Journal of Machine Learning Research 24 (2023) 1-32 Submitted 2/19; Revised 7/23; Published 10/23

Multi-view Collaborative Gaussian Process Dynamical Systems

Shiliang Sun SLSUN@CS.ECNU.EDU.CN
School of Computer Science and Technology,
East China Normal University, Shanghai 200062, P. R. China
Department of Automation,
Shanghai Jiao Tong University, Shanghai 200240, P. R. China

Jingjing Fei JINGJINGFEI16@163.COM

Jing Zhao JZHAO@CS.ECNU.EDU.CN

Liang Mao LMAO14@OUTLOOK.COM
School of Computer Science and Technology,
East China Normal University, Shanghai 200062, P. R. China

Editor: Massimiliano Pontil

Abstract

Gaussian process dynamical systems (GPDSs) have shown their effectiveness in many tasks of machine learning. However, when they address multi-view data, current GPDSs do not explicitly model the dependence between private and shared latent variables. Instead, they introduce structurally and intrinsically discrete segmentation in the latent space. In this paper, we propose the multi-view collaborative Gaussian process dynamical systems (McGPDSs) model, which assumes that the private latent variable for each view is controlled by its dynamical prior and the shared latent variable. The relevance between private and shared latent variables can be automatically learned by optimization in the Bayesian framework. The model is capable of learning an effective latent representation and generating novel data of one view given data of the other view. We evaluate our model on two-view data sets, and our model obtains better performance compared with the state-of-the-art multi-view GPDSs.

Keywords: Gaussian process, multi-view machine learning, dynamical system, variational inference, multi-output modeling

1. Introduction

A Gaussian process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution (Rasmussen and Williams, 2006). GPs are stochastic processes over real-valued functions and completely specified by mean functions and covariance functions (Rasmussen and Williams, 2006). Recently, GPs have been proved successful in various areas of machine learning (Lawrence and Jordan, 2005; Andreas and Carlos, 2007; Damianou et al., 2011; Lüthi et al., 2018; Feurer et al., 2018; Wei et al., 2019; Medina et al., 2019) because GPs can provide flexible function approximation. For example, to implement nonlinear dimensionality reduction, GP latent variables (GPLVMs) (Lawrence, 2004, 2005;

Titsias and Lawrence, 2010) have been presented, which use the global latent variables and assume the conditional independence among multiple outputs.

For modelling dynamics in sequential data, some Gaussian process dynamical systems (GPDSs) have been proposed, which extend GPLVMs by adding dynamical priors on the latent variables, such as GP dynamical models (GPDMs) (Wang et al., 2006), variational GPDSs (VGPDSs) (Damianou et al., 2011), variational dependent multi-output GPDSs (VDM-GPDSs) (Zhao and Sun, 2016) and collaborative Gaussian process dynamical systems (CGPDSs) (Zhao et al., 2018). Specifically, the GPDM models the dynamics by adding the Markov prior on the latent space and characterizes the variability among outputs via constructing the output variances with different parameters. The VGPDS employs the GP dynamical prior on the latent space, which is more flexible and can capture some specific dynamical information such as periodicity with specific kernels. The VDM-GPDS models the dependence among multiple outputs and employs convolution processes to capture the multi-output dependence explicitly. The VDM-GPDS obtains better performance than the GPDM and VGPDS, but the VDM-GPDS is time-consuming during training attributed to the introduced convolution processes. The CGPDS expresses each output as the sum of a global latent process and a local latent process, which can capture the universality and individuality of all outputs. Moreover, the CGPDS assumes that the latent processes are conditionally independent, which ensures the resulting evidence lower bound to be decomposed across dimensions and allows the stochastic optimization. We will detail CGPDSs in Section 2 in a self-contained form.

With the rapid development of information techniques, more and more data exhibit multi-view characteristics such as the URL link and text in a web document, the audio and image frames of a video, the surrounding words and image of a web image and so on. Data of different modalities often offer complementary information, and multi-view learning can exploit this information to learn representations, which are more comprehensive and expressive than that of single-view learning (Sun et al., 2019). More specifically, multi-view learning uses one function to model a view and optimizes all functions together during training. Consensus and complementarity are the two core principles of multi-view learning. The consensus principle maximizes the agreement on the representations of different views, and the complementarity principle exploits the complementary information contained in different views to represent multi-view data comprehensively (Li et al., 2018). Since multi-view learning can use the consensus and complementarity properties of multiple views and exploit the redundant views of the same input data, multi-view learning is often more natural and effective than single-view learning (Sun, 2013; Xu et al., 2013; Li et al., 2018; Ding et al., 2018).

