any regularization for manifold learning, cannot find such a nonlinear embedding space. PCA and NMF also do not consider manifold structure and thus are not effective for discovering discriminative low-dimensional representation. The performance of pLSA decreases as the number of topics increases beyond a certain point; it is well known that pLSA is prone to overfitting due to the large number of parameters, which grows proportionally with data size. PCA and NMF also demonstrate similar tendencies in our experiments.

6.4 Discussion

The time complexity of DTM is as follows. As defined previously, let K, M, and N be the number of topics, the size of the vocabulary, and the number of documents, respectively. The E-step, reestimation of ϕ, and the update of the log-likelihood of pLSA ($\mathcal{Q}_1$) in DTM are the same as the process of pLSA. Since, as known, the worst-case time complexity of pLSA is $O(KMN)$, the time complexity of these components of DTM is $O(KMN)$. The update of the regularization term ($\mathcal{Q}_2$) can be conducted by the matrix computation in Eqs. (23), (25), (28), and (29). Hence, the time complexity of this update is $O(KN^2)$. The number of binary search steps is bounded by a constant because the search is performed in the range [0, 1] (more precisely, between $\theta_1+0\cdot\delta$ and $\theta_1+1\cdot\delta$, where $\delta=\theta_2-\theta_1$) and the precision is constant. Therefore, the time complexity of reestimation θ after the updates of $\mathcal{Q}_1$ and $\mathcal{Q}_2$ is $O(KN^2)$ due to computation of the

