Similarity Gaussian Process Latent Variable Model for Multi-Modal Data Analysis

Guoli Song[1,2], Shuhui Wang[2], Qingming Huang[1,2], Qi Tian[3]

[1] Key Lab of Big Data Mining and Knowledge Management,
University of Chinese Academy of Sciences, Beijing, China
[2] Key Lab of Intell. Info. Process., Inst. of Comput. Tech, Chinese Academy of Sciences, Beijing, China
[3] Department of Computer Science, University of Texas at San Antonio, TX, 78249, USA
guoli.song@vipl.ict.ac.cn, wangshuhui@ict.ac.cn, qmhuang@jdl.ac.cn, qitian@cs.utsa.edu

Abstract

*Data from real applications involve multiple modalities representing content with the same semantics and deliver rich information from complementary aspects. However, relations among heterogeneous modalities are simply treated as observation-to-fit by existing work, and the parameterized cross-modal mapping functions lack flexibility in directly adapting to the content divergence and semantic complicacy of multi-modal data. In this paper, we build our work based on Gaussian process latent variable model (GPLVM) to learn the non-linear non-parametric mapping functions and transform heterogeneous data into a shared latent space. We propose **m**ulti-modal **Sim**ilarity **G**aussian **P**rocess latent variable model (m-SimGP), which learns the nonlinear mapping functions between the intra-modal similarities and latent representation. We further propose multi-modal regularized similarity GPLVM (m-RSimGP) by encouraging similar/dissimilar points to be similar/dissimilar in the output space. The overall objective functions are solved by simple and scalable gradient decent techniques. The proposed models are robust to content divergence and high-dimensionality in multi-modal representation. They can be applied to various tasks to discover the non-linear correlations and obtain the comparable low-dimensional representation for heterogeneous modalities. On two widely used real-world datasets, we outperform previous approaches for cross-modal content retrieval and cross-modal classification.*

1. Introduction

Data from real applications often involve multiple modalities representing content with the same semantics [4] and deliver rich information from complementary aspects. For example, in content-based image retrieval, the semantics in an image can be inferred from visual features, such as color or texture features, and from its associated textual descriptions such as user tags, paragraphs and user comments. The key problem for multi-modal data analytics is how to model the correlations across different modalities to facilitate content retrieval of heterogeneous modalities. This motivates latent variable modeling to discover the correlation information shared by different modalities.

As a possible solution, canonical correlation analysis (CCA) [11, 19, 20] projects multi-modal data into a shared subspace that guarantees different modalities are maximally correlated. However, CCA-based methods lack probabilistic interpretation on the intra-modal similarities. Topic models [3, 12, 25] learn latent topics to describe the intrinsic semantic correlations in multi-modal data. Based on Latent Dirichlet Allocation (LDA) [3], a variety of constraints are imposed. For example, mmLDA [2] enforces that all modalities share the same topic proportions, and CorrLDA [2] assumes one-to-one correspondence between the topics in each modality. These assumptions inherently restrain the flexibility of correlation models. In general, existing studies take intra-modal similarity and inter-modal similarity indiscriminately as the observations of multi-modal relation, thus the correlation learning problem is solved by fitting the mapping function outputs to the observations. However, the prior constraints are imposed to avoid non-smoothness in the functional space, lacking flexibility in directly adapting to the properties of multi-modal data.

Gaussian Process Latent Variable Model (GPLVM) [13, 8, 6] is a well-established generative approach for learning nonlinear low-dimensional embedding. Instead of specifying a set of deterministic (*e.g.* CCA-based [1, 11, 19, 20]) or parametric (*e.g.*, univariate Gaussian [27]) mapping func-

tions, a smooth non-parametric Gaussian process is defined in GPLVM on the probabilistic mapping from latent space to observation space. The flexibility of Gaussian process, determined by a variety of covariance functions, facilitates learning from real world data with content divergence and complicated semantic relations. Despite that GPLVM better adapts to different modalities, there has been very few studies in describing multi-modal relation using GPLVM.

We hereby address cross-modal correlation learning with multi-modal GPLVMs. In fact, GPLVM can effectively discover the non-linear relationship among multi-modal data by introducing additive priors over latent space [8] or modality-specific covariance functions [6]. Given the latent representation, multi-modal data is reconstructed by the learned Gaussian process parameterized by covariance kernels. However, two primary drawbacks of existing GPLVM [13, 8, 6] limit its application on real-world data.

First, the topological structure in the observation space is not guaranteed to be preserved in the function embedding process of GPLVM, which leads to model degradation in processing high-dimensional multi-modal data due to the *curse-of-dimensionality*. In existing study, the affinity structure is constructed to encode the modality-specific topological structure, which reflects the intra-modal content similarity [27, 24], context information [24] and semantic consistency [10]. It can be used as observation-to-fit [27] or mapping function regularizer [10] for learning a common latent space. To preserve the intra-modal topology, we propose to learn the nonlinear mapping functions between the intra-modal similarities and latent representation. From the maximum likelihood perspective, the nonlinear covariance matrices of multi-modal mapping functions are learned to maximize the consistency to the modality-specific intra-modal topologies. It better encodes the nonlinear semantic similarity in multi-modal data. Compared to existing correlation models, our non-parametric similarity-based GPLVM is robust to content divergence and high-dimensionality in multi-modal representation.

