Multi-Modal Distance Metric Learning: A Bayesian Non-parametric Approach

Behnam Babagholami-Mohamadabadi, Seyed Mahdi Roostaiyan, Ali Zarghami, Mahdieh Soleymani Baghshah

Department of Computer Engineering, Sharif University of Technology,Tehran, Iran
babagholami@alum.sharif.edu

Abstract. In many real-world applications (e.g. social media application), data usually consists of diverse input modalities that originates from various heterogeneous sources. Learning a similarity measure for such data is of great importance for vast number of applications such as *classification*, *clustering*, *retrieval*, etc.
Defining an appropriate distance metric between data points with multiple modalities is a key challenge that has a great impact on the performance of many multimedia applications. Existing approaches for multi-modal distance metric learning only offer point estimation of the distance matrix and/or latent features, and can therefore be unreliable when the number of training examples is small. In this paper we present a novel Bayesian framework for learning distance functions on multi-modal data through Beta Process, by which we embed data of different modalities into a single latent space. Moreover, using the flexible Beta process model, we can infer the dimensionality of the hidden space using training data itself. We also develop a novel Variational Bayes (VB) algorithm to compute the posterior distribution of the parameters that imposes the constraints (similarity/dissimilarity constraints) directly on the posterior distribution. We apply our framework to text/image data and present empirical results on retrieval and classification to demonstrate the effectiveness of the proposed model.

Keywords: Metric learning, Multi-modal data, Beta process, Variational inference, Gibbs sampling

1 Introduction

Recently, multi-modal data has been grown explosively thanks to the ubiquity of the social media (e.g. Facebook, Flicker, Youtube, iTuens, etc). In such data, information comes through multiple input channels (images contain tags and captions, videos are associated with audio signals and/or user comments). Hence, each modality can be characterized by different statistical features which reveals the importance of the fact that the modality corresponding to a distinct input source, carries different kinds of information.
In many applications including classification [6], retrieval [7], clustering [8], and recommendation systems [9], choosing a proper similarity measure between items

is a crucial task. To address this problem, a wide range of Distance Metric Learning (DML) methods have been proposed [1],[2],[3],[4],[5]. Although the performance of DML methods have been promising on similarity search problems, most existing DML algorithms are designed to work on single-modal data, hence, they are limited in that they do not effectively handle the distance measure of multimodal data which may originate from totally different resources.
Recently, the multi-modal distance metric learning has been received an increasing attention [11],[18]. In this paper, we propose a Bayesian framework for multi-modal distance metric learning based on the Beta process [19] that takes into account the distance supervision (similarity/dissimilarity constraints). Our method embeds data of arbitrary modalities into a single latent space with the ability to learn the dimensionality of the latent space by the data itself. Given supervisory information (labeled similar and dissimilar pairs), we develop a novel Variational Bayes (VB) algorithm which incorporates such information into the proposed Bayesian framework by imposing the constraints directly on the parameters of the posterior distribution of the latent features.
The rest of this paper is organized as follows. Section 2 introduces some related work in metric learning area. In Section 3, we briefly review the Beta process. We present the propose multi-modal distance metric learning framework based on Beta process model in Section 4. In Section 5, we introduce a novel VB algorithm to compute the posterior distribution of the parameters and the hidden variables. Experimental results are presented in Section 6. Finally, we conclude our work in Section 7.

2 Related Work

Metric Learning has become a very active research area over the last years [1],[2],[3],[4],[5]. In this problem, we intend to learn an appropriate dissimilarity measure from the data samples when some similarity and dissimilarity constraints on data points are available. Xing et al. [10] introduced a metric learning method by formulating the learning task into the following constrained convex optimization problem.

$$\begin{aligned} A^* =& \underset{A}{argmin} \sum_{(x_i,x_j)\in S} (x_i - x_j)^T A(x_i - x_j), \\ & s.t. \sum_{(x_i,x_j)\in D} (x_i - x_j)^T A(x_i - x_j) \geq 1, \; A \succeq 0, \end{aligned} \tag{1}$$

where A is a Mahalanobis distance matrix (A must be positive semidefinite matrix to satisfy the non-negativity and triangle inequality conditions), and S and D denote the set of positive and negative constraints respectively.

Some other well-known algorithms in this area include Relevant Component Analysis (RCA) [12], Discriminative Component Analysis (DCA) [13], Information-Theoretic Metric Learning (ITML) [14], Large Margin Nearest Neighbor (LMNN) [15], Regularized Metric Learning [16], Laplacian Regularized Metric Learning

(LRML) [17] that learn a Mahalanobis distance metric.
The problem with all of the above methods is that they are primarily designed for data with single modality and are not appropriate for multi-modal data. One very simple approach to remedy this problem is to join the features from different modalities into a single representation and learn an appropriate metric using that representation. Unfortunately, this naive solution does not consider the incompatibility of heterogeneous information sources and subsequently, ignores the dependency relationships between various modalities that could lead to suboptimal performance.
Attempting to address this issue, some researchers have introduced some metric learning methods for multi-modal data. McFee and Lanckriet [11] proposed a multi-modal distance metric learning method based on Multiple Kernel Learning that learns each kernel for a different modality of the data. This algorithm learns for features of each modality a Mahalanobis distance metric in the reproducing kernel Hilbert space (RKHS), that can be solved by semidefinite programming. Very recently, Xie and Xing [18] have combined multi-wing harmonium model (MVH) [20] for multimodel integration and the metric learning method introduced in [10] for incorporating supervisory information into the proposed model. More precisely, this method tries to embed data of different modalities into a single latent space by imposing the similarity/dissimilarity constraints on the latent features. This is done by minimizing the distance of similar pairs while separating dissimilar pairs with a certain margin in the latent space. Although the results of this algorithm is promising, there are two problems with this method. First, this algorithm only provides point estimation of the latent features which could be sensitive to the choice of training examples, hence the algorithm tends to be unreliable when the number of training examples is small. Second, the dimensionality of the latent space must be specified a priori that could be a hard assumption.
To address the above problems, in this paper, we present a Bayesian framework for Multi-Modal Distance Metric Learning (MMDML) based on the Beta process that targets tasks where the number of training examples is limited. Indeed, using the full Bayesian treatment, the proposed framework is better suited to dealing with a small number of training examples than the non-Bayesian approaches. Moreover, using the flexible Beta process, we are able to infer the number of latent features from the observed data.

