JMLR: Workshop and Conference Proceedings 25:395–410, 2012 Asian Conference on Machine Learning

Supervised dimension reduction with topic models

Khoat Than KHOAT@JAIST.AC.JP
Tu Bao Ho BAO@JAIST.AC.JP
Duy Khuong Nguyen KHUONGND@JAIST.AC.JP
Ngoc Khanh Pham KHANH@JAIST.AC.JP
Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan.

Editor: Steven C.H. Hoi and Wray Buntine

Abstract

We consider supervised dimension reduction (SDR) for problems with discrete variables. Existing methods are computationally expensive, and often do not take the local structure of data into consideration when searching for a low-dimensional space. In this paper, we propose a novel framework for SDR which is (1) general and flexible so that it can be easily adapted to various unsupervised topic models, (2) able to inherit scalability of unsupervised topic models, and (3) can exploit well label information and local structure of data when searching for a new space. Extensive experiments with adaptations to three models demonstrate that our framework can yield scalable and qualitative methods for SDR. One of those adaptations can perform better than the state-of-the-art method for SDR while enjoying significantly faster speed.

Keywords: supervised dimension reduction, topic models, scalability, local structure

1. Introduction

In supervised dimension reduction (SDR), we are asked to find a low-dimensional space which preserves the predictive information of the response variable. Projection on that space should keep the discrimination property of data in the original space. While there is a rich body of researches on SDR, our primary focus in this paper is on developing methods for discrete data. At least three reasons motivate our study: (1) current state-of-the-art methods for continuous data are really computationally expensive (Chen et al., 2012; Parrish and Gupta, 2012; Sugiyama, 2007), and hence can only deal with data of small size and low dimensions; (2) meanwhile, there are excellent developments which can work well on discrete data of huge size (Mimno et al., 2012; Smola and Narayanamurthy, 2010) and extremely high dimensions (Than and Ho, 2012a), but are unexploited for supervised problems; (3) further, continuous data can be easily discretized to avoid sensitivity and to effectively exploit certain algorithms for discrete data (Yang and Webb, 2009).

Topic modeling is a potential approach to dimension reduction. Recent advances in this new area can deal well with huge data of very high dimensions (Mimno et al., 2012; Than and Ho, 2012a; Smola and Narayanamurthy, 2010). However, due to their unsupervised nature, they do not exploit supervised information. Furthermore, because the local structure of data in the original space is not considered appropriately, the new space is not guaranteed to preserve the discrimination property and proximity between instances. These limitations make unsupervised topic models unappealing to supervised dimension reduction.

Investigation of local structure in topic modeling have been initiated by some previous researches (Wu et al., 2012; Huh and Fienberg, 2012; Cai et al., 2009). These are basically extensions of *probabilistic latent semantic analysis* (PLSA) by Hofmann (2001), which take local structure of data into account. Local structures are derived from nearest neighbors, and are often encoded in a graph. Those structures are then incorporated into the likelihood function when learning PLSA. Such an incorporation of local structures often results in very high complexity for learning. For instances, the complexity of each iteration of the learning algorithms by Wu et al. (2012) and Huh and Fienberg (2012) is quadratic in the size M of the training data; and that by Cai et al. (2009) is triple in M because of requiring a matrix inversion. Hence these developments, even though often being shown to work well, are very limited when the data size is large.

Some topic models (Blei and McAuliffe, 2007; Lacoste-Julien et al., 2008; Zhu et al., 2012) for supervised problems can do simultaneously two nice jobs. One job is derivation of a meaningful space which is often known as *"topical space"*. The other is that supervised information is explicitly utilized to find the topical space. Nonetheless, there are two common limitations of existing supervised topic models. First, the local structure of data is not taken into account. Such an ignorance can hurt the discrimination property in the new space. Second, current learning methods for those supervised models are often very expensive, which is problematic with large data of high dimensions.

In this paper, we approach to SDR in a novel way. Instead of developing new supervised models, we propose a framework which can inherit the scalability of recent advances for unsupervised topic models, and can exploit label information and local structure of the training data. The main idea behind the framework is that we first learn a unsupervised model to find an initial topical space; we next project documents on that space exploiting label information and local structure, and then reconstruct the final space. To this end, we employ the FW framework for doing projection/inference which is proposed by Than and Ho (2012b). Note that the FW framework is very scalable and flexible, and enables us to easily incorporate side information into inference.

Our framework for SDR is general and flexible so that it can be easily adapted to various unsupervised topic models. To provide some evidences, we adapt our framework to three models: *probabilistic latent semantic analysis* (PLSA) by Hofmann (2001), *latent Dirichlet allocation* (LDA) by Blei et al. (2003), and *fully sparse topic models* (FSTM) by Than and Ho (2012a). The resulting methods for SDR are respectively denoted as PLSA^c, LDA^c, and FSTM^c. Extensive experiments show that PLSA^c, LDA^c, and FSTM^c can perform substantially better than their unsupervised counterparts.[1] PLSA^c and LDA^c often perform comparably with the state-of-the-art supervised model, MedLDA by Zhu et al. (2012). FSTM^c can do consistently better than MedLDA, and reach comparable performance with SVM which works on the original space. Moreover, PLSA^c and FSTM^c consumed significantly less time than MedLDA to learn good low-dimensional spaces. These results suggest that our framework provides a competitive approach to supervised dimension reduction.

1. Note that due to being dimension reduction methods, PLSA, LDA, FSTM, PLSA^c, LDA^c, and FSTM^c themselves cannot directly do classification. Hence we use SVM for doing classification tasks on the low-dimensional spaces. MedLDA itself can do classification. Performance for comparison is the accuracy of classification.

ORGANIZATION: in the next section, we describe briefly some notations, results, and related unsupervised topic models. We present the proposed framework for SDR in Section 3. We also discuss in Section 4 the reasons why label information and local structure of data can be exploited well to result in good methods for SDR. Empirical evaluation is presented in Section 5. Finally, we discuss some open problems and conclusions in the last section.

2. Background

Consider a corpus $\mathcal{D} = \{\boldsymbol{d}_1, ..., \boldsymbol{d}_M\}$ consisting of M documents which are composed from a vocabulary of V terms. Each document $\boldsymbol{d}$ is represented as a vector of term frequencies, i.e. $\boldsymbol{d} = (d_1, ..., d_V) \in \mathbb{R}^V$, where d_j is the number of occurrences of term j in $\boldsymbol{d}$. Let $\{y_1, ..., y_M\}$ be the class labels assigned to those documents, respectively. The task of *supervised dimension reduction* (SDR) is to find a new space of K dimensions which preserves the predictiveness of the response/label variable Y. Loosely speaking, predictiveness preservation requires that projection of data points onto the new space should preserve separation (discrimination) between classes in the original space, and that proximity between data points is maintained. Once the new space is determined, we can work with projections in that low-dimensional space instead of the high-dimensional one.