Second, existing models exploit simple inter-modal relation in correlation learning. For example, CCA-based models [1, 11, 19, 20] assume that the inter-modal relation is expressed by co-occurrence of multi-modal data objects. The inter-modal relation is also encoded as binary observation matrix to be fit by the correlation models [3, 12, 25, 27]. By contrast, we directly impose two kinds of inter-modal relations (*i.e.*, semantic similarity and dissimilarity) as smooth priors on the output of multi-modal GPLVM. By using such regularization on the latent space, our model enforces that the semantically similar/dissimilar cross-modal observations are also similar/dissimilar in the latent space, which provides goal-oriented solution to maximize the cross-modal semantic consistency.

In summary, we propose the **m**ulti-modal **Sim**ilarity **G**aussian **P**rocess latent variable model (m-SimGP) for multi-modal data analysis, which learns the nonlinear mapping functions between the intra-modal similarities and latent representation. We further impose a cross-modal similarity/dissimilarity constraint as a smooth prior to the latent space. Accordingly, we develop a regularized model, called m-RSimGP, to learn the cross-modal correlation. The m-RSimGP model enforces that the semantically similar/dissimilar cross-modal observations are also similar/dissimilar in the latent space, which maximizes the cross-modal semantic consistency. The conditional dependency among latent space and muti-modal similarity observations can be easily constructed in a maximum a posteriori inference framework, while the overall objective functions can be solved by simple and scalable gradient decent techniques. The proposed models can be applied to various tasks to discover the non-linear correlations and obtain the comparable low-dimension representation for heterogeneous modalities. On two widely used real-world multi-modal datasets, we achieve at least 15% improvement over the existing approaches in cross-modal content retrieval task, and about 3% improvement over DS-SBP [10] in cross-modal classification task.

2. Preliminary

Gaussian process latent variable model (GPLVM) [13] is a probabilistic model for non-linear low dimensional embedding. It assumes that high dimensional data is generated from a low dimensional latent space, where the mapping from latent space to observation space is a Gaussian process (GP), as shown in Figure 1.a.

Let $Y = [y_1, \ldots, y_N]^\top \in \mathbb{R}^{N \times d}$ represent the training data set with N data points $y_i \in \mathbb{R}^d$, $i = 1, ..., N$. The goal is to learn the corresponding latent space $X = [x_1, \ldots, x_N]^\top \in \mathbb{R}^{N \times q}$, with $q \ll d$. The assumption is that the training data set is drawn from the latent space with a noisy process,

$$y_i = f(x_i) + \varepsilon, \tag{1}$$

where ε is additive Gaussian noise with zero mean. A Gaussian process prior is placed over the mapping function f:

$$f(x) \sim \mathrm{GP}(\mu(x), k(x, x')), \tag{2}$$

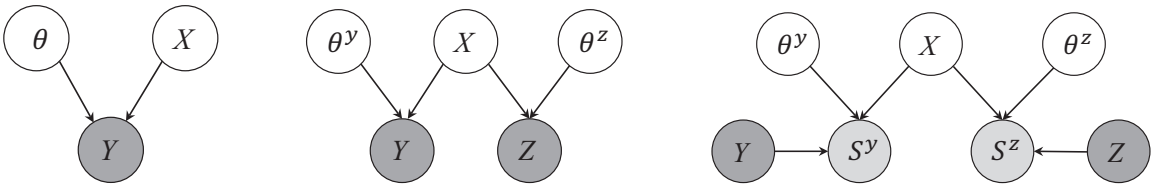
where the mean function $\mu(x)$ is typically taken to be zero for simplicity, and the covariance function $k(x, x')$ is necessarily constrained to positive definite matrices.

The marginal likelihood of the observation Y with respect to the latent space X can be formulated by integration over f,

$$p(Y|X) = \int p(Y|f)\, p(f|X)\, df. \tag{3}$$

Specifically, given a GP with covariance function $k(x, x')$, the likelihood of the data Y given the latent variable X is

$$p(Y|X, \theta) = \frac{1}{\mathcal{A}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(K^{-1} Y Y^\top\right)\right), \tag{4}$$



(a) GPLVM (b) Shared GPLVM (c) m-SimGP

Figure 1: GPLVMs *v.s.* the proposed m-SimGP. (a) is the original GPLVM proposed by Lawrence [13]. The observed data Y is assumed to be generated from a latent variable set X. (b) shows the shared GPLVMs for multi-modal data. Two observed data modalities Y and Z are assumed to share the common latent space X. (c) is the proposed m-SimGP in this paper. To preserve local structure of each data modality, we build a multi-modal latent variable model between the intra-modal similarities and latent space.

where the normalization factor $\mathcal{A} = \sqrt{(2\pi)^{Nd}|K|^d}$, and $K \in \mathbb{R}^{N\times N}$ is the kernel matrix defined on X, i.e., $K_{ij} = k(x_i, x_j)$. Any positive definite kernel can be used to construct a Gaussian process covariance function. Considering that RBF kernel is simpler and more effective for high dimensional data [13], we adopt RBF with white noise as the covariance function,

$$k(x, x') = \sigma^2_{\text{rbf}} \exp(-\frac{\|x - x'\|^2}{2l^2_{\text{rbf}}}) + \sigma^2_{\text{w}}\delta_{x,x'}, \quad (5)$$

where $\theta = \{\sigma^2_{\text{rbf}}, \sigma^2_{\text{w}}, l_{\text{rbf}}\}$ denotes the parameters of the covariance matrix, which govern the variance of the RBF kernel, the variance of additive noise, and the RBF bandwidth, respectively.