3 Beta Process

The Beta process $B \sim BP(c, B_0)$ is an example of completely random measures [22] which is defined as a distribution on positive random measures over a measurable space $(\Omega, \mathcal{F})$ [19]. It is parameterized by a base measure B_0 which is defined over Ω and a positive function $c(\omega)$ over Ω which is assumed constant for simplicity ($c(\omega) = c$). This process is an example of a Lévy process with the Lévy measure as

$$\nu(d\pi, d\omega) = c\pi^{-1}(1-\pi)^{c-1} d\pi B_0(d\omega). \tag{2}$$

For generating samples from $B \sim BP(c, B_0)$, first, a non-homogeneous Poisson process is defined on $\Omega \times \mathcal{R}^{+}$ with intensity function ν. Then, $Poisson(\lambda)$ number of points $(\pi_k, \omega_k) \in [0,1] \times \Omega$ are drawn from the Poisson process ($\lambda = \int_{[0,1]} \int_{\Omega} \nu(d\omega, d\pi) = \infty$). Finally, a draw from $B \sim BP(c, B_0)$ is constructed as

$$B_\omega = \sum_{k=1}^{\infty} \pi_k \delta_{\omega_k}, \tag{3}$$

where δ_{ω_k} is a unit point measure at ω_k (δ_{ω_k} equals one if $\omega = \omega_k$ and is zero otherwise) . It can be seen from equation 3, that B_ω is a discrete measure (with probability one), for which $B_\omega(A) = \sum_{k:\omega_k \in A} \pi_k$, for any set $A \subset \mathcal{F}$.

4 Proposed method

In this section, we present our Bayesian framework for multi-modal distance metric learning which directly imposes the constraints on the posterior distribution of the latent features.

4.1 Problem Formulation

Let $T = [\bar{X}, \bar{Y}]$ be the observed bi-modal data matrix (for simplicity, we assume we have two-modal data, but our method can be easily extended to multimodal data) where $\bar{X} = [X_1, X_2, ..., X_d]_{M \times d}$ denotes the first modality data matrix, where $X_i = [x_{1i}, ...x_{Mi}]^T$ is the i-th data point of the first modality, and $\bar{Y} = [Y_1, Y_2, ..., Y_d]_{N \times d}$ denotes the second modality data matrix, where $Y_i = [y_{1i}, ...y_{Ni}]^T$ is the i-th data point of the second modality. We also denote $\bar{H} = [H_1, H_2, ..., H_d]_{K \times d}$ as the latent feature matrix, where $H_i = [h_{1i}, ...h_{Ki}]^T$ is the latent feature for the i-th data point. We are also given two sets of pairwise constraints which are defined as

$$\begin{aligned} A &= \{(i,j) \mid (X_i, Y_i) \text{ and } (X_j, Y_j) \text{ are in the same class}\}, \\ D &= \{(i,j) \mid (X_i, Y_i) \text{ and } (X_j, Y_j) \text{ are in two different class}\}, \end{aligned}$$

where A is the set of similar pairwise constraints, and D is the set of dissimilar pairwise constraints. In order to utilize the Beta process in the proposed Bayesian framework, we model the latent feature matrix ($\bar{H}$) as an element-wise multiplication of a binary matrix ($\bar{Z} = [Z_1, ..., Z_d]_{K \times d}$) and a real weight matrix ($\bar{S} = [S_1, ..., S_d]_{K \times d}$). Hence, we have $\bar{H} = \bar{Z} \odot \bar{S}$, where $\odot$ is the element-wise multiplication operator. To be fully Bayesian, we must define appropriate prior and likelihood distributions for all observed $(\bar{X}, \bar{Y})$ and latent $(\bar{Z}, \bar{S})$ variables. Based on the above definitions, the proposed generative model goes as follows:

- For each data point $(X_i, Y_i)_{i=1}^{d}$, first draw corresponding features Z_i, S_i from the prior distributions $p(Z_i|\beta_z)$ and $p(S_i|\beta_s)$ respectively.
- For each drawn feature $(Z_i, S_i)_{i=1}^{d}$, draw the data point X_i and Y_i from the likelihoods $p(X_i|Z_i, S_i, \beta_x)$ and $p(Y_i|Z_i, S_i, \beta_y)$ respectively.

where $\beta_z, \beta_s, \beta_x$ and β_y are the free parameters of the proposed generative model (to be fully Bayesian, we put appropriate prior distribution on these parameters and infer the corresponding posterior distributions from observed data). It should be noted that using the above generative process, we assume that elements of each modality are independent of other modalities given the latent features. More precisely, we have:

$$p(\bar{X}, \bar{Y}|\bar{Z}, \bar{S}, \beta_x, \beta_y) = \prod_{i=1}^{d} \prod_{m=1}^{M} p(x_{mi}|Z_i, S_i, \beta_x) \prod_{n=1}^{N} p(y_{ni}|Z_i, S_i, \beta_y). \quad (4)$$

