Multimodal Gaussian Process Latent Variable Models with Harmonization

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

[1] 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@ucas.ac.cn, qi.tian@utsa.edu

Abstract

In this work, we address multimodal learning problem with Gaussian process latent variable models (GPLVMs) and their application to cross-modal retrieval. Existing GPLVM based studies generally impose individual priors over the model parameters and ignore the intrinsic relations among these parameters. Considering the strong complementarity between modalities, we propose a novel joint prior over the parameters for multimodal GPLVMs to propagate multimodal information in both kernel hyperparameter spaces and latent space. The joint prior is formulated as a harmonization constraint on the model parameters, which enforces the agreement among the modality-specific GP kernels and the similarity in the latent space. We incorporate the harmonization mechanism into the learning process of multimodal GPLVMs. The proposed methods are evaluated on three widely used multimodal datasets for cross-modal retrieval. Experimental results show that the harmonization mechanism is beneficial to the GPLVM algorithms for learning non-linear correlation among heterogeneous modalities.

1. Introduction

In real-world applications, we have access to rich data that involves multiple modalities, such as image with text [1, 2], or video with audio [3]. Better multimodal representations are required to describe the complementary information of heterogeneous modalities with intrinsic topic and semantic relations. In this work, we consider multimodal learning problem and its application to cross-modal retrieval [4–7] that has attracted much attention in computer vision community. Specifically, given queries from one modality (*e.g.*, image), the goal of cross-modal retrieval is to retrieve database entries from other modalities (*e.g.*, textual descriptions) that are semantically consistent or relevant to the queries.

Various methods have been proposed recently to model the correlation across different modalities. Among these, latent variable models are typically used to relate heterogeneous modalities to a latent space, where a joint representation is learned across content modalities. Canonical correlation analysis (CCA) [1, 8, 9] is one of the representative schemes for latent variable modeling, where the shared subspace is learned by maximizing the correlation between the projections of data modalities. However, the deterministic mappings in CCA-based methods generally lack probabilistic interpretation on the interactions between modalities and flexibilities to content divergence. In this work, we study generative non-linear and non-parametric model for cross-modal correlation learning.

As a probabilistic extension of PCA [10], a probabilistic Gaussian process mapping is defined in Gaussian process latent variable model (GPLVM) [10] from low dimensional latent space to high dimensional observation space. Further, multimodal GPLVMs [11–15] are proposed to learn a latent representation to capture the shared information among multimodal data. Due to the non-parametric nature, they can effectively learn low dimensional representations of heterogeneous data sources. The latent representation can be used for various tasks such as robotic imitation learning [11], facial recognition [15], tracking [16], and cross-modal retrieval [12].

In the context of multimodal GPLVM learning, how to learn the relation on heterogeneous data modalities is still a critical issue that is left to be further investigated. Typically, manifold alignment [17] is used to build connections between manifolds by fitting the point-wise related observations. By learning Gaussian process (GP) projections from each original data modality, alignment-based approaches [11, 13, 14, 18] attempt to discover a shared latent manifold

2380-7504/17 $31.00 © 2017 IEEE
DOI 10.1109/ICCV.2017.538

to align different modalities. Since multimodal GPLVMs are learned by unsupervised learning strategy, the learned latent representations of correlated data pairs might be quite different, and thus their structure and semantic correlations are not guaranteed to be well preserved.

A possible solution to this problem is to employ additive priors over the latent space by incorporating structure and semantic information of data observations [13, 14, 19] or enforcing data alignment [15, 20]. For example, latent points are back-constrained in [13] to be a smooth function of the data points for preserving affinity structure. A distance-preserved constraint is proposed for the latent space in [19] to maximally preserve the intra-modal global similarity structure. In [15], data-dependent GMRF prior is used to learn a discriminative shared manifold to align multiple views of a facial expression. For these GPLVM algorithms, the prior information for kernel hyperparameters is ignored, or only uninformative priors are imposed over kernel hyperparameters.

It is worth noticing that the learning of multimodal GPLVMs is not restricted to the set of latent representations for heterogeneous modalities. The GP mapping function, also to be learned, is generated conditionally on both the latent points and the kernel hyperparameters. Gaussian process kernels play an important role in pattern discovery [14, 21]. With more expressive kernels, one could use Gaussian processes to learn a better latent representation for multimodal data. Therefore, the fully Bayesian approaches [22, 23] for GPLVM are proposed by additionally placing priors on the kernel parameters. However, the latent points and the kernel hyperparameters are treated independently in these works. Such an individual learning scheme may be mutually incompatible on real world problems, and thus limited adaptation to content divergence and complex multimodal correlation may be achieved.

To address these concerns, we propose harmonized multimodal GPLVM, which includes a model-driven prior that goes beyond the individual design paradigm of latent priors and GP kernels. The harmonization is achieved by minimizing the divergence, measured by distance-induced loss functions, between modality-specific GP kernels and similarity matrix of the latent points. By building the model harmonization, the modality-specific structure information can be more sufficiently transferred among the kernel hyperparameters via the shared latent space, and in return the learned representations are endowed with better multimodal topology preservation. Furthermore, the additional information transfer pathway on the model parameter space can help to avoid the inappropriate solution brought by noise and correlation observation sparsity.