In practice, a maximum a posteriori (MAP) probability estimation is used to learn the latent space X. The posterior distribution can be written as

$$p(X, \theta | Y) \propto p(Y | X, \theta) p(X). \quad (6)$$

The latent space X is then obtained by minimizing the negative log posterior:

$$\arg\min_X \mathcal{L} - \log p(X), \quad (7)$$

where $\mathcal{L}$ is the negative logarithm associated with Eq. (4),

$$\mathcal{L} = \frac{d}{2}\ln|K| + \frac{1}{2}\text{tr}\left(K^{-1}YY^\top\right) \quad (8)$$

Different forms of prior knowledge can be easily introduced into the GPLVM to enhance the flexibility [13, 14, 22]. For example, the spherical Gaussian prior [13] is used over the latent variables to enforce the smoothness of X and prevent the GPLVM from placing latent positions infinitely far apart. To preserve local distance structure, the back-constrained GPLVM [14] is proposed, which preserves the local affinity structure in the original data space.

GPLVM can be generalized to multiple data spaces, which are assumed to share a common latent space [21, 9, 6], as shown in Figure 1.b. These models have achieved success in numerous applications, such as human pose estimation [8], tracking [23, 18], facial expression recognition [10], etc. We apply GPLVM to multi-modal data analysis for cross-modal correlation learning and classification.

3. The Proposed Approach

The goal of our work is to discover a general latent representation shared by observations from multiple modalities. To achieve this, GPLVM is constructed on the modalities for its flexibility in probabilistic modeling on the conditional dependency between observation and latent space. In Section 3.1, we introduce multi-modal similarity Gaussian process latent variable model, using similarity information in each data modality to preserve the intra-modal consistency. We further impose cross-modal similarity and dissimilarity constraints on the latent space, which further enhances the model generality of our multi-modal similarity GPLVM approach as shown in Section 3.2.

Specifically, we consider a set of bi-modal data objects $\mathcal{O} = \{o_i\}_{i=1}^N$, each comprising of observation from two modalities, *i.e.*, $o_i = \{y_i, z_i\}$. Let $Y \in \mathbb{R}^{N\times d_y}$ and $Z \in \mathbb{R}^{N\times d_z}$ represent two data modalities, respectively. The objective is to relate these two modalities to the same latent space. Gaussian kernel is used to measure the intra-modal similarities. Specifically, the similarity matrices $S^y \in \mathbb{R}^{N\times N}$ and $S^z \in \mathbb{R}^{N\times N}$ are defined as follows,

$$\begin{aligned} S^y(y_i, y_j) &= \exp\left(-d^2(y_i, y_j)/2\gamma_y\right), \\ S^z(z_i, z_j) &= \exp\left(-d^2(z_i, z_j)/2\gamma_z\right), \end{aligned} \quad (9)$$

where $d(y_i, y_j) = \|y_i - y_j\|_2$ and $d(z_i, z_j) = \|z_i - z_j\|_2$. $\gamma_y, \gamma_z > 0$ are bandwidth parameters.

3.1. Multi-modal Similarity GPLVM (m-SimGP)

As shown in Figure 1.c, we assume that the intra-modal similarity matrices S^y and S^z are generated from a shared q-dimensional latent manifold, where $q \ll \min(d_y, d_z)$. Each similarity matrix can be represented by the mappings with respect to a common latent space $X \in \mathbb{R}^{N\times q}$:

$$S^y_{ij} = f^y_{ij}(X) + \varepsilon^y_{ij}, \;\; S^z_{ij} = f^z_{ij}(X) + \varepsilon^z_{ij}, \quad (10)$$

where $f^y_{ij} = f^y_j(x_i)$ and $f^z_{ij} = f^z_j(x_i)$ map the latent variable to the corresponding similarity. Each x_i generates the

i-th row of S^y and S^z with $f_j, j = 1, \ldots, N$. The noise terms ε^y and ε^z are typically taken to be Gaussian with zero mean.

Similar as GPLVM, to find the latent representation X and the mappings $\{f_j^y\}_{j=1}^N$ and $\{f_j^z\}_{j=1}^N$, we place Gaussian process priors over the mappings:

$$\begin{aligned} f^y &\sim GP\left(\mu^y\left(X\right), K^y(X,X)\right), \\ f^z &\sim GP\left(\mu^z\left(X\right), K^z(X,X)\right). \end{aligned} \tag{11}$$

As in section 2, the mean functions are taken to be zero, and the covariance functions are generated by RBF kernel. The definition allows the mappings to be marginalized out analytically, and the marginal likelihood with respect to the latent variable can be computed as,

$$p\left(S^y, S^z \,|X, \theta^y, \theta^z\right) = p\left(S^y \,|X, \theta^y\right) p\left(S^z \,|X, \theta^z\right), \tag{12}$$

$$p(S^y \,|X, \theta^y) = \tfrac{1}{\mathcal{A}^y} \exp\left(-\tfrac{1}{2}\mathrm{tr}\left(K_y^{-1} S^y (S^y)^\top\right)\right), \tag{13}$$

$$p(S^z \,|X, \theta^z) = \tfrac{1}{\mathcal{A}^z} \exp\left(-\tfrac{1}{2}\mathrm{tr}\left(K_z^{-1} S^z (S^z)^\top\right)\right). \tag{14}$$

If Gaussian prior is used over the latent variable, the objective function can be written as:

$$\arg\min_X \mathcal{L}^y + \mathcal{L}^z + \sum_{i=1}^N \frac{1}{2}\|x_i\|^2, \tag{15}$$

where $\mathcal{L}^y$ and $\mathcal{L}^z$ are the negative log-likelihood associated with Eq. (13) and (14), respectively. When the observed data are coming from more than two modalities, the model can be readily extended by adding the modality-specific negative log-likelihood. The model optimization can be solved by scaled conjugate gradient (SCG) technique [17].