Recently, several models extended GPLVMs or GPDSs to the scenario of multi-view learning. The shared GPLVM assumed that each view has been generated from the same low-dimensional latent variable corrupted by additive Gaussian noise (Shon et al., 2006). Furthermore, a new version of the shared GPLVM, i.e., the subspace GPLVM, was proposed (Ek and Lawrence, 2009), in which the latent space for each view is factorized into a shared one, which captures the shared information across the views, and a private one, which explains the remaining variance. Salzmann et al. (2010) learned the dimensionality of the factorization by introducing regularizers. The manifold relevance determination (MRD) (Damianou et al., 2012) improved the "hard" segmentation between the private and shared

latent variables and employed the "soft" segmentation in the latent space. Concretely, the MRD employed learned scales in the automatic relevance determination (ARD) kernels and a pre-given threshold to determine whether a dimension is the private or shared latent variable. This threshold requires to be specified manually and often varies for different datasets, whose configuration thus needs expert knowledge and is time-consuming.

The above models do not explicitly model the correlation between private and shared latent variables (dimensions). This kind of model assumption in the latent space brings the structurally and intrinsically discrete segmentation between the shared and private latent variables. On many real-world data sets, it is quite difficult to clearly divide the latent space which generates multi-view observations into shared and private latent information because private and shared latent information is complexly coupled and interacts with each other. For example, in the multi-view data set which contains pictures of different faces under the same lighting condition, we can take the characteristics of faces as private information and the lighting condition as shared information. It is not difficult to figure out that intensely bright lighting conditions can affect the characteristics of the face, such as the color of the skin.

In this paper, we propose the multi-view collaborative Gaussian process dynamical systems (McGPDSs) model, which makes full use of the characteristics of multi-view data and the advantages of the CGPDSs. The proposed model relaxes the discrete structural segmentation in the latent space and automatically learns the relevance between private and shared latent variables through optimization. Since private latent variables are determined by their dynamical priors and the shared latent variable, McGPDSs can model more complex and abundant information of data. Experiments on the synthetic and real-world data sets also validate the superiority of our proposed McGPDSs.

The contributions of our model are summarized as follows: 1) Our model extends the CGPDS into multi-view learning, which possesses the advantages of the multi-view learning and the CGPDS to model high-dimensional multi-output data. 2) Our model explicitly models the relationship between shared and private latent variables and automatically learns their relevance. 3) All parameters in our model can be learned through optimization.

The remainder of the paper is structured as follows. Section 2 introduces the related work including multi-view learning, CGPDSs and several multi-view models based on GPLVMs and GPDSs. Section 3 presents the proposed model in detail. Section 4 describes the inference and learning techniques. Section 5 illustrates the procedure of prediction with McGPDSs. Section 6 provides extensive experimental evaluations to validate the effectiveness of our model, and Section 7 concludes the work and discusses future work.

2. Related Work

In this section, we first briefly review the related works on multi-view learning. Then we give an introduction to CGPDSs (Zhao et al., 2018) and several multi-view models based on GPLVMs and GPDSs (Shon et al., 2006; Ek and Lawrence, 2009; Damianou et al., 2012).

2.1 Multi-view Learning

Multi-view learning is concerned with learning from data represented by multiple views. It has received increasing attention and been applied widely. Wei et al. (2018) evaluated the

quality of community-based question answering through transductive multi-view learning. Hu et al. (2018) proposed a shareable and individual multi-view metric learning approach for visual recognition. Puyol et al. (2018) described a method of regional multi-view learning for cardiac motion analysis, and the method was applied to the identification of dilated cardiomyopathy patients. Jing et al. (2018) employed low-rank multi-view embedding learning to predict the popularity of the micro video. Tulsiani et al. (2018) considered multi-view consistency as the supervisory signal for learning shape and pose prediction.

In the literature, multi-view learning is closely related to other machine learning methods, such as active learning, domain adaptation, and representation learning. More specifically, Muslea et al. (2002) combined co-testing and co-EM where co-testing is a novel method for active learning with multiple views and co-EM is used to generate classifiers and select the unlabeled points with the largest amount of information for labeling. Muslea et al. (2006) improved co-testing by considering differences between strong and weak views and assuming strong views with more information. Domain adaptation solves the problem of adapting a model trained on the source domain to the target domain, where the data from the source and target domains are largely different. Domain adaptation can be applied in the cross-language text classification task where documents in different languages represent different views. Co-training (Wan, 2009) and multi-view co-classification (Amini and Goutte, 2010) have been proposed and successfully applied in the task.

Multi-view representation learning has been a promising research topic in recent years on account of the ability to provide abundant and complementary information for learning representations. Multi-view representation learning methods contain generative methods including multi-modal topic learning (Cohn and Hofmann, 2001; Barnard et al., 2003; Blei and Jordan, 2003), multi-view sparse coding (Jia et al., 2010; Cao et al., 2013; Liu et al., 2014) and multi-view latent space Markov networks (Xing et al., 2012; Chen et al., 2010), and deep neural methods including multi-modal autoencoders (Ngiam et al., 2011; Feng et al., 2014; Wang et al., 2015), multi-model Boltzmann machines (Srivastava and Salakhutdinov, 2012) and multi-modal recurrent neural networks (Karpathy and Fei-Fei, 2015; Mao et al., 2014; Donahue et al., 2015).

