

Large-Margin Predictive Latent Subspace Learning for Multiview Data Analysis

Ning Chen, Jun Zhu, *Member, IEEE*, Fuchun Sun, *Senior Member, IEEE*, and Eric Poe Xing, *Senior Member, IEEE*

Abstract—Learning salient representations of multiview data is an essential step in many applications such as image classification, retrieval, and annotation. Standard predictive methods, such as support vector machines, often directly use all the features available without taking into consideration the presence of distinct views and the resultant view dependencies, coherence, and complementarity that offer key insights to the semantics of the data, and are therefore offering weak performance and are incapable of supporting view-level analysis. This paper presents a statistical method to learn a predictive subspace representation underlying multiple views, leveraging both multiview dependencies and availability of supervising side-information. Our approach is based on a multiview latent subspace Markov network (MN) which fulfills a weak conditional independence assumption that multiview observations and response variables are conditionally independent given a set of latent variables. To learn the latent subspace MN, we develop a large-margin approach which jointly maximizes data likelihood and minimizes a prediction loss on training data. Learning and inference are efficiently done with a contrastive divergence method. Finally, we extensively evaluate the large-margin latent MN on real image and hotel review datasets for classification, regression, image annotation, and retrieval. Our results demonstrate that the large-margin approach can achieve significant improvements in terms of prediction performance and discovering predictive latent subspace representations.

Index Terms—Latent subspace model, large-margin learning, classification, regression, image retrieval and annotation

1 INTRODUCTION

MODERN data analytic problems in social media, information technology, and natural sciences often involve rich data consisting of multiple information modalities. For example, in a moment-sharing social network such as Instagram, a photo record would include image, text (status updates and viewer opinions), and various meta-information such as user demographics, geotags, time stamps, etc.; in a biomedical data repository, a clinical sample record may include gene expression intensity, protein activity status, clinical traits, and patient information with family history. These different modalities represent different angles to reveal the fundamental characteristics and properties of the study subjects and is often referred as *views* of the subjects.

Proper integration of multiple views present in multimodal data is of paramount importance for seeking accurate distillation of salient semantic representations of the study objects; therefore numerous efforts along this direction can be

found in the literature. To name a few, Blum and Mitchell [6] studied co-training scheme of a classification model for webpages based on both content and link anchor text; Xing et al. [44] proposed a dual view latent space model for video shot based on both color/shape of the keyframe and the corresponding closed captions; and this list continues to grow, under various contexts and addressing a diverse range of data forms [17], [11], [34], [35], [14]. However, most of these approaches for multiview integration and distillation do not go hand-in-hand with mainstream predictive methods such as support vector machines (SVMs) [8] or Boosting algorithms [19] to form a unified system that allows strongly *predictive* latent semantic representations of multiview data to be extracted. Typically, standard predictive methods would use one of the following strategies: 1) build a single classifier on observed features from all views, without taking into consideration the presence of distinct views; 2) build a set of classifiers defined on each view, regardless of the relationships among views; and 3) let a latent space model such as a multiview topic model to distill the latent representations of data without considering the predictive information,¹ and then apply a downstream classifier on such representations [44]. While offering many insights on how multiview data can be worked with, these approaches appear to enjoy limited practical benefits from the extra information present in multiview data in terms of predictive performance [7], computational cost [35], and power for *view-level analysis* [14] such as predicting tags for image annotation or analyzing the underlying relationships among views.

Moreover, with the rapid increase of free online information such as user tagging, ratings, etc., various forms of side-information that can potentially offer “free”

• N. Chen, J. Zhu, and F. Sun are with the Department of Computer Science and Technology, Tsinghua National Laboratory of Information Science and Technology, State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, FIT Building, Haidian District, Beijing 100084, China. E-mail: {ningchen, dcszj, fcsun}@tsinghua.edu.cn.

• E.P. Xing is with the School of Computer Science, Carnegie Mellon University, 8101 GHC, Pittsburgh, PA 15213. E-mail: epxing@cs.cmu.edu.

Manuscript received 20 Dec. 2010; revised 30 Sept. 2011; accepted 25 Jan. 2012; published online 28 Feb. 2012.

Recommended for acceptance by T. Jebara.

For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2010-12-0974.

Digital Object Identifier no. 10.1109/TPAMI.2012.64.

supervision over the media data have led to a need for new models and training schemes that can make effective use of such information to achieve better results, such as more discriminative latent representations of image contents and more accurate image classifiers. In this paper, we develop a new statistical framework that enables one to learn a *predictive* latent space representation shared by multiview data by leveraging supervising side information, and to perform both view-level analysis (e.g., image annotation) and response-level predictions (e.g., classification) based on the learned representation.

Our method builds on a probabilistic latent subspace model that relates features of perceivable entities (e.g., images) to abstract concepts (e.g., latent topics) in a probabilistic way which allows flexible and efficient statistical reasoning and inference. Our model is a generic *multiview latent space Markov network* (MN) that builds on a weak conditional independence assumption that the data from different views and the response variables are conditionally independent given a set of latent variables. This conditional independence is weaker than the typical assumption (e.g., in the seminal work of co-training [6]) that multiview data are conditionally independent given the very low dimensional response variables [18]. Although in principle a directed Bayesian network (BN) (e.g., latent Dirichlet allocation (LDA) [5], [4], [48]) can be extended to handle multiview data, its conditional dependency properties could make it hard to perform posterior inference because all latent variables are coupled given observed variables [41]. In contrast, undirected latent variable models, such as ours and those presented in [41], [32], [44], could be very efficient in inference because of the conditional independence.

One critical limitation of existing paradigms for learning probabilistic latent subspace models, as discussed in [48], is that the commonly used likelihood-based methods are often not discriminative enough to leverage the supervising side information accompanying the multiview data to extract a strongly predictive representation, and is prone to undesirable effects such as overfitting to small data [41], [44], [45], [28]. To overcome such limitations, we use a completely different and arguably more desirable learning paradigm based on the maximum margin principle. More specifically, we develop a new discriminative learning approach for the proposed latent space Markov network which jointly maximizes the likelihood of multiview data and minimizes a prediction loss on the labels from side information (e.g., hinge-loss for classification or ϵ -insensitive loss [33] for regression) to discover a strongly predictive subspace representation and learn a prediction model thereupon. The learning and inference problems are efficiently solved with an extension of the contrastive divergence method [40]. Extensive experiments show that the proposed large-margin approach can achieve significant improvements in terms of prediction performance and semantic saliency of the predictive latent subspace representations. Moreover, the inference in the latent space MN is much faster and easier compared with the directed counterpart models, e.g., MedLDA [48].

The remainder of the paper is organized as follows: Section 2 reviews related work. Section 3 presents the multiview latent subspace MN. Section 4 presents the large-margin training methods for both classification and regression.

Section 5 presents extensive empirical evaluation on various datasets. Finally, Section 6 concludes with future research directions discussed.

2 RELATED WORK

The literature of discovering latent representations from large collections of data consists of both deterministic (e.g., canonical correlation analysis (CCA) [24], [26], [1] and Fisher discriminant analysis (FDA) [15]) and probabilistic (e.g., directed LDA [5], [39], [50] and undirected Harmoniums [41], [32], [44]) methods. A deterministic method cannot be easily extended to perform view-level predictions, such as image annotation, and it would also need a density estimator in order to apply the information criterion [11] to detect view disagreement. Thus, we choose the probabilistic framework and base our approach on an undirected multiview latent space model, which enjoys nice properties, as discussed.

To consider supervising side information, supervised latent space models have been developed, including supervised LDA [4], [39], [48] and supervised Harmoniums [45], [28]. However, almost all these models are learned using likelihood-based estimation, which often involves dealing with an intractable normalization factor [39], [50] and may not yield improvements compared with the standard prediction tools based on purely discriminative ideas (e.g., SVM) [45]. The recent work of MedLDA [48] has shown a promising direction of applying the large-margin principle to learn predictive latent space representations which could be more suitable for prediction (e.g., classification). Other developments along this line include the large-margin upstream scene understanding models [49] and the conditional topic models with features [50]. However, these methods are all directed Bayesian networks, which may involve a hard inference problem, as we have discussed. The present work represents an important contribution of deploying the large-margin principle to learn undirected latent space models.

The large-margin principle has also been applied to learn Markov networks with latent variables [51], [16], [46]. However, their goals are mainly to use latent variables to capture residual and high-order dependency for improving prediction performance, essentially different from ours of learning predictive latent representations of the data. Our approach is also different from much of the existing research that has been done on exploring multiview information to alleviate semi-supervised learning [6], [14], [2], [18], [26], unsupervised clustering [9], and structured output problems [21]. Other work that relates to ours includes the hybrid generative/discriminative learning [31], which uses likelihood-based estimation, and the sufficient dimensionality reduction methods [20]. Finally, this paper is a systematic extension of the preliminary conference version [10].

3 MULTIVIEW LATENT SUBSPACE MNS

In this section, we present a multiview latent subspace Markov network by incorporating complex structures on each view. We will start with an unsupervised latent subspace MN and then present a supervised latent subspace MN based on maximum likelihood estimation (MLE).

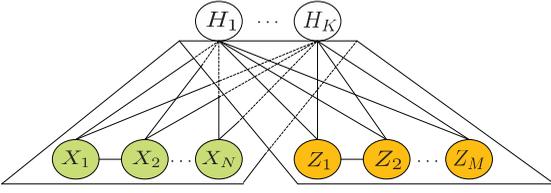


Fig. 1. An unsupervised two-view latent subspace MN.

3.1 Unsupervised Multiview Latent Subspace MNs

Fig. 1 shows the structure of a two-view latent subspace MN which consists of two types of input data $\mathbf{X} \triangleq \{X_i\}_{i=1}^N$ and $\mathbf{Z} \triangleq \{Z_j\}_{j=1}^M$, each corresponding to a view, and a set of latent variables $\mathbf{H} \triangleq \{H_k\}_{k=1}^K$, corresponding to the latent representations one desires to infer. We encode the structure of the variables on each view using a Markov network. Purely for simplicity of presentation, we focus on the case of pairwise interactions between variables within each view. We emphasize that our results easily extend to more general cases of higher order dependencies. Let E_x denote the set of edges² between the input variables \mathbf{X} , and likewise for E_z . We will use e to denote one individual edge and use \mathbf{X}_e to denote the variables associated with e .

A constructive way to define the joint distribution of a latent subspace MN is as follows: First, we define the distribution of the data on each view and the latent variables separately. For each view, we use an exponential family distribution:

$$\begin{aligned} p(\mathbf{x}) &= r(\mathbf{x}) \exp \left\{ \sum_{e \in E_x} \theta_e^\top \phi(\mathbf{x}_e) - A(\theta) \right\}, \\ p(\mathbf{z}) &= s(\mathbf{z}) \exp \left\{ \sum_{e \in E_z} \eta_e^\top \psi(\mathbf{z}_e) - B(\eta) \right\}, \end{aligned} \quad (1)$$

where ϕ and ψ are vectors of feature functions, θ and η are weights, and A and B are log partition functions. Like [41], we will treat $\log(r(\mathbf{x}))$ and $\log(s(\mathbf{z}))$ as additional features multiplied by a constant. For the latent variables \mathbf{H} , each component H_k has an exponential family distribution and

$$p(\mathbf{h}) = \prod_k p(h_k) = \prod_k \exp \{ \lambda_k^\top \varphi(h_k) - C_k(\lambda_k) \},$$

where $\varphi(h_k)$ is the vector of features of h_k . C_k is another log-partition function.