However, nothing in GPLVM encourages the semantically dissimilar observations to be far in the latent space, nor semantically similar observations to be close in the latent space [22]. The learned latent space may not appropriately reflect the true cross-modal correlation in the observation space in the context of multi-modal correlation learning. To address this problem, we propose a regularized similarity GPLVM in the following section.

3.2. Multi-modal Regularized Similarity GPLVM (m-RSimGP)

In the m-SimGP model, a simple spherical Gaussian prior is placed over the latent variable. Inspired by [26], to minimize the distance between similar data pairs and maximize the distance between dissimilar data pairs, we develop a multi-modal regularized similarity Gaussian process latent variable model (m-RSimGP), where the prior characterized by a cross-modal similarity matrix is placed over the latent space, as shown in Figure 2.

Given a set of data objects $\mathscr{O} = \{o_i\}_{i=1}^N$ with two feature modalities y_i and z_i, the cross-modal similarity matrix $S^{yz} \in \{0,1\}^{N\times N}$ is defined as follows:

$$\left(S^{yz}\right)_{ij} = \begin{cases} 1, & \text{if } (o_i, o_j) \in \mathscr{S} \\ 0, & \text{if } (o_i, o_j) \in \mathscr{D} \end{cases} \tag{16}$$

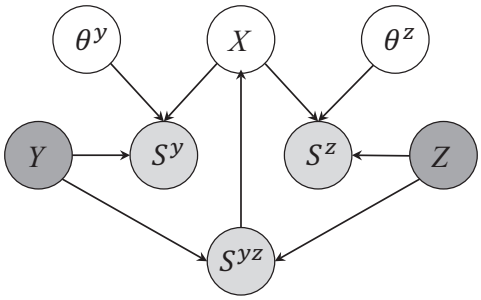


Figure 2: Multi-Modal Regularized Similarity GPLVM

where $i, j = 1, 2, \ldots, N$. $\mathscr{S} = \{(o_i, o_j)\}$ denotes the set of pairs with similar semantics, and $\mathscr{D} = \{(o_i, o_j)\}$ denotes the set of pairs with dissimilar semantics. To make sure that semantically similar observations are close to each other and semantically dissimilar observations are far from each other in the embedded latent space, we impose the similarity and dissimilarity priors on the latent representation. The corresponding learning problem with respect to the latent variable X is formulated as follows:

$$\begin{aligned} &\min_X \sum\nolimits_{(o_i,o_j)\in\mathscr{S}} \|x_i - x_j\|^2 \\ &\text{s.t.} \quad \|x_i - x_j\|^2 \geqslant 1, \forall\,(o_i, o_j) \in \mathscr{D} \end{aligned} \tag{17}$$

where x_i is the representation of the data point $o_i = \{y_i, z_i\}$ in the latent space. Euclidean distance is used as the distance measure for the embedded latent representation. The dissimilar points are separated by a margin of 1 in the latent space similar to [26].

The optimization problem in Eq. (17) can be interpreted as a prior over the latent variable and combined with the likelihood maximization problem, where the smooth Gaussian prior constraint in Eq. (15) is substituted with the cross-modal similarity and dissimilarity constraints. As a result, our proposed m-RSimGP model is formulated as:

$$\begin{aligned} &\min_X \mathcal{L}^y + \mathcal{L}^z + \sum\nolimits_{(o_i,o_j)\in\mathscr{S}} \|x_i - x_j\|^2 \\ &\text{s.t.} \quad \|x_i - x_j\|^2 \geqslant 1, \forall\,(o_i, o_j) \in \mathscr{D} \end{aligned} \tag{18}$$

The dissimilar constrains in Eq. (18) can be further relaxed with a convex hinge loss. Thus we obtain an unconstrained problem that is much easier to optimize:

$$\begin{aligned} \min_X\ &\mathcal{L}^y + \mathcal{L}^z + \lambda_1 \sum\nolimits_{(o_i,o_j)\in\mathscr{S}} \|x_i - x_j\|^2 + \\ &\lambda_2 \sum\nolimits_{(o_i,o_j)\in\mathscr{D}} \max\left(0, 1 - \|x_i - x_j\|^2\right) \end{aligned} \tag{19}$$

where λ_1 and λ_2 are the tradeoff parameters. They can be assigned with the same value, indicating equal importance of similar pairs and dissimilar pairs.

The m-RSimGP model can be easily extended and scaled to multi-modal data. For example, given data with three modalities, $o_i = (y_i, z_i, w_i), i = 1, \ldots, N$, the observations of different modalities share the common latent variables

x_i for each o_i. Therefore, the pair-wise semantic relation is still applied between o_i and o_j, $i,\ j = 1, \ldots, N$.

3.3. Optimization and Inference

By substituting the cross-modal similarity matrix S^{yz} into Eq. (19), the equivalent formulation is:

$$\min_X \ \mathcal{L}^y + \mathcal{L}^z + \lambda_1 \sum_{i,j} (S^{yz})_{ij} \cdot \|x_i - x_j\|^2 \qquad (20)$$

$$+\lambda_2 \sum_{i,j} \mathbb{1}\left(\|x_i - x_j\|^2 < 1\right)\left(1 - (S^{yz})_{ij}\right)\left(1 - \|x_i - x_j\|^2\right).$$

