Learning Harmonium Models with Infinite Latent Features

Ning Chen, Jun Zhu, *Member, IEEE,* Fuchun Sun, *Senior Member, IEEE,* Bo Zhang, *Senior Member, IEEE*

***Abstract*—Undirected latent variable models represent an important class of graphical models that have been successfully developed to deal with various tasks. One common challenge in learning such models is to determine the number of hidden units which is unknown a priori. Although Bayesian nonparametrics have provided promising results in bypassing the model selection problem in learning directed Bayesian Networks, very few efforts have been made towards applying Bayesian nonparametrics to learn undirected latent variable models. In this paper, we present the infinite exponential family Harmonium (iEFH), a bipartite undirected latent variable model that automatically determines the number of latent units from an unbounded pool. We also present two important extensions of iEFH to: 1) multi-view iEFH for dealing with heterogeneous data; and 2) infinite maximum-margin Harmonium (iMMH) for incorporating supervising side information to learn predictive latent features. We develop variational inference algorithms to learn model parameters. Our methods are computationally competitive due to the avoidance of selecting the number of latent units. Our extensive experiments on real image and text data sets appear to demonstrate the benefits of iEFH and iMMH inherited from Bayesian nonparametrics and max-margin learning. Such results were not available until now, and contribute to expand the scope of Bayesian nonparametrics to learn the structures of undirected latent variable models.**

***Index Terms*—Bayesian nonparametrics, exponential family Harmoniums, max-margin learning.**

I. Introduction

LEARNING probabilistic graphical models with latent variables can be useful for discovering latent semantic representations from large collections of data. Both directed Bayesian networks (e.g., latent Dirichlet allocation and its extensions [7][5]) and undirected Markov networks (MNs) (e.g., exponential family Harmoniums and its extensions [40][41]) have been extensively studied for such a purpose to discover latent representations based on the input features of unlabeled data, which can be text documents, images or even network entities. Besides the input contents, in many practical applications we can easily obtain useful side information. For instance, when online users post their reviews for products or restaurants, they usually associate each review with a rating score or a thumb-up/thumb-down opinion. Also, there is an increasing trend towards using online crowdsourcing services (such as Amazon Mechanical Turk) to collect large collections of labeled data with a reasonably low price. In order to leverage such supervising side information for discovering latent representations that could be more predictive for tasks such as classification and regression, many research efforts have been made to develop supervised latent variable models. Representative work under this trend includes both supervised Bayesian networks (e.g., supervised latent Dirichlet allocation (LDA) [6], maximum entropy discrimination LDA [43] and discriminative LDA [27]) and supervised Markov networks (e.g., discriminative restricted Boltzmann machine (RBM) [28], hierarchical Harmoniums [42] and max-margin latent space Markov networks [10]).

N. Chen, J. Zhu, F. Sun and B. Zhang are with the Department of Computer Science and Technology, National Lab of Information Science and Technology, State Key Lab of Intelligent Technology and Systems, Tsinghua University. Beijing, 100084 China. e-mail: ({ningchen, dcszj, fcsun, dcszb}@mail.tsinghua.edu.cn).

However, one common challenge in learning latent variable models (LVMs) is to determine the unknown number of latent units. A typical model selection procedure like cross-validation or likelihood ratio test [30] could be computationally expensive by enumerating, learning and comparing many candidate models. Alternatively, the recent fast-growing developments of Bayesian nonparametrics have shown promise on bypassing the model selection step. By imposing an appropriate stochastic process prior on a space of flexible models, such as models possessing an infinite number of components [2] or models having an infinite number of latent features [22], Bayesian nonparametric methods could automatically resolve the model complexity from empirical data and could further adapt the model complexity when the observed data changes, e.g., using more components or features to fit a larger data set.

Although much success on developing nonparametric latent variable models has been demonstrated in the context of directed Bayesian networks for both exploratory (e.g., discovering latent semantic representations) [3][1] and predictive (e.g., classification) [35][45][44] tasks, very few successful examples have been reported on utilizing Bayesian nonparametrics to solve the model selection problem in undirected Markov networks in the presence of latent variables, which represent an important class of LVMs and have complementary advantages (e.g., fast inference) compared to directed Bayesian networks. Various latent Markov networks, such as restricted Boltzmann machines (RBMs) [24] and exponential family Harmoniums (EFH) [40], have been successfully developed for image classification, retrieval and annotation tasks [41][10][42].

One challenge that has potentially lead to such slow progress on learning nonparametric undirected latent variable models is on dealing with a usually intractable partition function. Although exact algorithms do not exist, recent developments on approximate learning algorithms [39][24][31] have encouraged a systematical investigation of learning flexible undirected latent variable models, especially considering their broad practical applications.

This paper presents *infinite exponential family Harmoniums* (iEFH), a Bayesian EFH model having an unbounded number of latent units or latent features. To select a finite subset of features, we associate each latent feature with a data-specific binary variable and impose the sparsity-inducing Indian buffet process (IBP) [22] prior on the entire collection of binary variables (represented as a binary matrix). The resulting model is a

chain graph model, which has different Markov properties [21] from directed Bayesian networks. Moreover, we generalize iEFH to the *multi-view iEFH* model to deal with data from multiple sources or domains, and to the supervised *infinite max-margin Harmoniums* (iMMH), which directly regularizes the latent representations by imposing max-margin constraints with a linear expectation operator for discovering latent representations that are specialized for predictive tasks. The key insights to develop iMMH, which integrates max-margin learning and Bayesian nonparametric techniques, come from the recently developed regularized Bayesian inference [44], a Bayesian computational framework to consider regularization on the post-data posterior distribution. To learn the model parameters, we resort to approximate gradient methods by using the successful contrastive divergence [39][24] techniques. Finally, our experiments on a variety of real data sets demonstrate competitive results compared with a number of baseline methods.

The rest paper is structured as follows. Section II presents the related work. Section III introduces the infinite EFH model, together with approximate learning methods. Section IV and Section V present two extensions of iEFH to deal with multi-view data and learning predictive latent representations, respectively. Section VI presents empirical studies on image and document classification. Finally, Section VII concludes and discusses future research directions.

II. Related Work

Latent variable models (LVMs) can be divided into two categories — directed Bayesian network LVMs and undirected Markov network LVMs. Bayesian networks with one layer of observed variables and one/multiple layers of latent variables have been the dominating formalisms in many research fields. Most latent variable models fall into this category, including the mixture of Gaussian [4], probabilistic latent semantic indexing (pLSI) [25], probabilistic PCA [38], latent Dirichlet allocation (LDA) [7], etc. All these models are finite latent variable models and usually the number of latent classes/features needs to be artificially pre-specified. Recently, Bayesian nonparametric methods have shown great promise in learning flexible directed LVMs. For example, in latent class models, Dirichlet process (DP) [2] is often used as a prior to deal with the model selection problem (i.e., resolving the number of mixture components) for clustering, density estimation and supervised tasks [35]. In latent feature models, the Indian buffet process (IBP) [22] prior is often used to automatically resolve the number of latent features from an infinite pool of candidates for factor analysis or feature learning in support vector machines [44]. [1] uses the cascading IBP prior to learn the structure of layered deep belief network with an unbounded number of layers and an unbounded number of units at each layer.

The undirected analogue of the above directed families that enjoys nice properties (e.g., fast inference and easy interpretability) has been developed, including the exponential family of Harmoniums [40][39] and its special cases of restricted Boltzmann machine (RBM) [23] and influence combination model [20] that have discrete latent units. They are usually more efficient in posterior inference due to the weak independent assumption that observation variables are conditionally independent given a set of latent variables. Although Harmoniums have been successfully turned into practical methods for information retrieval [40], image classification, retrieval and annotation [41], to the best of our knowledge, very few attempts have been made towards using Bayesian nonparametric methods to bypass the model selection problem. To address this problem, in this paper, we first apply an IBP prior on the latent variables and develop an effective contrastive divergence approximation method to avoid the intractable normalization factor, which can bypass the model selection problem, as will be stated in Section III.

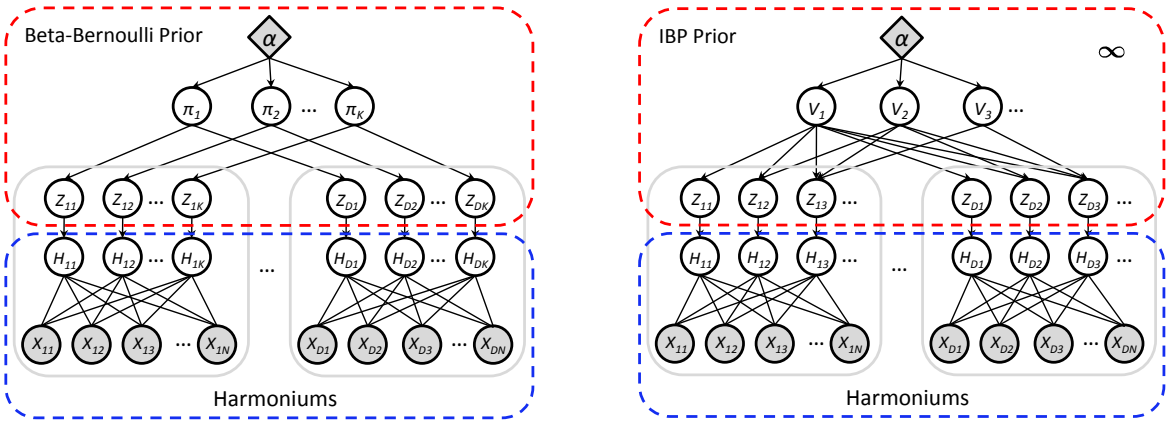
So far, most of the aforementioned latent variable models are unsupervised and unable to perform prediction tasks without using an additional classifier. Many supervised directed and undirected LVMs have been developed. In the directed supervising Bayesian LVMs, the supervised latent Dirichlet allocation (sLDA) [6] and discriminative latent Dirichlet allocation (discLDA) [27] are defined based on the joint/conditional distribution. The models are learned with maximum likelihood estimation (MLE). In order to learn more predictive supervised topic model, maximum entropy discrimination LDA (MedLDA) [43] is proposed by using max-margin learning but restricted to finite models. There are also several infinite Bayesian LVMs that have been developed for prediction tasks, including the DP mixture of generalized linear models [35] where a likelihood model is defined on the response variables that contains a normalization factor, and the recent work [45] that integrates Bayesian nonparametric methods with max-margin machines to obtain an infinite mixture of nonlinear large-margin kernel classifiers. And [44] uses the IBP prior to automatically determine the dimensionality of latent features for learning SVM classifiers and the latent projection matrix for multi-task learning. However, the undirected supervised latent variable models (e.g., supervised hierarchical Harmonium [42] and discriminative Restricted Boltzmann machines [28]) are all learned based on maximum likelihood estimation, which may not yield good prediction performance [42]. [10] proposes a max-margin Harmonium model that has shown superior performance in prediction but is also restricted to finite latent variable models; furthermore, its dimensionality is pre-specified and fixed. This has motivated us to develop the infinite max-margin Harmonium (iMMH) that can use Bayesian nonparametric techniques to automatically resolve the dimension of latent units.