We also make assumptions about the complete conditionals in the proposed model (a complete conditional is the conditional distribution of a latent variable given the other latent variables and the observations). We assume that these distributions are in the exponential family,

$$\begin{aligned} &p(Z_i|\bar{Z}_{-i}, \bar{S}, \bar{X}, \bar{Y}, \beta_z, \beta_x, \beta_y) \propto p(X_i|Z_i, S_i, \beta_x)p(Y_i|Z_i, S_i, \beta_y)p(Z_i|\beta_z) \\ &\quad \propto exp\big\{\eta_z(\bar{Z}_{-i}, S_i, X_i, Y_i, \beta_z, \beta_x, \beta_y)^T t_z(Z_i)\big\},\ i = 1, ..., d, \end{aligned} \quad (5)$$

$$\begin{aligned} &p(S_i|\bar{S}_{-i}, \bar{Z}, \bar{X}, \bar{Y}, \beta_s, \beta_x, \beta_y) \propto p(X_i|Z_i, S_i, \beta_x)p(Y_i|Z_i, S_i, \beta_y)p(S_i|\beta_s) \\ &\quad \propto exp\big\{\eta_s(\bar{S}_{-i}, Z_i, X_i, Y_i, \beta_s, \beta_x, \beta_y)^T t_s(S_i)\big\},\ i = 1, ..., d, \end{aligned} \quad (6)$$

where the notation $\bar{Z}_{-i}$ and $\bar{S}_{-i}$ refers to the set of columns of $\bar{Z}$ and $\bar{S}$ except the i-th column respectively, and the vector functions $\eta(.)$ and $t(.)$ are the *natural parameter* (the natural parameter is a function of the variables that are being conditioned on) and the *sufficient statistics* respectively. These assumptions on the complete conditionals imply a conjugacy relationship between the hidden variables and the observations that implies a specific form of the complete conditional for the latent features.

We put a prior distribution on the binary matrix $\bar{Z}$ using the extension of the Beta process which takes two scalar parameters a_π and b_π and was originally proposed by [21]. A sample from the extended Beta process $B \sim BP(a_\pi, b_\pi, B_0)$ with base measure B_0 may be represented as

$$B_\omega = \sum_{k=1}^{K} \pi_k \delta_{\omega_k}, \quad (7)$$

where,

$$\pi_k \sim Beta(a_\pi/K, b_\pi(K-1)/K), \quad \omega_k \sim B_0. \quad (8)$$

This sample will be a valid sample from the extended Beta process, if $K \to \infty$. B_ω can be considered as a vector of K probabilities that each probability π_k corresponds to the atom ω_k. Now, we consider each latent binary feature $Z_i (i = 1, ..., d)$ to be drawn from a Bernoulli process $Be(B_\omega)$ with B_ω defined as 7 where a sample from this process can be generated as

$$Z_i = \sum_{k=1}^{K} z_{ki} \delta_{\omega_k}, \quad i = 1, ..., d, \quad (9)$$

where z_{ki} is generated by $z_{ki} \sim Bernoulli(\pi_k)$. So, we set the free parameter $\beta_z = [\pi_1, ..., \pi_K]$. By letting $K \to \infty$, the number of the atoms K (the dimensionality of the latent feature space) can be learned from the training data.

Based on the Beta-Bernoulli process prior on binary latent features $\bar{Z}$, computing the posterior distribution of $\bar{Z}$ is tractable for any likelihood function. However, for the weight latent features $\bar{S}$, the prior and the likelihood distributions must be in the conjugate exponential family as

$$p(X_i, Y_i|S_i, Z_i, \beta_x, \beta_y) = \\ \prod_{m=1}^{M} \prod_{n=1}^{N} \Big(A(x_{mi})A(y_{ni})exp\{\eta(S_i, Z_i, \beta_x, \beta_y)^T t(x_{mi}, y_{ni}) - \phi(S_i, Z_i, \beta_x, \beta_y)\} \Big), \tag{10}$$

$$p(S_i|\beta_s) = \prod_{k=1}^{K} A(s_{ki})exp\{\beta_s^T t(s_{ki}) - \phi_s(\beta_s)\}, \tag{11}$$

where the scalar functions $A(.)$ and $\phi(.)$ are the *base measure* and *log-normalizer* respectively. Using the conjugacy relationship between Eqs. 10 and 11, the sufficient statistics for s_{ki} is

$$t(s_{ki}) = (\eta_{s_{ki}}(S_i, Z_i, \beta_x, \beta_y), -\phi_{s_{ki}}(S_i, Z_i, \beta_x, \beta_y)), \tag{12}$$

where $f_x(.)$ means that we consider x as the free parameter of the function f by considering all other parameters as constant. The parameter β_s has two components $\beta_s = (\beta_s^1, \beta_s^2)$. The first component β_s^1 is a vector of the same dimension as $\eta_{s_{ki}}(S_i, Z_i, \beta_x, \beta_y)$; the second component β_s^2 is a scalar. This form will be important when we derive constrained variational inference in Section 5.1.

To be fully Bayesian, we also put conjugate prior distributions on the free parameters β_s, β_x and β_y as

$$p(\beta_s|a_s) = A(\beta_s)exp\{a_s^T t(\beta_s) - \phi_{\beta_s}(a_s)\}, \tag{13}$$

$$p(\beta_x|a_x) = A(\beta_x)exp\{a_x^T t(\beta_x) - \phi_{\beta_y}(a_x)\}, \tag{14}$$

$$p(\beta_y|a_y) = A(\beta_y)exp\{a_y^T t(\beta_y) - \phi_{\beta_x}(a_y)\}, \tag{15}$$

where

$$t(\beta_s) = (\beta_s, -\phi_s(\beta_s)), \tag{16}$$

$$t(\beta_x) = (\eta_{\beta_x}(S_i, Z_i, \beta_x, \beta_y), -\phi_{\beta_x}(S_i, Z_i, \beta_x, \beta_y)), \tag{17}$$

$$t(\beta_y) = (\eta_{\beta_y}(S_i, Z_i, \beta_x, \beta_y), -\phi_{\beta_y}(S_i, Z_i, \beta_x, \beta_y)), \tag{18}$$

where $a_\pi, b_\pi, a_s, a_x, b_y$ are the hyper-parameters of the proposed model. The graphical representation of the proposed model is demonstrated in Fig. 1.