The scaled conjugate gradient can be applied to obtain the optimal latent representation X. Specifically, the gradients of $\mathcal{L}^y$ and $\mathcal{L}^z$ can be computed as follows:

$$\frac{\partial \mathcal{L}^y}{\partial x_i} = \frac{1}{2}\left(NK_y^{-1} - \left(K_y^{-1} S^y (S^y)^\top K_y^{-1}\right)\right)\frac{\partial K_y}{\partial x_i}, \qquad (21)$$

$$\frac{\partial \mathcal{L}^z}{\partial x_i} = \frac{1}{2}\left(NK_z^{-1} - \left(K_z^{-1} S^z (S^z)^\top K_z^{-1}\right)\right)\frac{\partial K_z}{\partial x_i}, \qquad (22)$$

where $\frac{\partial K_y}{\partial x_i}$ and $\frac{\partial K_z}{\partial x_i}$ can be easily obtained, since the RBF kernel is infinitely differentiable. The gradient of the third term in Eq. (20), denoted as $\mathcal{L}^s$, is computed as:

$$\frac{\partial \mathcal{L}^s}{\partial x_i} = 4\lambda_1 \sum_{j=1}^{N} (S^{yz})_{ij}\,(x_i - x_j). \qquad (23)$$

Since the hinge loss is convex and non-differential, we compute the subgradient with respect to the distances induced by the dissimilar pairs at each step. The subgradient of the last term in Eq. (20), denoted as $\mathcal{L}^d$, is computed as:

$$\frac{\partial \mathcal{L}^d}{\partial x_i} = 4\lambda_2 \sum_{j=1}^{N} \mathbb{1}\left(\|x_i - x_j\|^2 < 1\right)(1 - S^{yz})_{ij}\,(x_j - x_i)\,. \qquad (24)$$

The gradient of Eq. (20) with respect to the latent representation is the sum of these four terms.

After the optimization procedure, we obtain the Gaussian processes for generating multi-modal observations with the shared space X. When inferring the new set of observed test points, the inference procedure is straightforward in our solution framework. We take image observation as an example. The procedure for the text is similar. Given the image observation y_t, we need to learn the corresponding latent representation x_t by maximizing the posteriori probability. The inference is presented in Algorithm 1, where $(t, \cdot)$ indicate the index of the test sample and the index set of all the training samples, respectively. For a test observation with both modalities, we use similar way to obtain its latent representation. An approximation to the posterior $p(x_t|s^y_{t,\cdot}, s^z_{t,\cdot})$ is used to predict the latent representation x_t.

Based on different types of queries, our method can perform three kinds of cross-modal retrieval tasks: image retrieval from a text query, text retrieval from an image query, and multi-modal retrieval from a multi-modal query.

Algorithm 1: Inference for the latent space

Input: The image observation y_t.
Step 1: Compute the similarity matrix $s^y_{t,\cdot}$ between y_t and the training images Y according to Eq. (9).
Step 2: Find the corresponding latent position x_t by maximizing the posteriori probability $p(x_t|s^y_{t,\cdot})$.
Output: x_t.

By maximizing the posteriori probability of $p(x_t|s^y_{t,\cdot})$, $p(x_t|s^z_{t,\cdot})$ and $p(x_t|s^y_{t,\cdot}, s^z_{t,\cdot})$, the observations of different modalities can be projected into the unified latent space X. Then cross-modal retrieval is performed by measuring the distance between the latent representations.

4. Relation to Existing Models

Our model naturally extends GPLVM [13] on multiple modalities. To encode the intra-modal relation, we construct Gaussian processes on the mappings between latent representation and multi-modal observations at the similarity level rather than high-dimensional feature level [13]. It is robust to the notorious *curse-of-dimensionality* issue, which is even more jeopardizing for correlation modeling on heterogeneous high-dimensional data. It can be seen as a non-parametric generalization of existing subspace-learning-based correlation models, *e.g.*, canonical correlation analysis [11]. The conditional dependency among latent space and muti-modal observations can be easily constructed in a maximum a posteriori inference framework, while the overall objective functions can be solved by simple and scalable gradient decent techniques.

The mapping function of GPLVM does not necessarily guarantee that the similarity and dissimilarity in observation are preserved in the latent space. One solution to this problem is back-constraints [14], encoding the latent representation with the affinity information in the observation space. The discriminative shared-space prior [10], defined by a data-dependent weight matrix, enforces to preserve the topological structure. The topology preserving constraints regularize the latent space for multi-modal distance metric learning [26]. In the context of cross-modal learning, existing approaches [11, 12, 27] develop multi-modal projections to fit to observations of both intra-modal similarity and inter-modal relation among heterogeneous data objects. We directly impose cross-modal similarity and dissimilarity constraints on the output of our similarity-based GPLVM as smooth priors, which provides goal-oriented solution to maximize the cross-modal semantic consistency.

Most existing models assume that both inter-modal relation and intra-modal relation are independent. For example, CCA directly maximizes the correlation among a set of data object pairs that are assumed to be independently generated. MLBE [27] assumes that the intra-modal similarities and inter-modal relations are conditionally independent to each

other, and models the dependence between similarity observations and the latent variables by univariate Gaussians and binary latent factors. Markov random field is constructed on the topic models [12] that generate the mutually independent multi-modal similarities. Our method explains the multi-modal correlations from a new perspective of probabilistic interdependency. We assume that the intra-modal relation is conditionally independent of each other given the latent representation. The intra-modal similarities are sampled from multi-modal Gaussian processes determined by the modality-specific covariance functions on the latent representation. Such a non-parametric model better deals with content divergence and cross-modal correlation complicacy in real-world applications.

5. Experiments

5.1. Datasets and Experimental Settings

In our experiments, two popular multi-modal datasets listed below are used:

Wiki [19] is collected from Wikipedia consisting of 2,866 image-text documents. Each image is represented by a 128-dimensional bag-of-words based on SIFT descriptor and each text is represented by a 10-dimensional LDA feature. Totally 10 categories are considered and each document is labeled with one of them. A random 80/20 split of the dataset is used to produce a training set and a testing set.