III. Infinite Exponential Family Harmoniums

Now, we formally present the *infinite exponential family Harmonium* (iEFH) model, starting with a brief recapitulation of the basic Harmonium models.

A. Exponential Family Harmoniums

An exponential family Harmonium (EFH) model is a bipartite Markov network with two layers of variables, i.e., the input variables $\mathbf{X}$ and the latent variables $\mathbf{H} = \{H_1, \cdots, H_K\}$,



(a) (b)

Fig. 1. Graphical illustration of (a) Finite Bayesian spike & slab EFH model; and (b) Infinite EFH, using the IBP prior. The lower part in the blue dashed box is the basic Harmonium model, the upper part in the red dashed box is the Indian buffet process prior. Best viewed in color.

as shown in the lower part of Fig. 1(a). For each data $d \in \{1, \cdots, D\}$, $\mathbf{x}_d = \{x_{d1}, \cdots, x_{dN}\}$ denotes the set of observed input features. By using exponential family distributions, EFH allows both $\mathbf{X}$ and $\mathbf{H}$ to take discrete or continuous values. When $\mathbf{X}$ and $\mathbf{H}$ are continuous, the joint distribution can be defined as

$$p(\mathbf{x}_d, \mathbf{h}_d|\Theta) \propto \exp\left\{ \begin{array}{l} \boldsymbol{\eta}^\top \mathbf{x}_d - \frac{\mathbf{x}_d^\top \mathbf{x}_d}{2\sigma_{d1}^2} + \mathbf{x}_d^\top \mathbf{W} \mathbf{h}_d \\ + \boldsymbol{\beta}^\top \mathbf{h}_d - \frac{\mathbf{h}_d^\top \mathbf{h}_d}{2\sigma_{d2}^2} \end{array} \right\}, \quad (1)$$

where $\Theta = \{\boldsymbol{\eta}, \boldsymbol{\beta}, \mathbf{W}, \sigma_{d1}^2, \sigma_{d2}^2\}$ are model parameters. We can derive that the conditional distributions $p(\mathbf{x}_d|\mathbf{h}_d, \Theta)$ and $p(\mathbf{h}_d|\mathbf{x}_d, \Theta)$ are both well-defined isotropic Gaussian distributions with covariance matrices being $\sigma_{d1}^2 I$ and $\sigma_{d2}^2 I$, respectively. Thus, we can easily draw samples from such distributions or perform variational inference, a substep in a contrastive divergence learning method. Note that to make the normalization constant in Eq. (1) be finite, some constraints are needed, such as the upper bound constraint on the inputs [13] or the more sophisticated solutions in [14]. We adopted the former one for its simplicity. For discrete inputs, we set σ_{d1}^2 equal to ∞ and the quadratic term $\frac{\mathbf{x}_d^\top \mathbf{x}_d}{2\sigma_{d1}^2}$ vanishes.

Although EFH can be efficient in inference because of its conditional independence structure (i.e., $\mathbf{H}$ are conditionally independent given the observations) and has been applied to various applications [41][10], the number of latent units K is usually difficult to determine[1]. A general selection procedure that enumerates, learns and compares many different candidate models with various K values could be expensive. Below, we present iEFH as a nonparametric Bayesian technique to automatically resolve the unknown number K from empirical data. It is worth noting that our methods are applicable to restricted Boltzmann machines (RBM) [24], which are Harmonium models but typically with binary $\mathbf{X}$ and $\mathbf{H}$ variables [2].

[1]The paper [12] presented a case where increasing the number of latent features (within a particular range) boosts the performance. But we still need some mechanism to automatically determine an appropriate K.

[2]See [24] for a discussion of the RBM with non-binary units.

B. A Finite Beta-Bernoulli Bayesian Spike and Slab EFH

We first present a finite Bayesian EFH using binary variables to select subset from a potentially large but finite set of latent features. Fig. 1(a) illustrates the graphical structure of Beta-Bernoulli EFH, where the lower part in the blue dashed box is the basic Harmoniums, the upper part in the red dashed box shows the two layer Beta-Bernoulli prior. As we shall see, this finite model naturally generalizes to an infinite model. Formally, we introduce a set of binary variables $\mathbf{Z}_d$ for each data d, with Z_{dk} associated with the latent feature H_{dk}. Then, the *effective* latent feature for data d is

$$\widetilde{\mathbf{h}}_d = \mathbf{z}_d \circ \mathbf{h}_d,$$

where $\circ$ is the element-wise multiplication of two vectors. The binary element $z_{dk} = 1$ indicates that data d possesses feature k; otherwise, data d does not possess feature k. These binary variables $\mathbf{Z}_d$ are known as *spike* variables and the real-valued $\mathbf{H}_d$ are known as *slab* variables [13].

We assume that Z_{dk} follows a Beta-Bernoulli distribution

$$z_{dk}|\pi_k \sim \text{Bernoulli}(\pi_k), \quad \pi_k \sim \text{Beta}(\frac{\alpha}{K}, 1), \quad (2)$$

where α is a hyper-parameter. The Beta-Bernoulli prior has a nice property that the expected number of non-zero entries in the matrix $\mathbf{Z}$ is $N\alpha$ [22]. The finite Bayesian EFH has the joint distribution

$$p(\boldsymbol{\pi}, \{\mathbf{x}_d, \mathbf{h}_d, \mathbf{z}_d\}|\Theta) = p(\boldsymbol{\pi}) \prod_d p(\mathbf{z}_d|\boldsymbol{\pi}) p(\mathbf{x}_d, \mathbf{h}_d|\mathbf{z}_d, \Theta), \quad (3)$$

where $p(\mathbf{x}_d, \mathbf{h}_d|\mathbf{z}_d, \Theta)$ is an EFH model that uses binary variables $\mathbf{z}_d$ to select hidden features

$$p(\mathbf{x}_d, \mathbf{h}_d|\mathbf{z}_d, \Theta) \propto \exp\left\{ \begin{array}{l} \boldsymbol{\eta}^\top \mathbf{x}_d - \frac{\mathbf{x}_d^\top \mathbf{x}_d}{2\sigma_{d1}^2} + \mathbf{x}_d^\top \mathbf{W} \tilde{\mathbf{h}}_d \\ + \boldsymbol{\beta}^\top \tilde{\mathbf{h}}_d - \frac{\mathbf{h}_d^\top \mathbf{h}_d}{2\sigma_{d2}^2} \end{array} \right\}. \quad (4)$$

The above formulation leads to a hierarchical *Bayesian spike and slab exponential family Harmoniums*. It is worth noting that both the Beta-Bernoulli model and the iEFH to be presented can be reduced to use binary latent features by constraining that $\mathbf{H}$ are deterministic, i.e., taking a constant value such as 1.

C. An Infinite EFH

Now, we extend the above Bayesian spike and slab EFH to the infinite iEFH by letting $K \to \infty$. The insights are as follows. First, as shown in [22], the finite Beta-Bernoulli prior can be extended to the infinite case using the lof-equivalent classes of matrices and the resulting marginal distribution of $\mathbf{Z}$ is the well-defined IBP. Second, given the binary variables $\mathbf{z}_d$, which are sparse and have finite non-zeros in expectation under the IBP prior, we can define the EFH model $p(\mathbf{x}_d, \mathbf{h}_d|\mathbf{z}_d, \Theta)$ as in Eq. (4). Fig. 1(b) shows the graphical structure of iEFH, where we have used the stick-breaking representation of IBP [36] with intermediate variables $\mathbf{V} = \{V_1, V_2, \cdots\}$. Specifically, the IBP prior on $\mathbf{Z}$ can be described as a generative model

$$\forall d,\ z_{dk}|\pi_k \sim \text{Bernoulli}(\pi_k)$$

$$\pi_1 = v_1,\ \pi_k = v_k\pi_{k-1} = \prod_{i=1}^{k} v_i,\ v_i \sim \text{Beta}(\alpha, 1).$$

By the Markov property of a chain graph [21], the joint distribution of iEFH is

$$p(\mathbf{v}, \{\mathbf{x}_d, \mathbf{h}_d, \mathbf{z}_d\}|\Theta) = p(\mathbf{v}) \prod_d \big(p(\mathbf{z}_d|\mathbf{v}) p(\mathbf{x}_d, \mathbf{h}_d|\mathbf{z}_d, \Theta)\big) \quad (5)$$

where $p(\mathbf{x}_d, \mathbf{h}_d|\mathbf{z}_d, \Theta)$ is the same as in Eq. (4). Note that we have ignored the variables $\boldsymbol{\pi}$ in Eq. (5) because they are deterministic functions of $\mathbf{v}$.

1) Parameter learning by approximate gradient descent: We use the maximum likelihood estimation to learn the parameters Θ. Let $\mathcal{L}(\Theta)$ denote the negative log-likelihood, and let

$$\Delta\mathbb{E}\big[\,\cdot\,\big] \triangleq -\mathbb{E}_{p(\mathbf{v}, \{\mathbf{h}_d, \mathbf{z}_d\}|\{\mathbf{x}_d\})}\big[\,\cdot\,\big] + \mathbb{E}_{p(\mathbf{v}, \{\mathbf{h}_d, \mathbf{z}_d, \mathbf{x}_d\})}\big[\,\cdot\,\big],$$

where the distribution $p(\mathbf{v}, \{\mathbf{h}_d, \mathbf{z}_d\}|\{\mathbf{x}_d\})$ has $\mathbf{x}_d$ "clamped" to their input values, while the distribution $p(\mathbf{v}, \{\mathbf{h}_d, \mathbf{z}_d, \mathbf{x}_d\})$ has all variables free. Then, the likelihood gradients are $\nabla_{\eta_n}\mathcal{L} = \sum_d \Delta\mathbb{E}\big[x_{dn}\big]$, $\nabla_{\beta_k}\mathcal{L} = \sum_d \Delta\mathbb{E}\big[\widetilde{h}_{dk}\big]$ and $\nabla_{\mathbf{W}_{nk}}\mathcal{L} = \sum_d \Delta\mathbb{E}\big[x_{dn}\widetilde{h}_{dk}\big]$. For the variance parameters, we can fix them a priori or learn them using gradient descent in the log-space to avoid handling positive constraints. Let $t_{d1} = \log\sigma_{d1}^2$ and $t_{d2} = \log\sigma_{d2}^2$. Then we have

$$\nabla_{t_{d1}}\mathcal{L} = \frac{1}{2\sigma_{d1}^2}\Delta\mathbb{E}\big[\mathbf{x}_d^\top\mathbf{x}_d\big],\ \nabla_{t_{d2}}\mathcal{L} = \frac{1}{2\sigma_{d2}^2}\Delta\mathbb{E}\big[\mathbf{h}_d^\top\mathbf{h}_d\big].$$