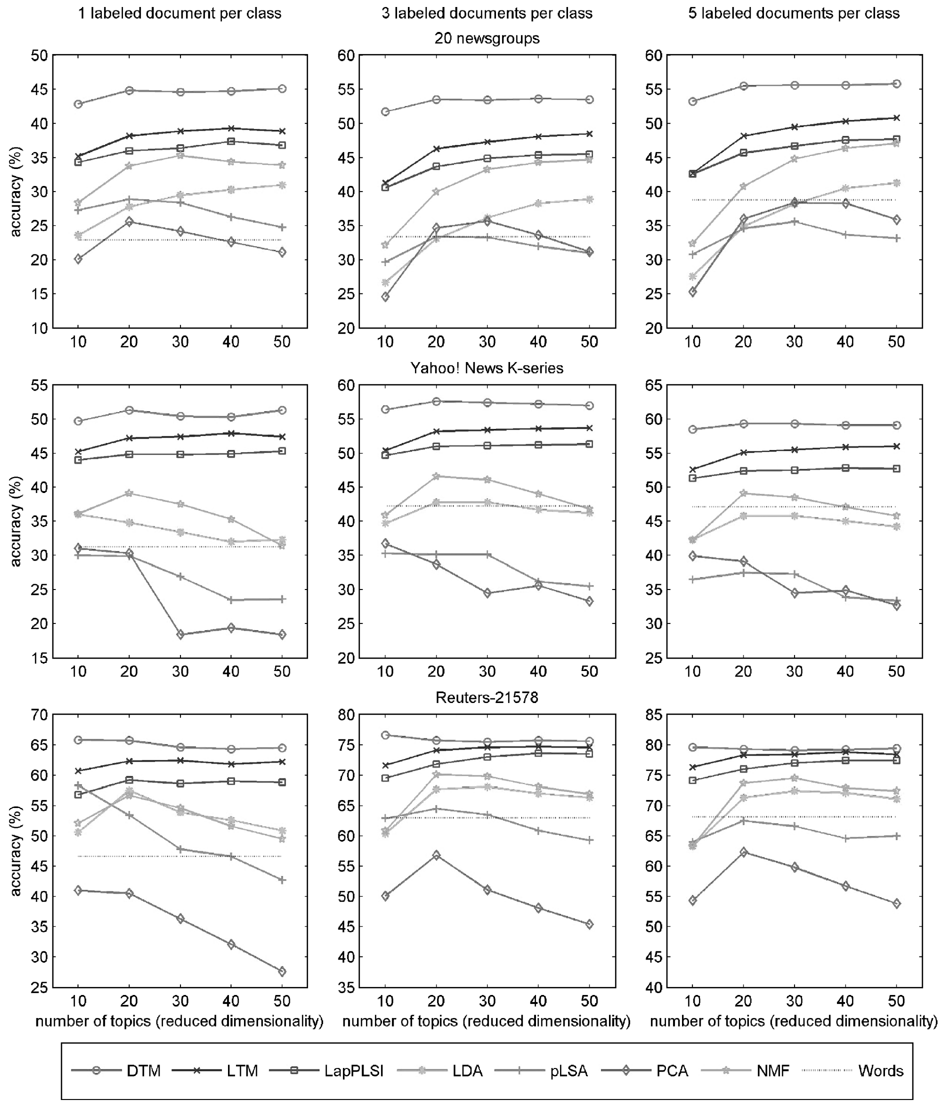

Fig. 4. Classification performance of the 1-Nearest Neighbor (1-NN) classifier after various topic models and dimensionality reduction methods are applied on three text corpora: 20 newsgroups (top), Yahoo! News K-series (middle), and Reuters-21578 (bottom) (best viewed in color). We selected a small number of documents among one (left), three (center), and five (right) from each category as labeled data for training. In each subplot, *x*-axis represents the number of topics or reduced dimensionality, and *y*-axis represents classification accuracy. "Words" indicates that no dimensionality reduction is applied; word histograms weighted by the td-idf weight scheme and subsequently normalized by L1-normalization are used as features.

directional derivative. By summing up all these time complexities, we arrive at the time complexity of DTM as $O(KMN + KN^2)$. In our experiments, the parameters of DTM converge in 300 iterations and clustering/classification accuracies reach their peaks in 30 to 50 iterations. The practical running time of DTM is comparable to that of LTM or LapPLSI.

DTM iteratively finds the Pareto improvement between $\mathcal{Q}_1$ and $\mathcal{Q}_2$. In the early iterations, $\mathcal{Q}_1$ and $\mathcal{Q}_2$ tend to move together because discovering a generative process of