In this work, three variants of multimodal GPLVMs are proposed for cross-modal retrieval. The Harmonized Multimodal GPLVM (hmGPLVM) enforces model harmonization in the learning process of standard multimodal GPLVM [11]. The Harmonized Similarity GPLVM (hmSimGP) introduces harmonization into the similarity-based GPLVM [12] to minimize the divergence among input similarity, latent similarity and GP covariance. The Harmonized m-RSimGP (hm-RSimGP) combines the harmonization constraint with the inter-modal (dis)similarity prior to enhance the cross-modal semantic consistency. The resulting low dimensional representations for heterogeneous modalities can be used to perform cross-modal retrieval by ranking their distances on the latent space. Significant improvement has been achieved over the existing approaches on three widely used real-world multimodal datasets.

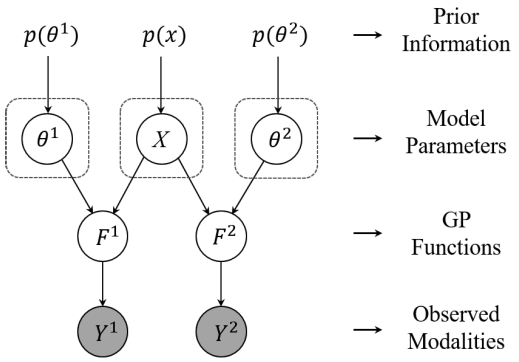


Figure 1. Multimodal GPLVM (mGPLVM): Independent prior constraints are imposed over the parameters (latent representations X and kernel hyperparameters θ^1, θ^2) of multimodal GPLVM.

2. Multimodal GPLVM

Multimodal Gaussian process latent variable model (mGPLVM) assumes that different data modalities are aligned in a shared manifold [11, 13, 14]. Without loss of generality, we discuss the multimodal learning on two data modalities in this paper. As shown in Figure 1, the objective of multimodal GPLVM is to relate two modalities $Y^1 \in \mathbb{R}^{N\times d_1}$ and $Y^2 \in \mathbb{R}^{N\times d_2}$ to the same latent space $X \in \mathbb{R}^{N\times q}$, where $q \ll \min(d_1, d_2)$. Data in each modality can be generated through the mapping functions (F^1, F^2) parameterized by two Gaussian processes. By marginalizing the non-linear mappings out, the joint marginal likelihood of Y^1 and Y^2 is given by,

$$\begin{aligned} p\left(Y^1, Y^2 \left| X, \theta\right.\right) &= p\left(Y^1 \left| X, \theta^1\right.\right) p\left(Y^2 \left| X, \theta^2\right.\right) \\ &= \int p\left(Y^1 \left| F^1\right.\right) p\left(F^1 \left| X, \theta^1\right.\right) dF^1 \\ &\quad \cdot \int p\left(Y^2 \left| F^2\right.\right) p\left(F^2 \left| X, \theta^2\right.\right) dF^2, \end{aligned} \tag{1}$$

where $\theta = \{\theta^1, \theta^2\}$ is the kernel or covariance hyperparameters for GP mapping functions. In the following, we denote $c \in \{1, 2\}$ in order to simplify our notation.

Different GPLVM approaches for multimodal learning

can be obtained by varying the assumption of the prior distribution over the model parameters. Generally, the shared representations X and the kernel hyperparameters θ are treated independently in existing works [22–24], and assumed to follow separate prior distributions $p(X)$, $p(\theta^1)$ and $p(\theta^2)$, as shown in Figure 1. By incorporating prior information on the problem at hand, we can learn the model parameters X and θ using maximum a posteriori (MAP) probability estimation. In practice, the learning of the model parameters is carried out by minimizing the negative log-posterior,

$$\mathcal{L} = \sum_c \mathcal{L}_c - \log p(\theta^c) - \log p(X), \tag{2}$$

where $\mathcal{L}_c$ is the corresponding negative log-likelihood of $p(Y^c|X, \theta^c)$, $c \in \{1, 2\}$, and is derived as

$$\mathcal{L}_c = \frac{d_c}{2} \ln |K_c| + \frac{1}{2} \text{tr}\left(K_c^{-1} Y^c (Y^c)^\top\right). \tag{3}$$

The covariance matrix $K_c = k_c(X, X)$ is defined by the kernel function k_c operating on the latent space X.

The original mGPLVM may encounter model degradation in processing high dimensional multimodal data, since the topological structure in the data space is not guaranteed to be preserved in the function embedding process. To solve this problem, similarity GPLVM (m-SimGP) for multimodal learning is proposed in [12], which learns latent space and multimodal mapping functions to maximize the consistency to the modality-specific topologies. The intra-modal similarities $S^1 \in \mathbb{R}^{N\times N}$ and $S^2 \in \mathbb{R}^{N\times N}$ are computed according to the Gaussian kernel, and they are assumed to be generated from a shared q-dimensional latent manifold $X \in \mathbb{R}^{N\times q}$. Similar as the generation procedure of multimodal GPLVM, the joint marginal likelihood of S^1 and S^2 given the model parameters can be computed as,

$$p\left(S^1, S^2 | X, \theta\right) = p\left(S^1 \middle| X, \theta^1\right) p\left(S^2 \middle| X, \theta^2\right), \tag{4}$$

$$p(S^c | X, \theta^c) = \frac{1}{\mathcal{A}^c} \exp\left(-\frac{1}{2} \text{tr}\left(K_c^{-1} S^c (S^c)^\top\right)\right), \tag{5}$$

where $\mathcal{A}^c = \sqrt{(2\pi)^{N^2} |K_c|^N}$, $c \in \{1, 2\}$. The shared latent space X and the kernel hyperparameters θ can be learned by minimizing the joint negative log-likelihood.