Contrastive divergence gradient approximation: With the above gradients, we can perform gradient descent to update the model parameters, e.g., using the very effective quasi-Newton method [29]. Now, the question is how to infer the distributions $p(\mathbf{v}, \{\mathbf{h}_d, \mathbf{z}_d\}|\{\mathbf{x}_d\})$ and $p(\mathbf{v}, \{\mathbf{x}_d, \mathbf{h}_d, \mathbf{z}_d\})$. Since exact inference is intractable, we must resort to approximate methods. One successful scheme to approximate the likelihood gradients is *contrastive divergence* [24]. Although MCMC sampling methods are widely used, we adopt the mean-field contrastive divergence [39], which derives a deterministic optimization problem that can be naturally extended for supervised learning as shown later. Specifically, let q_0 be a variational distribution to approximate $p(\mathbf{v}, \{\mathbf{h}_d, \mathbf{z}_d\}|\{\mathbf{x}_d\})$ via minimizing the KL-divergence $\text{KL}(q_0(\mathbf{v}, \{\mathbf{h}_d, \mathbf{z}_d\})\|p(\mathbf{v}, \{\mathbf{h}_d, \mathbf{z}_d\}|\{\mathbf{x}_d\}))$. We further constrain the feasible space of q_0 by making the truncated mean-field assumption

$$q_0(\mathbf{v}, \{\mathbf{h}_d, \mathbf{z}_d\}) = \Big(\prod_{k=1}^{T} q_0(v_k)\Big) \prod_{d=1}^{D} \Big(q_0(\mathbf{h}_d) \prod_{k=1}^{T} q_0(z_{dk})\Big),$$

where T is a sufficiently large truncation level, $q_0(v_k) = \text{Beta}(\gamma_{k1}, \gamma_{k2})$, and $q_0(z_{dk}) = \text{Bernoulli}(\nu_{dk})$. As shown in [18], the ℓ_1-distance truncation error of marginal distributions decreases exponentially as K increases. With this truncation assumption, we can efficiently perform the inference via an iterative procedure, as outlined below.

Inference: For $\mathbf{H}$, we have the mean-field update equations

$$q_0(\mathbf{h}_d) = \prod_{k=1}^{T} q_0(h_{dk}),$$

where $q_0(h_{dk}) = \mathcal{N}\big(\sigma_{d2}^2(\nu_{dk}\beta_k + \nu_{dk}\mathbb{E}_{q_0}[\mathbf{x}_d]^\top\mathbf{W}_{.k}), \sigma_{d2}^2\big)$ and $\mathbf{W}_{n.}$ ($\mathbf{W}_{.k}$) denotes the nth row (kth column) of $\mathbf{W}$. For variables $\mathbf{Z}$, the mean-field update equation for ν is:

$$\nu_{dk} = \text{Sigmoid}\Big(-\tau_1^k + \tau_2^k + \mathbb{E}_{q_0}[h_{dk}](\beta_k + \mathbb{E}_{q_0}[\mathbf{x}_d]^\top\mathbf{W}_{.k})\Big) \quad (6)$$

where $\tau_1^k = \mathbb{E}_{q_0}[\log(1 - \prod_{j=1}^{k} v_j)]$, $\tau_2^k = \sum_{j=1}^{k}\mathbb{E}_{q_0}[\log v_j]$. We can compute τ_2^k using the digamma function ψ as $\mathbb{E}_{q_0}[\log v_k] = \psi(\gamma_{k1}) - \psi(\gamma_{k1} + \gamma_{k2})$. For τ_1^k, we use the multinomial bound [18] when $k > 1$, which can be a good approximation and evaluated efficiently. For $k = 1$, we have $\mathbb{E}_{q_0}[\log(1 - \prod_{j=1}^{k} v_j)] = \psi(\gamma_{12}) - \psi(\gamma_{11} + \gamma_{12})$. The update equation for $\boldsymbol{\gamma}$ is the same as in [18].

After we have inferred q_0, one or several mean-field updates (initialized with q_0) are performed to reconstruct a distribution $q_1(\mathbf{v}, \{\mathbf{x}_d, \mathbf{h}_d, \mathbf{z}_d\})$, which approximates the model distribution $p(\mathbf{v}, \{\mathbf{x}_d, \mathbf{h}_d, \mathbf{z}_d\})$. We use one step of update in the experiments. For variables $\mathbf{X}$, we have the update equations

$$q_1(\mathbf{x}_d) = \prod_{n=1}^{N} q_1(x_{dn}),$$

where $q_1(x_{dn}) = \mathcal{N}\big(\sigma_{d1}^2(\eta_n + \mathbf{W}_{n.}(\boldsymbol{\nu}_d \circ \mathbb{E}_{q_1}[\mathbf{h}_d])), \sigma_{d1}^2\big)$. For the other variables, the mean-field update equations have the same forms as above, with q_0 replaced by q_1.

The above mean-field contrastive divergence method is in fact the gradient descent method to minimize the contrastive free energy (CF_1)[3]

$$CF_1(\Theta, q_0, q_1) \triangleq \text{KL}(q_0(\mathbf{v}, \{\mathbf{h}_d, \mathbf{z}_d\})\|p(\mathbf{v}, \{\mathbf{x}_d, \mathbf{h}_d, \mathbf{z}_d\})) - \text{KL}(q_1(\mathbf{v}, \{\mathbf{x}_d, \mathbf{h}_d, \mathbf{z}_d\})\|p(\mathbf{v}, \{\mathbf{x}_d, \mathbf{h}_d, \mathbf{z}_d\})),$$

which is an approximation of the negative log-likelihood $\mathcal{L}(\Theta)$. Note q_1 is reconstructed from q_0 as stated above.

IV. Multi-View iEFH

In this section and the next section, we present two important extensions of iEFH to deal with multi-view data analysis and learning predictive latent features by considering supervising side information.

[3]The free energy is finite because the IBP prior has exponentially decayed prior parameter π_k as k increases and the KL-divergence $\text{KL}(q_0\|p)$ is "zero forcing" [4, Chapter 10], i.e., when the prior distribution on feature k is small, the posterior distribution q_0 on feature k would be small too. As shown in [22], the expected number of active features is $N\alpha$ for IBP.

A. The Multi-view iEFH Model

Modern data analytic problems in social media, information technology and sciences often involve rich data consisting of multiple information modalities, which come from diverse sources or are extracted from different domains. For instance, web pages can be classified from their contents or link anchor texts [8], a video shot can be categorized from either color/shape of the keyframe or the corresponding closed captions [41], and many others [19][11][37][16], to name a few. These different modalities offer different angles to reveal the fundamental characteristics and properties of the study subjects, and is often referred to as different *views* of the subjects. Proper integration of multiple views presented 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, such as [8][41], and this list continues to grow, under various contexts and addressing diverse range of data forms [19][11][37][16].

To illustrate the basic idea, we consider two views of input features, denoted by $\mathbf{X} \triangleq \{X_i\}_{i=1}^N$ and $\mathbf{G} \triangleq \{G_j\}_{j=1}^M$, respectively, both of which are connected to the latent variables $\mathbf{H}$, which results in a similar chain graph model as in Fig. 1(b), using the stick-breaking representation of IBP prior. For simplicity, we consider a special case where $\mathbf{x}_d$ and $\mathbf{g}_d$ represent real-valued features and binary features, respectively. According to the Markov property of chain graphs, we have the joint distribution

$$p(\mathbf{v}, \{\mathbf{x}_d, \mathbf{g}_d, \mathbf{h}_d, \mathbf{z}_d\}) = p(\mathbf{v}) \prod_d p(\mathbf{z}_d|\mathbf{v}) p(\mathbf{x}_d, \mathbf{g}_d, \mathbf{h}_d|\mathbf{z}_d), \quad (7)$$

where

$$p(\mathbf{x}_d, \mathbf{g}_d, \mathbf{h}_d|\mathbf{z}_d) \propto \exp\Big\{E + \boldsymbol{\lambda}^\top \mathbf{g}_d + \mathbf{g}_d^\top \mathbf{U} \widetilde{\mathbf{h}}_d\Big\}, \quad (8)$$

where $E = \alpha^\top \mathbf{x}_d - \frac{\mathbf{x}_d^\top \mathbf{x}_d}{2\sigma_{d1}^2} + \beta^\top \widetilde{\mathbf{h}}_d - \frac{\mathbf{h}_d^\top \mathbf{h}_d}{2\sigma_{d2}^2} + \mathbf{x}_d^\top \mathbf{W} \widetilde{\mathbf{h}}_d$ is the potential function of the single-view iEFH and $\widetilde{\mathbf{h}}_d = \mathbf{z}_d \circ \mathbf{h}_d$, as defined before.

B. Optimization for multi-view iEFH

In order to deal with the intractable likelihood function, we resort to the similar approximate contrastive divergence method to obtain an approximation objective $\mathcal{L}'(\Theta, q_0, q_1)$ to the data likelihood, where q_0 is the variational distribution with input variables $\mathbf{x}$ and $\mathbf{g}$ clamped to the observations and q_1 is defined with all the variables free. We employ the structured mean field assumption on q (q_0 or q_1), that is, $q(\mathbf{v}, \{\mathbf{x}_d, \mathbf{g}_d, \mathbf{h}_d, \mathbf{z}_d\}) = \prod_k q(v_k) \prod_d q(\mathbf{x}_d) q(\mathbf{g}_d) q(\mathbf{h}_d) q(\mathbf{z}_d)$. Then we can compute each of the factored distribution separately. Specifically, for the variables $\mathbf{x}$ and $\mathbf{v}$, we have the same mean-field equations as in the single-view iEFH. For $\mathbf{g}$, $\mathbf{h}$ and $\mathbf{z}$, we can derive

$$\begin{aligned} q(g_{dm}) &= \text{Bernoulli}(g_{dm}; \zeta_{dm}) \\ q(h_{dk}) &= \mathcal{N}(h_{dk}; \mathbb{E}[h_{dk}], \sigma_{d2}^2) \\ q(z_{dk}) &= \text{Bernoulli}(z_{dk}; \nu'_{dk}) \end{aligned}$$

where

$$\begin{aligned} \mathbb{E}_q[h_{dk}] &= \nu'_{dk}\big(\beta_k + \mathbb{E}_q[\mathbf{x}_d]^\top \mathbf{W}_{\cdot k} + \mathbb{E}_q[\mathbf{g}_d]^\top \mathbf{U}_{\cdot k}\big) \\ \zeta_{dm} &= \text{Sigmoid}\big(\lambda_m + \mathbf{U}_{m\cdot} \mathbb{E}_q[\mathbf{h}_d] \circ \mathbb{E}_q[\mathbf{z}_d]\big) \\ \nu'_{dk} &= \text{Sigmoid}\big(E'_{dk} - \mathbb{E}_q[h_{dk}] \mathbb{E}_q[\mathbf{g}_d]^\top \mathbf{U}_{\cdot k}\big) \end{aligned}$$

and $E'_{dk} \triangleq \tau_2^k - \tau_1^k + \mathbb{E}_q[h_{dk}](\beta_k + \mathbb{E}_q[\mathbf{x}_d]^\top \mathbf{W}_{\cdot k}) = \log \frac{\nu_{dk}}{1-\nu_{dk}}$, where τ_1, τ_2, ν_{dk} are the same as in Eq. (6).