2.1. Unsupervised topic models

Probabilistic topic models often assume that a corpus is composed of K topics, and each document is a mixture of those topics. Example models includes PLSA (Hofmann, 2001), LDA (Blei et al., 2003), and FSTM (Than and Ho, 2012a). Under a model, each document has another latent representation, known as *topic proportion*, in the K-dimensional space. Hence topic models play a role as dimension reduction if $K < V$. Learning a low-dimensional space is equivalent to learning the topics of a model. Once such a space is learned, new documents can be projected onto that space via *inference*. Next, we describe briefly how to learn and to do inference for three models.

2.1.1. PLSA

Let $\theta_{dk} = P(z_k|\boldsymbol{d})$ be the probability that topic k appears in document $\boldsymbol{d}$, and $\beta_{kj} = P(w_j|z_k)$ be the probability that term j contributes to topic k. These definitions basically imply that $\sum_{k=1}^{K} \theta_{dk} = 1$ for each $\boldsymbol{d}$, and $\sum_{j=1}^{V} \beta_{kj} = 1$ for each topic k. The PLSA model assumes that document $\boldsymbol{d}$ is a mixture of K topics, and $P(z_k|\boldsymbol{d})$ is the proportion that topic k contributes to $\boldsymbol{d}$. Hence the probability of term j appearing in $\boldsymbol{d}$ is $P(w_j|\boldsymbol{d}) = \sum_{k=1}^{K} P(w_j|z_k)P(z_k|\boldsymbol{d}) = \sum_{k=1}^{K} \theta_{dk}\beta_{kj}$. Learning PLSA is to learn the topics $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K)$. Inference of document $\boldsymbol{d}$ is to find $\boldsymbol{\theta}_d = (\theta_{d1}, ..., \theta_{dK})$.

For learning, we use the EM algorithm to maximize the likelihood of training data:

$$\text{E-step:} \quad P(z_k|\boldsymbol{d}, w_j) = \frac{P(w_j|z_k)P(z_k|\boldsymbol{d})}{\sum_{l=1}^{K} P(w_j|z_l)P(z_l|\boldsymbol{d})}, \tag{1}$$

$$\text{M-step:} \quad \theta_{dk} = P(z_k|\boldsymbol{d}) \propto \sum_{v=1}^{V} d_v P(z_k|\boldsymbol{d}, w_v), \tag{2}$$

$$\beta_{kj} = P(w_j|z_k) \propto \sum_{\boldsymbol{d}\in\mathcal{D}} d_j P(z_k|\boldsymbol{d}, w_j). \tag{3}$$

Inference in PLSA is not explicitly derived. Hofmann (2001) proposed an adaptation from learning: keeping topics fixed, iteratively do the steps (1) and (2) until convergence. This algorithm is called *folding-in*.

2.1.2. LDA

Blei et al. (2003) proposed LDA as a Bayesian version of PLSA. In LDA, the topic proportions are assumed to follow a Dirichlet distribution. The same assumption is endowed over topics $\boldsymbol{\beta}$. Learning and inference in LDA are much more involved than those of PLSA. Each document $\boldsymbol{d}$ is independently inferred/projected by the variational method with the following updates:

$$\phi_{djk} \propto \beta_{kw_j} \exp \Psi(\gamma_{dk}), \tag{4}$$

$$\gamma_{dk} = \alpha + \sum_{d_j>0} \phi_{djk}, \tag{5}$$

where ϕ_{djk} is the probability that topic i generates the j^{th} word w_j of $\boldsymbol{d}$; γ_d is the variational parameters; Ψ is the digamma function; α is the parameter of the Dirichlet prior over $\boldsymbol{\theta}_d$.

Learning LDA is done by iterating the following two steps until convergence. The E-step does inference for each document. The M-step maximizes the likelihood of data w.r.t $\boldsymbol{\beta}$ by the following update:

$$\beta_{kj} \propto \sum_{\boldsymbol{d}\in\mathcal{D}} d_j \phi_{djk}. \tag{6}$$

2.1.3. FSTM

FSTM is a simplified variant of PLSA and LDA. It is the result of removing the endowment of Dirichlet distributions in LDA, and is a variant of PLSA when removing the observed variable associated with each document. Even though there is no explicit prior over topic proportions, Than and Ho (2012a) show that in fact an implicit prior exists. This interesting property is due to the sparse inference algorithm in FSTM. Learning of topics is simply a multiplication of the new and old representations of the training data.

$$\beta_{kj} \propto \sum_{\boldsymbol{d}\in\mathcal{D}} d_j \theta_{dk}. \tag{7}$$

2.2. The FW framework for inference

Inference is an integral part of probabilistic topic models. The main task of inference for a given document is to infer the topic proportion that maximizes a certain objective function. The most common objectives are likelihood and posterior probability. Most algorithms for inference are model-specific and are nontrivial to be adapted to other models. A recent study by Than and Ho (2012b) reveals that there exists a highly scalable algorithm for sparse inference that can be easily adapted to various models. That algorithm is very flexible so that an adaptation is simply a choice of an appropriate objective function. Details are presented in Algorithm 1, in which $\Delta = \{\boldsymbol{x} \in \mathbb{R}^K : ||\boldsymbol{x}||_1 = 1, \boldsymbol{x} \geq 0\}$ denotes the unit simplex in the K-dimensional space. The following theorem indicates some important properties.

Theorem 1 *(Clarkson, 2010) Let f be a continuously differentiable, concave function over Δ, and denote C_f be the largest constant so that $f(\alpha\boldsymbol{x}'+(1-\alpha)\boldsymbol{x}) \geq f(\boldsymbol{x})+\alpha(\boldsymbol{x}'-\boldsymbol{x})^t\nabla f(\boldsymbol{x})-\alpha^2 C_f, \forall \boldsymbol{x}, \boldsymbol{x}' \in \Delta$, $\alpha \in [0,1]$. After ℓ iterations, the Frank-Wolfe algorithm finds a point $\boldsymbol{\theta}_\ell$ on an $(\ell+1)-$dimensional face of Δ such that $\max_{\boldsymbol{\theta}\in\Delta} f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}_\ell) \leq 4C_f/(\ell+3)$.*

Algorithm 1 FW framework

Input: document $\boldsymbol{d}$ and topics $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K$.
Output: latent representation $\boldsymbol{\theta}$.
Step 1: select an appropriate objective function $f(\boldsymbol{\theta})$ which is continuously differentiable, concave over Δ.
Step 2: maximize $f(\boldsymbol{\theta})$ over Δ by the Frank-Wolfe algorithm.