3. Multimodal GPLVMs with Harmonization

Learning in the multimodal GPLVM consists of minimizing the log-posterior with respect to the latent space X and the hyperparameters θ. Beyond existing individual learning mechanism for the model parameters, *e.g.*, Eq. (2) assuming the three kinds of model parameters to be independent, we assume a joint prior distribution $p(\theta^c, x)$ over the hyperparameters θ^c and the latent space X for each data modality, $c \in \{1, 2\}$, as shown in Figure 2. The new

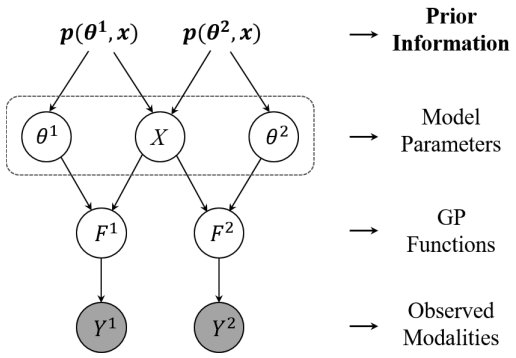


Figure 2. Harmonized multimodal GPLVM (hmGPLVM). We impose joint prior constraints on the parameters of multimodal GPLVM to harmonize X and kernel hyperparameters θ^1, θ^2.

negative log-posterior is given by

$$\mathcal{L} = \sum_c \mathcal{L}_c - \log p(\theta^c, X). \tag{6}$$

In this work, we define the joint prior distribution by a harmonization constraint over the modality-specific kernels and the similarity in the latent space. By building direct linkages between the model parameters, the proposed harmonization mechanism facilitates better GP learning, and enforces multimodal information to transfer across hyperparameter spaces via the latent space.

3.1. The harmonization constraint

For multimodal learning, we aim to learn a model in which the divergence between similarity in the data space and the latent space is small. In this work, the covariance kernel of the GP mapping is chosen to model the similarity among data. For structure preservation, we enforce the agreement between the modality-specific kernels and the similarity of latent points, and propose a harmonization constraint formulated as:

$$\mathcal{H}_c(K_c - S^x) \leq \rho_c, \quad c \in \{1, 2\}, \tag{7}$$

where $S^x \in \mathbb{R}^{N\times N}$ is the latent similarity matrix measured by the distances among latent points. K_1 and K_2 are the non-linear covariance matrices which depend on the respective kernel hyperparameters θ^1 and θ^2. For each $c \in \{1, 2\}$, $\mathcal{H}_c(\cdot, \cdot)$ is a convex sub-differentiable loss function operating on K_c and S^x. The constraint parameters $\rho_1, \rho_2 > 0$ are used to control the divergence between GP kernels and the similarity of latent points.

The proposed distance-based constraint Eq. (7) enforces the harmonization among the modality-specific kernels and the similarity of latent points, which enforces the agreement between GP kernels (K_1, K_2) for different modalities. Further, this will also enforce the agreement among the similarities in the data space across modalities. In our proposed models, we take the proposed distance-based harmonization

constraint Eq. (7) as joint priors over the latent space X and the kernel hyperparameter space θ. From this point, the mapping functions F^1 and F^2 are no longer conditionally independent given the latent points X. Therefore, different data modalities are more closely related in our models.

3.2. The proposed models

In this work, we propose three models for multimodal learning, *i.e.*, hmGPLVM, hm-SimGP, and hm-RSimGP, to evaluate the effectiveness of the proposed harmonization constraint on different multimodal GPLVM algorithms. We incorporate the constraint Eq. (7) respectively as prior information over model parameters into three multimodal GPLVM algorithms, *i.e.*, mGPLVM [11], m-SimGP [12], and m-RSimGP [12], which differ greatly in the model structure for GP learning and the function of multimodal data information.

3.2.1 Harmonized multimodal GPLVM (hmGPLVM)

The mGPLVM algorithm is the first multimodal generalization of the GPLVM that can handle multiple observation modalities, which assumes that the observable outputs Y^1 and Y^2 are generated from a common latent space, as described in Section 2. To learn the latent representation X shared by heterogeneous modalities, we minimize the negative log-posterior $\mathcal{L}$ given by Eq. (6), where the prior information over the model parameters is derived from the minimization of the loss function $\mathcal{H}_c$ in (7). Specifically, we replace the hard constraint in (7) by a penalty term on the loss function, and combine it with the negative log-likelihood of observed data modalities. The harmonized multimodal GPLVM is derived as follows,

$$\arg\min_{X,\theta} \sum_c \mathcal{L}_c + \mu_c \mathcal{H}_c(K_c - S^x), \tag{8}$$

where $c \in \{1, 2\}$. $\mathcal{L}_c$ is the negative log-likelihood function given by (3). μ_c is the tradeoff parameter. The regularization terms ensure that the solution for K_1 and K_2 should be in the vicinity of S^x, which enforces consistency of correlation structure among heterogeneous data.

3.2.2 Harmonized similarity GPLVM (hm-SimGP)

As mentioned in Section 2, the m-SimGP algorithm [12] is proposed for structure preservation, which learns a shared latent representation from the intra-modal similarities of multimodal data. We incorporate the harmonization constraint Eq. (7) into the similarity-based m-SimGP model to enforce the consistency among similarities on both the latent and the kernel hyperparameter spaces. The proposed hm-SimGP model is formulated as:

$$\arg\min_{X,\theta} \sum_c \mathcal{L}_c^s + \mu_c \mathcal{H}_c(K_c - S^x), \tag{9}$$

where $c \in \{1, 2\}$. $\mathcal{L}_c^s$ is the negative log-likelihood associated with Eq. (5). μ_1 and μ_2 are the tradeoff parameters.