After we have computed the approximate objective function $\mathcal{L}'(\Theta, q_0, q_1)$ and inferred the variational distributions q_0 and q_1, we can update the model parameters Θ by using approximate gradient descent, where the gradients for $\boldsymbol{\lambda}$ and $\mathbf{U}$ are

$$\nabla_{\lambda_m} \mathcal{L}' = \sum_d \Delta\mathbb{E}[g_{dm}], \;\; \nabla_{\mathbf{U}_{mk}} \mathcal{L}' = \sum_d \Delta\mathbb{E}[g_{dm} z_{dk} h_{dk}].$$

For other parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W}, \{\sigma_{d1}, \sigma_{d2}\})$, the gradients have the same forms as single-view iEFH.

V. Infinite maximum margin Harmoniums

The single-view and multi-view iEFH are unsupervised latent feature models. In some cases, we would like to infer latent representations that are more specialized for a particular task. For instance, if we use the latent representations for text/image classification, we would like them to be as discriminative among class categories as possible. One common approach to implementing classification using iEFH is a two-step procedure: 1) first learn a latent representation and 2) then do classification using a classifier such as Support Vector Machines (SVM) based on those latent representations [41]. However, the unsupervised iEFH does not distinguish various categories in learning and thus its inferred latent feature representations might not be optimal for classification. Various approaches can be developed to considering supervising side information. For example, a likelihood-based approach [42] can be built by defining a joint distribution on input features $\mathbf{X}$ and response variables Y. But, as shown in [10], such a likelihood based method could lead to unsatisfying results in terms of prediction performance and discovering semantically meaningful latent feature representations.

Following the suggestions in [10], this section introduces infinite max-margin Harmonium (iMMH), another extension of iEFH that could consider the widely available supervising information to discover predictive latent feature representations. In fact, the iMMH model to be presented is among the recent attempts towards uniting Bayesian nonparametrics and max-margin learning, which have been largely treated as isolated subfields in machine learning. Our key insights to develop iMMH come from the recent work of *regularized Bayesian inference* (RegBayes) [44][45], which has shown that Bayesian nonparametric methods and max-margin learning can be naturally integrated into one framework. iMMH contributes by expanding the scope of RegBayes to undirected latent variable models.

A. Regularized Bayesian Inference for undirected Latent Variable Models

1) Bayesian Inference as an Optimization Problem: We base our work upon the interpretation of Bayesian inference as

an optimization problem and its recent extension of regularized Bayesian inference [44][46]. Specifically, let $\mathbb{M}$ be a model space, containing any variables whose posterior distributions we are trying to infer. Bayesian inference starts with a prior distribution $\pi(\mathcal{M})$ and a likelihood function $p(\mathbf{x}|\mathcal{M})$ indexed by the model $\mathcal{M} \in \mathbb{M}$. We can show that the posterior distribution due to the Bayes' theorem is the solution of the problem

$$\min_{p(\mathcal{M})} \mathrm{KL}(p(\mathcal{M})\|\pi(\mathcal{M})) - \mathbb{E}_{p(\mathcal{M})}[\log p(\mathcal{D}|\mathcal{M})] \quad (9)$$
$$\text{s.t.} : p(\mathcal{M}) \in \mathcal{P}_{\text{prob}},$$

where $\mathrm{KL}(p(\mathcal{M})\|\pi(\mathcal{M}))$ is the Kullback-Leibler (KL) divergence, and $\mathcal{P}_{\text{prob}}$ is the space of valid probability distributions with an appropriate dimension. In order to apply this result to the undirected latent variable models, such as Harmonium models, we need two extensions.

The above formulation implicitly assumes that the model can be graphically drawn as a Bayesian network as illustrated in Fig. 2(a)[4]. Here, we extend the basic results to include undirected LVMs (e.g., Harmoniums), as well as the case where a model has unknown parameters Θ and needs an estimation procedure (e.g., MLE), besides posterior inference. The latter is also known as empirical Bayesian methods, which are frequently employed by practitioners.

Extension 1: Chain Graph: Compared to the directed latent variable models [44], special attention needs to be paid to the hybrid chain graph models (e.g., iEFH) because of their special Markov property. Fig. 2(b) illustrates the structure of a chain graph, where binary latent variable $\mathbf{Z}$ is connected to $\mathbf{H}$ and $\mathcal{D}$ via directed edges, $\mathbf{H}$ is connected to $\mathcal{D}$ via undirected edges. iEFH is a special case of chain graphs. In iEFH, since the hidden $\mathbf{H}$ and the observed $\mathbf{X}$ are forming a Markov network, we cannot write it in a Bayesian style by using priors and likelihood functions. The insights that we can generalize RegBayes [44] to undirected latent variable models come from the observation that the objective $\mathcal{L}_B(\Theta, p(\mathcal{M}|\Theta))$ of problem (11) is in fact an KL-divergence, namely, we can show that

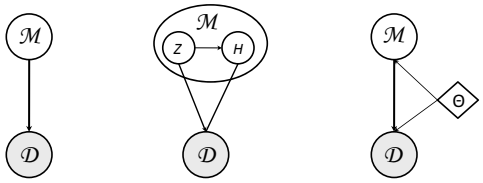
$$\mathcal{L}_B(\Theta, p(\mathcal{M}|\Theta)) = \mathrm{KL}(p(\mathcal{M}|\Theta)\|p(\mathcal{M}, \mathcal{D}|\Theta)), \quad (10)$$

where $p(\mathcal{M}, \mathcal{D})$ is the joint distribution. Then, we will have much freedom on defining $p(\mathcal{M}, \mathcal{D})$. For Bayesian networks [44], we naturally have $p(\mathcal{M}, \mathcal{D}) = \pi(\mathcal{M})p(\mathcal{D}|\mathcal{M})$. For the undirected iEFH model, we have $\mathcal{M} = \{\mathbf{v}, \mathbf{Z}, \mathbf{H}\}$ and again we can define the joint distribution as in Eq. (5).

Extension 2: Unknown Parameters: The hybrid iEFH model is not full Bayesian and we need some mechanisms to estimate the unknown parameters Θ, as illustrated in Fig. 2(c). We can perform the maximum likelihood estimation (MLE) as well as posterior inference jointly by solving $\mathcal{L}_B(\Theta, p(\mathcal{M}|\Theta))$

$$\min_{\Theta, p(\mathcal{M}|\Theta)} \mathrm{KL}(p(\mathcal{M}|\Theta)\|\pi(\mathcal{M})) - \mathbb{E}_{p(\mathcal{M}|\Theta)}[\log p(\mathcal{D}|\mathcal{M}, \Theta)]$$
$$\text{s.t.} : p(\mathcal{M}|\Theta) \in \mathcal{P}_{\text{prob}}(\Theta). \quad (11)$$

[4]The structure within $\mathcal{M}$ can be arbitrary, either a directed, undirected or hybrid chain graph.



(a) (b) (c)

Fig. 2. (a) Bayesian generative models; (b) chain graph models; and (c) empirical Bayesian models.

For problem (11), it is easy to show that the optimum solution of $p(\mathcal{M}|\Theta)$ is equivalent to the posterior distribution by Bayes' theorem for any Θ; and the optimum solution Θ^* is the MLE

$$\Theta^* = \operatorname*{argmax}_{\Theta} \log p(\mathcal{D}|\Theta).$$

2) Regularized Bayesian Inference for undirected Latent Variable Models: With the above discussions, the regularized Bayesian inference with estimating unknown model parameters can be generally formulated as

$$\min_{\Theta, p(\mathcal{M}|\Theta), \boldsymbol{\xi}} \mathcal{L}_B(\Theta, p(\mathcal{M}|\Theta)) + U(\boldsymbol{\xi}) \quad (12)$$
$$\text{s.t.} : p(\mathcal{M}|\Theta) \in \mathcal{P}_{\text{post}}(\Theta, \boldsymbol{\xi}),$$

where $\mathcal{P}_{\text{post}}(\Theta, \boldsymbol{\xi})$ is a subspace of distributions that satisfy a set of constraints, and $\mathcal{L}_B(\Theta, p(\mathcal{M}|\Theta))$ is the objective function of problem (11). The auxiliary parameters $\boldsymbol{\xi}$ are usually nonnegative and interpreted as slack variables. $U(\boldsymbol{\xi})$ is a convex function, which usually corresponds to a surrogate loss (e.g., hinge loss) of a prediction rule, as we shall see.

B. Infinite MMH for Classification

Now, we present the infinite MMH (iMMH) model for multi-class classification, where the response variable Y takes values from a finite set, e.g., $\mathcal{Y} = \{1, 2, \cdots, T\}$ without loss of generality. Binary classification and regression tasks could be easily developed following similar principles.

Specifically, to build a classifier, we use the effective latent features $\widetilde{\mathbf{h}}_d = \mathbf{z}_d \circ \mathbf{h}_d$ as the feature representation[5] of data d, for which we consider the case of having one type of features $\mathbf{x}_d$. Multiple types of features can be easily considered as in multi-view iEFH. When the latent variables $\mathbf{H}$ and $\mathbf{Z}$ are given, we define the linear discriminant function

$$F(y, \mathbf{h}, \mathbf{z}; \mathbf{x}, \Phi) \triangleq \Phi_y^\top \widetilde{\mathbf{h}} = \Phi^\top \mathbf{f}(y, \widetilde{\mathbf{h}}), \quad (13)$$

where $\mathbf{f}(y, \widetilde{\mathbf{h}})$ is a vector stacking T subvectors, of which the yth is $\widetilde{\mathbf{h}}^\top$ and all the others are 0; and Φ is the weight parameter that stacks T subvectors of Φ_y, each Φ_y corresponding to a class label y. Since both $\mathbf{h}$ and $\mathbf{z}$ are hidden random variables, we need to get rid of their uncertainty when performing prediction on input data, which are observed. We also treat Φ as random variables. Here, we follow the suggestions in [44] and define the *effective discriminant function* as an expectation

$$F(y; \mathbf{x}) \triangleq \mathbb{E}_{p(\mathbf{H}, \mathbf{Z}, \Phi)}[F(y, \mathbf{h}, \mathbf{z}; \mathbf{x}, \Phi)] \quad (14)$$
$$= \mathbb{E}_{p(\mathbf{H}, \mathbf{Z}, \Phi)}[\Phi^\top \mathbf{f}(y, \widetilde{\mathbf{h}})].$$

[5]We can consider observed features by concatenating them to $\widetilde{\mathbf{h}}_d$.