Algorithm 2 Frank-Wolfe algorithm

Input: objective function $f(\boldsymbol{\theta})$.
Output: $\boldsymbol{\theta}$ that maximizes $f(\boldsymbol{\theta})$ over Δ.
Pick as $\boldsymbol{\theta}_0$ the vertex of Δ with largest f value.
for $\ell = 0, ..., \infty$ **do**
  $i' := \arg\max_i \nabla f(\boldsymbol{\theta}_\ell)_i$;
  $\alpha' := \arg\max_{\alpha \in [0,1]} f(\alpha \boldsymbol{e}_{i'} + (1-\alpha)\boldsymbol{\theta}_\ell)$;
  $\boldsymbol{\theta}_{\ell+1} := \alpha' \boldsymbol{e}_{i'} + (1-\alpha')\boldsymbol{\theta}_\ell$.
end for

3. A two-steps framework for supervised dimension reduction

We now describe our framework for SDR. Existing methods for this problem often try to find directly a low-dimensional space that preserves separation of the data classes in the original space. For simplicity, we call that new space to be *discriminative space*. Different approaches have been employed such as maximizing the conditional likelihood (Lacoste-Julien et al., 2008), minimizing the empirical loss by max-margin principle (Zhu et al., 2012), or maximizing the joint likelihood of documents and labels (Blei and McAuliffe, 2007). Those are one-step algorithms to find the discriminative space, and bear resemblance to existing methods for continuous data (Parrish and Gupta, 2012; Sugiyama, 2007). Three noticeable drawbacks are that learning is very slow, that scalability of unsupervised models is not appropriately exploited, and more seriously, the inherent local structure of data is not taken into consideration.

To overcome those limitations of supervised topic models, we propose a novel framework which consists of two steps. Loosely speaking, the first step tries to find an initial topical space, while the second step tries to utilize label information and local structure of the training data to find the discriminative space. The first step can be done by employing a unsupervised topic model (Than and Ho, 2012a; Mimno et al., 2012), and hence inherits scalability of unsupervised models. Label information and local structure in the form of neighborhood will be used to guide projection of documents onto the initial space, so that inner-class local structure is preserved and inter-class margin is widen. As a consequence, the discrimination property is not only preserved, but likely made better in the final space.

Figure 1 depicts graphically this framework, and a comparison with other one-step methods. Note that we do not have to design entirely a learning algorithm as for existing approaches, but instead do one further inference step for the training documents. Details of our framework are presented in Algorithm 3. Each step from (2.1) to (2.4) will be detailed in the next subsections.

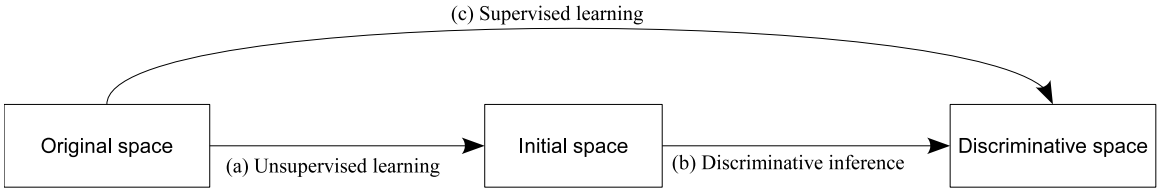

Figure 1: Sketch of approaches for SDR. Existing methods for SDR directly find the discriminative space, which is supervised learning (c). Our framework consists of two separate steps: (a) first find an initial space in a unsupervised manner; then (b) utilize label information and local structure of data to derive the final space.

Algorithm 3 Two-steps framework for supervised dimension reduction

Step 1: learn a unsupervised model to get K topics $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K$.
$\mathfrak{A} = span\{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K\}$ is the initial space.
Step 2: (finding discriminative space)
(2.1) for each class c, select a set S_c of topics which are potentially discriminative for c.
(2.2) for each document $\boldsymbol{d}$, select a set N_d of its nearest neighbors which are in the same class as $\boldsymbol{d}$.
(2.3) infer new representation $\boldsymbol{\theta}_d^*$ for each document $\boldsymbol{d}$ in class c by the FW framework with the objective function

$$f(\boldsymbol{\theta}) = \lambda.L(\widehat{\boldsymbol{d}}) + (1-\lambda).\frac{1}{|N_d|}\sum_{\boldsymbol{d}' \in N_d} L(\widehat{\boldsymbol{d}'}) + R.\sum_{j \in S_c} \sin(\theta_j),$$

where $L(\widehat{\boldsymbol{d}})$ is the log likelihood of document $\widehat{\boldsymbol{d}} = \boldsymbol{d}/||\boldsymbol{d}||_1$; $\lambda \in [0,1]$ and R are nonnegative constants.
(2.4) compute new topics $\boldsymbol{\beta}_1^*, ..., \boldsymbol{\beta}_K^*$ from all $\boldsymbol{d}$ and $\boldsymbol{\theta}_d^*$.
$\mathfrak{B} = span\{\boldsymbol{\beta}_1^*, ..., \boldsymbol{\beta}_K^*\}$ is the discriminative space.

3.1. Selection of discriminative topics

It is natural to assume that the documents in a class are talking about some specific topics which are little mentioned in other classes. Those topics are discriminative in the sense that they help us distinguish classes. Unsupervised models do not consider discrimination when learning topics, hence offer no explicit mechanism to see discriminative topics.

We use the following idea to find potentially discriminative topics: a topic that is discriminative for class c if its contribution to c is significantly greater than to other classes. The contribution of topic k to class c is approximated by

$$T_{ck} \propto \sum_{\boldsymbol{d} \in \mathcal{D}_c} \theta_{dk},$$

where $\mathcal{D}_c$ is the set of training documents in class c, $\boldsymbol{\theta}_d$ is the topic proportion of document $\boldsymbol{d}$ which had been inferred previously from a unsupervised model. We assume that topic k is discriminative for class c if

$$\frac{T_{ck}}{\min\{T_{1k}, ..., T_{Ck}\}} \geq \epsilon, \tag{8}$$

where C is the total number of classes, ϵ is a constant which is not smaller than 1.