Flickr [5] is a subset selected from NUS-WIDE, consisting of 5730 paired objects. Each pair includes an image represented by a 500-dimensional bag-of-words based on SIFT descriptor and 1000-dimensional tag text. The class labels of image-text pairs are selected as the classes with the top-10 largest numbers of images. We randomly choose 85% of the data for training and the remaining 15% for testing.

Unless specified, we use the optimal settings of the parameters tuned by a parameter validation process for all the experiments. The bandwidth parameters of similarity matrices and the tradeoff parameters in m-RSimGP are set to 1, *i.e.*, $\gamma_y = \gamma_z = \lambda_1 = \lambda_2 = 1$. CCA is used to obtain a low-dimensional initial representation of the latent space shared by two data modalities.

5.2. Image-Text Retrieval

Image-text retrieval is a typical cross-modal problem, consisting of two tasks: (1) image query vs. text database, (2) text query vs. image database. A retrieved result is considered correct if it belongs to the same class as the query. We use 11-point interpolated precision-recall (PR) curve and mean average precision (MAP) [16] to measure the retrieval performance.

We compare m-SimGP and m-RSimGP with canonical correlation analysis (CCA) [11], semantic correlation matching (SCM) [19], multi-modal latent binary embedding (MLBE[1]) [27] and back-constrained shared GPLVM (SGPLVM[2]) [8]. In SCM, the CCA modeling is first applied to learn two maximally correlated subspaces, and then Logistic regressors are learned in each of these subspaces. As a generative model, MLBE uses binary hash codes as latent variables to generate intra-modal and inter-modal similarities. The code length for MLBE is set to 8 in our experiments. SGPLVM uses back-constraints to learn a shared latent representation that captures the correlations among different modalities.

Methods \ Tasks	img-query	txt-query	Average
CCA [11]	0.2453	0.2010	0.2232
SCM [19]	0.2684	0.2276	0.2480
MLBE [27]	0.3787	0.4109	0.3948
SGPLVM [8]	0.1961	0.1546	0.1754
m-SimGP	0.4336	0.4188	0.4262
m-RSimGP	**0.4697**	**0.4418**	**0.4558**

Table 1: The MAP comparison on Wiki dataset. The results shown in boldface are the best.

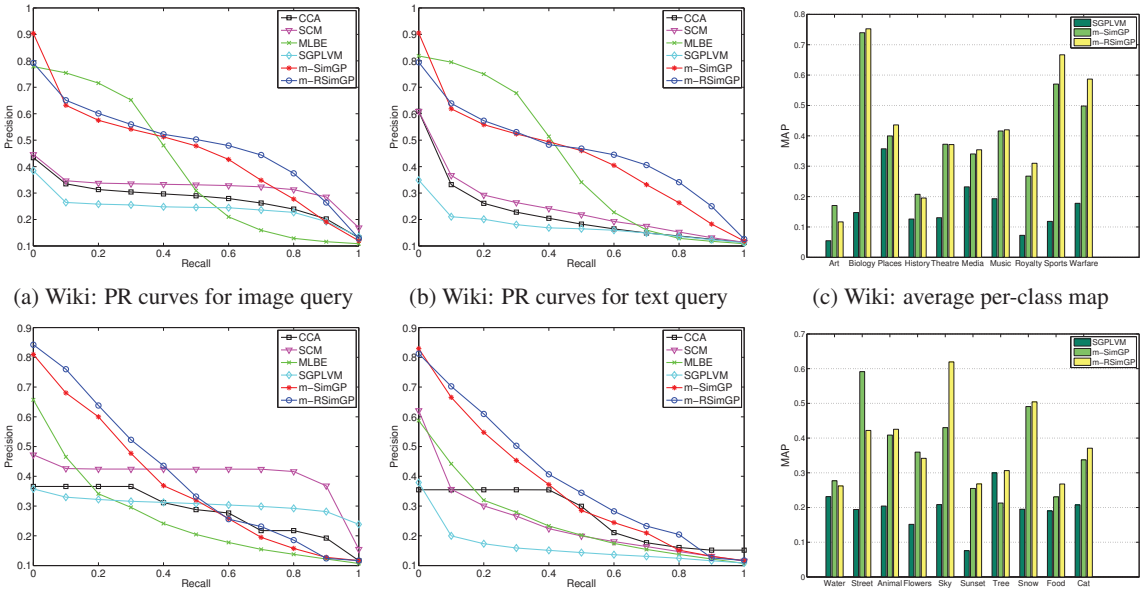
Methods \ Tasks	img-query	txt-query	Average
CCA [11]	0.2072	0.2003	0.2038
SCM [19]	0.3282	0.2187	0.2735
MLBE [27]	0.2533	0.3232	0.2883
SGPLVM [8]	0.2666	0.1429	0.2048
m-SimGP	0.3473	0.3380	0.3427
m-RSimGP	**0.3855**	**0.3719**	**0.3787**

Table 2: The MAP comparison on Flickr dataset. The results shown in boldface are the best.

On the Wiki dataset, Figure 3.a and 3.b show that our m-SimGP and m-RSimGP achieve significant improvement over CCA, SCM and SGPLVM in both retrieval tasks. Compared to SGPLVM, m-SimGP gains significant performance improvement, which indicates that similarity information is important in capturing the correlation structure of multi-modal data. Compared to our methods, MLBE is a parametric model pre-specified with the nearly optimal latent feature dimension and thus achieves better precision at low recall rates on small-scale dataset Wiki. Table 1 provides further comparison by measuring their MAP scores. It is clear that our nonparametric models outperform the parametric MLBE model. The best performance achieved by m-RSimGP outperforms MLBE by 15% higher MAP. It shows that the cross-modal similarity and dissimilarity

[1]MLBE: `https://bitbucket.org/zhenyisx`.