Then, the prediction rule for classification is naturally

$$y^* \triangleq \operatorname*{argmax}_{y \in \mathcal{Y}} F(y; \mathbf{x}). \quad (15)$$

Let $\mathcal{D} = \{(\mathbf{x}_d, y_d)\}_{d=1}^D$ be a set of labeled training data. With the above definitions, we can concertize the RegBayes framework and define iMMH as solving problem (12) with $U(\boldsymbol{\xi}) = C\sum_d \xi_d$ and

$$\mathcal{P}_{\text{post}}(\Theta, \boldsymbol{\xi}) = \left\{p(\{\mathbf{h}_d, \mathbf{z}_d\}, \Phi|\Theta)|\Delta F(y; \mathbf{x}_d) \geq \ell_d^\Delta(y) - \xi_d, \forall d, y\right\},$$

where $\Delta F(y; \mathbf{x}_d) = F(y_d; \mathbf{x}_d) - F(y; \mathbf{x}_d)$ is the expected margin favored by the true label y_d over an arbitrary label y, and $\ell_d^\Delta(y)$ is the nonnegative cost[6] of predicting the label to be y when the true label is y_d. It can be shown that minimizing $U(\boldsymbol{\xi})$ with the constrained subspace $\mathcal{P}_{\text{post}}$ is equivalent to minimizing the hinge loss

$$\mathcal{R}_h(p(\{\mathbf{h}_d, \mathbf{z}_d\}, \Phi|\Theta), \Theta) = C\sum_d \max_y \{\ell_d^\Delta(y) - \Delta F(y; \mathbf{x}_d)\}$$

of the expected prediction rule (15).

To complete the model, we define the joint distribution $p(\Phi, \mathbf{v}, \{\mathbf{x}_d, \mathbf{h}_d, \mathbf{z}_d\})$. For simplicity, we assume that Φ is independent from the other variables, that is $p(\Phi, \mathbf{v}, \{\mathbf{x}_d, \mathbf{h}_d, \mathbf{z}_d\}) = p_0(\Phi)p(\mathbf{v}, \{\mathbf{x}_d, \mathbf{h}_d, \mathbf{z}_d\})$, where $p(\mathbf{v}, \{\mathbf{x}_d, \mathbf{h}_d, \mathbf{z}_d\})$ is the same as in Eq. (5). One most common choice of $p_0(\Phi)$ is the zero-mean spherical normal distribution $\mathcal{N}(0, \sigma_0^2 I)$, although other choices [47] can be used.

C. Optimization with Contrastive Divergence

Again, due to the intractable normalization constant of the chain model, we cannot perform the inference exactly. Here, we present an approximation method, which can be effective in practice, as we shall see. Specifically, we first impose a mildly more strict assumption on the feasible distribution $p(\mathbf{v}, \{\mathbf{h}_d, \mathbf{z}_d\}, \Phi|\Theta)$, namely, we assume that $p(\mathbf{v}, \{\mathbf{h}_d, \mathbf{z}_d\}, \Phi|\Theta) = p(\Phi)p(\mathbf{v})\prod_d p(\mathbf{h}_d)p(\mathbf{z}_d)$. Then, the objective reduces to

$$\mathcal{L}_B(\Theta, p(\mathcal{M}|\Theta)) = \text{KL}(p(\Phi)\|p_0(\Phi)) + \text{KL}(p(\mathbf{v}, \{\mathbf{h}_d, \mathbf{z}_d\})\|p(\mathbf{v}, \{\mathbf{x}_d, \mathbf{h}_d, \mathbf{z}_d\})),$$

where minimizing the second term can be shown to be the wake-phase inference of iEFH. But it is still intractable to perform parameter estimation. Therefore, we need a second step to derive the approximation method. As in iEFH, we use contrastive divergence and approximate $\mathcal{L}_B$ as

$$\mathcal{L}_B(\Theta, q(\mathcal{M}|\Theta)) \approx \text{KL}(p(\Phi)\|p_0(\Phi)) + \mathcal{L}(\Theta, q_0, q_1), \quad (16)$$

where $\mathcal{L}(\Theta, q_0, q_1)$ is the contrastive divergence objective of iEFH, as defined before.

Then we develop an alternating procedure that iteratively infers (q_0, q_1) and estimates $(\Theta, p(\Phi))$ with an appropriate initialization. We omit the inference step, which is the same as in iEFH. The parameter estimation step involves minimizing the approximate objective subject to the posterior constraints. We perform this step with blockwise coordinate descent. For $p(\Phi)$, we solve the sub-problem

$$\min_{p(\Phi), \boldsymbol{\xi}} \ \text{KL}(p(\Phi)\|p_0(\Phi)) + C\sum_d \xi_d$$
$$\text{s.t.}: \ \mathbb{E}_{p(\Phi)}[\Phi]^\top \Delta\mathbf{f}_d(y, \mathbb{E}[\widetilde{\mathbf{h}}_d]) \geq \ell_d^\Delta(y) - \xi_d, \ \forall d, y$$

where $\Delta\mathbf{f}_d(y, \mathbb{E}[\widetilde{\mathbf{h}}_d]) = \mathbf{f}(y_d, \mathbb{E}[\widetilde{\mathbf{h}}_d]) - \mathbf{f}(y, \mathbb{E}[\widetilde{\mathbf{h}}_d])$. With a spherical Gaussian prior, the optimum solution is $p(\Phi) = \mathcal{N}(\boldsymbol{\mu}, \sigma_0^2 I)$, and we can solve for $\boldsymbol{\mu}$ using Lagrangian methods or solving a multi-class SVM primal problem [15]

$$\min_{\boldsymbol{\mu}, \boldsymbol{\xi}} \ \frac{1}{2\sigma_0^2}\|\boldsymbol{\mu}\|_2^2 + C\sum_d \xi_d$$
$$\text{s.t.}: \ \boldsymbol{\mu}^\top \Delta\mathbf{f}_d(y, \mathbb{E}[\widetilde{\mathbf{h}}_d]) \geq \ell_d^\Delta(y) - \xi_d, \ \forall d, y.$$

For Θ, using the equivalent unconstrained formulation with $\mathcal{R}_h$, we can do sub-gradient descent, where due to the properties of pointwise maximum function the sub-gradients are

$$\partial_{\beta_k} f = \nabla_{\beta_k}\mathcal{L} - C\sum_d (\boldsymbol{\mu}_{y_d} - \boldsymbol{\mu}_{\bar{y}_d})\kappa_{dk}$$
$$\partial_{\mathbf{W}_{nk}} f = \nabla_{\mathbf{W}_{nk}}\mathcal{L} - C\sum_d (\boldsymbol{\mu}_{y_d} - \boldsymbol{\mu}_{\bar{y}_d})\kappa_{dk}x_{dn}$$

where $\bar{y}_d = \text{argmax}_y(\ell_d^\Delta(y) + \boldsymbol{\mu}^\top \mathbf{f}(y, \mathbb{E}[\widetilde{\mathbf{h}}_d]))$ is the *loss-augmented prediction*; and $\kappa_{dk} = \nu_{dk}(1 - \nu_{dk})$ if $\mathbf{H}$ are held as constant, i.e., in binary iMMH; otherwise, $\kappa_{dk} = \nu_{dk}(1 - \nu_{dk})\mathbb{E}[h_{dk}] + \nu_{dk}^2$.

Note that there is an additional term (the second term) in the sub-gradients corresponding to $\boldsymbol{\beta}$ and $\mathbf{W}$, which will bias the model towards exploring a more discriminative latent representation when the estimated label $\bar{y}_d$ is different from the true label y_d. As we shall see, iMMH tends to learn not only sparse but also predictive latent representations from the data, which is more suitable for prediction tasks. We have adopted the effective method [15] to formulate our multi-class classifier. Alternative formulations exist, e.g., one-versus-one or one-versus-all with multiple binary classifiers. But an exhaustive comparison is beyond the scope of this paper.

VI. Experiments

We present qualitative and quantitative results on several text and image data sets for discovering semantically meaningful latent features and improving classification performance.

A. Results on image classification

1) Data Sets and Experiment Settings: The first set of experiments are done on two real image data sets. One is the TRECVID 2003 data set [42] containing 1078 video keyframes that belong to 5 categories. The data has two types of features, including 165-dim real-valued color histogram and 1894-dim discrete text features. We evenly partition the data for training and testing, following the same setting as [41] for result comparison. The other one is the Flickr image data set [10] containing 3411 images that belong to 13 classes of animals. This data set also has two types of features, including 634-dim

[6]It is normally assumed that $\ell_d^\Delta(y_d) = 0$, i.e., no cost for a correct prediction. Moreover, the constraint that $\xi_d \geq 0$ is not necessary for nonnegative cost, since $\xi_d \geq \ell_d^\Delta(y_d) = 0$.

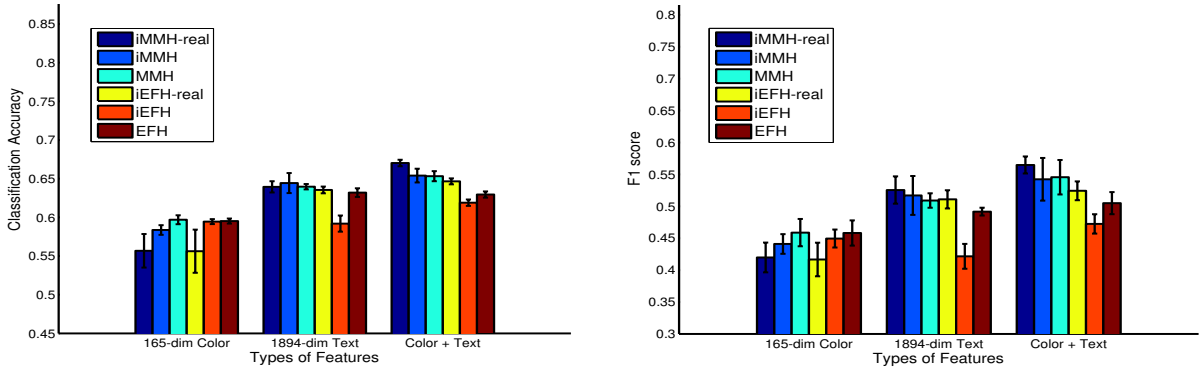


Fig. 3. Classification accuracy; and F1 score for different models on Trecvid 2003.