ϵ can be interpreted as the boundary to differentiate which classes a topic is discriminative for. For intuition, considering the problem with 2 classes, condition (8) says that topic k is discriminative for class 1 if its contribution to k is at least ϵ times the contribution to class 2. If ϵ is too large, there is a possibility that a certain class might not have any discriminative topic. On the other hand, a too small value of ϵ may yield non-discriminative topics. Therefore, a suitable choice of ϵ is necessary. In our experiments we find that $\epsilon = 1.5$ is appropriate and reasonable. We further constraint $T_{ck} \geq median\{T_{1k}, ..., T_{Ck}\}$ to avoid the topic that contributes equally to most classes.

3.2. Selection of nearest neighbors

The use of nearest neighbors in Machine Learning have been investigated by various researches (Wu et al., 2012; Huh and Fienberg, 2012; Cai et al., 2009). Existing investigations often measure proximity of data points by cosine or Euclidean distances. In contrast, we use the Kullback-Leibler divergence (KL). The reason comes from the fact that projection/inference of a document onto the topical space inherently uses KL divergence.[2] Hence the use of KL divergence to find nearest neighbors is more reasonable than that of cosine or Euclidean distances in topic modeling. Note that we find neighbors for a given document $\boldsymbol{d}$ within the class containing $\boldsymbol{d}$, i.e., neighbors are local and within-class. We use $KL(\boldsymbol{d}||\boldsymbol{d}')$ to measure proximity from $\boldsymbol{d}'$ to $\boldsymbol{d}$.

3.3. Inference for each document

Let S_c be the set of potentially discriminative topics of class c, and N_d be the set of nearest neighbors of a given document $\boldsymbol{d}$ which belongs to c. We next do inference for $\boldsymbol{d}$ again to find the new representation $\boldsymbol{\theta}_d^*$. At this stage, inference is not done by existing method of the unsupervised model in consideration. Instead, the FW framework is employed, with the following objective function to be maximized:

$$f(\boldsymbol{\theta}) = \lambda L(\widehat{\boldsymbol{d}}) + (1-\lambda)\frac{1}{|N_d|}\sum_{\boldsymbol{d}' \in N_d} L(\widehat{\boldsymbol{d}}') + R\sum_{j \in S_c}\sin(\theta_j), \tag{9}$$

where $L(\widehat{\boldsymbol{d}}) = \sum_{j=1}^{V} \hat{d}_j \log \sum_{k=1}^{K} \theta_k \beta_{kj}$ is the log likelihood of document $\widehat{\boldsymbol{d}} = \boldsymbol{d}/||\boldsymbol{d}||_1$; $\lambda \in [0, 1]$ and R are nonnegative constants.

It is worthwhile making some observations about implication of this choice of objective:

- First, note that function $\sin(x)$ monotonically increases as x increases from 0 to 1. Therefore, the last term of (9) implies that we are promoting contributions of the topics in S_c to document $\boldsymbol{d}$. In other words, since $\boldsymbol{d}$ belongs to class c and S_c contains the topics which are potentially discriminative for c, the projection of $\boldsymbol{d}$ onto the topical space should remain large contributions of the topics of S_c. Increasing the constant R implies heavier promotion of contributions of the topics in S_c.

2. For instance, consider inference of document $\boldsymbol{d}$ by maximum likelihood. Inference is the problem $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} L(\widehat{\boldsymbol{d}}) = \arg\max_{\boldsymbol{\theta}} \sum_{j=1}^{V} \hat{d}_j \log \sum_{k=1}^{K} \theta_k \beta_{kj}$, where $\hat{d}_j = d_j/||\boldsymbol{d}||_1$. Denoting $\boldsymbol{x} = \boldsymbol{\beta\theta}$, the inference problem is reduced to $\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} \sum_{j=1}^{V} \hat{d}_j \log x_j = \arg\min_{\boldsymbol{x}} KL(\widehat{\boldsymbol{d}}||\boldsymbol{x})$. This implies inference of a document inherently uses KL divergence.

- Second, the term $\frac{1}{|N_d|}\sum_{\boldsymbol{d}'\in N_d} L(\widehat{\boldsymbol{d}'})$ implies that the local neighborhood plays a role when projecting $\boldsymbol{d}$. The smaller the constant λ, the more heavily the neighborhood plays. Hence, this additional term ensures that the local structure of data in the original space should not be violated in the new space.

- In practice, we do not have to store all neighbors of a document in order to do inference. Indeed, storing the mean $\boldsymbol{v} = \frac{1}{|N_d|}\sum_{\boldsymbol{d}'\in N_d} \widehat{\boldsymbol{d}'}$ is sufficient, since $\frac{1}{|N_d|}\sum_{\boldsymbol{d}'\in N_d} L(\widehat{\boldsymbol{d}'}) = \frac{1}{|N_d|}\sum_{\boldsymbol{d}'\in N_d}\sum_{j=1}^{V} \hat{d}'_j \log\sum_{k=1}^{K}\theta_k\beta_{kj} = \sum_{j=1}^{V}\left(\frac{1}{|N_d|}\sum_{\boldsymbol{d}'\in N_d}\hat{d}'_j\right)\log\sum_{k=1}^{K}\theta_k\beta_{kj}$.

- It is easy to verify that $f(\boldsymbol{\theta})$ is continuously differentiable and concave over the unit simplex Δ if $\boldsymbol{\beta} > 0$. As a result, the FW framework can be seamlessly employed for inference. Theorem 1 guarantees that inference of each document is very fast and the inference error is provably good. The following corollary states formally that property.

Corollary 2 *Consider a document $\boldsymbol{d}$, and K topics $\boldsymbol{\beta} > 0$. Let C_f be defined as in Theorem 1 for the function $f(\boldsymbol{\theta}) = \lambda L(\widehat{\boldsymbol{d}}) + (1-\lambda)\frac{1}{|N_d|}\sum_{\boldsymbol{d}'\in N_d} L(\widehat{\boldsymbol{d}'}) + R\sum_{j\in S_c}\sin(\theta_j)$, where $\lambda \in [0,1]$ and R are nonnegative constants. Then inference by FW converges to the optimal solution with a linear rate. In addition, after L iterations, the inference error is at most $4C_f/(L+3)$, and the topic proportion $\boldsymbol{\theta}$ has at most $L+1$ non-zero components.*