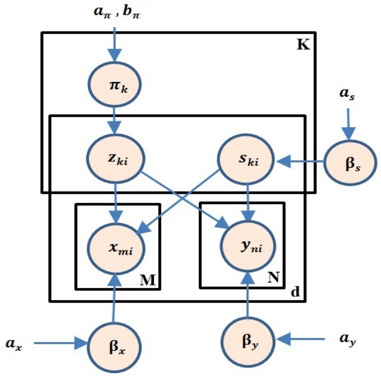

Fig. 1. The graphical representation of the proposed Bayesian model.

5 Posterior Inference

Due to the fact that computing the exact posterior distribution of the latent variables given the observations is intractable, in this section, we develop a new VB algorithm, to approximate that posterior distribution.
The goal of variational inference is to approximate the true posterior distribution over the latent variables with a variational distribution which is closest in KL divergence to the true posterior distribution. A brief review of the VB algorithm for the exponential family distributions provided in the supplementary Material.
In our variational inference framework, we use the finite Beta-Bernoulli approximation, in which the dimensionality of the latent space (K) is truncated and set to a finite but large number. If K is large enough, the analyzed multi-modal data using this number of latent features, will reveal less than K components.
In the following section, we introduce our VB method which incorporates the information of similarity/dissimilarity constraints into inferring the posterior distributions.

5.1 Constrained Variational Inference

In the proposed Bayesian MMDML model, the latent variables are $\varXi = \Big\{ \varPi = [\pi_1, \pi_2, ..., \pi_K], \bar{Z}, \bar{S}, \beta_s, \beta_x, \beta_y \Big\}$, and the hyper-parameters are $\varPhi = \{a_x, a_y, a_s, a_\pi, b_\pi\}$. So, the joint probability of data and unknown variables are

$$P(\bar{X}, \bar{Y}, \varXi \mid \varPhi) = \prod_{i=1}^{d} \Bigg(\prod_{m=1}^{M} P(x_{mi} \mid Z_i, S_i, \beta_x) \prod_{n=1}^{N} P(y_{ni} \mid Z_i, S_i, \beta_y)$$

$$\prod_{k=1}^{K} P(z_{ki} \mid \pi_k) P(s_{ki} \mid \beta_s) \Bigg) \prod_{k=1}^{K} P(\pi_k \mid a_\pi, b_\pi) P(\beta_x | a_x) P(\beta_y | a_y) P(\beta_s | a_s). \quad (19)$$

We use a fully factorized variational distribution for the hidden variables as

$$q(\mathit{\Pi}, \bar{Z}, \bar{S}, \beta_s, \beta_x, \beta_y) = \prod_{k=1}^{K} q_{\pi_k}(\pi_k) \prod_{i=1}^{d} \prod_{k=1}^{K} q_{z_{ki}}(z_{ki}) q_{s_{ki}}(s_{ki}) q_{\beta_s}(\beta_s) q_{\beta_x}(\beta_x) q_{\beta_y}(\beta_y).$$

Since all the distributions belong to the conjugate exponential families, we can determine the form of the approximate posterior distributions, so we have:

$$\begin{aligned}
&q_{\pi_k}(\pi_k) = Beta(\pi_k; a_\pi^k, b_\pi^k), \;\; k = 1, ..., K,\\
&q_{z_{ki}}(z_{ki}) = Bernoulli(z_{ki}; \nu_{ki}), \;\; k = 1, ..., K, i = 1, ..., d,\\
&q_{s_{ki}}(s_{ki}) = A(s_{ki}) exp\{(\hat{\beta}_s^{ki})^T t(s_{ki}) - \phi_s(\hat{\beta}_s^{ki})\}, \;\; k = 1, ..., K, i = 1, ..., d,\\
&q_{\beta_s}(\beta_s) = A(\beta_s) exp\{\hat{a}_s^T t(\beta_s) - \phi_{\beta_s}(\hat{a}_s)\},\\
&q_{\beta_x}(\beta_x) = A(\beta_x) exp\{\hat{a}_x^T t(\beta_x) - \phi_{\beta_y}(\hat{a}_x)\},\\
&q_{\beta_y}(\beta_y) = A(\beta_y) exp\{\hat{a}_y^T t(\beta_y) - \phi_{\beta_x}(\hat{a}_y)\}.
\end{aligned}$$

If we denote the set of posterior parameters by $\mathit{\Omega}$, the proposed constrained variational inference then involves maximizing the lower bound on the marginal likelihood ($P(\bar{X}, \bar{Y} \mid \mathit{\Phi})$ by enforcing similar/dissimilar pairs to have similar/dissimilar posterior distributions over their latent features. This is equivalent to solving the following optimization problem

$$\begin{aligned}
\hat{\mathit{\Omega}} =& \underset{\mathit{\Omega}}{argmin} \; - \mathbb{E}_q(\log P(\bar{X}, \bar{Y}, \mathit{\Xi} \mid \mathit{\Phi})) - H[q] + \frac{1}{|A|} \sum_{(i,j)\in A} \sum_{k=1}^{K} \big(\lambda_z(\nu_{ki} - \nu_{kj})^2 \\
&+ \lambda_s(\hat{\beta}_s^{ki} - \hat{\beta}_s^{kj})^2\big), \;\; s.t. \; \forall (i,j) \in D, \sum_{k=1}^{K} (\nu_{ki} - \nu_{kj})^2 \geq 1, \sum_{k=1}^{K} (\hat{\beta}_s^{ki} - \hat{\beta}_s^{kj})^2 \geq 1,
\end{aligned} \tag{20}$$