Then, the joint model distribution is defined by combining the above components in the log-domain and introducing additional terms that couple the random variables \mathbf{X} , \mathbf{Z} , and \mathbf{H} . Specifically, we have

$$\begin{aligned} p(\mathbf{x}, \mathbf{z}, \mathbf{h}) &\propto \exp \left\{ \sum_{e \in E_x} \theta_e^\top \phi(\mathbf{x}_e) + \sum_{e \in E_z} \eta_e^\top \psi(\mathbf{z}_e) + \sum_k \lambda_k^\top \varphi(h_k) \right. \\ &\quad \left. + \sum_{e \in E_{x,k}} \phi(\mathbf{x}_e)^\top \mathbf{W}_e^k \varphi(h_k) + \sum_{e \in E_{z,k}} \psi(\mathbf{z}_e)^\top \mathbf{U}_e^k \varphi(h_k) \right\}, \end{aligned} \quad (2)$$

2. We treat a singleton vertex as a degenerate edge.

where \mathbf{W} and \mathbf{U} are feature weights. From the joint distribution, we can derive the conditional distributions on each view with shifted parameters $(\hat{\theta}, \hat{\eta}, \hat{\lambda})$

$$\begin{aligned} p(\mathbf{x}|\mathbf{h}) &= \exp \left\{ \sum_{e \in E_x} \hat{\theta}_e^\top \phi(\mathbf{x}_e) - A(\hat{\theta}) \right\}, \\ p(\mathbf{z}|\mathbf{h}) &= \exp \left\{ \sum_{e \in E_z} \hat{\eta}_e^\top \psi(\mathbf{z}_e) - B(\hat{\eta}) \right\}, \\ p(\mathbf{h}|\mathbf{x}, \mathbf{z}) &= \prod_k \exp \{ \hat{\lambda}_k^\top \varphi(h_k) - C_k(\hat{\lambda}_k) \}, \end{aligned}$$

where $\hat{\theta}_e = \theta_e + \sum_k \mathbf{W}_e^k \varphi(h_k)$, $\hat{\eta}_e = \eta_e + \sum_k \mathbf{U}_e^k \varphi(h_k)$, and $\hat{\lambda}_k = \lambda_k + (\sum_{e \in E_x} \phi(\mathbf{x}_e)^\top \mathbf{W}_e^k + \sum_{e \in E_z} \psi(\mathbf{z}_e)^\top \mathbf{U}_e^k)^\top$. We can see that conditioned on the latent variables, both $p(\mathbf{x}|\mathbf{h})$ and $p(\mathbf{z}|\mathbf{h})$ define a Markov network, which is known as conditional random fields (CRFs) [27], where \mathbf{h} correspond to global conditions and \mathbf{x} or \mathbf{z} correspond to structured prediction variables in CRFs.

Reversely, one can start with defining the local conditional distributions as above and directly write the compatible joint distribution, which is of the log-linear form as in (2). In the sequel, we use Θ to denote all the parameters $(\theta, \eta, \lambda, \mathbf{W}, \mathbf{U})$. It is worth noting that both the exponential family Harmonium (EFH) [41] and its extension of dual-wing Harmonium (DWH) [44] are special cases of multiview latent subspace MNs when the generalized edge sets E_x and E_z contain only singleton vertices. Therefore, it is not surprising to see that multiview MNs inherit the widely advocated property of EFH that the model distribution can be constructively defined based on local conditionals on each view.

We briefly introduce DWH here as it sets up the ground for our experiments in Section 5. As in [44], DWH has a two-view structure, where \mathbf{X} is a vector of discrete word features (e.g., image tags) and \mathbf{Z} is a vector of real-valued features (e.g., color histograms). We assume that each X_i is a Bernoulli variable that denotes whether the i th term of a dictionary appears or not in an image, and each Z_j is a real number that denotes the normalized color histogram of an image. Each real-valued H_k follows a univariate Gaussian distribution. Therefore, the conditional distributions can be defined as

$$\begin{aligned} p(x_i = 1|\mathbf{h}) &= \text{Logistic}(\alpha_i + \mathbf{W}_i \mathbf{h}), \\ p(z_j|\mathbf{h}) &= \mathcal{N}(z_j | \sigma_j^2(\beta_j + \mathbf{U}_j \mathbf{h}), \sigma_j^2), \\ p(h_k|\mathbf{x}, \mathbf{z}) &= \mathcal{N}(h_k | \mathbf{x}^\top \mathbf{W}_{\cdot k} + \mathbf{z}^\top \mathbf{U}_{\cdot k}, 1), \end{aligned}$$

where \mathbf{W}_i and $\mathbf{W}_{\cdot k}$ denote the i th row and k th column of \mathbf{W} , respectively. Likewise for \mathbf{U}_j and $\mathbf{U}_{\cdot k}$.

To learn the unsupervised multiview latent subspace MNs, a natural method is the maximum likelihood estimation, which has been widely used to train directed [39], [47] and undirected latent variable models [41], [32], [44], [45]. To deal with the intractable log-likelihood $\log p(\mathbf{x}, \mathbf{z})$, an approximation method such as mean field or contrastive divergence [44] is usually applied. More details will be provided, along with the algorithm development for large-margin learning.

To use the unsupervised multiview MN for prediction (e.g., classification), a naive method is a two-stage procedure: 1) using the latent subspace MN to discover latent representations, and 2) feeding the latent representations into a downstream prediction model (e.g., SVM). This two-step procedure can be rather suboptimal for prediction because supervising information is ignored in discovering the latent representations. Moreover, as we have stated, supervising side information can be almost “free” to obtain; thus, it is desirable to develop new models and learning methods to consider such information for improving performance. Below, we present supervised latent subspace MNs which incorporate supervising side information into the procedure of discovering latent subspace representations. As we shall see, if learned appropriately, e.g., using large-margin training, a supervised latent subspace MN can achieve significant improvements in discovering predictive latent subspace representations and prediction performance.

3.2 Supervised Multiview Latent Subspace MNs

Similar to learning an unsupervised latent subspace MN, MLE is the natural method to learn a supervised latent subspace MN. In this section, we present the MLE-based supervised latent subspace MN, which would motivate our development of a large-margin approach.

In order to perform MLE, we need to define a likelihood model for observed data, including input features and response variables in the supervised case. Specifically, let Y be the response variable and \mathbf{V} be the parameters of a response variable model. Then, we need to define the joint distribution $p(\mathbf{x}, \mathbf{z}, \mathbf{h}, y)$. We consider univariate prediction, where Y can be a discrete variable for classification or a continuous variable for regression. Based on the constructive definition, we need to specify the conditional distribution of Y given \mathbf{H} in order to define $p(\mathbf{x}, \mathbf{z}, \mathbf{h}, y)$. For the general multiclass classification, where $y \in \{1, \dots, T\}$, we define the conditional distribution using a softmax function:³

$$p(y|\mathbf{h}) = \frac{\exp\{\mathbf{V}^\top \mathbf{f}(y, \mathbf{h})\}}{\sum_{y'} \exp\{\mathbf{V}^\top \mathbf{f}(y', \mathbf{h})\}}, \quad (3)$$

where $\mathbf{f}(y, \mathbf{h})$ is the feature vector whose elements from $(y-1)K+1$ to yK are those of \mathbf{h} and all others are 0. \mathbf{V} is a stacking parameter vector of T subvectors $\mathbf{V}_{y'}$, of which each one corresponds to a class label y' . Then, the joint distribution $p(\mathbf{x}, \mathbf{z}, \mathbf{h}, y)$ has the same form as in (2), but with an additional term of $\mathbf{V}^\top \mathbf{f}(y, \mathbf{h}) = \mathbf{V}_y^\top \mathbf{h}$ in the exponential. For regression, where $y \in \mathbb{R}$, we define the conditional distribution as a normal distribution

$$p(y|\mathbf{h}) = \mathcal{N}(y|\mathbf{V}^\top \mathbf{h}, \sigma^2), \quad (4)$$

where \mathbf{V} is now a K -dim vector. Then, the joint distribution $p(\mathbf{x}, \mathbf{z}, \mathbf{h}, y)$ has a similar form as in (2) with an additional term of $-\frac{1}{2\sigma^2}(y^2 - y\mathbf{V}^\top \mathbf{h})$ in the exponential.

Note that the supervised hierarchical (or tri-wing) Harmonium (TWH) [45] is a special case of the supervised latent subspace MN for classification. With the above joint

likelihood function, we can perform standard MLE by using contrastive divergence or mean field approximation to learn the parameters.⁴ The procedure is generally similar to that in learning TWH [45]. The major difference lies in posterior inference, which will be clear after we have presented the large-margin learning.

4 LARGE-MARGIN SUPERVISED MULTIVIEW LATENT SUBSPACE MNS

As stated above, the MLE-based supervised latent subspace MN requires defining a normalized distribution as in (3), of which the normalization factor could make the inference hard, especially in directed models [39], [50]. Moreover, as shown in [45] and our empirical studies, the MLE-based model may not obtain improvements over the naive two-step method discussed at the end of Section 3.1. These motivate us to develop a more discriminative procedure for learning supervised latent subspace MNs. In this section, we present a large-margin supervised latent subspace MN for discovering predictive latent subspace representations from multiview data by incorporating the widely available supervising side information, which can be discrete for classification or continuous for regression.

4.1 Classification

We first present the classification model. For brevity, we consider the general multiclass classification. The binary case can be similarly derived.

4.1.1 Problem Definition

Similar to the log-linear model in (3), we define the *latent discriminant function* $F(y, \mathbf{h}; \mathbf{V})$ as linear when latent variables \mathbf{H} are given, that is, $F(y, \mathbf{h}; \mathbf{V}) = \mathbf{V}^\top \mathbf{f}(y, \mathbf{h})$, where \mathbf{f} and \mathbf{V} are defined the same as in (3). Now, the problem is how to consider the uncertainty of \mathbf{H} in the deterministic large-margin principle. Here, we take the expectation (i.e., first moment) of the latent variables \mathbf{H} and define the *expected prediction rule*:

$$\begin{aligned} y^* &\triangleq \arg \max_y \mathbb{E}_{p(\mathbf{h}|\mathbf{x}, \mathbf{z})} [F(y, \mathbf{h}; \mathbf{V})] \\ &= \arg \max_y \mathbf{V}^\top \mathbb{E}_{p(\mathbf{h}|\mathbf{x}, \mathbf{z})} [\mathbf{f}(y, \mathbf{h})], \end{aligned} \quad (5)$$

where the expectation can be efficiently computed with the factorized form of $p(\mathbf{h}|\mathbf{x}, \mathbf{z})$ when \mathbf{x} and \mathbf{z} are fully observed. If missing values exist in \mathbf{x} or \mathbf{z} , an inference procedure is needed to compute the expectation of the missed components, as detailed below in (7).