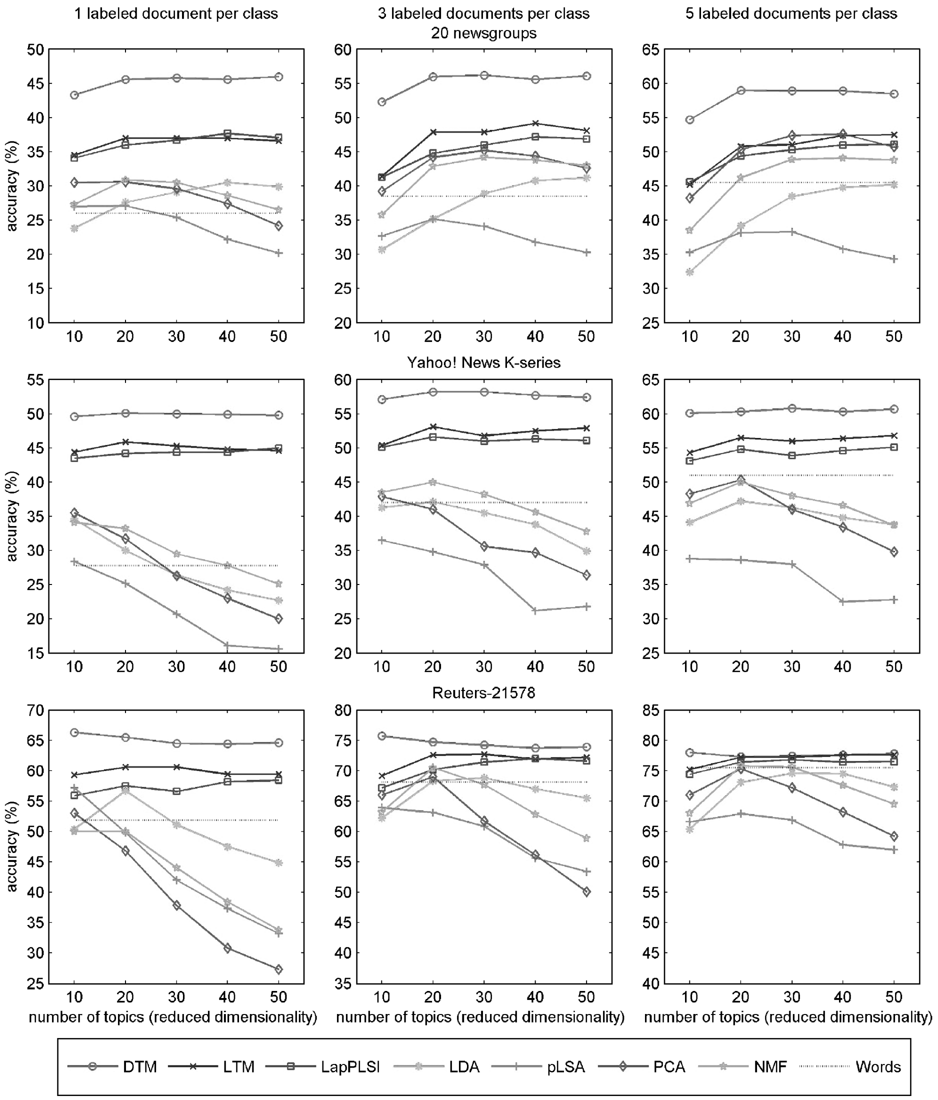

Fig. 5. Classification performance of the Support Vector Machine (SVM) classifier after topic models and dimensionality reduction methods are applied on three text corpora: 20 newsgroups (top), Yahoo! News K-series (middle), and Reuters-21578 (bottom) (best viewed in color). We selected a small number of documents among one (left), three (center), and five (right) from each category as labeled data for training. In each subplot, x-axis represents the number of topics or reduced dimensionality, and y-axis represents classification accuracy. “Words” indicates that no dimensionality reduction is applied; word histograms weighted by the td-idf weight scheme and subsequently normalized by L1-normalization are used as features.

data contributes to categorizing them. As the iteration increases, discriminative power is decoupled from generative power so that the improvement on $\mathcal{Q}_1$ is restricted by the regularization effect of $\mathcal{Q}_2$. As a result, DTM avoids modeling the details of the generative process of data that undermine discriminative power. This Pareto optimization strategy is different from general regularization schemes in that the optimization of $\mathcal{Q}_1$ and $\mathcal{Q}_2$ is constrained by current $\mathcal{Q}_1$ and $\mathcal{Q}_2$ at each iteration. In other words, the improvement on $\mathcal{Q}_1$ or $\mathcal{Q}_2$ is irreversible.

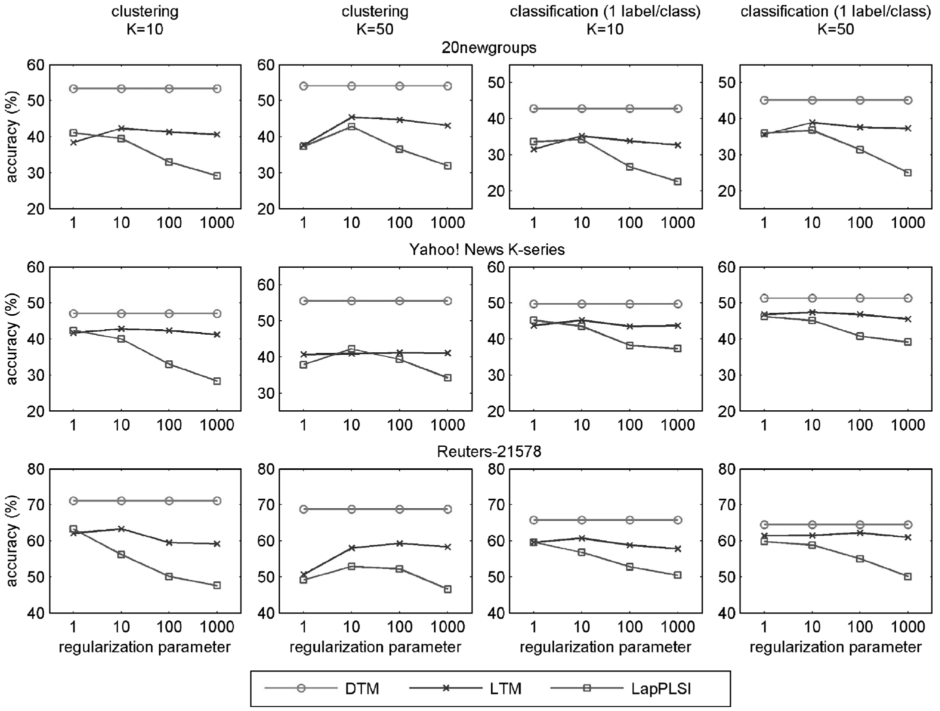

Fig. 6. Performance change of LTM and LapPLSI as the regularization parameter varies, on three text corpora: 20 newsgroups (top), Yahoo! News K-series (middle), and Reuters-21578 (bottom) (best viewed in color). K denotes the number of topics or reduced dimensionality. In each subplot, x-axis represents the value of the regularization parameter and y-axis represents clustering or classification accuracy.