where $H[.]$ denotes the Entropy operator, and $|A|$ denotes the cardinality of the set A. In [20], the similarity and dissimilarity constraints are directly imposed on the latent features. In our Bayesian framework, instead of imposing the constraints directly on latent features $(\bar{Z}, \bar{S})$, we impose them on the parameters of the posterior distributions of the latent features. It should be noted that both $P(\bar{X}, \bar{Y}, \mathit{\Xi} \mid \mathit{\Phi})$ and $H[q]$ are functions of posterior parameters $(\mathit{\Omega})$. The VB method simply tries to minimize the above objective function using the Coordinate Descent method. $E_q(\log P(\bar{X}, \bar{Y}, \mathit{\Xi} \mid \mathit{\Phi}))$ can be decomposed as

$$\begin{aligned}
&\mathbb{E}_q(\log P(\bar{X}, \bar{Y}, \mathit{\Xi} \mid \mathit{\Phi})) = \sum_{k=1}^{K} \mathbb{E}_q \log P(\pi_k \mid a_\pi, b_\pi) + \sum_{i=1}^{d} \sum_{k=1}^{K} \mathbb{E}_q \log P(z_{ki} \mid \pi_k) + \\
&\sum_{i=1}^{d} \Bigg(\sum_{m=1}^{M} \mathbb{E}_q \log P(x_{mi} \mid Z_i, S_i, \beta_x) + \sum_{n=1}^{N} \mathbb{E}_q \log P(y_{ni} \mid Z_i, S_i, \beta_y) \\
&+ \sum_{k=1}^{K} E_q \log P(s_{ki} \mid \beta_s) \Bigg) + \mathbb{E}_q \log P(\beta_s \mid a_s) + \mathbb{E}_q \log P(\beta_x \mid a_x) + \mathbb{E}_q \log P(\beta_y \mid a_y).
\end{aligned} \tag{21}$$

The update equation for each distribution is as follows (due to the conjugacy relationship for $\mathit{\Pi}, \beta_s, \beta_x, \beta_y$ and the fact that these variables do not appear in the constraints, updating posterior distribution of these variables is straightforward and is omitted due to the lack of space).

Update for $\bar{Z} = [Z_1, Z_2, ..., Z_d]$:
Due to the fact that there are some constraints on the posterior parameters of the binary feature matrix $\bar{Z}$, we cannot derive the update equation for the posterior parameters of $\bar{Z}$ in the closed form. Hence, for updating the parameters in our coordinate descent framework, we reformulate the objective function of Eq. 20 as a function of the posterior parameters of $\bar{Z}$ and directly solve the obtained optimization problem (it should be noted that in expanding the objective function, we consider all parameters fixed but the parameters of the binary feature matrix $\bar{Z}$). For updating each posterior parameter $\nu_{ki}(k = 1, ..., K; i = 1, ..., d)$, first, we define function $F(\nu_{ki})$ as:

$$F(\nu_{ki}) = -\mathbb{E}_q(\log P(\bar{X}, \bar{Y}, \mathit{\Xi} \mid \mathit{\Phi})) - H[q] + \frac{\lambda_z}{|A|} \sum_{j \in \{j|(i,j) \in A\}} (\nu_{ki} - \nu_{kj})^2 + c, \quad (22)$$

where c is the summation of all terms which are independent of ν_{ki}.

$$\begin{aligned}
\mathbb{E}_q(\log P(\bar{X}, \bar{Y}, \mathit{\Xi} \mid \mathit{\Phi})) &= \sum_{m=1}^{M} \mathbb{E}_q \log P(x_{mi} \mid Z_i, S_i, \beta_x) \\
&+ \sum_{n=1}^{N} \mathbb{E}_q \log P(y_{ni} \mid Z_i, S_i, \beta_y) + \mathbb{E}_q \log P(z_{ki} \mid \pi_k) + c \\
&= \sum_{m=1}^{M} f_{mi}(\nu_{ki}) + \sum_{n=1}^{N} f_{ni}(\nu_{ki}) + \langle \log \pi_k \rangle \nu_{ki} + \langle \log(1 - \pi_k) \rangle (1 - \nu_{ki}) + c, \quad (23)
\end{aligned}$$

where $f_{mi}(\nu_{ki}) = \mathbb{E}_q \log P(x_{mi} \mid Z_i, S_i, \beta_x)$ and $f_{ni}(\nu_{ki}) = \mathbb{E}_q \log P(y_{ni} \mid Z_i, S_i, \beta_y)$ and $\langle . \rangle$ indicates the expectation operator. For the entropy, we have:

$$H[q] = -\mathbb{E}_q \log q_{z_{ki}}(z_{ki}) + c = -\nu_{ki} \log \nu_{ki} - (1 - \nu_{ki}) \log(1 - \nu_{ki}) + c. \quad (24)$$

We can update the parameter ν_{ki} by solving the following optimization problem

$$\begin{aligned}
\hat{\nu}_{ki} = &\underset{\nu_{ki}}{argmin} \sum_{m=1}^{M} f_{mi}(\nu_{ki}) + \sum_{n=1}^{N} f_{ni}(\nu_{ki}) + \langle \log \pi_k \rangle \nu_{ki} + \langle \log(1 - \pi_k) \rangle (1 - \nu_{ki}) \\
&\quad + \nu_{ki} \log \nu_{ki} + (1 - \nu_{ki}) \log(1 - \nu_{ki}) \\
&s.t. \quad (\nu_{ki} - \nu_{kj})^2 \geq 1 \quad \forall j \in \{j|(i,j) \in D\}. \quad (25)
\end{aligned}$$

It is worth noting that the above optimization problem is a one dimensional problem that can be solved efficiently. Similarly, we can update the posterior parameters $(\{\hat{\beta}_s^{ki}\}(i = 1, ..., d; k = 1, ..., K))$ of the feature matrix $\bar{S}$ using the same procedure.