Then, learning is to find an optimal \mathbf{V}^* that minimizes a loss function. Here, we minimize the hinge loss, as used in the very successful large-margin SVMs. Specifically, given training data $\mathcal{D} = \{(\mathbf{x}_d, \mathbf{z}_d, y_d)\}_{d=1}^D$, the hinge loss of the expected predictive rule (5) is

$$\mathcal{R}_{\text{hinge}}(\mathbf{V}) \triangleq \sum_d \max_y [\Delta \ell_d(y) - \mathbf{V}^\top \mathbb{E}_{p(\mathbf{h}|\mathbf{x}_d, \mathbf{z}_d)} [\Delta \mathbf{f}_d(y)]],$$

3. For notation simplicity, we omit the offset parameters in both classification and regression models. Offset parameters can be easily included by adding one unit dimension to \mathbf{h} .

4. A discriminative method that maximizes the conditional likelihood $p(y|\mathbf{x}, \mathbf{z})$ could be developed as in [28], but it could be inferior to a hybrid generative/discriminative method.

where $\Delta\ell_d(y)$ is a loss function (e.g., 0/1-loss) that measures how different a candidate prediction y is compared to the true label y_d , and $\mathbb{E}_{p(\mathbf{h}|\mathbf{x}_d, \mathbf{z}_d)}[\Delta\mathbf{f}_d(y)] = \mathbb{E}_{p(\mathbf{h}|\mathbf{x}_d, \mathbf{z}_d)}[\mathbf{f}(y_d, \mathbf{h})] - \mathbb{E}_{p(\mathbf{h}|\mathbf{x}_d, \mathbf{z}_d)}[\mathbf{f}(y, \mathbf{h})]$. It can be proven that the hinge loss is an upper bound of the empirical error $\mathcal{R}_{emp} \triangleq \sum_d \Delta\ell_d(y_d^*)$. Applying the principle of *regularized risk minimization*, we define the joint problem of learning a prediction model \mathbf{V} and a likelihood model Θ for fitting the input data as solving

$$P1: \min_{\Theta, \mathbf{V}} L(\Theta) + \frac{1}{2}C_1\|\mathbf{V}\|_2^2 + C_2\mathcal{R}_{hinge}(\mathbf{V}), \quad (6)$$

where $L(\Theta) \triangleq -\sum_d \log p(\mathbf{x}_d, \mathbf{z}_d)$ is the negative data likelihood and C_1 and C_2 are nonnegative constants which can be selected via cross validation. Note that \mathcal{R}_{hinge} is also a function of Θ .

Since problem (6) jointly maximizes the data likelihood and minimizes a training loss, it can be expected that by solving this problem we can find a predictive latent subspace representation (i.e., $\mathbb{E}_{p(\mathbf{h}|\mathbf{x}, \mathbf{z})}[\mathbf{h}]$) and a prediction model (represented by the parameter \mathbf{V}), which on one hand tend to predict as accurately as possible on the training data, while on the other hand tend to explain the data well. More insights will be provided in the next section, along with the algorithm development.

4.1.2 Optimization with Contrastive Divergence

Since the data likelihood $L(\Theta)$ is generally intractable to compute, we use an efficient variational inference method (i.e., contrastive divergence) [23], [40], [41], [44] to approximate the joint likelihood. Specifically, we derive a variational approximation $\mathcal{L}^v(q_0, q_1)$ to represent the negative log-likelihood $L(\Theta)$:

$$\mathcal{L}^v(q_0, q_1) \triangleq R(q_0(\mathbf{x}, \mathbf{z}, \mathbf{h}), p(\mathbf{x}, \mathbf{z}, \mathbf{h})) - R(q_1(\mathbf{x}, \mathbf{z}, \mathbf{h}), p(\mathbf{x}, \mathbf{z}, \mathbf{h})),$$

where $R(q, p)$ is the relative entropy between distributions q and p , q_0 is a variational distribution with \mathbf{x} and \mathbf{z} clamped to their observed values, while q_1 is a distribution with all the variables free. For q (either q_0 or q_1), we employ the *structured* mean field assumption [43] that⁵ $q(\mathbf{x}, \mathbf{z}, \mathbf{h}) = q(\mathbf{x})q(\mathbf{z})q(\mathbf{h})$.

Substituting the variational approximation $\mathcal{L}^v(q_0, q_1)$ into problem (6), we get an approximate objective function $\mathcal{L}(\Theta, \mathbf{V}, q_0, q_1)$. Then, we can develop an alternating minimization method which iteratively minimizes $\mathcal{L}(\Theta, \mathbf{V}, q_0, q_1)$ over (q_0, q_1) and (Θ, \mathbf{V}) . The problem of solving q_0 and q_1 is *posterior inference*. Specifically, for a variational distribution q (can be q_0 or q_1), we keep (Θ, \mathbf{V}) fixed and update each marginal as

$$\begin{aligned} q(\mathbf{x}) &= p(\mathbf{x}|\mathbb{E}_{q(\mathbf{h})}[\mathbf{h}]), & q(\mathbf{z}) &= p(\mathbf{z}|\mathbb{E}_{q(\mathbf{h})}[\mathbf{h}]), \\ q(\mathbf{h}) &= \prod_k p(h_k|\mathbb{E}_{q(\mathbf{x})}[\mathbf{x}], \mathbb{E}_{q(\mathbf{z})}[\mathbf{z}]). \end{aligned} \quad (7)$$

For q_0 , (\mathbf{x}, \mathbf{z}) are clamped at their observed values and only $q_0(\mathbf{h})$ is updated, which can be very efficiently done due to its factorized form. The distribution q_1 is achieved by

5. The parametric form assumptions of q , as employed in previous work [44], [45], are not needed.

performing the above updates starting from q_0 , and several iterations (e.g., 5 used in our experiments) can yield a good q_1 . Note that (7) holds for exponential family models where \mathbf{h} enters linearly in $\ln p(\mathbf{x}, \mathbf{z}|\mathbf{h})$. Please see [40] for more details. Again, we can observe that both $q(\mathbf{x})$ and $q(\mathbf{z})$ are CRFs, with the expectation of \mathbf{H} as the condition. Therefore, for linear-chain models, we can use a message passing scheme [27] to infer their marginal distributions as needed for parameter estimation and view-level prediction (e.g., image annotation), as we shall see. For generally structured models, approximate inference techniques [37] can be applied.

After we have inferred q_0 and q_1 , parameter estimation can be solved with coordinate descent by alternating the following two steps: 1) estimating \mathbf{V} with Θ fixed: this problem is learning a multiclass SVM [13], which can be efficiently done with existing solvers; and 2) estimating Θ with \mathbf{V} fixed: this can be solved with subgradient descent. By defining $\Delta\mathbb{E}[\cdot] \triangleq \mathbb{E}_{q_1}[\cdot] - \mathbb{E}_{q_0}[\cdot]$, we can compute the subgradient as follows: For θ , we have $\forall e \in E_x, \partial\theta_e = \Delta\mathbb{E}[\phi(\mathbf{x}_e)]$; for η , we have $\forall e \in E_z, \partial\eta_e = \Delta\mathbb{E}[\psi(\mathbf{z}_e)]$; for λ , we have $\forall k, \partial\lambda_k = \Delta\mathbb{E}[\varphi(h_k)]$; and for \mathbf{W} and \mathbf{U} , we have

$$\begin{aligned} \partial\mathbf{W}_e^k &= \Delta\mathbb{E}[\phi(\mathbf{x}_e)\varphi(h_k)^\top] - C_2 \sum_d (\mathbf{V}_{y_{dk}} - \mathbf{V}_{\bar{y}_{dk}}) \frac{\partial\mathbb{E}_{q_0}[h_k]}{\partial\mathbf{W}_e^k}, \\ \partial\mathbf{U}_e^k &= \Delta\mathbb{E}[\psi(\mathbf{z}_e)\varphi(h_k)^\top] - C_2 \sum_d (\mathbf{V}_{y_{dk}} - \mathbf{V}_{\bar{y}_{dk}}) \frac{\partial\mathbb{E}_{q_0}[h_k]}{\partial\mathbf{U}_e^k}, \end{aligned}$$

where $\bar{y}_d = \arg\max_y [\Delta\ell_d(y) + \mathbf{V}^\top \mathbb{E}_{q_0}[\mathbf{f}(y, \mathbf{h})]]$ is the *loss-augmented prediction*. The expectation $\mathbb{E}_{q_0}[\phi(\mathbf{x}_e)]$ is actually the count frequency of $\phi(\mathbf{x}_e)$ on the training data \mathcal{D} ; likewise for $\mathbb{E}_{q_0}[\psi(\mathbf{z}_e)]$. With the above subgradients, we apply L-BFGS [29], which uses line search to choose a step size, to iteratively solve for the optimum Θ until convergence.

Note that in our integrated large-margin formulation, the subgradients corresponding to \mathbf{W} and \mathbf{U} contain an additional term (i.e., the third term) compared to the standard DWH [44] with contrastive divergence approximation. This additional term introduces a regularization effect to the latent subspace model. If the loss-augmented prediction \bar{y}_d differs from the true label y_d , this term will be nonzero and it will bias the model toward discovering a better representation for prediction. As we shall see, this bias term will make the large-margin-based multiview latent subspace model tend to discover a latent representation that is more predictive.

4.2 Regression

In this section, we present the large-margin latent subspace MN for regression.

4.2.1 Problem Definition

Similarly to the classification model, we define the linear *expected* prediction rule for regression as

$$y^* \triangleq \mathbf{V}^\top \mathbb{E}_{p(\mathbf{h}|\mathbf{x}, \mathbf{z})}[\mathbf{h}], \quad (8)$$

where \mathbf{V} is a K -dim vector. To learn the prediction model \mathbf{V} , we need to devise a loss function that integrates the large-margin principle for prediction with latent subspace discovery. Here, for prediction, we choose to minimize

the ϵ -insensitive loss, which is used in standard support vector regression (SVR) [33]:

$$\mathcal{R}_\epsilon(\mathbf{V}) \triangleq \sum_d \max(0, |y_d - \mathbf{V}^\top \mathbb{E}[\mathbf{h}_d]| - \epsilon),$$

where $\epsilon \in \mathbb{R}_+$ is the precision parameter, which is usually small, and we have defined $\mathbb{E}[\mathbf{h}_d] = \mathbb{E}_{p(\mathbf{h}|\mathbf{x}_d, \mathbf{z}_d)}[\mathbf{h}]$ for notation simplicity. Similarly, following the *regularized risk minimization* principle, we learn the entire model for regression and fitting the observed input data by solving the joint optimization problem:

$$\text{P2: } \min_{\Theta, \mathbf{V}} L(\Theta) + \frac{1}{2} C_1 \|\mathbf{V}\|_2^2 + C_2 \mathcal{R}_\epsilon(\mathbf{V}), \quad (9)$$

where $L(\Theta)$ is the negative log likelihood of input data as we have defined in classification.