DTM does not involve the tuning of the regularization parameter that balances between $\mathcal{Q}_1$ and $\mathcal{Q}_2$ because the updates are conducted to maximize $\mathcal{Q}_2$ among Pareto improvements between $\mathcal{Q}_1$ and $\mathcal{Q}_2$, which implies maximizing discriminative power without sacrificing generative power already obtained. Therefore, the balance between $\mathcal{Q}_1$ and $\mathcal{Q}_2$ is not fixed, but is determined by current $\mathcal{Q}_1$ and $\mathcal{Q}_2$ at each iteration. On the other hand, LapPLSI and LTM require setting the parameter before optimization, but it is not clear how to perform the tuning particularly with little label information available, e.g., unsupervised clustering or classification with one label per class. Figure 6 shows the clustering and classification performance of LTM and LapPLSI with four different regularization parameters compared to the performance of DTM, which does not vary due to the irrelevance of the parameter. As can be seen, the performance changes of LTM and LapPLSI due to the parameter are sometimes not negligible and the best parameter is not constant but varies according to the text corpus, the amount of labels, and the number of topics or reduced dimensionality.

Out-of-sample data can be handled in two ways as with other semisupervised learning methods: inclusive and exclusive approaches [Trosset et al. 2008]. The former approach reconstructs similarity and dissimilarity matrices with new data in addition to in-sample data. Based on the matrices, a new model is then trained. Since this approach repeats the entire modeling process, it is inefficient, but it is effective when the distribution of out-of-sample data is not consistent with that of in-sample data. The latter approach maintains the model trained with in-sample data and extrapolates the topic distribution for all new data. One way to perform extrapolation is to apply the formalized out-of-sample extension for spectral clustering presented in Bengio et al. [2004] and normalize the result.

The name of our model, discriminant topic model (DTM), may bring attention to other topic models whose names imply a similar idea, e.g., supervised LDA (sLDA) [Blei and McAuliffe 2008], discriminatively trained LDA (DiscLDA) [Lacoste-Julien et al. 2008], maximum entropy discrimination LDA (MedLDA) [Zhu et al. 2009]. DTM essentially differs from these supervised models in that it does not require label information. In other words, DTM is an unsupervised topic model developed for clustering or semisupervised classification, while the others are supervised topic models for supervised classification or regression.

7. CONCLUSIONS

In this article, we have proposed a topic model that incorporates the information from the manifold structure of data by considering unfavorable relationships in addition to favorable ones; the former have been ignored in previous work. We have also presented an efficient model-fitting algorithm, based on generalized EM and Pareto improvement, which enables reliable discovery of the low-rank hidden structures by minimizing the sensitivity to parameters. We empirically demonstrated that our approach outperforms previous topic models in terms of unsupervised clustering and semisupervised classification accuracies on three popularly used text corpora. We envision other applications of the approach in this article when we combine text with other data elements and structures, such as references or links.

APPENDIX

In this appendix, we provide the proof of Theorem 2.

A. Proof of Theorem 2

We reintroduce the concept of auxiliary function [Lee and Seung 2000; Sha et al. 2003].

Definition A.1. $G(x, x')$ is an auxiliary function for $F(x)$ if the two following conditions are satisfied.

$$G(x, x') \leq F(x), \quad G(x, x) = F(x) \tag{37}$$

This definition is useful with the following Lemma.