5.2 Latent feature prediction

After computing the posterior distribution of the latent features of the training data, in order to compute the posterior distribution of the latent feature for a new instance (X_t, Y_t), we must compute $P(H_t|X_t, Y_t, T)$ by integrating out the variables $\beta_x, \beta_y, \beta_s$, and $\mathit{\Pi}$ as

$$\begin{aligned} P(H_t|X_t, Y_t, T) &= \iiiint P(Z_t, S_t, \beta_x, \beta_y, \beta_s, \mathit{\Pi}|X_t, Y_t, T) d\beta_x d\beta_y d\beta_s d\mathit{\Pi} \\ &= \iiiint P(Z_t, S_t|\beta_x, \beta_y, \beta_s, \mathit{\Pi}, X_t, Y_t, T) P(\beta_x|T, X_t, Y_t) P(\beta_y|T, X_t, Y_t) \times \\ &\quad P(\beta_s|T, X_t, Y_t) P(\mathit{\Pi}|T, X_t, Y_t) d\beta_x d\beta_y d\beta_s d\mathit{\Pi}. \end{aligned} \tag{26}$$

Since the above expression cannot be computed in closed form, we resort to Gibbs sampling to approximate it. In other words, we estimate $P(Z_t, S_t|X_t, Y_t, T)$ as

$$P(Z_t, S_t|X_t, Y_t, T) \approx \frac{1}{L} \sum_{l=1}^{L} \delta_{z,s}(Z_t^l, S_t^l), \tag{27}$$

where L and r^l denote the number of samples and the l-th sample of the latent variable r. To sample from $P(Z_t, S_t|X_t, Y_t, T)$, we sample from $P(Z_t, S_t, \beta_x, \beta_y, \beta_s, \mathit{\Pi}|X_t, Y_t, T)$ based on the Gibbs sampling method [23]. Then, we simply ignore the values for $\beta_x, \beta_y, \beta_s, \mathit{\Pi}$ in each sample (it is worth noting that for generating samples for $\beta_x, \beta_y, \beta_s, \mathit{\Pi}$, we consider the approximate posterior distributions $q(\beta_x), q(\beta_y), q(\beta_s), q(\mathit{\Pi})$ as the prior distributions for these variables respectively).

Due to assumption that the posterior distribution of the latent features belong to the exponential famiy (Eqs. 5,6,13,14,15), deriving Gibbs sampling equations is straightforward.

In order to compare a test data point (X_t, Y_t) with a training data point (X_j, Y_j) based on their latent features, we first generate L samples $(\{Z_t^l, S_t^l\}_{l=1}^L)$ for the latent features Z_t, S_t based on the Gibbs sampling method. Then, we simply use Euclidean distance between the empirical mean of the generated samples $(\frac{1}{L}\sum_{l=1}^L Z_t^l \odot S_t^l)$ and the mean of the posterior distribution of the latent feature of (X_j, Y_j):

$$d\big((X_t, Y_t), (X_j, Y_j)\big) = \|\frac{1}{L}\sum_{l=1}^{L} H_t^l - \mathbb{E}_q[H_j]\|_2 = \|\frac{1}{L}\sum_{l=1}^{L} Z_t^l \odot S_t^l - \mathbb{E}_q[Z_j] \odot \mathbb{E}_q[S_j]\|_2,$$

where $d(.,.)$ and $\mathbb{E}_q[x]$ denote the distance function and the mean of the posterior distribution of the random variable x respectively.

6 Experimental Results

In this section, we verify the performance of the proposed Bayesian framework on tagged images data (images are associated with user textual tags such as title, description, comments, etc) which is ubiquitous in many photo sharing websites such as Instagram and Flickr.

Following [18], for image data modality, we first extract SIFT features from images. Then, we represent each image (X_i) using normalized bag-of-words on SIFT features. We also consider a discrete bag-of-words representation for text (Y_i) data [18].

To specialize our general Bayesian framework to tagged images data, we propose the following model for this bi-modal data:

$$\begin{aligned}
&P(z_{ki} \mid \pi_k) \sim Bernoulli(\pi_k), \quad k=1,...,K,\ i=1,...,d,\\
&P(\pi_k; a_\pi, b_\pi) \sim Beta(a_\pi/K, b_\pi(K-1)/K), \quad k=1,...,K,\\
&P(s_{ki} \mid \gamma_s) \sim \mathcal{N}(0, \gamma_s^{-1}), \quad k=1,...,K,\ i=1,...,d,\\
&P(x_{mi} \mid Z_i, S_i, W_m, \gamma_x) \sim \mathcal{N}(W_m^T(S_i \odot Z_i), \gamma_x^{-1}) \ \ m=1,...,M,\ i=1,...,d,\\
&P(y_{ni} \mid Z_i, S_i, U_n, \theta_n) = \frac{1}{1+exp(-y_{ni}(U_n^T(S_i \odot Z_i)+\theta_n))} \ \ n=1,...,N,\ i=1,...,d,\\
&P(W_m \mid \gamma_w) \sim \mathcal{N}(0, \gamma_w^{-1}I), \quad m=1,...,M,\\
&P(U_n \mid \gamma_u) \sim \mathcal{N}(0, \gamma_u^{-1}I), \quad n=1,...,N,\\
&P(\gamma_s; a_s, b_s) \sim Gamma(a_s, b_s),\ P(\gamma_x; a_x, b_x) \sim Gamma(a_x, b_x),\\
&P(\gamma_w; a_w, b_w) \sim Gamma(a_w, b_w),\ P(\gamma_u; a_u, b_u) \sim Gamma(a_u, b_u),\\
&P(\theta_n) \sim \mathcal{N}(0,1), \quad n=1,...,N,
\end{aligned}$$

where we assume that each element of Y_i is a binary random variable with logistic function distribution ($y_{ni} = +1$ if the n-th term of a tag dictionary appears around the i-th image and $y_{ni} = -1$ otherwise). We also assume that each element x_{mi} of X_i is a Gaussian variable denoting the normalized bag-of-words representation based on the SIFT feature.