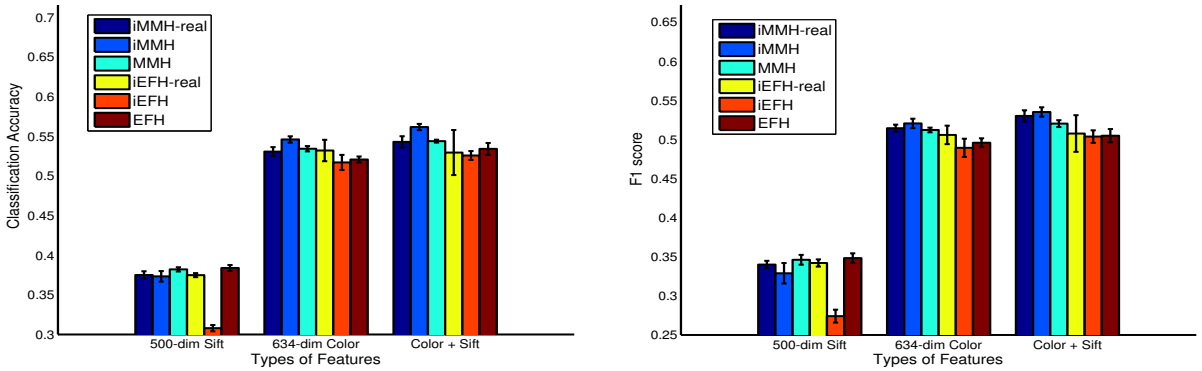


Fig. 4. Classification accuracy; and F1 score for different models on Flickr data set.

real color features and 500-dim discrete SIFT features. We use 2054 for training and 1357 for testing.

In all the experiments, we use two different configurations of iEFH and iMMH. The first one is the models that use binary latent features (i.e., fixing the variables $\mathbf{H}$ at unit value 1), denoted by iEFH and iMMH respectively. The second type is the models that use real-valued features (i.e., automatically inferring the values of $\mathbf{H}$), denoted by iEFH-real and iMMH-real, respectively. Furthermore, we report the results when using a single view of features, as well as the results when using two views of features on both data sets. As discussed in Section IV, both iEFH and iMMH can be extended to consider multi-view features, where multiple types (each type corresponds to one view) of input features share the same latent feature representations. For real-valued features, we use Gaussian units; and for the discrete text/SIFT features, we use binary units for simplicity. We compare with the finite EFH and max-margin EFH (MMH) [10], which were shown to outperform many other competitors[7] by selecting a good number of hidden units. We learn the multi-class SVM classifier (i.e., Φ) using the package[8] in the supervised iMMH and MMH. For iEFH and EFH, a downstream multi-class SVM classifier is trained based on the discovered latent features. C is selected via cross-validation during training. Both iEFH and iMMH are robust to α, as we shall see.

2) Prediction performance: Fig. 3 and Fig. 4 show the classification accuracy and F1 scores on the TRECVID and Flickr data sets. We can observe that: **1**) In both single-view and multi-view settings, iEFH performs comparable with (or even better than) the finite EFH, and the supervised iMMH model performs comparable with (or even better than) the finite MMH. The results demonstrate the effectiveness of Bayesian nonparametrics to bypass model selection; **2**) The supervised iMMH and MMH generally give better prediction results than the unsupervised models (i.e., iEFH and EFH), especially when using multi-view features; **3**) Using multi-view features could generally improve the performance. For example, on the TRECVID data set, the multi-view iMMH based on both color and text features outperforms the single-view iMMH using only text features. This is because the visual features tend to offer complementary information to the text features and provide incremental effect towards yielding better prediction performance than using only the text features; and **4**) Comparing the prediction performance of using real-valued hidden units (i.e., iMMH-real and iEFH-real) and binary hidden units (i.e., iMMH and iEFH), we can see that no one is dominating the other. On the TRECVID data, using real-valued units tends to improve the performance, while on the Flickr data, using binary latent units produces better results. But from the computational point of view, using binary latent units can save some machine time. Moreover, using binary units results in non-negative latent features, while real units can lead to positive or negative features.

[7]The linear multi-class SVM [15] on the raw multiview input features achieves about 0.5 in accuracy for Flickr and about 0.6 for TRECVID.

[8]http://svmlight.joachims.org/svm multiclass.html

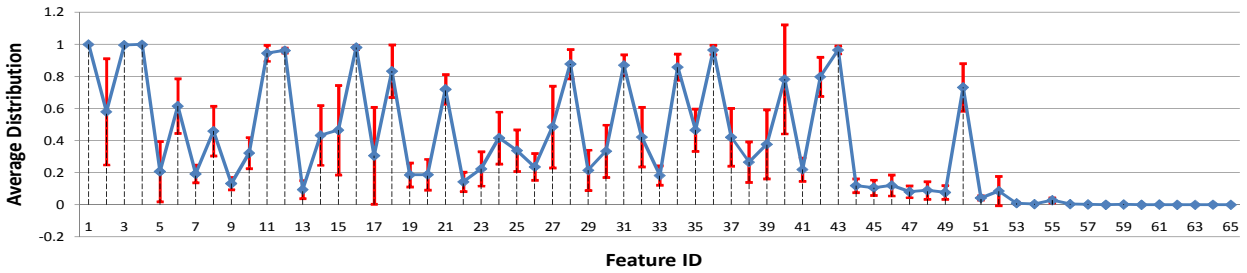


Fig. 5. Average posterior expectation of each latent feature by iMMH on the TRECVID data.

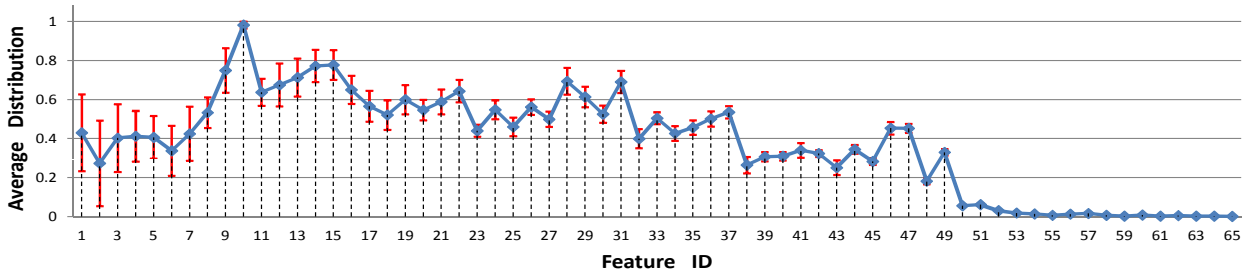


Fig. 6. Average posterior expectation of each latent feature by iMMH on the Flickr data.

3) Discriminative Latent Features: To give a holistic view of the discriminative power of latent features by iMMH, Fig. 5 and Fig. 6 show the average posterior expectations of latent feature variables[9] on TRECVID and Flickr data sets, where we only show the first 65 features; all the other features have almost zero probabilities to be active. We can see that the IBP prior is very effective in inferring a sparse subset of features. The standard deviation among the per-class average posterior expectations generally represents the discriminative power of that particular feature among all the images (i.e., features with larger std values tend to be good at distinguishing images from different classes).

To further carefully examine the discriminative power of each latent features, Fig. 7 shows 5 example features discovered by iMMH on the Flickr data set. For each feature, we show the top-8 images which produce a high activation value at that feature unit and also the average probabilities that the feature represents each of the 13 categories. We can see that the automatically learned features could have some semantic meanings. For instance, feature F1 is about "whales" and feature F3 is about "snake". In general, the features have good discriminative ability. For example, F1 is very good at discriminating "whales" from the other animals, and F4 is good at distinguishing "hawk" from "tiger", "zebra", "whales", and etc. However, we also have some features that do not have a strong discriminative power. For example, even though "lion" is ranked as the most likely animal that is described by F5, the difference from other animals like "wolf", "tiger" and "zebra" is very small.

4) Time Efficiency: Fig. 8 shows the training time of MMH and iMMH on TRECVID and Flickr data set. We can observe that the total time of running a single iMMH (using one set of α and C values) using both text and color features on the TRECVID data set is about $1,400$ seconds. In comparison, the single run of MMH (using one set K and C values)

[9]For feature k, the average is $1/D \sum_{d=1}^{D} \mathbb{E}[z_{dk} h_{dk}]$.

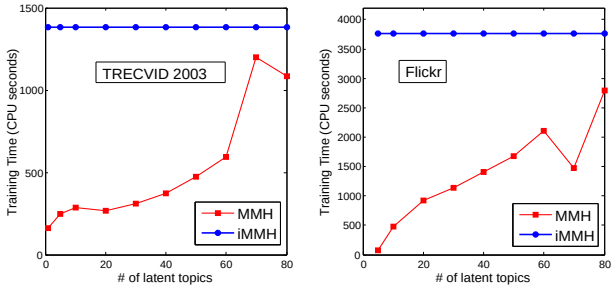


Fig. 8. Training time on (L) Trecvid 2003; and (R) Flickr data set.

needs less than 500 seconds when $K < 60$, about $1,000$ seconds when $K = 70$ and $K = 80$. Since iMMH is robust with respect to α and does not need to select K, its total running time is competitive. We have similar observations on the Flickr data set. For testing, all these models are efficient due to the conditional independence of the latent variables given observations.

We also show the proportion of the training time spent on solving the SVM subproblems for both iMMH and MMH in Table I. We can see that in iMMH, solving SVM subproblems are very efficient, which learning the other parameters (i.e., Θ) is much more expensive because of the variational inference algorithms. For MMH, solving SVM subproblems is also more efficient than learning the other parameters. Therefore, the key step to further improve the time efficiency is to improve the posterior inference algorithms.

5) Sensitivity to Hyper-parameters: Fig. 9 and Fig. 10 show the sensitivity of the infinite Harmonium models, including iMMH and iEFH using different input features, with respect to the hyper-parameters C, α and ℓ. Rather than analyzing the huge product space of these three parameters, we restrict ourselves to analyze each of them with all others approximately optimized via a grid search. We can observe that the changes of α and ℓ do not affect the prediction performance on both