3.4. Computing new topics

One of the most involved parts in our framework is to find the final space from the old and new representations of documents. PLSA and LDA do not provide a direct way to compute topics from $\boldsymbol{d}$ and $\boldsymbol{\theta}^*_d$, while FSTM provides a natural one. We use (7) to find the discriminative space for FSTM,

$$\text{FSTM:} \qquad \beta^*_{kj} \propto \sum_{\boldsymbol{d}\in\mathcal{D}} d_j\theta^*_{dk}; \tag{10}$$

and use the following adaptations to compute topics for PLSA and LDA:

$$\text{PLSA:} \qquad \tilde{P}(z_k|\boldsymbol{d}, w_j) \propto \theta^*_{dk}\beta_{kj}, \tag{11}$$

$$\beta^*_{kj} \propto \sum_{\boldsymbol{d}\in\mathcal{D}} d_j\tilde{P}(z_k|\boldsymbol{d}, w_j); \tag{12}$$

$$\text{LDA:} \qquad \phi^*_{djk} \propto \beta_{kw_j}\exp\Psi(\theta^*_{dk}), \tag{13}$$

$$\beta^*_{kj} \propto \sum_{\boldsymbol{d}\in\mathcal{D}} d_j\phi^*_{djk}. \tag{14}$$

Note that we use the topics of the unsupervised models which had been learned previously in order to find the final topics. As a consequence, this usage provides a chance for unsupervised topics to affect discrimination of the final space. In contrast, using (10) to compute topics for FSTM does not encounter this drawback, and hence can inherit discrimination of $\boldsymbol{\theta}^*$. For LDA, the new representation $\boldsymbol{\theta}^*_d$ is temporarily considered to be variational parameter in place of $\boldsymbol{\gamma}_d$ in (4), and is smoothed by a very small constant to make sure the existence of $\Psi(\theta^*_{dk})$. Other adaptations are possible to find $\boldsymbol{\beta}^*$, nonetheless, we observe that our proposed adaptation is very reasonable. The reason is that computation of $\boldsymbol{\beta}^*$ uses as

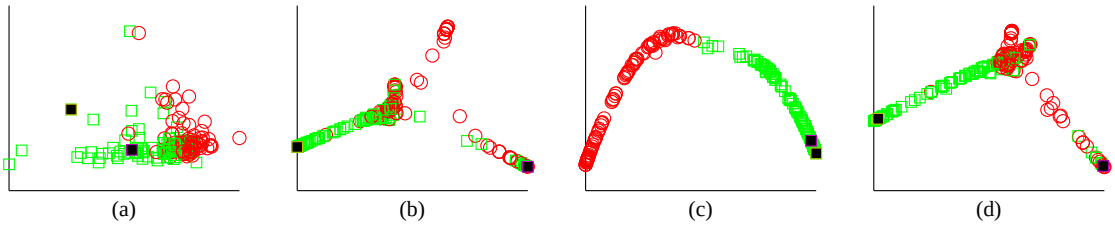

Figure 2: Laplacian embedding in 2D space. (a) data in the original space, (b) unsupervised projection, (c) projection when neighborhood is taken into account, (d) projection when topics are promoted. These projections onto the 60-dimensional space were done by FSTM and experimented on 20Newsgroups. The two black squares are documents in the same class.

little information from unsupervised models as possible, whereas inheriting label information and local structure encoded in $\boldsymbol{\theta}^*$, to reconstruct the final space $\mathfrak{B} = span\{\boldsymbol{\beta}_1^*, ..., \boldsymbol{\beta}_K^*\}$. This reason is further supported by extensive experiments as discussed later.

4. Why the framework is good?

We next theoretically elucidate the main reasons for why our proposed framework is reasonable and can result in a good method for SDR. In our observations, the most important reason comes from the choice of the objective (9) for inference. Inference with that objective plays two crucial roles to preserve the discrimination property of data in the topical space.

The first role is to preserve inner-class local structure of data. This is a result of the use of the additional term $\frac{1}{|N_d|}\sum_{\boldsymbol{d}' \in N_d} L(\widehat{\boldsymbol{d}'})$. Remember that projection of document $\boldsymbol{d}$ onto the unit simplex Δ is in fact a search for the point $\boldsymbol{\theta}_d \in \Delta$ that is closest to $\boldsymbol{d}$ in a certain sense.[3] Hence if $\boldsymbol{d}'$ is close to $\boldsymbol{d}$, it is natural to expect that $\boldsymbol{d}'$ is close to $\boldsymbol{\theta}_d$. To respect this nature and to keep the discrimination property, projecting a document should take its local neighborhood into account. As one can realize, the part $\lambda L(\widehat{\boldsymbol{d}}) + (1-\lambda)\frac{1}{|N_d|}\sum_{\boldsymbol{d}' \in N_d} L(\widehat{\boldsymbol{d}'})$ in the objective (9) serves well our needs. This part interplays goodness-of-fit and neighborhood preservation. Increasing λ means goodness-of-fit $L(\widehat{\boldsymbol{d}})$ can be improved, but local structure around $\boldsymbol{d}$ is prone to be broken in the low-dimensional space. Decreasing λ implies better preservation of local structure. Figure 2 demonstrates sharply these two extremes, $\lambda = 1$ for (b), and $\lambda = 0.1$ for (c). Projection by unsupervised models ($\lambda = 1$) often results in pretty overlapping classes in the topical space, whereas exploitation of local structure significantly helps us separate classes.

The second role is to widen the inter-class margin, owing to the term $R\sum_{j \in S_c} \sin(\theta_j)$. As noted before, function $\sin(x)$ is monotonically increasing for $x \in [0, 1]$. It implies that the term $R\sum_{j \in S_c} \sin(\theta_j)$ promotes contributions of the topics in S_c when projecting document $\boldsymbol{d}$. In other words, the projection of $\boldsymbol{d}$ is encouraged to be close to the topics which are potentially discriminative for class c. Hence projection of class c is preferred to distributing around the discriminative topics of c. Increasing the constant R implies forcing projections

3. More precisely, the vector $\sum_k \theta_{dk}\boldsymbol{\beta}_k$ is closest to $\boldsymbol{d}$ in terms of KL divergence.

Table 1: Statistics of data for experiments

Data	Training size	Testing size	Dimensions	Classes
20Newsgroups	15935	3993	62061	20
Emailspam	3461	866	38729	2

to distribute more densely around the discriminative topics, and therefore making classes farther from each other. Figure 2(d) illustrates the benefit of this second role.

5. Evaluation

This section is dedicated to investigation of effectiveness and efficiency of our framework in practice. We investigate three methods, PLSA^c, LDA^c, and FSTM^c, which are the results of adapting our framework to unsupervised models, PLSA (Hofmann, 2001), LDA (Blei et al., 2003), and FSTM (Than and Ho, 2012a), respectively. To see advantages of our framework, we take MedLDA (Zhu et al., 2012) as the state-of-the-art method for SDR into comparison.[4] Two benchmark data sets were used in our investigations: 20Newsgroups consisting of 19396 postings in 20 categories; Emailspam consisting of 4327 emails.[5] After preprocessing and removing stopwords and rare terms, the final corpora are detailed in Table 1.