[2]SGPLVM: `https://github.com/SheffieldML/SGPLVM`.



(a) Wiki: PR curves for image query (b) Wiki: PR curves for text query (c) Wiki: average per-class map

(d) Flickr: PR curves for image query (e) Flickr: PR curves for text query (f) Flickr: average per-class map

Figure 3: Precision-Recall curves of cross-modal retrieval using both image ((a)(d)) and text ((b)(e)) queries. Average per-class MAP scores across image and text queries also shown in (c) and (f).

constraints in Eq. (17) over the latent variables contribute significantly to the cross-modal correlation learning. Figure 3.c also shows the per-class MAP scores of our methods compared to SGPLVM. On all the classes of Wiki, both m-SimGP and m-RSimGP have higher MAP scores than SGPLVM, a GPLVM-based baseline. The m-RSimGP model outperforms m-SimGP on most of the Wiki classes.

Figure 3.d and 3.e show the PR curves on the Flickr dataset. For the image retrieval task with textual queries, we can see that m-SimGP and m-RSimGP perform much better than any other methods. m-RSimGP further achieves the best performance. For the text retrieval task with image queries, the PR curves of our methods dominate other methods but the classification-based SCM. Our methods achieve higher precision than SCM at low levels of recall, which is more applicable in practice. Table 2 further provides solid evidence demonstrating the superior performance of our methods. Specifically, m-RSimGP achieves 31% higher MAP compared to MLBE. The per-class MAP scores on the Flickr dataset are also compared to SGPLVM, shown in Figure 3.f. Similar to the Wiki dataset, m-RSimGP has the best overall performance.

Our methods consistently achieve promising performance on both retrieval tasks, which verifies the effectiveness of our methods in reducing the semantic gap between modalities. As shown in Table 1 and 2, other latent variable models either achieve better MAP performance of image query (*e.g.*, SGPLVM) or better MAP of text query (*e.g.*, MLBE). For our methods, the MAP scores of both retrieval tasks are pretty close to each other. Therefore, our models can better achieve the semantic consistency among cross-modal data and the learned latent representation can better reflect the cross-modal correlation in the observation space.

5.3. Classification

Our work aims to discover a general latent representation shared by multi-modal observations. Therefore, the resulting posterior of our framework is the latent space instead of the class information. In other words, the classification problem is not directly modeled in our methods. To obtain the class prediction, we apply a classifier to the learned latent space. In our experiments, classification is accomplished by using the k-nearest neighbor (k-NN) classifier to find the closest latent representations to the test data.

The proposed m-SimGP and m-RSimGP are compared to 1-NN, CCA, SGPLVM, Discriminative GPLVM (D-GPLVM) [22] and Discriminative Shared GPLVM (DS-GPLVM[3]) [10]. As a single-view method, D-GPLVM restricts the latent space with a prior based on Linear Discriminant Analysis (LDA). In our experiments, we extend it to learn from multi-modal observations. DS-GPLVM generalizes the Gaussian Markov Random Field (GMRF) prior for single view to multi-view learning. In [10], DS-GPLVM is performed in two scenarios for inference. In the first, each modality is independently back-projected to the latent s-

[3]DS-GPLVM: `https://ibug.doc.ic.ac.uk/resources/ele_TIP15`.

Datasets	Methods							
	1-NN	CCA [11]	SGPLVM [8]	D-GPLVM [22]	DS-IBP [10]	DS-SBP [10]	m-SimGP	m-RSimGP
Wiki	0.1746	0.1948	0.1457	0.1934	0.1499	**0.6921**	0.5959	0.6472
Flickr	0.1956	0.1713	0.2072	0.1562	0.1988	0.6898	0.6765	**0.7095**

Table 3: Average classification accuracy on both Wiki and Flickr datasets. The results shown in boldface are the best.

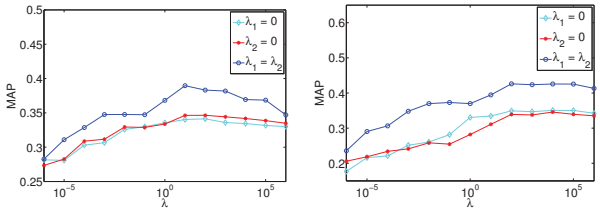
pace, where the back-constraints are defined on each modality separately. In the second, a single back-projection to the latent space is performed for all the modalities, where the back-constraint is defined on the set of all the modalities. We denote the former as DS-IBP and the latter as DS-SBP. We build 1-NN classifier baseline in the image feature space. In the testing stage, we apply 1-NN classifier to the learned latent space to obtain the prediction results.

Table 3 presents the average classification accuracy on both the Wiki and Flickr datasets. The results show that our methods can effectively learn a discriminative latent space from multi-modal observations. We can see that our methods are either comparable or better than other methods. DS-SBP and our m-RsimGP achieve the best classification accuracy on Wiki and Flickr, respectively. Though DS-IBP and D-GPLVM are designed for the classification problem, their poor performance reflects that these two methods lack the ability to effectively capture the semantically consistent representation of multi-modal data. Since the major difference between DS-IBP and DS-SBP is the pattern of back-projection, we attribute the superior performance of DS-SBP to the fact that DS-SBP back-projects complementary information from all the data modalities during the inference process. Different from DS-SBP, the inference procedure of our methods is much simpler, where only the image information is used to estimate a posteriori to infer the testing latent representations.