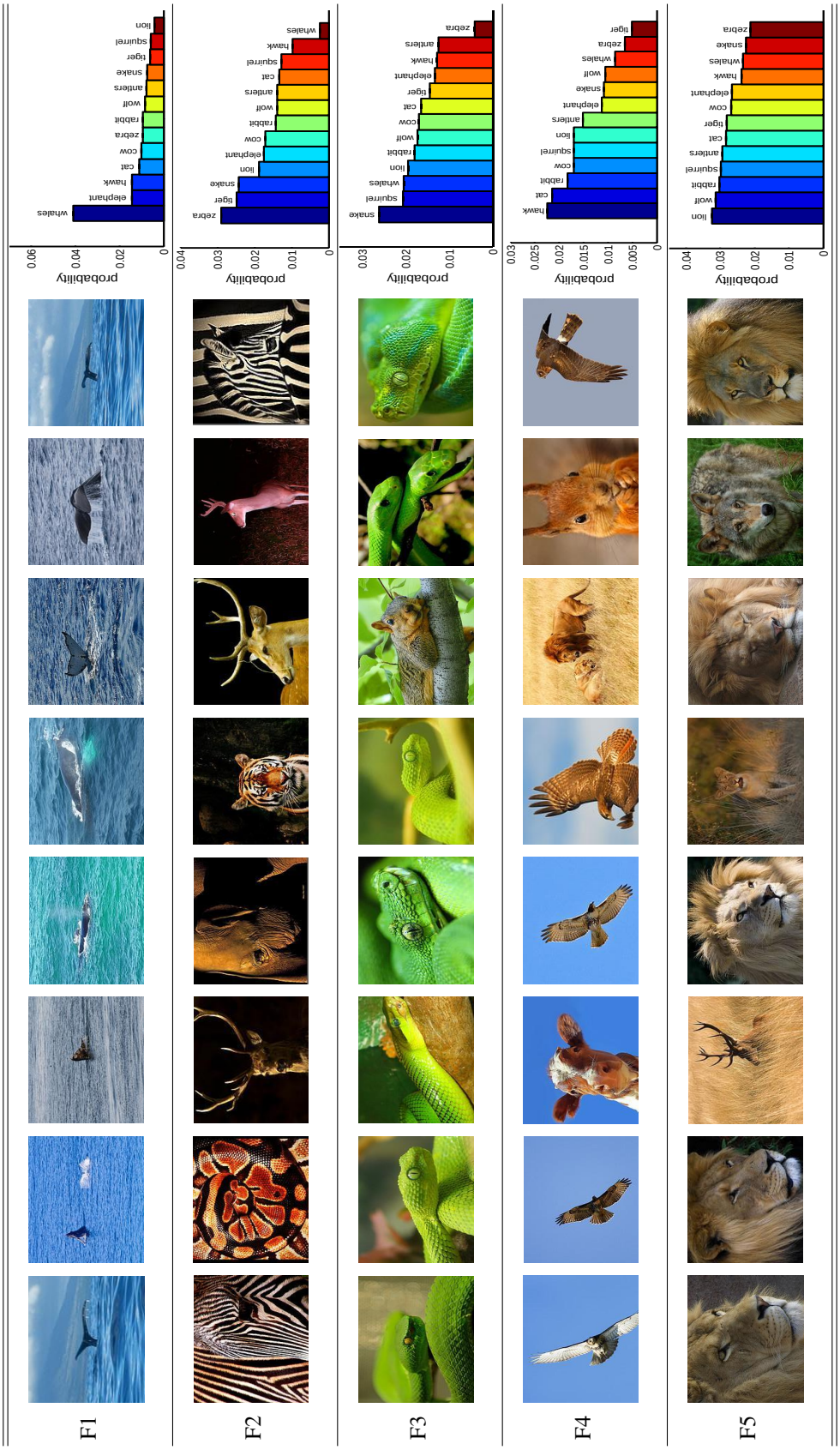
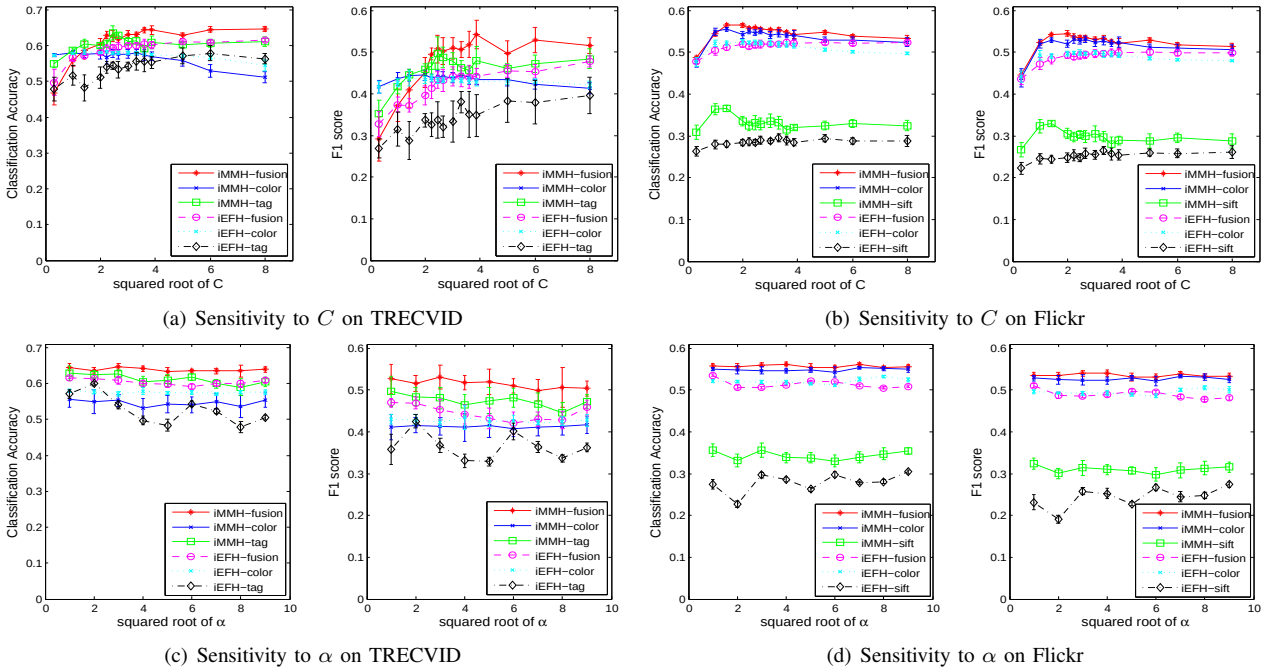


Fig. 7. Example features discovered by iMMH on the Flickr animal data set. For each feature, we show 8 topic-ranked images as well as the average probabilities of that topic on representing images from the 13 categories.



(a) Sensitivity to C on TRECVID (b) Sensitivity to C on Flickr

(c) Sensitivity to α on TRECVID (d) Sensitivity to α on Flickr

Fig. 9. Sensitivity analysis to C and α on the TRECVID and Flickr data sets.

TABLE I
SPLIT OF TRAINING TIME FOR IMMH AND MMH.

	iMMH		MMH	
TRECVID	1386.63 (seconds)		1099.52 (seconds)	
	SVM	Others	SVM	Others
	5.86 (0.4%)	1380.26 (99.6%)	18.54 (1.8%)	1079.29 (98.2%)
Flickr	3745.23 (seconds)		2516.09 (seconds)	
	SVM	Others	SVM	Others
	363.36 (9.7%)	3381.37 (90.3%)	711.10 (28.3%)	1804.99 (71.7%)

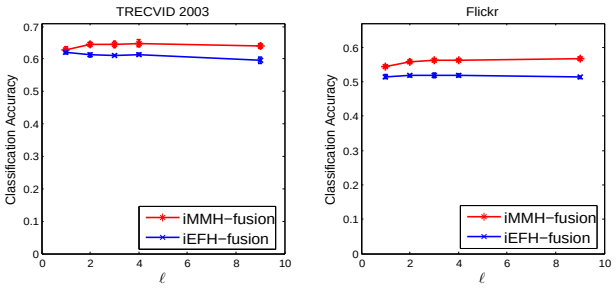


Fig. 10. Sensitivity to ℓ on (L) TRECVID 2003; and (R) Flickr data sets.

data sets, especially when using multi-view input features. The performance oscillates when C is less than 20 but turns smooth afterwards. Generally, the highest performance with respect to each setting (e.g, iEFH-fusion that uses both color and text features) are comparable with the results in Fig. 3 and Fig. 4, which are achieved via cross-validation. Finally, note that for the finite models EFH and MMH, the sensitivity to K has been extensively demonstrated in [10]; and for the regularization parameter C, its sensitivity is similar to that of iEFH and iMMH. We provide further results of MMH in Appendix.

B. Results on document classification

Now we further examine the model performance on document classification problems.

1) Data Sets and Experiment Settings: The document classification experiments are done on the standard 20 Newsgroups data set, which consists of about 20K documents in 20 categories. To compare with discriminative RBM (DRBM) [28], we use the same vocabulary with 5000 most frequent words and the same train/test split. For all models in this task, we treat $\mathbf{X}$ as binary and infer binary latent features. We set the truncation level $T = 200$ and compare with SVM, NNet (neural networks), RBM, RBM+NNet, and DRBM, whose results were reported in [28].

TABLE II
CLASSIFICATION ERRORS ON THE 20 NEWSGROUPS DATA SET.

Model	Classification Error	Train Time (seconds)
SVM [28]	0.328 ± 0.0000	-
NNet [28]	0.282 ± 0.0000	-
RBM (K=1000) [28]	**0.249** ± 0.0000	-
DRBM (K=50) [28]	0.276 ± 0.0000	-
RBM+NNet [28]	0.268 ± 0.0000	-
EFH+SVM (K=50)	0.375 ± 0.0000	3875.6
EFH+SVM (K=200)	0.305 ± 0.0000	20406.0
EFH+SVM (K=700)	0.273 ± 0.0000	42066.4
MMH (K=50)	0.264 ± 0.0000	8560.4
MMH (K=200)	**0.250** ± 0.0000	34523.6
MMH (K=700)	0.255 ± 0.0000	68305.6
iEFH+SVM	0.283 ± 0.0022	19825.4
iMMH	**0.252** ± 0.0029	86511.5

2) Prediction performance: Table II shows the document classification errors and training time of different models. We can observe that: **1**) Using the same number of hidden units, the max-margin MMH outperforms the maximum likelihood estimation based DRBM; **2**) Even with only a few (e.g., 200)

hidden units, the discriminatively trained MMH can perform very competitively, in fact comparable to the generatively trained RBM which uses much more (e.g., 1000) units; **3**) Using Bayesian nonparametric techniques, iEFH and iMMH can avoid the model selection problem without sacrificing classification performance. For iEFH, it even outperforms the finite EFH if the number of hidden units in EFH is not set appropriately; and **4**) Using supervising information, the discovered latent representations by MMH or iMMH are generally more discriminative than those discovered by an unsupervised method (e.g., EFH or iEFH). Finally, the time efficiency of iMMH and iEFH is still competitive since they don't need to choose K.

3) Example features: To illustrate the semantics of the learned latent features, Table III shows some examples discovered by iMMH. For each feature, we present the most frequent words appearing in the top-15 documents that have high activation values at that particular feature dimension. We can see that in general the learned features could have some semantic meanings. For instance, features F3 and F37 (with the words "god", "jesus", "christian", etc) are strongly associated with the category "religion.christian", features F40 and F51 (with words "orbit", "moon", "space") are very likely to represent the category "sci.space", and features F7, F11 and F21 (with the words "bus", "controller", "drive", "windows") tend to represent the category "os. ms-windows. misc". For all other features, we have associated them with the closest class categories according to their semantics.

TABLE III
EXAMPLE FEATURES AND THEIR ASSOCIATED NEWSGROUPS BY IMMH ON THE 20 NEWSGROUPS DATA SET.

