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, geo-tags, 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 multi-view data can be worked with, these approaches appear to enjoy limited practical benefits from the extra information present in multi-view 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
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).
3.1 Unsupervised Multi-view Latent Subspace MNs

Fig. 1 shows the structure of a two-view latent subspace MN which consists of two types of input data $X = \{X_j\}_{j=1}^N$ and $Z = (Z_j)_{j=1}^M$, each corresponding to a view, and a set of latent variables $H = \{H_k\}_{k=1}^K$ 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 edges between the input variables $X$, and likewise for $E_z$. We will use $e$ to denote one individual edge and use $X_e$ 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:

$$p(x) = r(x) \exp \left\{ \sum_{e \in E_x} \theta_e^x \phi(x_e) - A(\theta) \right\},$$

$$p(z) = s(z) \exp \left\{ \sum_{e \in E_z} \eta_e^z \psi(z_e) - B(\eta) \right\},$$

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

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

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 $X$, $Z$, and $H$. Specifically, we have

$$p(x, z, h) \propto \exp \left\{ \sum_{e \in E_x} \theta_e^x \phi(x_e) + \sum_{e \in E_z} \eta_e^z \psi(z_e) + \sum_k \lambda_k^h \varphi(h_k) ight. \left. + \sum_{e \in E_x,k} \phi(x_e)^T W_{e,k} \varphi(h_k) + \sum_{e \in E_z,k} \psi(z_e)^T U_{e,k} \varphi(h_k) \right\},$$

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

$$p(x|h) = \exp \left\{ \sum_{e \in E_x} \tilde{\theta}_e^x \phi(x_e) - A(\tilde{\theta}) \right\},$$

$$p(z|h) = \exp \left\{ \sum_{e \in E_z} \tilde{\eta}_e^z \psi(z_e) - B(\tilde{\eta}) \right\},$$

$$p(h|x, z) = \prod_k \exp \left\{ \tilde{\lambda}_k^h \varphi(h_k) - C_k(\tilde{\lambda}_k) \right\},$$

where $\tilde{\theta}_e = \theta_e + \sum_k W_{e,k} \varphi(h_k)$, $\tilde{\eta}_e = \eta_e + \sum_k U_{e,k} \varphi(h_k)$, and $\tilde{\lambda}_k = \lambda_k + (\sum_{e \in E_x} \phi(x_e)^T W_{e,k} + \sum_{e \in E_z} \psi(z_e)^T U_{e,k})$. We can see that conditioned on the latent variables, both $p(x|h)$ and $p(z|h)$ define a Markov network, which is known as conditional random fields (CRFs) [27], where $h$ correspond to global conditions and $x$ or $z$ 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 $\Theta$ to denote all the parameters $(\theta, \eta, \lambda, W, U)$. 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 $X$ is a vector of discrete word features (e.g., image tags) and $Z$ 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

$$p(x_i = 1|h) = \text{Logistic}(a_i + W_e h),$$

$$p(z_j|h) = N(z_j | \sigma^2(\beta_j + U_j h), \sigma^2_j),$$

$$p(h_k|x, z) = N(h_k | x^T W_k + z^T U_k, 1),$$

where $W_e$ and $W_k$ denote the $i$th row and $k$th column of $W$, respectively. Likewise for $U_j$ and $U_k$.

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 $p(x,z)$, 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 $V$ be the parameters of a response variable model. Then, we need to define the joint distribution $p(x, z, h, y)$. 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 $H$ in order to define $p(x, z, h, y)$. For the general multiclass classification, where $y \in \{1, \ldots, T\}$, we define the conditional distribution using a softmax function:

$$p(y|h) = \frac{\exp(V^T f(y, h))}{\sum_y \exp(V^T f(y', h))},$$

where $f(y, h)$ is the feature vector whose elements from $(y - 1)K + 1$ to $yK$ are those of $h$ and all others are 0. $V$ is a stacking parameter vector of $T$ subvectors $V_y$, of which each one corresponds to a class label $y$. Then, the joint distribution $p(x, z, h, y)$ has the same form as in (2), but with an additional term of $V^T f(y, h) = V_y^T h$ in the exponential. For regression, where $y \in \mathbb{R}$, we define the conditional distribution as a normal distribution