By introducing harmonization, we build the interaction among three different kinds of similarities in different manifolds for hm-SimGP, *i.e.*, the latent similarity (S^x) in shared space, the intra-modal similarities (S^1, S^2) in multimodal data spaces, and the kernels of mapping functions (K_1, K_2). The harmonization mechanism encourages the divergence between these similarities to be small, and thus brings a more consistent representation for multimodal data.

3.2.3 Harmonized m-RSimGP (hm-RSimGP)

The m-RSimGP algorithm [12] incorporates semantic information of multimodal data into the m-SimGP model, where the inter-modal semantic relation is used as a smooth prior over the latent space to maximize the cross-modal semantic consistency. By incorporating the harmonization constraint Eq. (7) into the m-RSimGP model, we arrive at the following minimization problem:

$$\begin{aligned} \arg\min_{X,\theta} & \sum_c \mathcal{L}_c^s + \mu_c \mathcal{H}_c(K_c - S^x) \\ & + \lambda_1 \sum\nolimits_{(o_i,o_j)\in\mathscr{S}} \|x_i - x_j\|^2 \\ & + \lambda_2 \sum\nolimits_{(o_i,o_j)\in\mathscr{D}} \max\left(0, 1 - \|x_i - x_j\|^2\right), \end{aligned} \tag{10}$$

where $\mathcal{L}_c^s, c \in \{1, 2\}$, is the negative log-likelihood associated with Eq. (5). The data object $o_i = \{y_i^1, y_i^2\}$ is represented by the point x_i in the low dimensional latent space, where $i = 1, 2, \ldots, N$. $\mathscr{S} = \{(o_i, o_j)\}$ denotes the set of pairs with similar semantics, and $\mathscr{D} = \{(o_i, o_j)\}$ denotes the set of pairs with dissimilar semantics. μ_1 and μ_2 are the tradeoff parameters for the harmonization regularization terms. λ_1 and λ_2 are the tradeoff parameters for the cross-modal semantic regularization terms. The new model is denoted as hm-RSimGP.

3.3. Optimization and inference

The problems to be solved in Section 3.2 are highly nonlinear functions of the latent variable X and kernel parameters θ, and there are no closed form solutions. Note that the regularization terms, $\mathcal{H}_c(K_c - S^x)$, are chosen to be convex sub-differentiable functions, and the log-likelihood functions, *e.g.*, $\mathcal{L}_1$ and $\mathcal{L}_2$ in Eq. (3), are differentiable as long as the gradients of kernel functions can be computed with respect to the model parameters. Therefore, we can use a non-linear gradient-based optimizer such as scaled conjugate gradients [25] to obtain the low dimensional embedding for multimodal data. As described in [10], the optimization process for the latent variable X and the kernel parameters is accelerated through sparsification of the model, *i.e.*, optimization in an active subset of M points ($M \ll N$) selected from all points in the dataset. In our model, we use an active block matrix selection strategy on the kernel

matrices in the harmonization constraints, which is an improved optimization scheme of fast GPLVM [10] by reducing the gradient computation complexity on the harmonization constraints. For learning, the dominant complexity of our methods reduces to $O(NM^2)$. Consequently, our methods are efficient in processing large training datasets.

Once we have learned the Gaussian processes on the training multimodal data, the inference procedure is straightforward. Given a new observed point (*e.g.*, a test image y_t^1), we obtain the corresponding latent representation x_t by maximizing the posterior probability $p(x_t|y_t^1)$. Then we can perform cross-modal retrieval to discover the non-linear correlations among latent representations of multimodal observations. Specifically, given an image query, retrieval from the other modality (*e.g.*, text) is accomplished by ranking the retrieved data according to the distance measured in the shared latent space.

4. Experiments

In this section, we conduct experimental evaluation on three datasets for multimodal learning to demonstrate the advantages of our harmonized models.

4.1. Datasets

Cross-modal retrieval is a typical multimodal application that requires a common representation of heterogeneous data objects. The experiments are performed on three publicly available image/text datasets, *i.e.*, PASCAL Sentence [26], Wikipedia [1], and TVGraz [27].

PASCAL Sentence [26] is probably the first dataset aligning images with captions. The dataset contains a total of 1000 images collected from 20 categories of PASCAL 2008. For each of the categories, 50 images are randomly selected. Each image is annotated with 5 sentences via Amazon Mechanical Turk. We use the same feature representation as in [28]. After SIFT features are extracted, each image is represented as an 1024-dim feature vector with bag-of-visual-words (BoVW) model. The text representation is based on Latent Dirichlet Allocation (LDA) model with 100 topics. A random $70\%/30\%$ split of the dataset is used for training/testing.

Wikipedia [1] is a widely used benchmark for cross-modal retrieval. It consists of 2,866 image-text pairs which are collected from Wikipedia articles. 2,173 pairs are randomly chosen for training and the remaining 693 pairs are used for testing. Each image is represented by an 128-dim BoVW feature with SIFT descriptors, and each text is represented by a 10-dim LDA feature. All of the image-text documents cover 10 semantic categories, and each document is categorized as one of them.

TVGraz [27] contains 2,058 image-text pairs from 10 visual object categories of the Caltech-256 dataset. It is collected from webpages retrieved by Google image search with keywords of the 10 categories. We use the same data provided by [28], where each image is represented by an 1024-dim BoVW vector based on SIFT, and the text is represented by the 100-dim LDA feature. The dataset is randomly divided into a training set of 1,558 document pairs and a test set of 500 document pairs.