Categories	Latent features by iMMH
religion. christian	F3: god, people, jesus, life, christian, christ, christians, hell F73: sandvik, christians, god, people, law, christian
alt. atheism	F83: god, atheism, quadra, mac, problem, strong, belief, atheists F92: god, atheism, exist, belief, atheists, people, strong, islam
comp. graphics	F35: work, graphics, windows, information, cylinder F97: graphics, windows, linux, image, file, gif, find, program, ftp
os. ms-windows. misc	F7: bus, controller, ide, drive, windows, card, mouse, driver F11: key, keys, win, system, time, team, windows, files F21: drive, hard, drives, disk, pc, cd, columbia, card
talk. politics. guns	F84: pitt, gordon, banks, geb, surrender, intellect, skepticism F86: people, control, firearms, law, gun, weapons, public, guns F94: people, militia, arms, amendment, state, government, gun
sports. hockey	F79: ca, team, hockey, game, play, year, montreal, cup, playoffs F85: team, gm, murray, win, good, hockey
sports. baseball	F55: year, runs, team, pitching, games, game, baseball, season F81: team, apr, game, games, baseball, series, people, win
misc. forsale	F53: sale, computer, price, drive, things, forsale, pc, misc F71: mark, sale, optilink, file, case, email, price, shipping
sci. space	F40: orbit, moon, sun, lunar, thing, years, made, space, earth F51: nasa, henry, orbit, comet, baalke, moon, kelvin, space, earth
sci. crypt	F11: key, keys, win, system, time, team, windows, files F24: key, encryption, chip, make, don, public, keys, clipper, secure
rec. autos	F34: car, cars, bike, apr, make, back, ve, speed, cso, ride, cso F47: car, email, find, windows, craig, graphics, great, problem, dealer
rec. motorcycles	F89: bike, dod, dog, ride, turn, bikes, riders, left F90: bike, dog, dod, apr, riders, riding, ride, bmw, mot, rider

VII. CONCLUSIONS AND DISCUSSIONS

This paper presents an infinite exponential family Harmonium model (iEFH) that has an infinite number of latent units a *priori* and automatically infers an appropriate subset from empirical data by leveraging the recent advances in Bayesian nonparametric techniques. We further extend iEFH to multi-view iEFH for dealing with data from multiple sources or domains, and to the supervised infinite max-margin Harmonium (iMMH) which directly regularizes the properties of latent features for improving their discriminative abilities by utilizing the max-margin principle. Experiments show compelling results compared with extensive state-of-the-art models, which verify the effectiveness of the proposed methods.

We present a preliminary attempt towards expanding the scope of Bayesian nonparametrics to solve the challenging problem of learning nonparametric undirected latent variable models, and a lot of room still remains to further improve. For future work, we are interested in developing more accurate and efficient inference algorithms, such as using multi-core or multi-machine architectures to do parallel inference. Furthermore, we are also interested in further broadening the use of Bayesian nonparametrics to infer the structures of more sophisticated undirected models, such as deep Boltzmann machines [34] and conditional random fields (CRFs) [9]. Also, using EFH or RBM in particular to deal with relational network data [32] or text documents introduces additional challenges, such as high dimensional inputs. Sophisticated representation and inference methods [17][33] are needed to make the model scalable to large scale data sets. We plan to systematically investigate such methods. Finally, extreme learning machines (ELM) [26] provide alternative solutions of building max-margin classifiers, and it is interesting to investigate how Bayesian nonparametrics can help them.

ACKNOWLEDGMENT

This work is supported by National Key Project for Basic Research of China (Grant Nos: 2013CB329403, 2012CB316301), National Natural Science Foundation of China (Nos: 61210013, 61273023), Tsinghua Self-innovation Project (Grant Nos: 20121088071, 20111081111), and China Postdoctoral Science Foundation Grant (Grant Nos: 2013T60117, 2012M520281).

REFERENCES

[1] R.P. Adams, H. Wallach, and Z. Ghahramani. Learning the structure of deep sparse graphical models. In *International Conference on Artificial Intelligence and Statistics*, 2010.

[2] C.E. Antoniak. Mixture of Dirichlet process with applications to Bayesian nonparametric problems. *Annals of Statistics*, (273):1152–1174, 1974.

[3] M.J. Beal, Z. Ghahramani, and C.E. Rasmussen. The infinite hidden Markov model. In *Advances in Neural Information Processing Systems*, 2002.

[4] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] D. Blei and J. Lafferty. Correlated topic models. In *Advances in Neural Information Processing Systems*, 2006.

[6] D. Blei and J.D. McAuliffe. Supervised topic models. In *Advances in Neural Information Processing Systems*, 2007.

[7] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, (3):993–1022, 2003.

[8] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-trainnig. In *Conference on Computational Learning Theory*, 1998.

[9] K. Bousmalis, S. Zafeiriou, L. Morency, and M. Pantic. Infinite hidden conditional random fields for human behavior analysis. *IEEE Trans. on Neural Networks and Learning Systems*, 24(1):170–177, 2013.

[10] N. Chen, J. Zhu, F. Sun, and E.P. Xing. Large-margin predictive latent subspace learning for multiview data analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(12):2365–2378, 2012.

[11] C.M. Christoudias, R. Urtasun, and T. Darrell. Multi-view learning in the presence of view disagreement. In *Conference on Uncertainty in Artificial Intelligence*, 2008.

[12] A. Coates, H. Lee, and A.Y. Ng. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, 2011.

[13] A. Courville, J. Bergstra, and Y. Bengio. A spike and slab restricted Boltzmann machine. In *International Conference on Artificial Intelligence and Statistics*, 2011.

[14] A. Courville, J. Bergstra, and Y. Bengio. Unsupervised models of images by spike-and-slab RBMs. In *International Conference on Machine Learning*, 2011.

[15] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, (2):265–292, 2001.

[16] M. Culp, G. Michailidis, and K. Johnson. On multi-view learning with additive models. *Annals of Applied Statistics*, 3(1):292–318, 2009.

[17] G.E. Dahl, R.P. Adams, and H. Larochelle. Training restricted Boltzmann machines on word observations. In *International Conference on Machine Learning*, 2012.

[18] F. Doshi-Velez, K.T. Miller, J. Van Gael, and Y.W. Teh. Variational inference for the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, 2009.

[19] V. Ferrari, T. Tuytelaars, and L. V. Gool. Integrating multiple model views for object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[20] Y. Freund and D. Haussler. Unsupervised learning of distributions of binary vectors using 2-layer networks. In *Advances in Neural Information Processing Systems*, 1992.

[21] M. Frydenberg. The chain graph Markov property. *Scandinavian Journal of Statistics*, 17:333–353, 1990.

[22] T.L. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems*, 2006.

[23] G. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.

[24] G. Hinton. A practical guide to training restricted Boltzmann machines. *UTML TR 2010-003, Version 1, University of Toronto*, 2010.

[25] T. Hofmann. Probabilistic latent semantic analysis. In *Conference on Uncertainty in Artificial Intelligence*, 1999.

[26] G. Huang, H. Zhou, X. Ding, and R. Zhang. Extreme learning machine for regression and multiclass classification. *IEEE Trans. on Systems, Man, and Cybernetics*, 42(2):513–529, 2012.

[27] S. Lacoste-Jullien, F. Sha, and M.I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Advances in Neural Information Processing Systems*, 2008.

[28] H. Larochelle and Y. Bengio. Classification using discriminative restricted Boltzmann machines. In *International Conference on Machine Learning*, 2008.

[29] D.C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, (45):503–528, 1989.

[30] X. Liu and Y. Shao. Asymptotics for likelihood ratio tests under loss of identifiability. *Annals of Statistics*, 31(3):807–832, 2003.

[31] I. Murray and Z. Ghahramani. Bayesian learning in undirected graphical models: Approximate MCMC algorithms. In *Conference on Uncertainty in Artificial Intelligence*, 2004.

[32] C. H. Nguyen and H. Mamitsuka. Latent feature kernels for link prediction on sparse graphs. *IEEE Trans. on Neural Networks and Learning Systems*, 23(11):1793–1804, 2012.

[33] R. Salakhutdinov and G. Hinton. Replicated softmax: an undirected topic model. In *Advances in Neural Information Processing Systems*, 2009.

[34] R. Salakhutdinov and H. Larochelle. Efficient learning of deep Boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, 2011.

[35] B. Shahbaba and R. Neal. Nonlinear models using Dirichlet process mixtures. *Journal of Machine Learning Research*, 10:1829–1850, 2009.

[36] Y.W. Teh, D. Gorur, and Z. Ghahramani. Stick-breaking construction of the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, 2007.

[37] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. Van Gool. Towards multi-view object class detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[38] M.E. Tipping and C.M. Bishop. Probabilistic principle component analysis. *Journal of the Royal Statistical Society: Series B*, 61(3):611–622, 1999.

[39] M. Welling and G. Hinton. A new learning algorithm for mean field Boltzmann machines. In *In International Conference on Artificial Neural Networks*, 2001.

[40] M. Welling, M. Rosen-Zvi, and G. Hinton. Exponential family Harmoniums with an application to information retrieval. In *Advances in Neural Information Processing Systems*, 2004.

[41] E.P. Xing, R. Yan, and A.G. Hauptmann. Mining associated text and images with dual-wing Harmoniums. In *Conference on Uncertainty in Artificial Intelligence*, 2005.

[42] J. Yang, Y. Liu, E.P. Xing, and A.G. Hauptmann. Siam conference on data mining. In *SIAM Conference on Data Mining*, 2007.

[43] J. Zhu, A. Ahmed, and E.P. Xing. MedLDA: Max margin supervised topic models for classification and regression. In *International Conference on Machine Learning*, 2009.

[44] J. Zhu, N. Chen, and E.P. Xing. Infinite latent SVM for classification and multi-task learning. In *Advances in Neural Information Processing Systems*, 2011.

[45] J. Zhu, N. Chen, and E.P. Xing. Infinite SVM: a Dirichlet process mixture of large margin kernel machines. In *International Conference on Machine Learning*, 2011.

[46] J. Zhu, N. Chen, and E.P. Xing. Bayesian inference with posterior regularization and infinite latent support vector machines. *arXiv: 1210.1766*, 2012.

[47] J. Zhu and E.P. Xing. Maximum entropy discrimination markov networks. *Journal of Machine Learning Research*, (10):2531–2569, 2009.

Ning Chen received her BS from China Northwestern Polytechnical University, and PhD degree in the Department of Computer Science and Technology at Tsinghua University, China, where she is currently a post-doc fellow. She was a visiting researcher in the Machine Learning Department of Carnegie Mellon University. Her research interests are primarily in machine learning, especially probabilistic graphical models, Bayesian Nonparametrics with applications on data mining and computer vision.

Jun Zhu received his BS, MS and PhD degrees all from the Department of Computer Science and Technology (CS&T) in 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 on latent variable models, max-margin learning and Bayesian nonparametrics. He is a member of the IEEE.

Fuchun Sun received his BS and MS degrees from China Naval Aeronautical Engineering Academy and 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.

Bo Zhang graduated from the Department of Automatic Control, Tsinghua University, Beijing, China, in 1958. Currently, he is a Professor in the Department of Computer Science and Technology, Tsinghua University and a Fellow of Chinese Academy of Sciences, Beijing, China. His main interests are artificial intelligence, pattern recognition, neural networks, and intelligent control. He has published over 150 papers and four monographs in these fields.