$$p(y|h) = \mathcal{N}(y|V^T h, \sigma^2),$$

where $V$ is now a $K$-dim vector. Then, the joint distribution $p(x, z, h, y)$ has a similar form as in (2) with an additional term of $-\frac{1}{2\sigma^2}(y^2 - yV^T h)$ 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. 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 $F(y, h; V)$ as linear when latent variables $H$ are given, that is, $F(y, h; V) = V^T f(y, h)$, where $f$ and $V$ are defined the same as in (3). Now, the problem is how to consider the uncertainty of $H$ in the deterministic large-margin principle. Here, we take the expectation (i.e., first moment) of the latent variables $H$ and define the expected prediction rule:

$$y^* = \arg\max_y \mathbb{E}_{p(h|x,z)}[F(y, h; V)]$$

$$= \arg\max_y V^T \mathbb{E}_{p(h|x,z)}[f(y, h)],$$

where the expectation can be efficiently computed with the factorized form of $p(h|x, z)$ when $x$ and $z$ are fully observed. If missing values exist in $x$ or $z$, 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 $V^*$ that minimizes a loss function. Here, we minimize the hinge loss, as used in the very successful large-margin SVMs. Specifically, given training data $D = \{(x_d, z_d, y_d)\}_{d=1}^{D}$, the hinge loss of the expected predictive rule (5) is

$$R_{\text{hinge}}(V) = \sum_d \max_y [\Delta f_d(y) - V^T \mathbb{E}_{p(h|x_d, z_d)}[\Delta f_d(y)]]$$

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 $h$.

4. A discriminative method that maximizes the conditional likelihood $p(y|x, z)$ could be developed as in [28], but it could be inferior to a hybrid generative/discriminative method.
where $\Delta f_d(y)$ 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 $E_p[h|x,z] \Delta f_d(y) = E_p[h|x,z] [f(y_d,h)] - E_p[h|x,z] [f(y,h)]$. It can be proven that the hinge loss is an upper bound of the empirical error $R_{emp} \geq \sum_d \Delta f_d(y_d)$. Applying the principle of regularized risk minimization, we define the joint problem of learning a prediction model $V$ and a likelihood model $\Theta$ for fitting the input data as solving

$$ P_1: \min_{\Theta,V} L(\Theta) + \frac{1}{2}C_1 ||V||^2 + C_2 R_{hinge}(V), \quad (6) $$

where $L(\Theta) \triangleq -\sum_d \log p(x_d,z_d)$ is the negative data likelihood and $C_1$ and $C_2$ are nonnegative constants which can be selected via cross validation. Note that $R_{hinge}$ is also a function of $\Theta$.

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., solving this problem) and approximate the joint likelihood. Specifically, we derive a variational approximation $L^*(q_0,q_1)$ to represent the negative log-likelihood $L(\Theta)$:

$$ L^*(q_0,q_1) \triangleq R(q_0(x,z,h), p(x,z,h)) - R(q_1(x,z,h), p(x,z,h)), $$

where $R(q,p)$ is the relative entropy between distributions $q$ and $p$, $q_0$ is a variational distribution with $x$ and $z$ 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] that $q(x,z,h) = q(x)q(z)q(h)$.

Substituting the variational approximation $L^*(q_0,q_1)$ into problem (6), we get an approximate objective function $L(\Theta,V,q_0,q_1)$. Then, we can develop an alternating minimization method which iteratively minimizes $L(\Theta,V,q_0,q_1)$ over $(q_0,q_1)$ and $(\Theta,V)$. 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 $(\Theta,V)$ fixed and update each marginal as

$$ q(x) = p(x|E_q[h]|h), \quad q(z) = p(z|E_q[h]|h), $$

$$ q(h) = \prod_k p(h_k|E_q(x)|x), \quad E_q(x)|z). \quad (7) $$

For $q_0$, $(x,z)$ are clamped at their observed values and only $q_0(h)$ is updated, which can be very efficiently done due to its factored form. The distribution $q_1$ is achieved by 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 $h$ enters linearly in $\ln p(x,z|h)$. Please see [40] for more details.