Similarly to the classification model, by jointly minimizing the negative log likelihood and the regression loss, we can expect to learn a latent subspace representation as well as a prediction model which, on the one hand, tends to predict the data accurately, while, on the other hand, attempts to interpret the data well.

4.2.2 Optimization with Contrastive Divergence

Although in principle we can use a similar procedure as in the classification model to solve problem P2 by employing a subgradient descent method to learn the parameters Θ , here we use a Lagrangian method to solve an equivalent constrained formulation of P2. One reason is that the loss \mathcal{R}_ϵ is a bit more complicated than $\mathcal{R}_{\text{hinge}}$ because of the nondifferentiable absolute operator within the max function. Specifically, problem P2 can be equivalently written as

$$\begin{aligned} \text{P2}' : \min_{\Theta, \mathbf{V}, \xi_d, \xi_d^*} & L(\Theta) + \frac{1}{2} C_1 \|\mathbf{V}\|_2^2 + C_2 \sum_d (\xi_d + \xi_d^*) \\ \text{s.t. } \forall d : & \begin{cases} y_d - \mathbf{V}^\top \mathbb{E}[\mathbf{h}_d] \leq \epsilon + \xi_d \\ -y_d + \mathbf{V}^\top \mathbb{E}[\mathbf{h}_d] \leq \epsilon + \xi_d^* \\ \xi_d, \xi_d^* \geq 0, \end{cases} \end{aligned} \quad (10)$$

where ξ_d and ξ_d^* are slack variables.

The constrained problem P2' is generally intractable because the likelihood $L(\Theta)$ is intractable to evaluate. As in the classification model, we approximate $L(\Theta)$ with the contrastive divergence approximation $\mathcal{L}^v(q_0, q_1)$. Then, we introduce Lagrange multipliers $\mu_d, \mu_d^*, v_d, v_d^*$ for the four constraints associated with data d , and define the Lagrangian function L with the approximate likelihood $\mathcal{L}^v(q_0, q_1)$:

$$\begin{aligned} L = & \mathcal{L}^v(q_0, q_1) + \frac{1}{2} C_1 \|\mathbf{V}\|_2^2 + C_2 \sum_d (\xi_d + \xi_d^*) \\ & - \sum_d (v_d \xi_d + v_d^* \xi_d^*) - \sum_d \{ \mu_d (\epsilon + \xi_d - y_d + \mathbf{V}^\top \mathbb{E}[\mathbf{h}_d]) \\ & + \mu_d^* (\epsilon + \xi_d^* + y_d - \mathbf{V}^\top \mathbb{E}[\mathbf{h}_d]) \}. \end{aligned}$$

Now, we optimize the Lagrangian function L by alternatively performing the following steps:

1. Inferring q_0 and q_1 . This step is the same as in the classification model.

2. Estimating Θ with μ_d and μ_d^* fixed. This can be solved with gradient descent (e.g., using L-BFGS [29] as in the classification model), where the gradients for (θ, η, λ) are as before and for (\mathbf{W}, \mathbf{U}) we have

$$\begin{aligned} \partial \mathbf{W}_e^k &= \Delta \mathbb{E}[\phi(\mathbf{x}_e) \varphi(h_k)^\top] \\ &\quad - \sum_d (\mu_d - \mu_d^*) \mathbf{V}_k \frac{\partial \mathbb{E}_{q_0}[h_k]}{\partial \mathbf{W}_e^k}, \\ \partial \mathbf{U}_e^k &= \Delta \mathbb{E}[\psi(\mathbf{z}_e) \varphi(h_k)^\top] \\ &\quad - \sum_d (\mu_d - \mu_d^*) \mathbf{V}_k \frac{\partial \mathbb{E}_{q_0}[h_k]}{\partial \mathbf{U}_e^k}. \end{aligned} \quad (11)$$

3. Estimating the Lagrange multipliers $\{\mu_d, \mu_d^*\}$. By setting $\partial L / \partial \xi_d, \partial L / \partial \xi_d^*, \partial L / \partial \mathbf{V} = 0$ and exploring the KKT conditions, we can get

$$\mathbf{V} = \frac{1}{C_1} \sum_d (\mu_d - \mu_d^*) \mathbb{E}[\mathbf{h}_d]. \quad (12)$$

Plugging (12) into the Lagrangian function L , we get the dual problem

$$\begin{aligned} \max_{\mu, \mu^*} & -\frac{1}{2C_1} \left\| \sum_d (\mu_d - \mu_d^*) \mathbb{E}[\mathbf{h}_d] \right\|_2^2 \\ & - \sum_d [\epsilon (\mu_d + \mu_d^*) - y_d (\mu_d - \mu_d^*)] \\ \text{s.t. } \forall d : & \mu_d, \mu_d^* \in [0, C_2], \end{aligned}$$

which can be solved using an existing algorithm like SVM-light [25] to obtain μ_d and μ_d^* .

Again, we can see that in this integrated large-margin formulation for regression, the gradients of \mathbf{W} and \mathbf{U} contain an additional term encoded with μ_d and μ_d^* . Similarly as in the classification model, this additional term introduces a regularization effect to the latent subspace model. If the prediction $\mathbf{V}^\top \mathbb{E}[\mathbf{h}_d]$ differs from the true value y_d with the absolute gap larger than ϵ , the Lagrangian multipliers μ_d or μ_d^* (at most one is nonzero because of the KKT conditions) will be nonzero and will bias the model toward discovering a better representation for prediction.

4.3 Special Case: Maximum Margin Harmonium

We have developed the large-margin learning framework on a general multiview latent subspace MN for classification and regression. In order to fully examine the basic learning principle and compare with existing Harmonium models [41], [44], [45], we introduce a specialized but very rich instantiation of our supervised latent subspace MN where the data on each view are not structured. We denote the specialized model by max-margin Harmonium (MMH). We emphasize that this simplification does not restrict our ability to demonstrate the generability of the framework because both the problem definition and optimization algorithm are general to any structured input data, as we have presented. Specifically, MMH uses the DWH model detailed at the end of Section 3.1 as the probabilistic likelihood model to fit the input data (\mathbf{x}, \mathbf{z}) , where \mathbf{x} is a vector of discrete word features (e.g., image tags) and \mathbf{z} is a vector of real-valued features (e.g., color histograms). We

can follow the same procedure as above to do parameter estimation. For inferring q_0 and q_1 , the distributions of \mathbf{x} , \mathbf{z} , and \mathbf{h} are all fully factorized. Therefore, the subgradients in classification or gradients in regression can be easily computed. Details are deferred to Appendix A.1, which can be found in the Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/TPAMI.2012.64>.

4.4 Time Complexity on Testing

Before ending this section, we discuss the time complexity of applying the supervised latent subspace MN on various applications, including classification, regression, image retrieval and annotation.

The commonality among using a latent subspace MN for classification, regression, and retrieval is that all these applications rely on inferring the latent representations (i.e., $\mathbb{E}_{p(\mathbf{h}|\mathbf{x},\mathbf{z})}[\mathbf{h}]$) only without missing information on the input data. For the large-margin latent subspace MN, since it defines a *partial* likelihood function, that is, the likelihood on input data (\mathbf{x}, \mathbf{z}) only, we can infer these latent representations in a single-round manner. More precisely, the latent representation is $\mathbb{E}_{p(\mathbf{h}|\mathbf{x},\mathbf{z})}[\mathbf{h}] = \Upsilon$, where $\Upsilon_k = \sum_{e \in E_x} \phi(\mathbf{x}_e)^\top \mathbf{W}_e^k + \sum_{e \in E_z} \psi(\mathbf{z}_e)^\top \mathbf{U}_e^k, \forall 1 \leq k \leq K$, which can be very efficiently computed in a linear complexity in terms of the dimensionality of the input features. In contrast, for the MLE-based supervised latent subspace MN, it defines a *full* likelihood function $p(\mathbf{x}, \mathbf{z}, y)$ over the input data (\mathbf{x}, \mathbf{z}) and response value y . In testing where Y is not observed, the inference involves an interactive procedure which iteratively infers the (approximate) posterior distribution of Y and the latent representation. Therefore, the testing time of using the MLE-based supervised latent subspace MN is typically a constant times more expensive than that of the large-margin-based method. But, in general, these undirected models are much more efficient than their directed counterpart models, as we will show in Section 5.5.

For image annotation, let us use \mathbf{x} to represent tags which are observed in training. In testing, we infer the posterior distribution $p(\mathbf{x}|\mathbf{z})$, which can be approximately computed by running the update (7) with \mathbf{z} clamped at its observed values. Then, tags with high probabilities are selected as the annotation results. As we can see, the inference procedure is similar as the iterative one of using a MLE-based latent subspace MN for classification. Therefore, the time complexities of the unsupervised and supervised (both large-margin and MLE-based) multiview MNs are almost the same.

5 EXPERIMENTS

Now we present qualitative as well as quantitative evaluation on three real datasets to demonstrate the advantages (e.g., effectiveness and time efficiency) of large-margin supervised multiview latent subspace MNs. We first extensively evaluate the specialized but rich MMH model and compare with extant latent subspace models for classification, regression, image annotation, and retrieval in Section 5.3. Then, we present a structured latent subspace

MN for modeling paragraph ordering information on hotel review data in Section 5.4.

5.1 Data Sets and Features

The datasets⁶ are TRECVID2003 [44], 13class-animal Flickr image data, and hotel review data [50]. These datasets are quite rich and diverse in terms of feature types and dimensionality, as detailed below.

TRECVID2003 contains 1,078 manually labeled video shots that belong to five categories. Each shot is represented as a 1,894-dim vector of text features and a 165-dim vector of HSV color histogram which is extracted from the associated keyframe. We evenly split this dataset into training and testing sets.

The Flickr dataset is a subset selected from NUS-WIDE [12], which is constructed from Flickr web images. This dataset contains 3,411 images of 13 animals—*squirrel, cow, cat, zebra, tiger, lion, elephant, whales, rabbit, snake, antlers, hawk, and wolf*. See Fig. 8 for example images from each category. For each image, six types of low-level features [12] are extracted, including 634-dim real-valued features (i.e., 64-dim color histogram, 144-dim color correlogram, 73-dim edge direction histogram, 128-dim wavelet texture, and 225-dim blockwise color moments) and 500-dim bag-of-word SIFT [30] features. We randomly select 2,054 images for training and use the rest for testing. The 1,000-dim online tags are also downloaded for evaluating image annotation.