Since the non-conjugacy of sigmoid function and Gaussian function violates our conjugacy assumption of posterior distribution over latent features $\bar{S}$, we use the local lower bound to the sigmoid function [24], which states for any $x \in R$ and $\xi \in [0, +\infty]$

$$\frac{1}{1+exp(-x)} \geq \sigma(\xi) exp\Big((x-\xi)/2 - \lambda(\xi)(x^2-\xi^2)\Big), \tag{28}$$

where,

$$\lambda(\xi) = \frac{-1}{2\xi}\big(\frac{1}{1+exp(-\xi)} - \frac{1}{2}\big). \tag{29}$$

So, we replace each sigmoid factor with the above lower bound, then we optimize the factorized variational distributions and free parameters ($\{\xi_{ni}\}_{n=1,i=1}^{N,d}$) using the EM algorithm (the constrained VB algorithm and the Gibbs sampling equations for this model is available in the Supplementary Material).

6.1 Experimental setup

We report the results of the proposed method (PM) on NUS-WIDE-1.5K: a subset selected from NUS-WIDE dataset which was used in [18]. The images of this

Table 1. k-NN classification accuracy on NUS-WIDE-1.5K dataset.

Method	O-Xing	O-ITML	MVH-Xing	MVH-ITML	MKE	SMVH-Xing	PM
1-NN	87.34	89.74	89.60	92.67	81.20	92.80	**95.34**
3-NN	82.26	68.27	87.47	89.07	70.94	90.54	**93.07**
5-NN	67.46	49.87	84.53	84.94	57.60	88.13	**90.14**
10-NN	46.27	26.14	74.40	71.74	46.14	84.93	**88.26**
20-NN	13.74	7.07	60.53	46.80	19.07	71.86	**77.74**

dataset is from Flicker and each image is associated with more than one user tag. For this dataset, we selected 30 classes and choose 50 images for each class (the total number of images is 1500). The 30 classes are **food**, **glacier**, **bridge**, **buddha**, **cliff**, **clouds**, **building**, **car**, **cathedral**, **leaf**, **monks**, **forest**, **computers**, **desert**, **flag**, **mushrooms**, **flowers**, **hills**, **lake**, **moon**, **motorcycle**, **actor**, **butterfly**, **camels**, **airplane**, **bicycle**, **ocean**, **police**, and **pyramid**. We randomly choose half of the images for training and the other half for testing. For the text modality, 1000 tags with top frequency are selected to form the tag dictionary. For image modality, we extract SIFT based bag-of-words representation with a codebook of size 1024. We need to generate side information in the forms of pairwise training instances. Following [1], we sample "similar" pairs by picking up two instances from the same class and "dissimilar" pairs by choosing two instances from different classes. We randomly sample about 10K "similar" pairs and 10K "dissimilar" pairs from the training set.

For comparison purposes, we compare our method with the **O-Xing** (we concatenate original feature vectors of text modality and image modality into a single representation and subsequently learn a Mahalanobis distance using the metric learning method proposed in [10]), **O-ITML** (we combine features of text and image into a whole and feed it to the ITML [1] method), **MVH-Xing** (we use the unsupervised MWH model to embed data from text and image modalities to the latent space and learn distance measure on the latent representations using the method proposed in [10]), **MVH-ITML** (we use ITML [1] to learn distance on the latent feature vectors obtained from MWH model), **MKE** (We compare with the multiple kernel embedding method proposed in [11]), and **SMVH-Xing** (we use the supervised MWH model based on "similar" and "dissimilar" pairs to embed data from text and image modalities to the latent space and learn distance measure on the latent representations using the method proposed in [18].

In the experiment, all Gamma priors are set as Gamma $(10^{-6}, 10^{-6})$ to make the prior distributions uninformative. The parameters a_π, b_π of the Beta distribution are set to $a_\pi = K$ and $b_\pi = K/2$. A preset large dimensionality of the latent features $K = 120$ is used for this dataset. The regularization parameters are also set as $\lambda_z = 1000$; $\lambda_s = 1000$. For the Gibbs sampling inference, we discard the initial 200 samples (burn-in period), and collect the next 300 samples to present the posterior distribution over the latent feature of a test instance.

Table 2. Average precision (AP) of image retrieval on NUS-WIDE-1.5K dataset.

Method	O-Xing	O-ITML	MVH-Xing	MVH-ITML	MKE	SMVH-Xing	PM
AP	52.24	42.48	74.85	68.89	51.26	84.42	**88.61**

6.2 Classification and Retrieval Experiments

We apply the learned distance measure for k-nearest neighbor (k-NN) classification on the dataset. Table 1 summarizes the classification accuracy for $k = 1, 3, 5, 10, 20$.

This table shows three major points. First, the proposed method (PM), SMVH-Xing, MVH-Xing, and MVH-ITML significantly outperform the two other methods, because these methods capture the correlation and complementary relationships between the modalities by transferring two different modalities into a shared single modality (latent space).

Second, MVH-Xing, and MVH-ITML are less accurate than SMVH-Xing, and PM. The reason is that in MVH-Xing, and MVH-ITML, feature embedding and metric learning are performed separately, while SMVH-Xing, and PM embed multi-modal data into the latent space and learn distance metric simultaneously to achieve the overall optimality that leads to better performance.

Third, our method has better performance than the SMVH-Xing due to the fact that the number of the training data points are small. More precisely, SMVH-Xing uses Maximum Likelihood (MAP estimation from probabilistic point of view) which can overfit to small-size training data. In contrast, the proposed method uses Bayesian learning that is relatively immune to overfitting.

In order to demonstrate the ability of the proposed method to learn the dimensionality of the latent space as well as the latent features, we plot the sorted values of $\langle \Pi \rangle$ for the NUS-WIDE-1.5K dataset, inferred by the algorithm (Fig. 2). As it can be seen, the algorithm inferred approximately 83 number of features, fewer than the 120 initially provided.