LEMMA A.2. *If $G(x, x')$ is an auxiliary function, then $F(x)$ is nonincreasing under the update*

$$x^{t+1} = \arg\max_x G(x, x') \tag{38}$$

PROOF. $F(x^{t+1}) \geq G(x^{t+1}, x^t) \geq G(x^t, x^t) = F(x^t)$. □

We define $\hat{\tau}$ for topic id p and document id i as

$$\hat{\tau} = \frac{\overline{D}_{ii} P(z_p|d_i) + \alpha \sum_{j=1}^{N} W_{ij} P(z_p|d_j)}{\sum_{j=1}^{N} \overline{W}_{ij} P(z_p|d_j) + \alpha D_{ii} P(z_p|d_i)}, \tag{39}$$

and also define

$$\mathcal{R}(\theta) = \sum_{i,j=1}^{N} \sum_{k=1}^{K} \overline{W}_{ij} \big(P(z_k|d_i) - P(z_k|d_j)\big)^2 - \alpha \sum_{i,j=1}^{N} \sum_{k=1}^{K} W_{ij} \big(P(z_k|d_i) - P(z_k|d_j)\big)^2. \tag{40}$$

LEMMA A.3. *$\mathcal{R}(\theta)$ is nondecreasing after reestimation of $[P(z_1|d_i), \cdots, P(z_K|d_i)]$ by the following equations with $\tau = \hat{\tau}$.*

$$P(z_k|d_i) = \begin{cases} \tau P(z_p|d_i), & \textit{if } k = p \\ \frac{1-\tau P(z_p|d_i)}{1-P(z_p|d_i)} P(z_k|d_i), & \textit{otherwise} \end{cases}. \tag{41}$$

PROOF. Let $F(\tau)$ be the value of $\mathcal{R}(\theta)$ at $\theta = \tilde{\theta}^{(t+1)}$ that is obtained by applying the update in Eq. (41) to the current parameters $\theta^{(t)} = \{P(z|d)\}$. Then, the first order derivative of $F(\tau)$ is

$$\begin{aligned}\frac{\partial F(\tau)}{\partial \tau} &= 2\sum_{j=1}^{N}(\overline{W}_{ij} - \alpha W_{ij})\Big[\big(\tau P(z_p|d_i) - P(z_p|d_j)\big)P(z_p|d_i) \\ &- \sum_{k\neq p}\Big(\frac{1-\tau P(z_p|d_i)}{1-P(z_p|d_i)}P(z_k|d_i) - P(z_k|d_j)\Big)\frac{P(z_p|d_i)}{1-P(z_p|d_i)}\Big].\end{aligned} \tag{42}$$

Since $\sum_{k\neq p} P(z_k|d) = 1 - P(z_p|d)$,

$$\begin{aligned}&\sum_{k\neq p}\Big(\frac{1-\tau P(z_p|d_i)}{1-P(z_p|d_i)}P(z_k|d_i) - P(z_k|d_j)\Big) = \frac{1-\tau P(z_p|d_i)}{1-P(z_p|d_i)}\sum_{k\neq p}P(z_k|d_i) - \sum_{k\neq p}P(z_k|d_j) \\ &= \frac{1-\tau P(z_p|d_i)}{1-P(z_p|d_i)}\big(1-P(z_p|d_i)\big) - \big(1-P(z_p|d_j)\big) = \big(1-\tau P(z_p|d_i)\big) - \big(1-P(z_p|d_j)\big) \\ &= -\tau P(z_p|d_i) + P(z_p|d_j)\end{aligned} \tag{43}$$

Putting Eq. (43) into Eq. (42) yields

$$\begin{aligned}\frac{\partial F(\tau)}{\partial \tau} &= 2\sum_{j=1}^{N}(\overline{W}_{ij} - \alpha W_{ij})\big(\tau P(z_p|d_i) - P(z_p|d_j)\big)\Big(P(z_p|d_i) + \frac{P(z_p|d_i)}{1-P(z_p|d_i)}\Big) \\ &= 2c\sum_{j=1}^{N}(\overline{W}_{ij} - \alpha W_{ij})\big(\tau P(z_p|d_i) - P(z_p|d_j)\big) \\ &= 2c\Big(\sum_{j=1}^{N}\overline{W}_{ij}P(z_p|d_i) - \alpha\sum_{j=1}^{N}W_{ij}P(z_p|d_i)\Big)\tau - 2c\Big(\sum_{j=1}^{N}\overline{W}_{ij}P(z_p|d_j) - \alpha\sum_{j=1}^{N}W_{ij}P(z_p|d_j)\Big) \\ &= 2c\big(\overline{D}_{ii}P(z_p|d_i) - \alpha D_{ii}P(z_p|d_i)\big)\tau - 2c\Big(\sum_{j=1}^{N}\overline{W}_{ij}P(z_p|d_j) - \alpha\sum_{j=1}^{N}W_{ij}P(z_p|d_j)\Big),\end{aligned} \tag{44}$$

where $c = \Big(P(z_p|d_i) + \frac{P(z_p|d_i)}{1-P(z_p|d_i)}\Big)$ and, D_{ii} and $\overline{D}_{ii}$ are defined in Eq. (19).