For parameter tuning, we further randomly choose 30% of all the training subsets of the three datasets as the validation sets in subsequent experiments on parameter sensitivity analysis, and the remaining 70% data pairs in the training subsets are used as the training data in the parameter validation processes.

4.2. Cross-modal retrieval

We evaluate the performance of our methods for two cross-modal retrieval tasks, *i.e.*, image retrieval with text query and text retrieval with image query.

4.2.1 Experimental settings

Our models, hmGPLVM, hm-SimGP and hm-RSimGP, are compared with mGPLVM [11], m-SimGP [12] and m-RSimGP [12], respectively. In the experiments, log-likelihood functions of the two baseline models, mGPLVM and m-SimGP, are penalized by a Gaussian prior on the latent space X, *i.e.*, $p(X) = \prod_{n=1}^{N} N(x_n|0, \mathrm{I})$. The m-RSimGP method imposes inter-modal relations (*i.e.*, semantic similarity and dissimilarity) as smooth priors over the latent space X. For all these baselines, there are no informative prior over the kernel hyperparameters θ.

Notice that our algorithms work for any convex loss function. In the experiments, we use the popular Frobenius norm to define the harmonization constraint, *i.e.*, $\mathcal{H}_c(K_c - S^x) = \|K_c - S^x\|_F^2$, $c \in \{1, 2\}$. The choice of kernel functions is also arbitrary in our algorithms. For simplicity, we use an exponential kernel (RBF) to define the non-linear covariance matrices K_1 and K_2. Gaussian kernel is used to compute the similarity S^x on the latent representation.

We also present the performances of several state-of-the-art approaches for multimodal learning. The probabilistic model MLBE [6] uses binary hash codes as latent variables to generate intra-modal and inter-modal similarities. DCCAE [9] is a DNN-based multimodal feature learning algorithm which combines CCA-based [30] and autoencoder-based [31] terms. LGCFL [29] is a supervised cross-modal matching approach, which utilizes class labels to learn consistent feature representations from heterogeneous modalities, and introduces a local group-based priori for better utilizing block-based image features.

In all experiments, we use a consistent setting of the parameters. The tradeoff parameters μ_1 and μ_2 are assigned with the same value, indicating equal importance of two data modalities. We use CCA to initialize the shared latent space with the nearly optimal latent feature dimension. The retrieval performance is measured with mean average preci-

Methods	PASCAL			Wiki			TVGraz		
	Image query	Text query	Average	Image query	Text query	Average	Image query	Text query	Average
MLBE [6]	0.2543	0.2215	0.2379	0.3787	0.4109	0.3948	0.3468	0.3849	0.3659
DCCAE [9]	0.1988	0.1670	0.1829	0.2542	0.1916	0.2229	0.3879	0.3736	0.3808
LGCFL [29]	0.2570	0.2379	0.2475	0.2736	0.2241	0.2489	0.4366	0.4140	0.4253
mGPLVM [11]	0.1507	0.1318	0.1413	0.2054	0.1628	0.1841	0.2645	0.2784	0.2715
hmGPLVM	0.1755	0.1471	0.1613	0.2392	0.1826	0.2109	0.3572	0.3227	0.3400
m-SimGP [12]	0.2761	0.2724	0.2743	0.4336	0.4188	0.4262	0.4467	0.4453	0.4460
hm-SimGP	0.2993	0.3074	0.3034	0.4557	0.4391	0.4474	0.4647	0.4633	0.4640
m-RSimGP [12]	0.3301	0.3275	0.3288	0.4697	0.4418	0.4558	0.5102	0.5079	0.5091
hm-RSimGP	**0.3538**	**0.3514**	**0.3526**	**0.4861**	**0.4791**	**0.4826**	**0.5435**	**0.5351**	**0.5393**

Table 1. The mAP comparison for cross-modal retrieval task on three datasets.

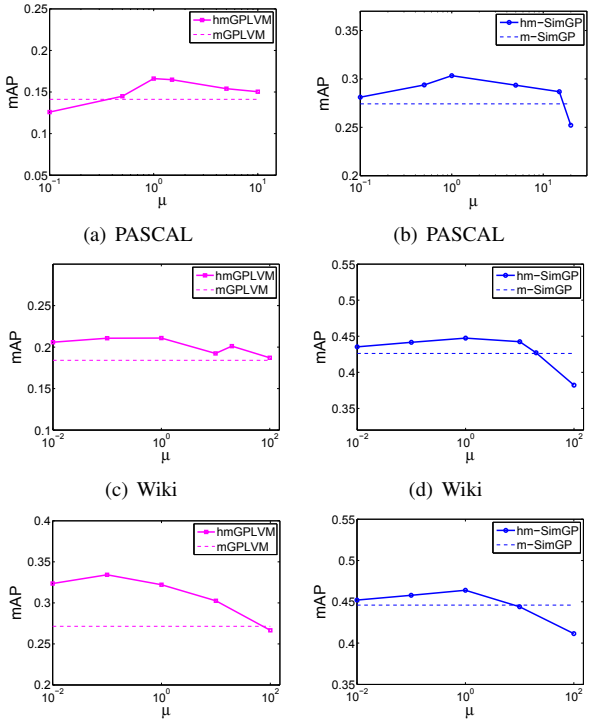
sion (mAP) [32], *i.e.*, average precision at the ranks where recall changes.

4.2.2 Experimental results

Table 1 summarizes the experimental results on all the datasets. We see that the proposed harmonization mechanism is well adapted to multimodal GPLVM framework. Our method hmGPLVM achieves significant improvement over the baseline mGPLVM in all cases, which indicates that the joint prior is powerful in enhancing the consistency of the latent representation for multimodal data.