Again, we can observe that both $q(x)$ and $q(z)$ are CRFs, with the expectation of $H$ 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 $V$ with fixed $\Theta$; this problem is learning a multiclass SVM [13], which can be efficiently solved with existing solvers; and 2) estimating $\Theta$ with $V$ fixed: this can be solved with subgradient descent. By defining $\Delta E[\cdot] \triangleq E_q[\cdot] - E_{q_0}[\cdot]$, we can compute the subgradient as follows: For $\Theta$, we have $\forall e \in E_x, \partial \lambda_e = \Delta E[\phi(x_e)];$ for $\eta$, we have $\forall e \in E_z, \partial \eta_e = \Delta E[\psi(z_e)];$ for $\lambda$, we have $\forall k, \partial \lambda_k = \Delta E[\phi(h_k)];$ and for $W$ and $U$, we have

$$ \partial W_e = \Delta E[\phi(x_e)\phi(h_e)^T] - C_2 \sum_d (V_{ydk} - V_{ydk}) \frac{\partial E_q[h_k]}{\partial W_e}, $$

$$ \partial U_e = \Delta E[\psi(z_e)\phi(h_e)^T] - C_2 \sum_d (V_{ydk} - V_{ydk}) \frac{\partial E_q[h_k]}{\partial U_e}, $$

where $\partial W_e = \arg \max_y \Delta f_d(y) + V^T E_q[h] [f(y,h)]$ is the loss-augmented prediction. The expectation $E_q[\phi(x_e)]$ is actually the count frequency of $\phi(x_e)$ on the training data $D$; likewise for $E_q[\psi(z_e)]$. With the above subgradients, we apply L-BFGS [29], which uses line search to choose a step size, to iteratively solve for the optimum $\Theta$ until convergence.

Note that in our integrated large-margin formulation, the subgradients corresponding to $W$ and $U$ 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 $y_d$ 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

$$ y^* \triangleq V^T E_p[h|x,z]|h, \quad (8) $$

where $V$ is a $K$-dim vector. To learn the prediction model $V$, 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
Now, we optimize the Lagrangian function \( L \), written as

\[
\mathcal{R}_c(V) \triangleq \sum_d \max(0, |y_d - V^T \mathbb{E}[h_d]| - \epsilon),
\]

where \( \epsilon \in \mathbb{R}_+ \) is the precision parameter, which is usually small, and we have defined \( \mathbb{E}[h_d] = \mathbb{E}_{\theta(z, x, z')}[h] \) 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:

\[
P_2 : \min_{\theta, V} L(\theta) + \frac{1}{2} C_1 \| V \|_2^2 + C_2 \mathcal{R}_c(V),
\]

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 \( P_2 \) by employing a subgradient descent method to learn the parameters \( \theta \), here we use a Lagrangian method to solve an equivalent constrained formulation of \( P_2 \). One reason is that the loss \( \mathcal{R}_c \) is a bit more complicated than \( \mathcal{R}_{\text{hinge}} \) because of the nondifferentiable absolute operation within the max function. Specifically, problem \( P_2 \) can be equivalently written as

\[
P_2' : \min_{\theta, V, \xi, \zeta} L(\theta) + \frac{1}{2} C_1 \| V \|_2^2 + C_2 \sum_d (\xi_d + \zeta_d)
\]

s.t. \( \forall d : \)

\[
\begin{align*}
y_d - V^T \mathbb{E}[h_d] &\leq \epsilon + \xi_d \\
y_d - V^T \mathbb{E}[h_d] &\leq \epsilon + \zeta_d, \\
\xi_d, \zeta_d &\geq 0,
\end{align*}
\]

where \( \xi_d \) and \( \zeta_d \) are slack variables.

The constrained problem \( P_2' \) 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 \( \mathcal{L}''(q_0, q_1) \). 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 \( \mathcal{L}''(q_0, q_1) : \)

\[
L = \mathcal{L}''(q_0, q_1) + \frac{1}{2} C_1 \| V \|_2^2 + C_2 \sum_d \xi_d + \zeta_d \]

\[- \sum_d (v_d \xi_d + v'_d \zeta_d) - \sum_d \{ \mu_d (\epsilon + \xi_d - y_d + V^T \mathbb{E}[h_d]) \\
+ \mu'_d (\epsilon + \zeta_d + y_d - V^T \mathbb{E}[h_d]) \}. \]

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 \( \Theta \) with \( \mu_d \) and \( \mu'_d \) fixed. This can be solved with gradient descent (e.g., using L-BFGS [29] as in the classification model), where the gradients for \( (\theta, \eta, \lambda) \) are as before and for \( (W, U) \) we have

\[
\frac{\partial W^k_c}{\partial \epsilon} = \Delta \mathbb{E}\{\phi(x_k)\varphi(h_k)^T\} \quad \frac{\partial W^k_c}{\partial \epsilon} = \Delta \mathbb{E}\{\psi(z_k)\varphi(h_k)^T\}
\]

\[
\frac{\partial U^k_c}{\partial \epsilon} = \Delta \mathbb{E}\{\phi(h_k)\varphi(h_k)^T\} - \sum_d (\mu_d - \mu'_d) V_k \frac{\partial \mathbb{E}_q[h_k]}{\partial W^k_c}
\]