In addition, the second order derivative of $F(\tau)$ is

$$\frac{\partial^2 F(\tau)}{\partial \tau^2} = 2c\big(\overline{D}_{ii}P(z_p|d_i) - \alpha D_{ii}P(z_p|d_i)\big). \tag{45}$$

We define G as an auxiliary function of $F(\tau)$ by replacing the second order derivative in the Taylor series expansion of $F(\tau)$ at $\tau = 1$.

$$G(\tau, 1) = F(1) + \frac{\partial F(\tau)}{\partial \tau}\Big|_{\tau=1}(\tau - 1) - c\Big(\sum_{j=1}^{N}\overline{W}_{ij}P(z_p|d_j) + \alpha D_{ii}P(z_p|d_i)\Big)(\tau-1)^2 \tag{46}$$

Since $c \geq 0$ and all elements of $\overline{W}$ are nonnegative,

$$G(\tau, 1) - F(\tau) = -c\big(\sum_{j=1}^{N} \overline{W}_{ij} P(z_p|d_j) + \overline{D}_{ii} P(z_p|d_i)\big)(\tau - 1)^2 \leq 0. \tag{47}$$

Hence, G is indeed an auxiliary function of F. As nonnegativity of the elements of W and $\overline{W}$ ensures that $G(\tau, 1)$ is concave with respect to τ, solving $\frac{\partial G(\tau,1)}{\partial \tau} = 0$ yields $\hat{\tau}$ in Eq. (39) that maximizes $G(\tau, 1)$. Therefore, by Lemma A.2,

$$\mathcal{R}(\tilde{\theta}^{(t+1)}) = F(\hat{\tau}) \geq G(\hat{\tau}, 1) \geq G(1, 1) = F(1) = \mathcal{R}(\theta^{(t)}). \tag{48}$$

□

LEMMA A.4. *$\mathcal{R}(\theta)$ is nondecreasing by the updates in Eq.* (20) *with β_{pi} in Eq.* (18).

PROOF. For any μ such that $0 \leq \mu \leq 1$,

$$G(1, 1) = (1 - \mu)G(1, 1) + \mu G(1, 1) \leq (1 - \mu)G(1, 1) + \mu G(\hat{\tau}, 1). \tag{49}$$

Since $G(\tau, 1)$ is concave,

$$(1 - \mu)G(1, 1) + \mu G(\hat{\tau}, 1) \leq G((1 - \mu) + \mu\hat{\tau}, 1). \tag{50}$$

Thus, $G(1, 1) \leq G(\nu, 1)$ for any ν between 1 and $\hat{\tau}$ (either $1 \leq \nu \leq \hat{\tau}$ or $\hat{\tau} \leq \nu \leq 1$).

Let $\theta^{(t+1)}$ result from applying the updates in Eq. (20) to $\theta^{(t)}$. Since β_{pi} is always between 1 and $\hat{\tau}$,

$$\mathcal{R}(\theta^{(t+1)}) = F(\beta_{pi}) \geq G(\beta_{pi}, 1) \geq G(1, 1) = F(1) = \mathcal{R}(\theta^{(t)}). \tag{51}$$

□

Proof of Theorem 2

PROOF. Since $\alpha = \mathcal{Q}_2(\theta^{(t)})$, $\mathcal{R}(\theta^{(t)}) = 0$. By Lemma A.4,

$$\begin{aligned}\mathcal{R}(\theta^{(t+1)}) = &\sum_{i,j=1}^{N}\sum_{k=1}^{K} \overline{W}_{ij}\big(P(z_k|d_i) - P(z_k|d_j)\big)^2\Big|_{\theta=\theta^{(t+1)}} \\ &- \alpha \sum_{i,j=1}^{N}\sum_{k=1}^{K} W_{ij}\big(P(z_k|d_i)^{(t+1)} - P(z_k|d_j)\big)^2\Big|_{\theta=\theta^{(t+1)}} \geq 0.\end{aligned} \tag{52}$$