The hotel review dataset consists of 5,000 hotel reviews randomly collected from TripAdvisor.⁷ Each review document is associated with two-view features (i.e., 12,000-dim bag-of-word features and 14-dim contextual features [50]) as well as a global rating score and five aspect rating scores. The global ratings rank from 1 to 5. In our experiment, we predict the global rating scores for reviews and uniformly partition the dataset into training and testing sets. Note that the bag-of-words features (e.g., text or SIFT) are treated as binary and modeled using the Bernoulli view.

5.2 Predictive Latent Subspace Representations

To demonstrate the power of our method in discovering predictive subspace representations, in this section we examine various characteristics of the latent subspace representations for modeling both image and text.

5.2.1 Image Modeling

We first take a holistic view of the entire latent representations. Fig. 2 shows the 2D embedding of the discovered 10-dim latent representations by MMH, DWH, and TWH on the video keyframes in the TRECVID dataset. Here, we use the t-SNE stochastic neighborhood embedding algorithm [36] to embed the latent representations in a 2D space. The results clearly show that the latent subspace representations discovered by MMH exhibit a strong grouping pattern for the images belonging to the same category, while images from different categories tend to be separated from each other on the 2D embedding space. In contrast, the latent subspace representations discovered by the likelihood-based DWH and TWH do not show a clear

6. <http://www.cs.cmu.edu/~junzhu/data.htm>.

7. <http://www.tripadvisor.com>.

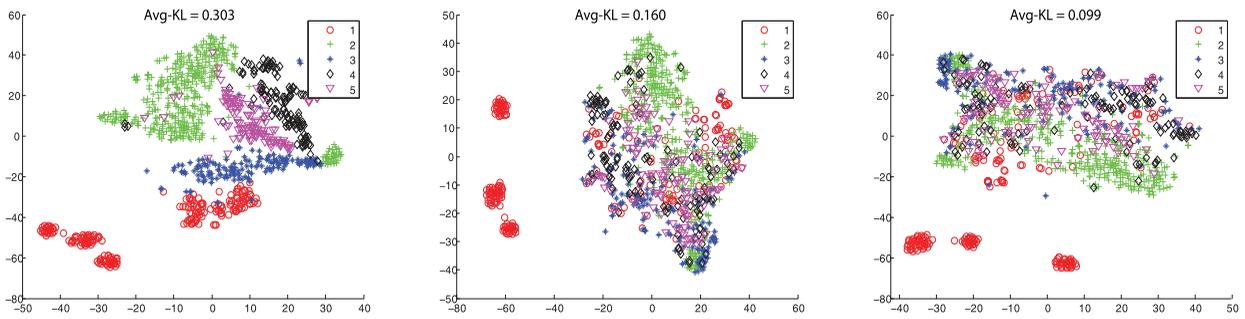


Fig. 2. t-SNE 2D embedding of the discovered latent subspace representation by (left) MMH, (middle) DWH, and (right) TWH on the TRECVID video dataset (better viewed in color).

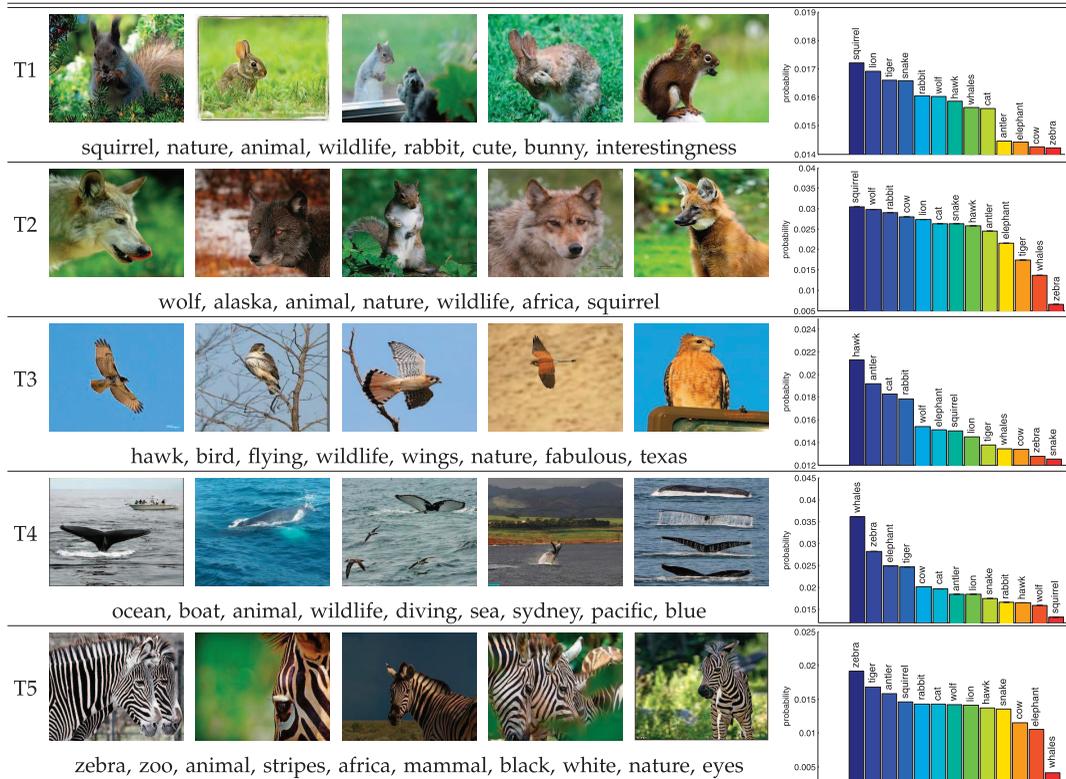


Fig. 3. Example topics discovered by a 60-topic MMH on the Flickr animal dataset. For each topic, we show five topic-ranked images as well as the average probabilities of that topic on representing images from the 13 categories.

grouping pattern, except for the first category, and images from different categories tend to mix together. These observations suggest that the large-margin-based MMH can discover more discriminative latent subspace representations, which will result in better prediction performance, as we shall see. We have similar observations on the Flickr dataset.

Now, we take a closer examination of each dimension in the discovered latent subspace. We take the Flickr data as an example. Fig. 3 shows five example topics (each topic corresponds to one dimension in the latent subspace) discovered by the large-margin MMH on the Flickr image data. Due to space limitation, for each topic T_k , we show the five top-ranked images that yield a high expected value of H_k , together with the associated tags. Please see Fig. 11 in Appendix A.3, available in the online supplemental material, for the five bottom-ranked images for each topic. Also, to qualitatively visualize the discriminative power of each topic among the 13 categories, we show the average

probability⁸ of each category distributed on the particular topic, as shown in the right part of Fig. 3. From the results, we can see that many of the discovered topics are predictive for one or several categories. For example, topics T3 and T4 are discriminative in predicting the categories *hawk* and *whales*, respectively. Similarly, topics T1 and T5 are good at predicting *squirrel* and *zebra*, respectively. We also have some topics which are good at discriminating a subset of categories against another subset. For example, topic T2 is good at discriminating $\{squirrel, wolf, rabbit\}$ against $\{tiger,$

8. To compute the distribution, we first turn the expected value of \mathbf{H} to be nonnegative by subtracting each element by the smallest value and then normalize it into a distribution over the K topics. The per-class average is computed by averaging the topic distributions of the images within the same class. Then, we show the topic distribution on 13 categories specified by different topic. Note that our transformation (i.e., shift-normalization) doesn't affect our interpretation of the discriminative power, which is visually reflected as the (normalized) difference of the average values between categories. Using the raw values of $\mathbb{E}[\mathbf{H}]$ will produce the similar visualization patterns.

TABLE 1
Average Distributions over the Topics for Documents with Different Rating Scores by a 5-Topic MMH and 5-Topic DWH

Max Margin Harmonium (Avg-KL: 3.568)									
Average $\mathbb{E}_{p(\mathbf{h} x,z)}[\mathbf{h}]$ in 5-level Rating Score examples					T1	T2	T3	T4	T5
					room hotel n't time stay day night <i>good</i> staff pool back rooms food area <i>nice</i>	<i>great</i> <i>loved</i> arrived <i>enjoyed</i> <i>fantastic</i> bit <i>wonderful</i> <i>lovely</i> pool trip toilet beach <i>fun</i> <i>happy</i> pools <i>perfect</i>	small worst dirty shower broken smell paying bathroom poor toilet staying refund breakfast hotel walls	worst small dirty shower broken smell paying bathroom poor toilet refund staying walls hotel carpet	worst dirty small shower broken smell paying poor toilet manager bathroom walls carpet paid
Dual-Wing Harmonium (Avg-KL: 0.038)									
Average $\mathbb{E}_{p(\mathbf{h} x,z)}[\mathbf{h}]$ in 5-level Rating Score examples					T1	T2	T3	T4	T5
					room hotel n't time stay day night <i>good</i> staff pool back rooms food area <i>nice</i>	beach food pool resort <i>great</i> restaurants bar drinks restaurant lunch sea <i>beautiful</i> entertainment pools view	food told asked holiday reception day bar staff back manager people evening entertainment arrived hotel	breakfast reception bathroom bed shower holiday coffee evening <i>small</i> <i>clean</i> bar hotel <i>good</i> main tea	belize brett cam canapes canoeing caracol hosts nadege underway wineries adopted amanda aurora begun boasted

whales, zebra}, but it is not very discriminative between squirrel and wolf.

To quantitatively evaluate the predictiveness of the discovered latent subspace representations, we compute the pair-wise average KL-divergence between the per-class average distribution over latent topics.⁹ As shown on the top of each plot in Fig. 2, the large-margin-based MMH obtains a larger average KL-divergence score than likelihood-based methods. This again suggests that the latent subspace representations by MMH are more discriminative or predictive. We obtain similar observations on the Flickr dataset (see Fig. 3 for some example topics), where the average KL-divergence scores of 60-topic MMH, DWH, and TWH are 1.62, 1.28, and 0.232, respectively. This is consistent with our intuitive observations that the latent subspace representations (see Fig. 3) by MMH are more discriminative.

5.2.2 Text Modeling

Now, we examine the properties of latent subspace MN on text modeling. Again, we present both holistic and topic-wise close examinations. Table 1 shows the topics discovered by 5-topic MMH and DWH on the hotel review data. As in [50], we denote the five rating scores from small to large by R_1, R_2, \dots, R_5 . We also show the per-rating average distributions over topics in the left part, which are

9. We first turn the expected value of \mathbf{H} into a distribution over the K topics. The per-class average is computed by averaging the topic distributions of the images within the same class. For a pair of distributions p and q , the average KL-divergence is $1/2(R(p, q) + R(q, p))$.

computed in a similar way as the per-class average distributions in the above section. The right side of Table 1 shows the top 15 words for each topic T_k .