3. Estimating the Lagrange multipliers \( \{\mu_d, \mu'_d\} \). By setting \( \partial L/\partial \xi_d, \partial L/\partial \zeta_d, \partial L/\partial V = 0 \) and exploring the KKT conditions, we can get

\[
V = \frac{1}{C_1} \sum_d (\mu_d - \mu'_d) \mathbb{E}[h_d].
\]

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

\[
\max \left\{ \frac{1}{2C_1} \| \sum_d (\mu_d - \mu'_d) \mathbb{E}[h_d] \|_2^2 - \sum_d (\epsilon (\mu_d + \mu'_d) - y_d (\mu_d - \mu'_d)) \right\}
\]

s.t. \( \forall d : \mu_d, \mu'_d \in [0, C_2] \),

which can be solved using an existing algorithm like SVM-light [25] to obtain \( \mu_d \) and \( \mu'_d \).

Again, we can see that in this integrated large-margin formulation for regression, the gradients of \( W \) and \( U \) contain an additional term encoded with \( \mu_d \) and \( \mu'_d \). Similarly as in the classification model, this additional term introduces a regularization effect to the latent subspace model. If the prediction \( V^T \mathbb{E}[h_d] \) differs from the true value \( y_d \) with the absolute gap larger than \( \epsilon \), the Lagrangian multipliers \( \mu_d \) or \( \mu'_d \) (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 \( (x, z) \), where \( x \) is a vector of discrete word features (e.g., image tags) and \( z \) is a vector of real-valued features (e.g., color histograms).
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 datasets are TRECVID2003 [44], 13-class-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. 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
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 probability 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, ...
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. 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 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 5-topic DWH on the hotel review data. As in [50], we denote the five rating scores from small to large by R1, R2, ..., R5. We also show the per-rating average distributions over topics in the left part, which are 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 Tk.

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., R5 and R4), but has smaller probabilities (drops to near zero) on documents with lower rating scores (e.g., R1 and R2). 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 R1 to R5. Topic T1 is kind of neutral, which has the highest probability on representing the documents with a neutral rating score (e.g., R3 or R4) 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,”

9. We first turn the expected value of 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)).
“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\textsuperscript{multiclass10} 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 + SVM\textsuperscript{11}) 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. 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).
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—$X$ for tag and $Z$ 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 $P \times C$ observation matrix $x$, where $P$ is the number of paragraphs in this document and $C$ is the vocabulary size. Each row $x_p$ is a vector, of which each element $x_{pi}$ denotes whether word $i$ appears in paragraph $p$. Each column $x_i$ 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 $x_i$, while assuming that different $x_i$s

<table>
<thead>
<tr>
<th></th>
<th>MMH</th>
<th>DWH</th>
<th>TWH</th>
<th>sLDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F1_{15}$</td>
<td>0.245</td>
<td>0.202</td>
<td>0.218</td>
<td>0.146</td>
</tr>
<tr>
<td>$F1_{16}$</td>
<td>0.258</td>
<td>0.208</td>
<td>0.228</td>
<td>0.159</td>
</tr>
<tr>
<td>$F1_{15}$</td>
<td>0.262</td>
<td>0.210</td>
<td>0.236</td>
<td>0.169</td>
</tr>
<tr>
<td>$F1_{16}$</td>
<td>0.259</td>
<td>0.208</td>
<td>0.240</td>
<td>0.171</td>
</tr>
<tr>
<td>$F1_{17}$</td>
<td>0.256</td>
<td>0.206</td>
<td>0.239</td>
<td>0.175</td>
</tr>
</tbody>
</table>

12. The cosine similarity between vectors $x_1$ and $x_2$ is $\frac{x_1 \cdot x_2}{\|x_1\| \|x_2\|}$.
are conditional independent. More formally, we define the conditional distribution
\[ p(x|h) = \prod_{i=1}^N p(x_i|h), \]
where each
\[ p(x_i|h) \]
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 \textit{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 \textit{structMMH} outperforms the unstructured MMH and the two-stage method (denoted by \textit{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. 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.](image-url)

![Fig. 9. (a) Classification accuracy of structured MMH and DWH models, and (b) training and testing time on hotel review data [50] for regression.](image-url)
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 identifiability 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-0546954, and an Alfred P. Sloan Research Fellowship. Ning Chen and Jun Zhu contributed equally to this paper.

REFERENCES