In our experiments, we used the same criteria for topic models: relative improvement of the log likelihood (or objective function) is less than 10^{-4} for learning, and 10^{-6} for inference; at most 1000 iterations are allowed to do inference. The same criterion was used to do inference by FW in Step 2 of Algorithm 3. MedLDA is a supervised topic model and is trained by minimizing a hinge loss. We used the best setting as studied by Zhu et al. (2012) for some other parameters: cost parameter $\ell = 32$, and 10-fold cross-validation for finding the best choice of the regularization constant C in MedLDA. These settings are to avoid a biased comparison.

It is worth noting that our framework plays the main role in searching for the discriminative space $\mathfrak{B}$. Hence, other works aftermath such as projection/inference new documents are done by unsupervised models. For instances, FSTM^c works as follows: we first train FSTM in a unsupervised manner to get an initial space $\mathfrak{A}$; we next do Step 2 of Algorithm 3 to find the discriminative space $\mathfrak{B}$; projection of documents onto $\mathfrak{B}$ then is done by the inference method of FSTM.

5.1. Class separation, quality, and time

Separation of classes in low-dimensional spaces is our first concern. A good method for SDR should preserve inter-class separation of data in the original space. Figure 3 depicts an illustration of how good different methods are. In this experiment, 60 topics were used to

4. MedLDA was retrieved from http://www.ml-thu.net/~jun/code/MedLDAc/medlda.zip
LDA was taken from http://www.cs.princeton.edu/~blei/lda-c/
FSTM was taken from http://www.jaist.ac.jp/~s1060203/codes/fstm/
PLSA was written by ourselves with the best effort.
5. 20Newsgroups was taken from http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/. Emailspam was taken from http://csmining.org/index.php/spam-email-datasets-.html

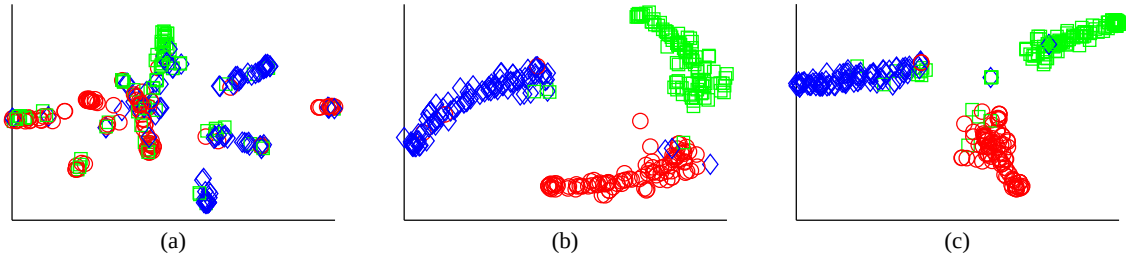

Figure 3: Projection of three classes of 20newsgroups onto the topical space by (a) FSTM, (b) FSTM^c, and (c) MedLDA. FSTM did not provide a good projection in the sense of class separation, since label information was ignored. FSTM^c and MedLDA actually found good discriminative topical spaces, and provided a good separation of classes.

train FSTM and MedLDA.[6] One can observe that projection by FSTM can maintain separation between classes to some extent. Nonetheless, because of ignoring label information, a large number of documents have been projected onto incorrect classes. On the contrary, FSTM^c and MedLDA exploited seriously label information for projection, and hence the classes in the topical space separate very cleanly. The good preservation of class separation by MedLDA is mainly due to the training algorithm by max margin principle. Each iteration of the algorithm tries to widen the expected margin between classes. Hence such an algorithm implicitly inherits the discrimination property in the topical space. FSTM^c can separate the classes well owing to the fact that projecting documents has taken local neighborhood into account seriously, which very likely keeps inter-class separation of the original data. Furthermore, it also tries to widen the margin between classes as discussed in Section 4.

Classification quality: we next use classification as a means to quantify the goodness of the considered methods for SDR. The main role of methods for SDR is to find a low-dimensional space so that projection of data onto that space preserves or even makes better the discrimination property of data in the original space. In other words, predictiveness of the response variable is preserved or improved. Classification is a good way to see this preservation or improvement.

For each method, we projected the training and testing data ($\boldsymbol{d}$) onto the topical space, and then used the associated projections ($\boldsymbol{\theta}$) as inputs for multi-class SVM (Keerthi et al., 2008) to do classification.[7] MedLDA does not need to be followed by SVM since it can do classification itself. We also included SVM which worked on the original space to see clearly the advantages of our framework. Keeping the same setting as described before and varying the number of topics, the results are presented in Figure 4.

6. For our framework, we set $N_d = 20, \lambda = 0.1, R = 1000$. This setting basically says that local neighborhood plays a heavy role when projecting documents, and that classes are very encouraged to be far from each other in the topical space.
7. This classification method is included in Liblinear package which is available at http://www.csie.ntu.edu.tw/∼cjlin/liblinear/

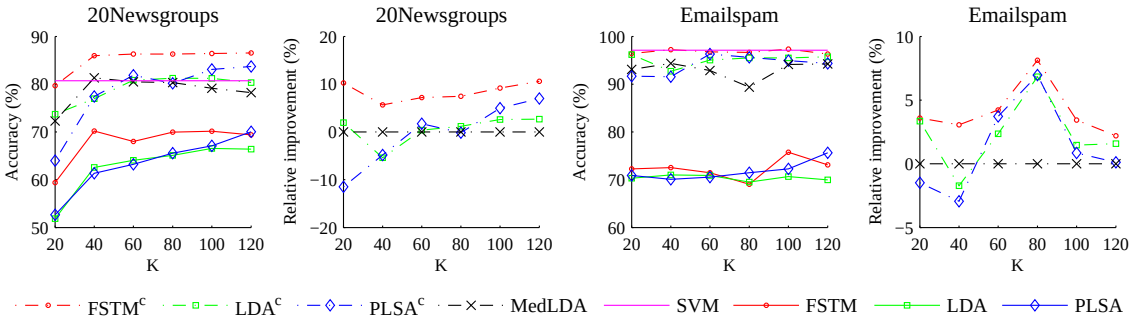

Figure 4: Accuracy of 8 methods as the number K of topics increases. Relative improvement is improvement of a method (A) over the-state-of-the-art MedLDA, and is defined as $\frac{accuracy(A)-accuracy(MedLDA)}{accuracy(MedLDA)}$. SVM worked on the original space.