It is clear from Table 1 that the harmonization constraint is still a great boost to GPLVM modeling with similarity outputs. Our harmonization mechanism is able to enhance the performance of the original similarity-based GPLVMs, which significantly outperform the rest of the methods, *e.g.*, the probabilistic MLBE, the DNN-based DCCAE, the supervised LGCFL. Specifically, when combined with cross-modal semantic constraint in the m-RSimGP model, another kind of priors over the latent space, our method gain further improvements on the ability of preserving inter-modal semantic relations for multimodal GP models. On the whole, we can conclude that the proposed harmonization constraint has strong generalization ability as a joint prior over the model parameters for multimodal GPLVM learning.

Some examples of cross-modal retrieval on Wiki dataset are shown in Figure 6. A retrieved result is considered correct if it belongs to the same class as the query [1]. We use a textual query from the "biology" class as shown on the left of Figure 6. As can be seen, all the top retrieved results by hm-SimGP are from the "biology" category, the same as the query text, while some of the top retrieved images by other methods are from different categories. For example, the 4th result of hmGPLVM is incorrect while the first retrieved result of mGPLVM and the second of mSimGP are incorrect. Therefore, we can see that the harmonized multimodal GPLVM models achieve better cross-modal retrieval performance, especially on the top retrieved documents.



(a) PASCAL (b) PASCAL

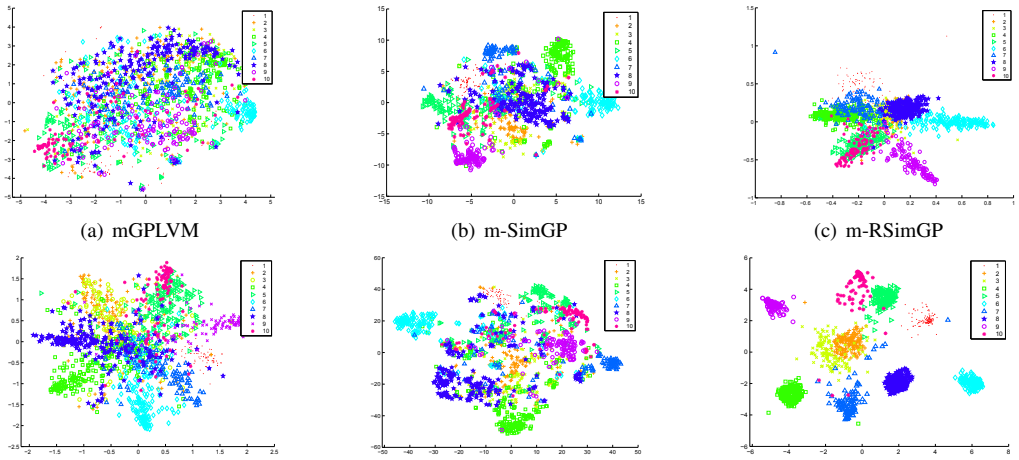
(c) Wiki (d) Wiki

(e) TVGraz (f) TVGraz

Figure 3. Sensitivity test on the tradeoff parameters w.r.t. the performance of cross-modal retrieval where $\mu_1 = \mu_2 = \mu$.

4.3. Parameter sensitivity analysis

The harmonization parameters μ_1 and μ_2 control the extent to which we enforce the agreement between the GP kernel matrices (K_1 and K_2) and the latent similarity matrix (S^x). In our experiments, they are assigned with the same value, *i.e.*, $\mu_1 = \mu_2 = \mu$, indicating equal importance of the observation modalities. We conduct sensitivity analysis on them to test how they impact on the cross-modal correlation learning performance. Figure 3 shows the curves of average mAP scores of image-to-text and text-to-image retrieval with different setting on the tradeoff parameter μ.



(a) mGPLVM (b) m-SimGP (c) m-RSimGP

(d) hmGPLVM (e) hm-SimGP (f) hm-RSimGP

Figure 4. Visualization of the discovered latent representation on the TVGraz dataset (Better viewed in color).

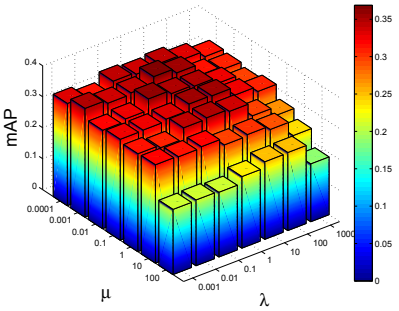


Figure 5. Sensitivity test on the tradeoff parameters in hm-RSimGP *w.r.t.* the performance of image-text retrieval: an example on the PASCAL dataset.

As seen in Figure 3, the average mAP is improved as μ is increased and achieves the best performance around 1 for most cases. However, the further increase of μ leads to a drop on the retrieval performance. For larger μ, the performance of the harmonized methods is even worse than those of the baselines without the harmonization mechanism (as shown in the dashed line in all the subfigures in Figure 3). These phenomenons are possibly due to the fact that a very large μ will increase the risk of co-adaptation and cause the model to become trapped in a local minima. For consistency, we set $\mu_1 = \mu_2 = 1$ for hmGPLVM and hm-SimGP in our experiments.