Similarly to the observations in image modeling, we can see that the latent subspace representations discovered by MMH are much more discriminative than those discovered by DWH, as reflected from the much higher pairwise average KL-divergence score and the quite different average distributions over topics, and the individual dimensions (i.e., topics) of the latent subspace learned by MMH are very expressive and discriminative, too. For example, topic T2 for MMH has larger probabilities on representing documents with high rating scores (e.g., R_5 and R_4), but has smaller probabilities (drops to near zero) on documents with lower rating scores (e.g., R_1 and R_2). Moreover, the probability of topic T2 shows a smooth increasing trend on representing documents with rating scores from low to high. If we look at the top words of T2 (e.g., “great,” “fantastic,” “wonderful,” “perfect,” etc.) as shown in the right part of Table 1, we can see that T2 represents a positive aspect of a hotel. Therefore, it is more likely to appear in representing a positive review. In contrast, the negative topics T3, T4, and T5 (e.g., with negative words “worst,” “dirty,” “poor,” etc.) show a smooth decreasing trend on probabilities in representing documents with rating scores from R_1 to R_5 . Topic T1 is kind of neutral, which has the highest probability on representing the documents with a neutral rating score (e.g., R_3 or R_4) and overall T1 has a much larger probability than any other topics on representing a document. This is reasonable on the hotel review data because most of the words in a review are about the basic hotel information (e.g., “room,” “hotel,”

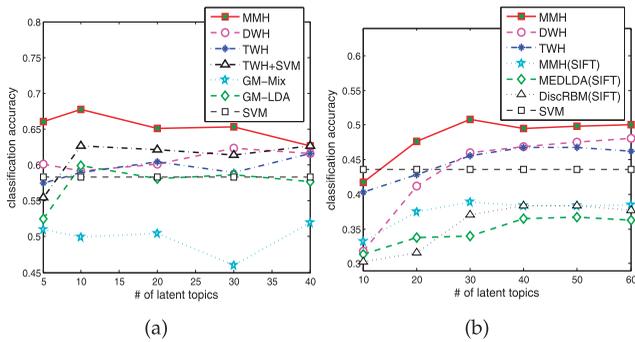


Fig. 4. Classification accuracy on the (a) TRECVID and (b) Flickr image datasets.

“food,” and “area”). For DWH, the topics are not very discriminative, as demonstrated by the comparable probabilities on representing documents with different rating scores. Table 2 in Appendix A.3, available in the online supplemental material, also shows the results on TWH, which is comparable to DWH.

5.3 Prediction Performance

In this section, we provide quantitative results on classification, regression, image annotation, and retrieval.

5.3.1 Classification

We first compare MMH with SVM, DWH, TWH, Gaussian Mixture (GM-Mix), Gaussian Mixture LDA (GM-LDA), and Correspondence LDA (CorrLDA) on the TRECVID dataset. See [3] for the details of the last three models. We use $SVM^{multiclass}$ ¹⁰ to solve the substep of learning V in MMH and build the SVM baseline that uses all the available features without distinguishing them in different views. For the unsupervised models (i.e., DWH, GM-Mix, GM-LDA, and CorrLDA), a downstream SVM classifier is built based on the discovered latent representations. Fig. 4a shows the classification accuracy of different models, where CorrLDA is omitted because of its too low performance. We can see that the max-margin-based multiview MMH performs consistently better than any other competitors. In contrast, the MLE-based TWH does not show any conclusive improvements compared to the unsupervised DWH. If we train a downstream SVM classifier using the representations by TWH, the classification performance (denoted by TWH + SVM¹¹) will be improved, but still inferior to that of MMH. These results show that supervising side information can help in discovering predictive latent subspace representations that are more suitable for prediction if the model is appropriately learned, e.g., using the large-margin method. The superior performance of MMH compared to the flat SVM demonstrates the usefulness of modeling multiview inputs for prediction. The reasons for the inferior performance of other models (e.g., CorrLDA and GM-Mix) are analyzed in [44], [45].

Fig. 4b shows the accuracy on the Flickr dataset. For brevity, we compare MMH with the best performing DWH, TWH, and SVM. We use the 500-dim SIFT and 634-dim real

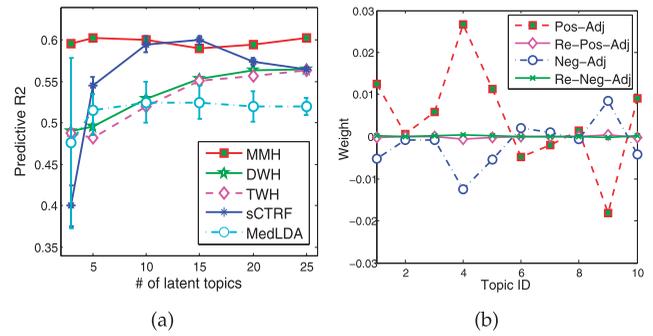


Fig. 5. (a) Prediction R2. (b) Feature weights in MMH with 10 topics.

features as two views of inputs for MMH, DWH, and TWH. Also, we compare with the single-view MedLDA [48] and discriminative restricted Boltzmann machine (DiscRBM) [28], which use SIFT features only. To be fair, we also evaluate a version of MMH that uses SIFT features, and denote it by MMH (SIFT). Again, we can see that the large-margin-based multiview MMH performs much better than any other methods, including SVM, which ignores the presence of multiview features. For the single-view MMH (SIFT), it performs comparably with DiscRBM and the large-margin MedLDA, which is a directed BN. As we have stated, MMH represents an important extension of MedLDA to the undirected latent subspace models and for multiview data analysis. For DiscRBM, since it performs discriminative training (i.e., maximizing the conditional likelihood of Y given input features) and doesn’t estimate the model for generating input features, it can’t perform view-level analysis (e.g., predicting image tags). In [28], a hybrid generative/discriminative likelihood objective was also discussed to learn RBM, which outperforms DiscRBM. MMH is different from the hybrid method in three aspects: 1) Our generative likelihood $L(\theta)$ doesn’t include Y , 2) our discriminative part is a hinge-loss instead of a conditional log-likelihood, and 3) we use an explicit regularization instead of the implicit regularization (i.e., early stopping) used in [28].

5.3.2 Regression

Following [50], we treat the problem of predicting rating scores on the hotel review dataset as a regression problem. We compare the MMH regression model with DWH, TWH, sCTRF (i.e., supervised conditional topic random fields) [50], and MedLDA [48]. For DWH, we build a linear SVR as the downstream regression model. Many other baselines (e.g., supervised LDA) are not included because they are inferior to sCTRFs, as reported in [50]. We consider two views for each document, where one view X denotes the bag-of-word features and the other view Z represents the 14 types of contextual features [50].

Fig. 5a shows the predictive R2 scores (please see [4] for the definition). We can observe that by exploring supervising side information (i.e., rating score) in learning the latent subspace model, MMH consistently outperforms the decoupled two-step procedure that is adopted in unsupervised DWH, but the MLE-based TWH does not show improvements over the unsupervised DWH. This again verifies that the large-margin learning plays a significant role in discovering predictive latent subspace representations that are suitable for prediction tasks (e.g., regression).

10. http://svmlight.joachims.org/svm_multiclass.html.

11. This naive combination uses supervision twice and is not an elegant model compared to MMH. The similar naive combination MMH+SVM wouldn’t outperform MMH in theory because both methods build SVM classifiers using the same latent representations. See [48] for similar studies.

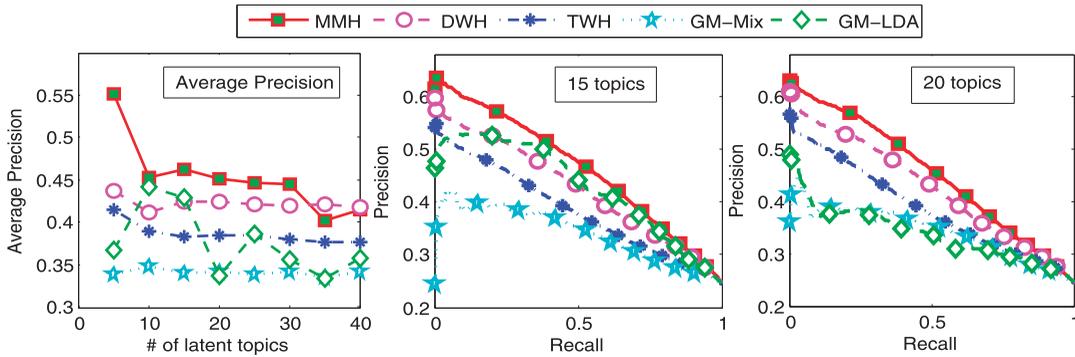


Fig. 6. The average precision curve and the two precision-recall curves for image retrieval on TRECVID data.

In addition, the reason why MMH achieves superior performance than the single-view MedLDA is that MMH can use multiview features simultaneously, which again demonstrates the benefits of modeling multiview instead of single-view input for prediction. In fact, the second-view features play an important role in finding a predictive latent subspace. We show the weights of four features on the 10 topics discovered by a 10-topic MMH in Fig. 5b, where the four features as studied in [50] are: “Pos-Adj”—positive adjective, “Re-Pos-Adj”—positive adjective that has a denying word before it, “Neg-Adj”—negative adjective, and “Re-Neg-Adj”—negative adjective that has a denying word before it, respectively. We can see that both the positive and negative adjective features tend to discover topics that are more discriminative for rating prediction (e.g., T4 and T9). The best performance of MMH is comparable to that of sCTRF, which is a directed model. As we shall see in Section 5.5, MMH is much more efficient in training and testing.

5.3.3 Image Retrieval

We apply MMH for image retrieval on the TRECVID and Flickr datasets. Each test image is a query and training images are ranked based on their cosine similarity¹² with the given query, which is computed based on the inferred latent subspace representations using the learned models. An image is considered relevant to the query if they belong to the same category. We evaluate the performance by drawing precision-recall curves and computing the average precision (AP) score [44], [45].

Fig. 6 compares MMH with four other models when the topic number K changes. Here, we show the precision-recall curves when K is set at 15 and 20. Interestingly, although MMH does not directly optimize a ranking-based loss measure, the latent representations discovered by MMH can result in higher retrieval performance than all other methods in most cases. On the Flickr dataset, we have similar observations. For instance, the AP scores of the 60-topic MMH, DWH, and TWH are 0.163, 0.153, and 0.158, respectively.

5.3.4 Image Annotation

We also report the annotation results on the Flickr dataset, with a dictionary of 1,000 unique tags. The average number of tags per image is about 4.5. We compare MMH with DWH

and TWH with two views— \mathbf{X} for tag and \mathbf{Z} for all the 634-dim real-valued features. We also compare with the sLDA annotation model [39], which uses SIFT features. We use the top- N F1-measure [39], denoted by $F1@N$. With 60 latent topics, the top- N F-measure scores are shown in Fig. 7. Again, we can see that although not directly minimizing an annotation loss measure, the large-margin MMH outperforms other competitors, mainly because of its good latent representations. Fig. 8 shows example images from all 13 categories, where for each category the left image is generally of good annotation quality and the right one is relatively worse.