Observing the figure, one easily realizes that the supervised methods consistently performed substantially better than the unsupervised ones. This suggests that FSTM^c, LDA^c, PLSA^c, and MedLDA exploited well label information when searching for a topical space. Sometimes, they even performed better than SVM which worked on the original high-dimensional space. FSTM^c, LDA^c, and PLSA^c performed better than MedLDA when the number of topics is relatively large (≥ 60). FSTM^c consistently achieved the best performance amongst topic-model-based methods, and sometimes reached 10% improvement over the-state-of-the-art MedLDA. In our observations, this improvement is mainly due to the fact that FSTM^c had taken seriously local structure of data into account whereas MedLDA did not. Ignoring local structure in searching for a topical space could harm or break the discrimination property of data. This could happen with MedLDA even though learning by max margin principle is well-known to keep good classification quality. Besides, FSTM^c even significantly outperformed SVM on 20Newsgroups, while performed comparably on Emailspam. These results support further our analysis in Section 4.

Why FSTM^c often performs best amongst three adaptations including FSTM^c, LDA^c, and PLSA^c? This question is natural, since our adaptations for three topic models use the same framework and settings. In our observations, the key reason comes from the way of deriving the final space in Step 2 of our framework. As noted before, deriving topical spaces by (12) and (14) directly requires unsupervised topics of PLSA and LDA, respectively. Such adaptations implicitly allow some chances for unsupervised topics to have direct influence on the final topics. Hence the discrimination property may be affected heavily in the new space. On the contrary, using (10) to recompute topics for FSTM does not allow a direct involvement of unsupervised topics. Therefore, the new topics can inherit almost the discrimination property encoded in $\boldsymbol{\theta}^*$. This helps the topical space found by FSTM^c is more likely discriminative than those by PLSA and by LDA. Another reason is that the inference method of FSTM is provably good (Than and Ho, 2012a), and is often more accurate than that of LDA and PLSA (Than and Ho, 2012b).

Learning time: the final measure for comparison is how quickly the methods do? We mostly concern methods for SDR including FSTM^c, LDA^c, PLSA^c, and MedLDA. Note

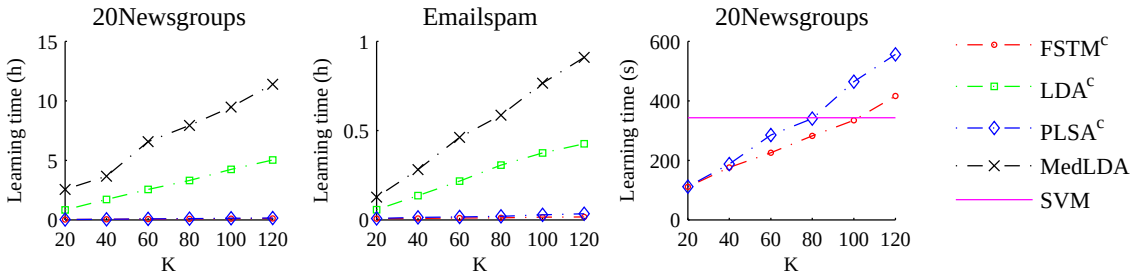

Figure 5: Necessary time to learn a discriminative space, as the number K of topics increases. SVM is included for reference, where we recorded the time for learning a classifier from the given training data.

that the time for learning a discriminative space by FSTM^c is the time to do 2 steps of Algorithm 3 which includes time to learn a unsupervised model, FSTM. The same holds for PLSA^c and LDA^c. Figure 5 summarizes the overall time for each method. Observing the figure, we find that MedLDA and LDA^c consumed intensive time, while FSTM^c and PLSA^c did substantially more speedily. One of the main reasons for slow learning of MedLDA and LDA^c is that inference by variational methods of MedLDA and LDA is often very slow. Inference in those models requires various evaluation of Digamma and gamma functions which are expensive. Further, MedLDA requires a further step of learning a classifier at each EM iteration, which is empirically slow in our observations. All of these contributed to the slow learning of MedLDA and LDA^c.

In contrast, FSTM has a linear time inference algorithm and requires simply a multiplication of two sparse matrices for learning topics, while PLSA has a very simple learning formulation. Hence learning in FSTM and PLSA is unsurprisingly very fast (Than and Ho, 2012a). The most time consuming part of FSTM^c and PLSA^c is to search nearest neighbors for each document. A modest implementation would requires $O(V.M^2)$ arithmetic operations, where M is the data size. Such a computational complexity will be problematic when the data size is large. Nonetheless, as empirically shown in Figure 5, the overall time of FSTM^c and PLSA^c was significantly less than that of MedLDA and LDA^c. Even for 20Newsgroups of average size, learning time of FSTM^c and PLSA^c is very competitive compared with MedLDA.

Summarizing, the above investigations demonstrate that the proposed framework can result in very competitive methods for SDR. Three methods, FSTM^c, LDA^c, and PLSA^c, have been observed to significantly outperform their corresponding unsupervised models. LDA^c and PLSA^c reached comparable performance with the state-of-the-art method, MedLDA, when the number of topics is not small. Amongst three adaptations, FSTM^c behaved superior in both classification performance and learning speed. Classification in the low-dimensional space found by FSTM^c is often comparable or better than that in the original high-dimensional space.

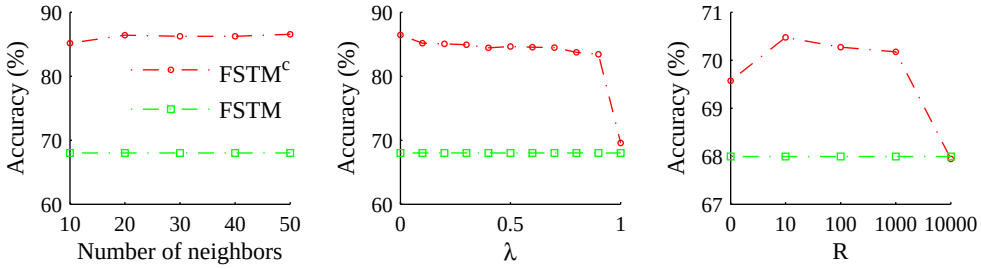

Figure 6: Impact of the parameters on the success of our framework. (left) Change the number of neighbors, while fixing $\lambda = 0.1, R = 0$. (middle) Change λ the extent of seriousness of taking local structure, while fixing $R = 0$ and using 10 neighbors for each document. (right) Change R the extent of promoting topics, while fixing $\lambda = 1$. Note that the interference of local neighborhood played a very important role, since it consistently resulted in significant improvements.