We also conduct sensitivity analysis on the tradeoff parameters in hm-RSimGP (Eq. (10)) to evaluate how the harmonization constraint affects multimodal GPLVM learning with other kinds of latent priors. The cross-modal retrieval performance is tested on the PASCAL dataset. In the experiments, both similar and dissimilar semantic information are used in the hm-RSimGP model, and the tradeoff parameters λ_1 and λ_2 are also assigned with the same value, *i.e.*, $\lambda_1 = \lambda_2 = \lambda$. Seen from Figure 5, our hm-RSimGP can achieve consistently good performance as long as the value of μ is not too large. Specifically, for all the given λs, the performance decreases significantly when the value of μ is larger than 10. However, our hm-RSimGP performs much better when the values of the parameters μ and λ are limited to $[0.001, 0.1]$ and $[0.1, 10]$, respectively. Overall, the proposed harmonization constraint can improve the performance of the m-RSimGP model with cross-modal (dis)similarity constraint. In our experiments, we fix the tradeoff parameters of hm-RSimGP and set $\mu = 0.1$ and $\lambda = 1$ for all the datasets.

4.4. Latent space visualization

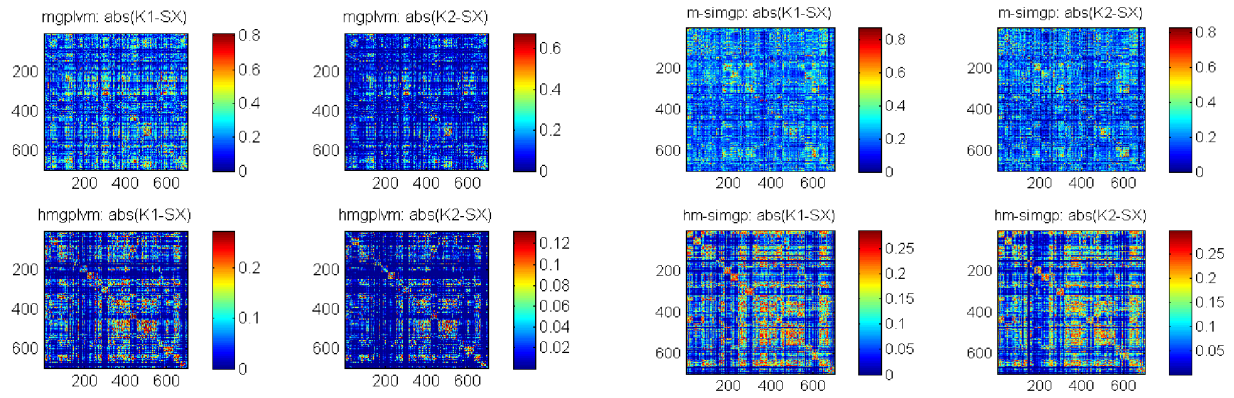
We visualize the discovered latent space to evaluate the learning quality. The experiment is performed on the TVGraz dataset with 10 categories. For visualization, the 10-dim latent representations are embedded into a 2-dim space using the t-SNE algorithm [33]. As shown in Figure 4, our harmonized methods perform much better in producing a low dimensional embedding compared to the original GPLVM-based methods. For example, the latent representations discovered by mGPLVM provide little information on the category structure of the data objects. In contrast, the latent representations discovered by our hmGPLVM exhibit a more clear grouping pattern for the data from the same category. Therefore, our harmonized GPLVMs can learn a more discriminative latent space from multimodal data.

4.5. Analysis of the harmonization mechanism

The multimodal GPLVMs learn a shared latent representation for multimodal data to bridge heterogeneous modalities. In order to preserve structure and propagate the semantic information among different modalities, similar representations for correlated data pairs should be guaranteed. The harmonization constraint in Eq. (7) forces the divergence between the similarities in different modalities and

(a) Text query (b) mGPLVM (c) hmGPLVM (d) m-SimGP (e) hm-SimGP

Figure 6. Text-to-image retrieval on Wiki. Here we present top four retrieved images. Red rectangle indicates a false positive example.



(a) mGPLVM *v.s.* hmGPLVM (b) m-SimGP *v.s.* hm-SimGP

Figure 7. Visualization of the absolute element-wise difference between modality-specific GP kernels (K_1, K_2) and similarity matrix of the latent representations (S^x) on PASCAL (Better viewed in color).

F-Norm / Method	$K_1 - S^x$	$K_2 - S^x$	Total
mGPLVM	173.371	104.595	277.966
hmGPLVM	66.559	25.124	91.683
m-SimGP	181.413	164.016	345.439
hm-SimGP	114.880	109.031	223.911

Table 2. The Frobenius-norm of the difference between modality-specific GP kernels (K_1, K_2) and the similarity matrix of the latent representations (S^x) on PASCAL.

the latent space to be small. With Frobenious-norm, the minimization of the loss function forces them to be element-wise "closer" to each other, and thus the structure consistency is achieved by a gradual "resonance" effect between the two during the model learning process. Here we show some qualitative and quantitative comparisons between different models. Figure 7 shows the difference between modality-specific GP kernels (K_1, K_2) and the similarity matrix (S^x) of the latent representations on PASCAL data. It is obvious that the divergences between the two components of hmGPLVM and hm-SimGP are much smaller than those of mGPLVM and m-SimGP, which is also validated by the quantitative results in Table 2. The results show that the difference between GP kernels and the similarity of latent points is reduced in our proposed models with harmonization, and thus better structure consistency is achieved among GP kernels and the similarity in the latent space.