Therefore,

$$\begin{aligned}\mathcal{Q}_2(\theta^{(t+1)}) &= \frac{\sum_{i,j=1}^{N}\sum_{k=1}^{K} \overline{W}_{ij}\big(P(z_k|d_i) - P(z_k|d_j)\big)^2\big|_{\theta=\theta^{(t+1)}}}{\sum_{i,j=1}^{N}\sum_{k=1}^{K} W_{ij}\big(P(z_k|d_i) - P(z_k|d_j)\big)^2\big|_{\theta=\theta^{(t+1)}}} \\ &\geq \alpha = \mathcal{Q}_2(\theta^{(t)}).\end{aligned} \tag{53}$$

□

ACKNOWLEDGMENTS

We thank the anonymous reviewers of an earlier version of this article for their insightful feedback.

REFERENCES

Barr, N. 2004. *Economics of the Welfare State*. Oxford University Press, Oxford, UK.

Belkin, M. and Niyogi, P. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*. 586–691.

Belkin, M., Niyogi, P., and Sindhwani, V. 2006. Mainfold regularization: A geometric framework for learning from examples. *J. Mach. Learn. 7*, 2399–2434.

Bengio, Y., Paiement, J., Vincent, P., Delalleau, O., Le Roux, N., and Ouimet, M. 2004. Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*.

Blei, D. and McAuliffe, J. 2008. Supervised topic models. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*. 121–128.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent Dirichlet allocation. *J. Mach. Learn. 3*, 993–1022.

Boley, D. L. 1998. Principal direction divisive partitioning. *Data Mining Knowl. Discov. 2*, 4, 325–344.

Cai, D., Mei, Q., Han, J., and Zhai, C. 2008. Modeling hidden topics on document manifold. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*. 911–920.

Cai, D., Wang, X., and He, X. 2009. Probabilistic dyadic data analysis with local and global consistency. In *Proceedings of the International Conference on Machine Learning (ICML)*. 105–112.

Chung, F. R. K. 1997. *Spectral Graph Theory.* American Mathematical Society.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. 1990. Indexing by latent semantic analysis. *J. Amer. Soc. Inf. Sci. 41*, 391–407.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. Series B* (Methodological) *39*, 1–38.

Hinton, G. and Roweis, S. 2002. Stochastic neighbor embedding. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*. 833–840.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 50–57.

Huh, S. and Fienberg, S. E. 2010. Discriminative topic modeling based on manifold learning. In *Proceedings of the ACM SIGKDD Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)*. 653–661.

Jolliffe, I. T. 2002. *Principal Component Analysis* 2nd Ed. Springer Series in Statistics, Vol. 29, Springer, NY.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval Res. Logist. Quarterly*, 83–97.

Lacoste-Julien, S., Sha, F., and Jordan, M. I. 2008. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*. 897–904

Lee, D. D. and Seung, H. S. 2000. Algorithms for non-negative matrix factorization. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*. 556–562.

Roweis, S. and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science 290*, 2323–2326.

Salton, G. and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Inform. Proc. Manage. 24*, 513–523.

Sha, F., Saul, L. K., and Lee, D. D. 2003. Multiplicative updates for nonnegative quadratic programming in support vector machines. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*. 1041–1048.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science 290*, 2319–2323.

Trosset, M. W., Priebe, C. E., Park, Y., and Miller, M. I. 2008. Semisupervised learning from dissimilarity data. *Comput. Statist. Data Anal. 52*, 10, 4643–4657.

van der Maaten, M. and Hilton, G. 2008. Visualizing data using T-SNE. *J. Mach. Learn. Res. 9*, 2579–2605.

Xu, W., Liu, X., and Gong, Y. 2003. Document clustering based on non-negative matrix factorization. In *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR)*. 267–273.

Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q., and Lin, S. 2007. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Trans. Pattern Anal. Mach. Intell. 29*, 1, 40–51.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, B. 2003. Learning with local and global consistency. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*. 321–328.

Zhu, J., Ahmed, A., and Xing, E. P. 2009. MedLDA: Maximum margin supervised topic models for regression and classification. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Zhu, X., Ghahramani, Z., and Lafferty, J. D. 2003. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the International Conference on Machine Learning (ICML)*. 912–919.

Received March 2011; revised August 2011; accepted November 2011