5.4. Parameter Sensitivity Analysis

We conduct sensitivity analysis on the tradeoff parameters λ_1 and λ_2 to test how they impact the cross-modal correlation learning. Without loss of generality, we perform sensitivity experiments on the training sets of the Wiki and Flickr datasets, respectively. Figure 4 shows the curves of average MAP scores of image-text retrieval with different tradeoff parameters. We consider three different settings: (1) λ_1 is fixed with 0, (2) λ_2 is fixed with 0, (3) λ_1 and λ_2 are set to the same value. For simplicity, the parameter variables are denoted as λ , as shown in Figure 4.

We can observe that both similar and dissimilar semantic information have great impact on the performance of m-RSimGP. The curves of λ_1 and λ_2 are pretty similar to each other, which indicates their equal importance in cross-modal correlation learning. In all the settings, the average MAP is low for small λ, *e.g.*, 10^{-4}, and it is improved by increasing λ. When λ is increased to 10^1 and 10^2, the performance is much better. Figure 4.a clearly shows that the



(a) Wiki (b) Flickr

Figure 4: Sensitivity test on the tradeoff parameters w.r.t. the performance of image-text retrieval

average MAP on Wiki decreases when λ reaches 10^3, and Figure 4.b also shows the average MAP on Flickr is on a downward trend. These phenomenons are possibly due to the over-fitting problem caused by an overly large λ.

6. Conclusions

We address cross-modal correlation learning problem based on non-parametric GPLVM. We propose m-SimGP, which learns the nonlinear mapping functions between the intra-modal similarities and latent representation. We further propose m-RSimGP by forcing similar/dissimilar points to be similar/dissimilar in the output space. The proposed models are robust to content divergence and high-dimensionality in multi-modal representation, and can be applied to various tasks to obtain the comparable low-dimensional representation. The effectiveness of our models has been shown on cross-modal retrieval and classification. In future work, we will investigate on constructing hierarchical/deep structure for latent variable model [15, 7] to better capture the intrinsic semantic consistency of heterogeneous modalities. We will also address the problems of correspondence missing and information imbalance in real-world data based on our similarity-based GPLVM.

Acknowledgement This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400 and 2015CB351800, National Natural Science Foundation of China: 61025011, 61332016, 61390511, 61303160 and 61429201, 863 program of China: 2014AA015202, and Postdoctoral Science Foundation of China: 2014T70126. This work was supported in part to Dr. Qi Tian by ARO grants W911NF-15-1-0290, W911NF-12-1-0057 and Faculty Research Awards by NEC Laboratories of America.

References

[1] F. R. Bach and M. I. Jordan. A probabilistic interpretation of canonical correlation analysis. 2005.

[2] D. M. Blei and M. I. Jordan. Modeling annotated data. In *ACM SIGIR*, pages 127–134, 2003.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

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

[5] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *CIVR*, 2009.

[6] A. C. Damianou, C. H. Ek, M. K. Titsias, and N. D. Lawrence. Manifold relevance determination. In *ICML*, 2012.

[7] A. C. Damianou and N. D. Lawrence. Deep gaussian processes. In *AISTATS 2013*, pages 207–215, 2013.

[8] C. H. Ek, J. Rihan, P. H. S. Torr, G. Rogez, and N. D. Lawrence. Ambiguity modeling in latent spaces. In *MLMI*, pages 62–73, 2008.

[9] C. H. Ek, P. H. S. Torr, and N. D. Lawrence. Gaussian process latent variable models for human pose estimation. In *MLMI*, pages 132–143, 2007.

[10] S. Eleftheriadis, O. Rudovic, and M. Pantic. Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *IEEE Transactions on Image Processing*, 24(1):189–204, 2015.

[11] D. R. Hardoon, S. Szedmák, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[12] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In *ICCV*, pages 2407–2414, 2011.

[13] N. D. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 2005.

[14] N. D. Lawrence and J. Q. Candela. Local distance preservation in the GP-LVM through back constraints. In *ICML*, pages 513–520, 2006.

[15] N. D. Lawrence and A. J. Moore. Hierarchical gaussian process latent variable models. In *ICML*, pages 481–488, 2007.

[16] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[17] M. F. Møller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6(4):525–533, 1993.

[18] V. A. Prisacariu and I. Reid. Nonlinear shape manifolds as shape priors in level set segmentation and tracking. In *CVPR*, pages 2185–2192, 2011.

[19] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM Multimedia*, pages 251–260, 2010.

[20] A. Sharma, A. Kumar, H. D. III, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *IEEE CVPR*, pages 2160–2167, 2012.

[21] A. P. Shon, K. Grochow, A. Hertzmann, and R. P. N. Rao. Learning shared latent structure for image synthesis and robotic imitation. In *NIPS*, pages 1233–1240, 2005.

[22] R. Urtasun and T. Darrell. Discriminative gaussian process latent variable models for classification. In *ICML*, 2007.

[23] R. Urtasun, D. J. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. In *IEEE ICCV*, pages 403–410, 2005.

[24] X. Wang and E. Grimson. Spatial latent dirichlet allocation. In *NIPS*, pages 1577–1584. 2008.

[25] Y. Wang, F. Wu, J. Song, X. Li, and Y. Zhuang. Multi-modal mutual topic reinforce modeling for cross-media retrieval. In *ACM Multimedia*, pages 307–316, 2014.

[26] P. Xie and E. P. Xing. Multi-modal distance metric learning. In *IJCAI*, 2013.

[27] Y. Zhen and D. Yeung. A probabilistic model for multimodal hash function learning. In *ACM KDD*, pages 940–948, 2012.