For the retrieval task, we treat each test image as query and we rank the other images of the test set according to their distances with the given query. We consider an image relevant to query if both images share the same class label.

We use the standard Average Precision (AP) [18] to evaluate the retrieval result. The AP value is the area under precision-recall curve for a query. The **recall** is the ratio of the relevant examples retrieved over the total relevant examples, and the precision value is the ratio of relevant examples over the total retrieved examples in the database.

The AP result is summarized in Table 2 from which, we can see that our methods have better performance than the other methods.

6.3 Sensitivity Analysis

We test the sensitivity of the proposed method to different choices of the parameter λ_z. Fig. 2 shows the variation of average precision (AP) with varying λ_z

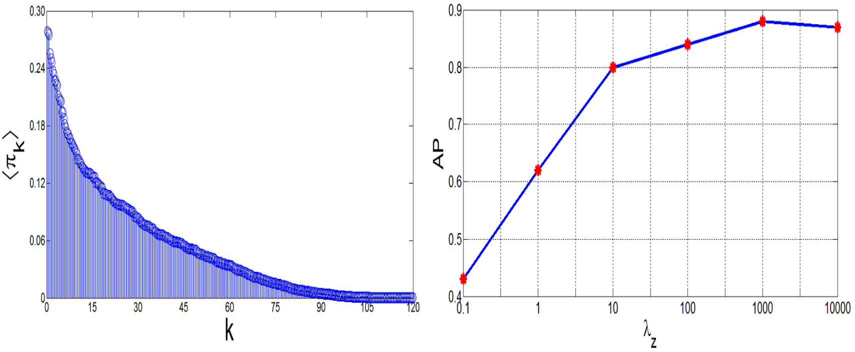

Fig. 2. Left: Inferred $\langle \Pi \rangle$ for the NUS-WIDE-1.5K dataset, Right: Retrieval performance sensitivity with respect to λ_z.

(while evaluating λ_z, the parameter λ_s is fixed). As can be seen, by increasing λ_z from 0.1 to 1000, AP is improved. Moreover, further increasing λ_z reduces the average score mildly (the AP drops from 88.61 to 87.78). The possible reason is that using Bayesian learning prevents the model from overfitting to the training data.

7 Conclusion

In this paper, we propose a general Bayesian framework of multi-modal distance metric learning. This framework embed arbitrary number of data modalities into a single latent space with the ability of learning the dimensionality of the latent space from observed data itself. Moreover, a new Varitional Inference algorithm is introduced that is capable of encoding distance supervision of data points. Empirical results on tagged image retrieval and classification applications demonstrated the benefits inherited from the proposed fully-Bayesian method.

References

1. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: ICML (2007)
2. Globerson, A., Roweis, S.: Metric learning by collapsing classes. In: NIPS (2006)
3. Guillaumin, M., Verbeek, J., Schmid, C.: Is that you? Metric learning approaches for face identification. In: ICCV (2009).
4. McFee, B., Lanckriet, G.: Metric learning to rank. In: ICML (2010)
5. Nguyen, N., Guo, Y.: Metric Learning: A Support Vector Approach. In: ECML/PKDD (2008)
6. Nishida, K., Hoshide, T., Fujimura, K.: Improving tweet stream classification by detecting changes in word probability. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (2012)
7. Zhen, Y., Yeung, D.Y.: A probabilistic model for multi-modal hash function learning. In: KDD (2012)

8. Qi, G.J., Aggarwal, C.C., Huang, T.S.: On clustering heterogeneous social media objects with outlier links. In: Proceedings of the fifth ACM international conference on Web search and data mining (2012)
9. Aizenberg, N, Koren, Y., Somekh, O.: Build your own music recommender by modeling internet radio streams. In: Proceedings of the 21st international conference on World Wide Web (2012)
10. Xing, E., Ng, A.Y., Jordan, M.I., Russell, S.: Distance metric learning with application to clustering with side information. In: NIPS (2003)
11. McFee, B., Lanckriet, G.: Learning multi-modal similarity. The Journal of Machine Learning Research (2011)
12. Bar-Hillel, A., Hertz, T., Shental, N., Weinshall, D.: Learning a mahalanobis metric from equivalence constraints. Journal of Machine Learning Research (2005)
13. Hoi, S.C.H., Liu, W., Lyu, M.R., Ma, W.Y.: Learning distance metrics with contextual constraints for image retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition (2006)
14. Jain, P., Kulis, B., Davis, J.V., Dhillon, I.S.: Metric and kernel learning using a linear transformation. Journal of Machine Learning Research (2012)
15. Weinberger, K.,Blitzer, J., Saul, L.: Distance metric learning for large margin nearest neighbor classification. In: NIPS (2006)
16. Si, L., Jin, R., Hoi, S.C.H, Lyu, M.R.: Collaborative image retrieval via regularized metric learning. ACM Multimedia Systems Journal (2006)
17. Hoi, S.C.H., Liu, W., Chang, S.F.: Semi-supervised distance metric learning for collaborative image retrieval and clustering. ACM Trans. Multimedia Comput. Commun. Appl (2010)
18. Xie, P., Xing, E.P.: Multi-Modal Distance Metric Learning. In: IJCAI (2013)
19. Hjor, N.L.: Nonparametric bayes estimators based on beta processes in models for life history data. Annals of Statistics, (1990)
20. Xing, E.P., Yan, R., Hauptmann, A: Mining associated text and images with dual-wing harmoniums (2005)
21. Paisley, J., Carin, L.: Nonparametric factor analysis with beta process priors. In: ICML (2009)
22. Kingman, J. F. C.: Completely random measures. Pacific Journal of Mathematics (1967)
23. Robert, C.P., Casella, G.: Monte carlo statistical methods. Springer Verlag (2004)
24. Jaakkola, T., Jordan, M.I.: Bayesian parameter estimation via variational methods. Statistics and Computing (2000)