5.4 Structured Latent Subspace MN on Modeling Paragraph Ordering Information

We have extensively evaluated the advantages of large-margin learning based on a specialized dual-wing model (i.e., MMH). In this section, we present a structured latent subspace MN for modeling paragraph ordering information on hotel review data. As we mentioned, on TripAdvisor, there are five predefined aspects (e.g., Location), which could guide the users to compose their review contents. Since these aspects are displayed in a particular order to users, we can expect that the composed contents about each aspect would present a similar ordering regularity. Although other possible treatments (e.g., sentence-level ordering) exist, we consider such ordering information between paragraphs and design the structured latent subspace MN, as follows.

We represent a document as a $P \times N$ observation matrix \mathbf{x} , where P is the number of paragraphs in this document and N is the vocabulary size. Each row \mathbf{x}_p is a vector, of which each element x_{pi} denotes whether word i appears in paragraph p . Each column \mathbf{x}_i represents the appearance pattern of word i in all paragraphs. To consider the paragraph ordering information, we define a first-order Markov chain on each \mathbf{x}_i while assuming that different \mathbf{x}_i s

	MMH	DWH	TWH	sLDA
$F1@3$	0.245	0.202	0.218	0.146
$F1@4$	0.258	0.208	0.228	0.159
$F1@5$	0.262	0.210	0.236	0.169
$F1@6$	0.259	0.208	0.240	0.171
$F1@7$	0.256	0.206	0.239	0.175

Fig. 7. Top- N F1-measure.

12. The cosine similarity between vectors \mathbf{x}_1 and \mathbf{x}_2 is $\frac{\mathbf{x}_1^\top \mathbf{x}_2}{\|\mathbf{x}_1\|_2 \|\mathbf{x}_2\|_2}$.

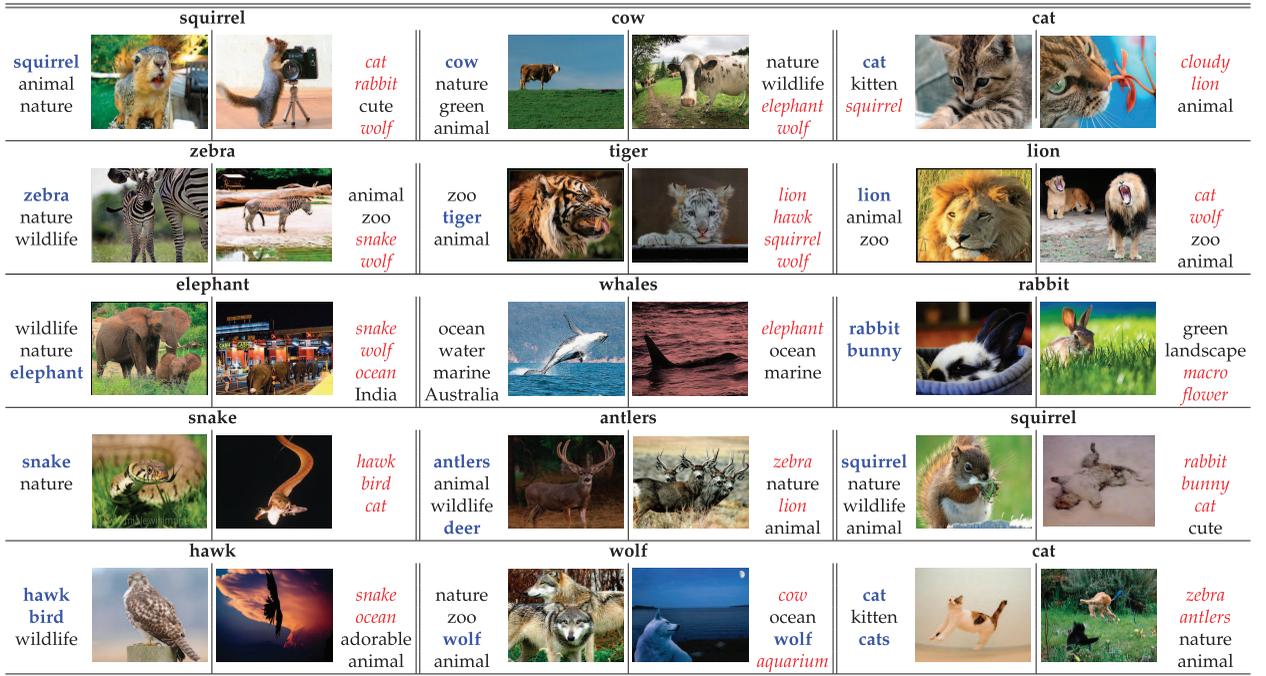


Fig. 8. Example images from the 13 categories on the Flickr animal dataset with predicted annotations. Tags in blue and bold are correct annotations, while red and italic ones are wrong predictions. The other tags are neutral. We have repeated the categories “squirrel” and “cat” in the right corner to fill the empty space.

are conditional independent. More formally, we define the conditional distribution $p(\mathbf{x}|\mathbf{h}) = \prod_{i=1}^N p(\mathbf{x}_i|\mathbf{h})$, where each $p(\mathbf{x}_i|\mathbf{h})$ is a linear chain CRF [27]. For this structured model, which is in fact an N -view latent subspace Markov network, we can perform efficient inference with message passing whose complexity is also linear in terms of N . The details are deferred to Appendix A.2, available in the online supplemental material.

To evaluate the structured model, denoted by *structMMH*, we build another dataset from the hotel reviews on TripAdvisor, which contains 600 reviews for each of the five rating scores. We randomly choose half as training and test on the rest. The reason why we didn’t use the dataset [50] is that it contains many reviews that have only one paragraph. Here, while regression can be performed too, we report the classification accuracy in Fig. 9a. We observe that the large-margin *structMMH* outperforms the unstructured MMH and the two-stage method (denoted by *structDWH*) that uses a structured MN

as defined above to infer the latent representations and learn a downstream SVM for classification. This observation demonstrates that the paragraph ordering information is helpful to discover more predictive latent subspace representations for the hotel review data.

5.5 Running Time and Sensitivity Analysis

Fig. 9b compares the time efficiency of MMH with TWH and directed models, including MedLDA and sCTRF [50], on the hotel review dataset [50] for regression. For testing, we can see that: 1) the undirected MMH and TWH are much more efficient than the directed MedLDA, which requires a relatively expensive iterative procedure to infer the distributions of latent variables; 2) TWH is about several times slower than MMH because of the reasons as we have discussed in Section 4.4; and 3) sCTRF is about 10 times slower than MedLDA or about 10,000 times slower than MMH. The main reason for such slowness is that sCTRF models every sentence in a document using a Markov chain. Therefore, it spends most of the time on performing

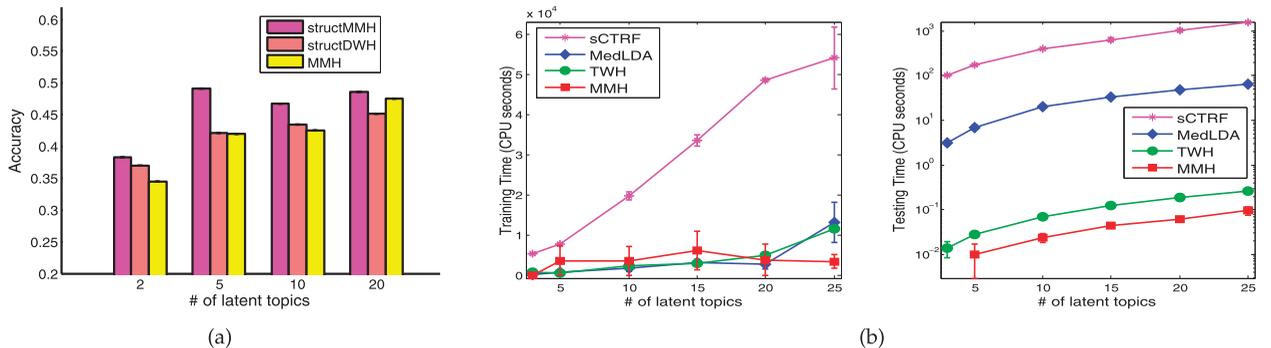


Fig. 9. (a) Classification accuracy of structured MMH and DWH models, and (b) training and testing time on hotel review data [50] for regression.

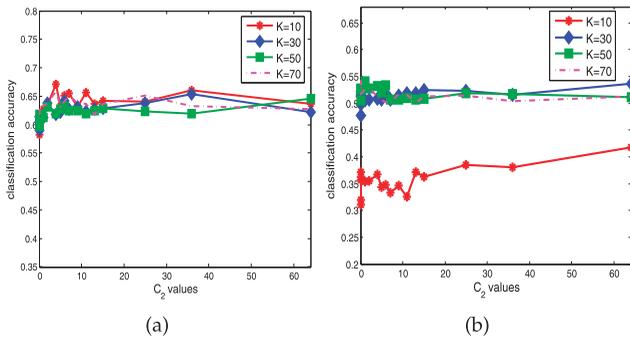


Fig. 10. Sensitivity to C_2 on the (a) TRECVID and (b) Flickr datasets.

message-passing. See [50] for more details. For training, we can see that MMH takes comparable time as TWH and MedLDA, and is much more efficient than sCTRF, whose inference is much slower as shown in testing times.

Finally, as shown in Fig. 10, MMH is not very sensitive to the regularization constant C_2 on either the TRECVID or Flickr dataset when the topic number K is set appropriately. In all the above experiments, we fixed C_1 at 0.5 and chose C_2 using cross validation during training.

6 CONCLUSIONS AND DISCUSSIONS

We have presented a large-margin learning framework for discovering predictive latent subspace representations shared by multiview data. Besides the proposed multiview latent subspace Markov networks, the large-margin learning method is generally applicable for the broad family of undirected latent subspace models. The inference and learning can be efficiently done with contrastive divergence methods. Finally, we present extensive evaluation results on various types of real datasets including both image and text data to demonstrate the advantages of large-margin learning for both predictive latent subspace discovery and prediction.

Compared to directed topic models, one drawback of undirected latent subspace models is that their interpretation is generally hard because of the unidentifiability issue [41]. Although our transformation retains the discriminative power, more elegant methods (e.g., imposing nonnegative constraints on parameter weights) are needed to improve the interpretability. Another potential limitation of such latent subspace models is that they do not have an explicit control on the sparsity of the discovered latent representations. Sparsity is desirable for large-scale applications where the dimensionality of the latent representations can be tens of thousands. We plan to do systematic studies along these lines. We are also interested in large-scale image annotation and classification, as motivated by the very exciting work [42], where dealing with noisy labeling information is important and challenging in order to learn a robust large-margin model. Finally, we plan to perform more investigation of the large-margin learning method on structured multiview data analysis, e.g., on text mining [38] and computer vision [22] applications.