5. Conclusion

We have introduced a harmonization constraint as a joint prior over the model parameters for multimodal GPLVMs. Three harmonized extensions of multimodal GPLVMs, *i.e.*, hmGPLVM, hm-SimGP and hm-RSimGP, have been proposed for multimodal correlation learning. Compared to existing models, we build the additional information transfer pathway on the model parameter space, so that the intra-modal and inter-modal information can be more sufficiently transferred among the kernel hyperparameters via the shared latent space. In return, a more semantically consistent latent representation can be obtained with better multimodal topology preservation. In future work, we will investigate more complex and flexible prior distributions of the model parameters in multimodal GPLVMs for multimodal correlation learning.

Acknowledgement This work was supported in part by the National Natural Science Foundation of China: 61672497, 61332016, 61620106009, 61650202, U1636214, and 61429201, National Basic Research Program of China (973 Program): 2015CB351802, and the Key Research Program of Frontier Sciences of CAS under Grant QYZDJ-SSW-SYS013. This work was also supported in part to Dr. Qi Tian by ARO under Grant W911NF-15-1-0290, and Faculty Research Gift Awards by NEC Laboratories of America and Blippar.

References

[1] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM Multimedia*, pages 251–260, 2010.

[2] Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *CVPR*, pages 3441–3450, 2015.

[3] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *ICML*, pages 689–696, 2011.

[4] Abhishek Sharma, Abhishek Kumar, Hal Daumé III, and David W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*, pages 2160–2167, 2012.

[5] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 106(2):210–233, 2014.

[6] Yi Zhen and Dit-Yan Yeung. A probabilistic model for multimodal hash function learning. In *ACM KDD*, pages 940–948, 2012.

[7] Jianfeng He, Bingpeng Ma, Shuhui Wang, Yugui Liu, and Qingming Huang. Cross-modal retrieval by real label partial least squares. In *ACM Multimedia*, pages 227–231, 2016.

[8] Harold Hotelling. Relations between two sets of variates. *Biometrika*, pages 321–377, 1936.

[9] Weiran Wang, Raman Arora, Karen Livescu, and Jeff A. Bilmes. On deep multi-view representation learning. In *ICML*, pages 1083–1092, 2015.

[10] Neil D. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *JMLR*, 6:1783–1816, 2005.

[11] Aaron Shon, Keith Grochow, Aaron Hertzmann, and Rajesh P Rao. Learning shared latent structure for image synthesis and robotic imitation. In *NIPS*, pages 1233–1240, 2005.

[12] Guoli Song, Shuhui Wang, Qingming Huang, and Qi Tian. Similarity gaussian process latent variable model for multimodal data analysis. In *ICCV*, pages 4050–4058, 2015.

[13] Carl Henrik Ek, Jon Rihan, Philip HS Torr, Grégory Rogez, and Neil D. Lawrence. Ambiguity modeling in latent spaces. In *MLMI*, pages 62–73. Springer, 2008.

[14] Andreas C. Damianou, Carl Henrik Ek, Michalis K. Titsias, and Neil D. Lawrence. Manifold relevance determination. In *ICML*, 2012.

[15] Stefanos Eleftheriadis, Ognjen Rudovic, and Maja Pantic. Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *IEEE TIP*, 24 (1):189–204, 2015.

[16] Raquel Urtasun, David J. Fleet, and Pascal Fua. 3d people tracking with gaussian process dynamical models. In *CVPR*, pages 238–245, 2006.

[17] Jihun Ham, Daniel D. Lee, and Lawrence K. Saul. Semisupervised alignment of manifolds. In *AISTAT*, 2005.

[18] Carl Henrik Ek, Philip H. S. Torr, and Neil D. Lawrence. Gaussian process latent variable models for human pose estimation. In *MLMI*, pages 132–143, 2007.

[19] Guoli Song, Shuhui Wang, Qingming Huang, and Qi Tian. Multimodal similarity gaussian process latent variable model. *IEEE TIP*, 26(9):4168–4181, 2017.

[20] Stefanos Eleftheriadis, Ognjen Rudovic, and Maja Pantic. Multi-conditional latent variable model for joint facial action unit detection. In *ICCV*, pages 3792–3800, 2015.

[21] Andrew Gordon Wilson and Ryan Prescott Adams. Gaussian process kernels for pattern discovery and extrapolation. In *ICML*, pages 1067–1075, 2013.

[22] Michalis K. Titsias and Miguel Lázaro-Gredilla. Variational inference for mahalanobis distance metrics in gaussian process regression. In *NIPS*, pages 279–287, 2013.

[23] Andreas C. Damianou, Michalis K. Titsias, and Neil D. Lawrence. Variational inference for latent variables and uncertain inputs in gaussian processes. *JMLR*, 2, 2015.

[24] Zhenwen Dai, Andreas Damianou, Javier González, and Neil Lawrence. Variational auto-encoded deep Gaussian processes. *ICLR*, 2016.

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

[26] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *NAACL HLT Workshop*, pages 139–147, 2010.

[27] Inayatullah Khan, Amir Saffari, and Horst Bischof. Tvgraz: Multi-modal learning of object categories by combining textual and visual features. In *AAPR Workshop*, pages 213–224, 2009.

[28] Jose Costa Pereira and Nuno Vasconcelos. On the regularization of image semantics by modal expansion. In *CVPR*, pages 3093–3099, 2012.

[29] Cuicui Kang, Shiming Xiang, Shengcai Liao, Changsheng Xu, and Chunhong Pan. Learning consistent feature representation for cross-modal multimedia retrieval. *IEEE Trans. Multimedia*, 17(3):370–381, 2015.

[30] Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML*, pages 1247–1255, 2013.

[31] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *ICML*, pages 689–696, 2011.

[32] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[33] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(2579-2605):85, 2008.