5.2. Sensitivity

There are three parameters that influence the success of our framework, including the number of nearest neighbors, λ, and R. This subsection investigates the impact of each. 20Newsgroups was selected for experiments, since it has average size which is expected to exhibit clearly and accurately what we want to see.

We varied the value of a parameter while fixed the others, and then measured the accuracy of classification. Figure 6 presents the results of these experiments. It is easy to realize that when taking local neighbors into account, the classification performance was very high and significant improvements can be achieved. We observed that very often, 25% improvement were reached when local structure was used, even with different settings of λ. These observations suggest that the use of local structure plays a very crucial role for the success of our framework. It is worth remarking that one should not use too many neighbors for each document, since performance may be worse. The reason is that using too many neighbors likely break local structure around documents. We have experienced with this phenomenon when setting 100 neighbors in Step 2 of Algorithm 3, and got worse results.

Changing the value of R implies changing promotion of topics. In other words, we are expecting projections of documents in the new space to distribute more densely around discriminative topics, and hence making classes farther from each other. As shown in Figure 6, an increase in R often leads to better results. However, too large R can deteriorate the performance of the SDR method. The reason may be that such large R can make the term $R\sum_{j\in S_c}\sin(\theta_j)$ to overwhelm the objective (9), and thus worsen the goodness-of-fit of inference by FW. Setting $R \in [10, 1000]$ is reasonable in our observation.

6. Conclusion and discussion

We have proposed a framework for doing dimension reduction of supervised discrete data. The framework was demonstrated to exploit well label information and local structure of the

training data to find a discriminative low-dimensional space. Generality and flexibility of our framework was evidenced by adaptation to three unsupervised topic models, resulted in PLSA^c, LDA^c, and FSTM^c. These methods for supervised dimension reduction (SDR) can perform qualitatively comparably with the state-of-the-art method, MedLDA. In particular, FSTM^c performed significantly best and can often achieve more than 5% improvement over MedLDA. Working on the low-dimensional space found by FSTM^c is often comparable or better than working on the original space of data. Meanwhile, FSTM^c consumes substantially less time than MedLDA does. These results show that our framework can inherit scalability of unsupervised models to yield qualitatively competitive methods for SDR.

There is a number of possible extensions to our framework. First, one can easily modify the framework to deal with multilabel data. Second, the framework can be modified to deal with semi-supervised data. A key to these extensions is an appropriate utilization of labels to search for nearest neighbors, which is necessary for our framework. Other extensions can encode more prior knowledge into the objective function for inference. In our framework, label information and local neighborhood are encoded into the objective function and have been observed to work well. Hence, we believe that other prior knowledge can be used to derive good methods.

Of the most expensive steps in our framework is the search for nearest neighbors. By a modest implementation, it requires $O(k.V.M)$ to search k nearest neighbors for a document. Overall, finding all k nearest neighbors for all documents requires $O(k.V.M^2)$. This computational complexity will be problematic when the number of training documents is large. Hence, a significant extension would be to reduce running time for this search. It is possible to reduce the complexity to $O(k.V.M.\log M)$ as suggested by Arya et al. (1998). Furthermore, because our framework use local neighborhood to guide projection of documents onto the low-dimensional space, we believe that approximation to local structure can still provide good result. However, this assumption should be studied further. A positive point of using approximation of local neighborhood is that computational complexity of a search for neighbors can be done in linear time $O(k.V.M)$ (Clarkson, 1983).

References

Sunil Arya, David M. Mount, Nathan S. Netanyahu, Ruth Silverman, and Angela Y. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM*, 45(6):891–923, 1998. ISSN 0004-5411. doi: 10.1145/293347.293348.

David Blei and Jon McAuliffe. Supervised topic models. In *Neural Information Processing Systems (NIPS)*, 2007.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(3):993–1022, 2003.

Deng Cai, Xuanhui Wang, and Xiaofei He. Probabilistic dyadic data analysis with local and global consistency. In *ICML*, pages 105–112. ACM, 2009.

Minhua Chen, W. Carson, M. Rodrigues, R. Calderbank, and L. Carin. Communication inspired linear discriminant analysis. In *Proceedings of the 29th Annual International Conference on Machine Learning (ICML)*, 2012.

Kenneth L. Clarkson. Fast algorithms for the all nearest neighbors problem. *FOCS*, pages 226–232, 1983. doi: http://doi.ieeecomputersociety.org/10.1109/SFCS.1983.16.

Kenneth L. Clarkson. Coresets, sparse greedy approximation, and the frank-wolfe algorithm. *ACM Trans. Algorithms*, 6:63:1–63:30, September 2010.

Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196, 2001.

Seungil Huh and Stephen Fienberg. Discriminative topic modeling based on manifold learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4):20, 2012.

S.S. Keerthi, S. Sundararajan, K.W. Chang, C.J. Hsieh, and C.J. Lin. A sequential dual method for large scale multi-class linear svms. In *KDD*, pages 408–416. ACM, 2008.

S. Lacoste-Julien, F. Sha, and M.I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *NIPS*, volume 21, pages 897–904. MIT, 2008.

David Mimno, Matthew D. Hoffman, and David M. Blei. Sparse stochastic inference for latent dirichlet allocation. In *ICML*, 2012.

Nathan Parrish and Maya R. Gupta. Dimensionality reduction by local discriminative gaussian. In *ICML*, 2012.

Alexander Smola and Shravan Narayanamurthy. An architecture for parallel topic models. *Proceedings of the VLDB Endowment*, 3(1-2):703–710, 2010.

Masashi Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *The Journal of Machine Learning Research*, 8:1027–1061, 2007.

Khoat Than and Tu Bao Ho. Fully sparse topic models. In *ECML PKDD*. LNCS vol. 7523, pages 490–505. Springer, 2012a.

Khoat Than and Tu Bao Ho. Managing sparsity, time, and quality of inference in topic models. Technical report, 2012b.

H. Wu, J. Bu, C. Chen, J. Zhu, L. Zhang, H. Liu, C. Wang, and D. Cai. Locally discriminative topic modeling. *Pattern Recognition*, 45(1):617–625, 2012.

Ying Yang and Geoffrey Webb. Discretization for naive-bayes learning: managing discretization bias and variance. *Machine learning*, 74(1):39–74, 2009.

Jun Zhu, Amr Ahmed, and Eric P. Xing. Medlda: maximum margin supervised topic models. *The Journal of Machine Learning Research*, 13:2237–2278, 2012.