ACKNOWLEDGMENTS

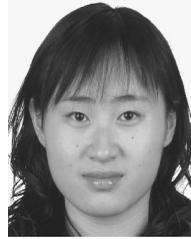
The authors thank the reviewers for their helpful comments. Part of this work was done while Ning Chen was a

visiting researcher at Carnegie Mellon University (CMU) under a CSC fellowship from China. Ning Chen, Jun Zhu, and Fuchun Sun are supported by National Key Project for Basic Research of China (Grant Nos. 2013CB329403, 2012CB316301), Tsinghua Self-innovation Project (Grant Nos. 20121088071, 20111081111), NNSF of China (Grant Nos: 60820304 and 91120011). Jun Zhu was supported by US Office of Naval Research (ONR) N000140910758 at CMU. Eric Poe Xing is supported by ONR N000140910758, US National Science Foundation (NSF) IIS-0713379, NSF Career DBI-0546594, and an Alfred P. Sloan Research Fellowship. Ning Chen and Jun Zhu contributed equally to this paper.

REFERENCES

- [1] S. Akaho, "A Kernel Method for Canonical Correlation Analysis," *Proc. Int'l Meeting on Psychometric Soc.*, 2001.
- [2] K. Ando and T. Zhang, "Two-View Feature Generation Model for Semi-Supervised Learning," *Proc. Int'l Conf. Machine Learning*, 2007.
- [3] D.M. Blei and M.I. Jordan, "Modeling Annotated Data," *Proc. ACM Int'l Conf. Information Retrieval*, pp. 127-134, 2003.
- [4] D.M. Blei and J.D. McAuliffe, "Supervised Topic Models," *Proc. Advances in Neural Information Processing Systems*, 2007.
- [5] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research*, vol. 3, pp. 993-1022, 2003.
- [6] A. Blum and T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training," *Proc. Ann. Conf. Learning Theory*, 1998.
- [7] U. Brefeld and T. Scheffer, "Co-EM Support Vector Learning," *Proc. Int'l Conf. Machine Learning*, 2004.
- [8] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.
- [9] K. Chaudhuri, S.M. Kakade, K. Livescu, and K. Sridharan, "Multi-View Clustering via Canonical Correlation Analysis," *Proc. Int'l Conf. Machine Learning*, 2009.
- [10] N. Chen, J. Zhu, and E.P. Xing, "Predictive Subspace Learning for Multi-View Data: A Large Margin Approach," *Proc. Advances in Neural Information Processing Systems*, 2010.
- [11] C.M. Christoudias, R. Urtasun, and T. Darrell, "Multi-View Learning in the Presence of View Disagreement," *Proc. Conf. Uncertainty in Artificial Intelligence*, 2008.
- [12] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.T. Zheng, "NUS-WIDE: A Real-World Web Image Database from National University of Singapore," *Proc. Int'l Conf. Image and Video Retrieval*, 2009.
- [13] K. Crammer and Y. Singer, "On the Algorithmic Implementation of Multiclass Kernel-Based Vector Machines," *J. Machine Learning Research*, vol. 2, pp. 265-292, 2001.
- [14] M. Culp, G. Michailidis, and K. Johnson, "On Multi-View Learning with Additive Models," *Annals of Applied Statistics*, vol. 3, no. 1, pp. 292-318, 2009.
- [15] T. Diethe, D.R. Hardoon, and J. Shawe-Taylor, "Multiview Fisher Discriminant Analysis," *Proc. NIPS Workshop Learning from Multiple Sources*, 2008.
- [16] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part Based Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627-1645, Sept. 2010.
- [17] V. Ferrari, T. Tuytelaars, and L.V. Gool, "Integrating Multiple Model Views for Object Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004.
- [18] D. Foster, S. Kakade, and T. Zhang, "Multi-View Dimensionality Reduction via Canonical Correlation Analysis," Technical Report TR-2008-4, TTI-Chicago, 2008.
- [19] Y. Freund, "An Adaptive Version of the Boost by Majority Algorithm," *Machine Learning*, vol. 43, no. 3, pp. 293-318, 2001.
- [20] K. Fukumizu, F. Bach, and M. Jordan, "Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces," *J. Machine Learning Research*, vol. 5, pp. 73-99, 2004.

- [21] K. Ganchev, J.V. Graca, J. Blitzer, and B. Taskar, "Multi-View Learning over Structured and Non-Identical Outputs," *Proc. Conf. Uncertainty in Artificial Intelligence*, 2008.
- [22] D. Gökulp and S. Aksoy, "Scene Classification Using Bag-of-Regions Representations," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007.
- [23] G.E. Hinton, "Training Products of Experts by Minimizing Contrastive Divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771-1800, 2002.
- [24] H. Hotelling, "Relations between Two Sets of Variates," *Biometrika*, vol. 28, nos. 3/4, pp. 321-377, 1936.
- [25] T. Joachims, "Making Large-Scale SVM Learning Practical," *Advances in Kernel Methods-Support Vector Learning*, pp. 169-184, MIT press, 1999.
- [26] S.M. Kakade and D.P. Foster, "Multi-View Regression via Canonical Correlation Analysis," *Proc. Ann. Conf. Learning Theory*, 2007.
- [27] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," *Proc. Int'l Conf. Machine Learning*, 2001.
- [28] H. Larochelle and Y. Bengio, "Classification Using Discriminative Restricted Boltzmann Machines," *Proc. Int'l Conf. Machine Learning*, 2008.
- [29] D. Liu and J. Nocedal, "On the Limited Memory BFGS Method for Large Scale Optimization," *Math. Programming*, vol. 45, pp. 503-528, 1989.
- [30] D.G. Lowe, "Object Recognition from Local Scale-Invariant Features," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1999.
- [31] A. McCallum, C. Pal, G. Druck, and X. Wang, "Multi-Conditional Learning: Generative/Discriminative Training for Clustering and Classification," *Proc. Nat'l Conf. Artificial Intelligence*, 2006.
- [32] R. Salakhutdinov and G.E. Hinton, "Replicated Softmax: An Undirected Topic Model," *Proc. Advances in Neural Information Processing Systems*, 2009.
- [33] A.J. Smola and B. Scholkopf, "A Tutorial on Support Vector Regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199-222, 2003.
- [34] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. VanGool, "Towards Multi-View Object Class Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006.
- [35] A. Torralba, K.P. Murphy, and W.T. Freeman, "Sharing Visual Features for Multiclass and Multiview Object Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 854-869, May 2007.
- [36] L. van der Maaten and G.E. Hinton, "Visualizing Data Using t-SNE," *J. Machine Learning Research*, vol. 9, pp. 2579-2605, 2008.
- [37] M.J. Wainwright and M.I. Jordan, "Graphical Models, Exponential Families, and Variational Inference," *Foundations and Trends in Machine Learning*, vol. 1, nos. 1/2, pp. 1-305, 2008.
- [38] H.M. Wallach, "Topic Modeling: Beyond Bag-of-Words," *Proc. Int'l Conf. Machine Learning*, 2006.
- [39] C. Wang, D.M. Blei, and L. Fei-Fei, "Simultaneous Image Classification and Annotation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009.
- [40] M. Welling and G.E. Hinton, "A New Learning Algorithm for Mean Field Boltzmann Machines," *Proc. Int'l Conf. Artificial Neural Networks*, 2001.
- [41] M. Welling, M. Rosen-Zvi, and G.E. Hinton, "Exponential Family Harmoniums with an Application to Information Retrieval," *Proc. Advances in Neural Information Processing Systems*, pp. 1481-1488, 2004.
- [42] J. Weston, S. Bengio, and N. Usunier, "Large Scale Image Annotation: Learning to Rank with Joint Word-Image Embeddings," *Proc. European Conf. Machine Learning*, 2010.
- [43] E.P. Xing, M.I. Jordan, and S. Russell, "A Generalized Mean Field Algorithm for Variational Inference in Exponential Families," *Proc. Conf. Uncertainty in Artificial Intelligence*, 2003.
- [44] E.P. Xing, R. Yan, and A.G. Hauptmann, "Mining Associated Text and Images with Dual-Wing Harmoniums," *Proc. Conf. Uncertainty in Artificial Intelligence*, 2005.
- [45] J. Yang, Y. Liu, E.P. Xing, and A.G. Hauptmann, "Harmonium Models for Semantic Video Representation and Classification," *Proc. SIAM Conf. Data Mining*, 2007.
- [46] C. Yu and T. Joachims, "Learning Structural SVMs with Latent Variables," *Proc. Int'l Conf. Machine Learning*, 2009.
- [47] J. Zhang, Z. Ghahramani, and Y. Yang, "Flexible Latent Variable Models for Multi-Task Learning," *Machine Learning*, vol. 73, no. 3, pp. 221-242, 2008.
- [48] J. Zhu, A. Ahmed, and E.P. Xing, "MedLDA: Maximum Margin Supervised Topic Models for Regression and Classification," *Proc. Int'l Conf. Machine Learning*, 2009.
- [49] J. Zhu, L. Li, L. Feifei, and E.P. Xing, "Large Margin Training of Upstream Scene Understanding Models," *Proc. Advances in Neural Information Processing Systems*, 2010.
- [50] J. Zhu and E.P. Xing, "Conditional Topic Random Fields," *Proc. Int'l Conf. Machine Learning*, 2010.
- [51] J. Zhu, E.P. Xing, and B. Zhang, "Partially Observed Maximum Entropy Discrimination Markov Networks," *Proc. Advances in Neural Information Processing Systems*, pp. 1977-1984, 2008.



Ning Chen received the PhD degree from the Department of Computer Science and Technology, Tsinghua University, China. She is currently a postdoctoral researcher in the Department of Computer Science and Technology, Tsinghua University. She was a visiting student in the Machine Learning Department, Carnegie Mellon University. Her research interests are primarily in machine learning, especially in probabilistic graphical models, max margin learning, and Bayesian nonparametrics.



from various fields. He is a member of the IEEE.

Jun Zhu received the BS, MS, and PhD degrees, all from the Department of Computer Science and Technology (CS&T), Tsinghua University, China, where he is currently an associate professor. He was a project scientist and postdoctoral fellow in the Machine Learning Department, Carnegie Mellon University. His research interests are primarily in developing statistical machine learning methods to understand scientific and engineering data arising



Fuchun Sun received the BS and MS degrees from the China Naval Aeronautical Engineering Academy and the PhD degree from the Department of Computer Science and Technology, Tsinghua University. He is currently a professor in the Department of Computer Science and Technology, Tsinghua University. His research interests include neural-fuzzy systems, variable structure control, networked control systems, and robotics. He is a senior member of the IEEE.



involving automated learning, reasoning, and decision making in high-dimensional, multimodal, and dynamic possible worlds in social and biological systems. He is a senior member of the IEEE.

Eric Poe Xing received the BS degree from Tsinghua University, China. He received the PhD degree in molecular biology from Rutgers University, and another PhD degree in computer science from the University of California, Berkeley. He is currently an associate professor in the School of Computer Science, Carnegie Mellon University. His research interests lie in the development of machine learning and statistical methodology, especially for solving problems

► For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.