2.2 CGPDS

CGPDSs aim to model multi-output sequential data. As a multi-output model, the CGPDS supposes that each output is the sum of a global latent process and a designed local latent process to capture dependence among multiple outputs and maintain the unique characteristics of each output. Since standard Bayesian inference is analytically intractable, CGPDSs adopt variational inference and introduce inducing points to learn the model. Moreover, the evidence lower bound can be decomposed regarding dimensions attributed to the conditional independence of outputs, which allows optimizing parameters in a stochastic optimization framework. Figure 1 shows the graphical model of CGPDSs.

Given a multi-output sequential data $Y \in \mathbb{R}^{N\times D}$ with $\mathbf{y}_n \in \mathbb{R}^D$ be the observation at time $t_n \in \mathbb{R}^+$, the CGPDS assumes that there are low-dimensional latent variables $X \in \mathbb{R}^{N\times Q}$ (with $Q \ll D$) that generate the observations. Moreover, a GP prior on the low-dimensional latent variables is used to model the dynamics, as in Damianou et al. (2011). Specifically, the CGPDS is defined as a four-layer GPDS through the following generative

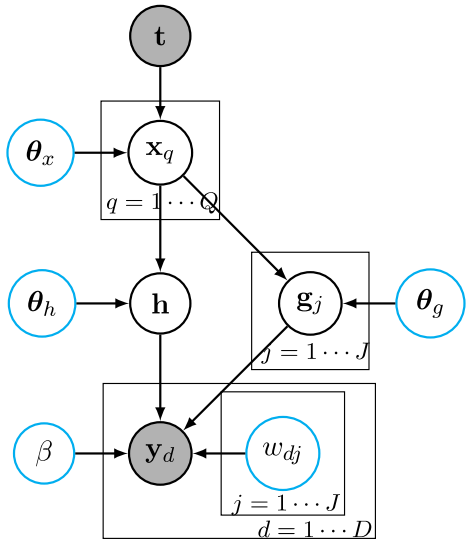


Figure 1: The graphical model for the CGPDSs. The gray solid circles represent observations. The black hollow circles represent latent variables. The cyan hollow circles represent parameters.

process:

$$p(X|\mathbf{t}) = \prod_{q=1}^{Q} N(\mathbf{x}_q|0, \mathbf{K}_{\mathbf{t},\mathbf{t}}),$$

where the $\mathbf{x}_q \in R^N$ is the qth row of X and $\mathbf{K}_{\mathbf{t},\mathbf{t}}$ is the covariance matrix computed by $\kappa(t, t')$.

$$\begin{aligned} p(\mathbf{h}|X) &= \mathcal{N}(\mathbf{h}|\mathbf{0}, \mathbf{H}_{X,X}), \\ p(\mathbf{g}|X) &= \prod_{j=1}^{J} \mathcal{N}(\mathbf{g}_j|\mathbf{0}, \mathbf{G}^j_{X,X}), \end{aligned}$$

where latent processes $\mathbf{h}$ and $\{\mathbf{g}_j\}_{j=1}^J$ are both GPs with input $\mathbf{x}$, and the $\mathbf{H}_{X,X}$ and $\mathbf{G}^j_{X,X}$ are covariance matrices computed by $\kappa_h(\mathbf{x}, \mathbf{x}')$ and $\kappa^j_g(\mathbf{x}, \mathbf{x}')$, respectively.

The CGPDS introduces latent processes $\mathbf{h}$ and $\{\mathbf{g}_j\}_{j=1}^J$, which is entirely different from the previous GPDSs such as the VGPDS and VDM-GPDS. The VGPDS uses a single GP mapping from X to F (the noise-free version of the output Y), which can only learn the common information among multiple outputs, but cannot learn the unique information of each output. The VDM-GPDS employs convolution processes to explicitly model the dependence among multiple outputs. In the VDM-GPDS, the mapping from X to F contains an $ND \times ND$ matrix, which increases the computational complexity of the model and prevents the model from scaling to large datasets. The CGPDS can capture the dependence and differences among multiple outputs with a relatively simple model structure.

$$p(\mathbf{y}_d|\mathbf{g}, \mathbf{h}) = \mathcal{N}(\mathbf{y}_d|\boldsymbol{\ell}_d + \mathbf{h}, \beta^{-1}I)$$

$$= \mathcal{N}(\mathbf{y}_d | \sum_{j=1}^{J} w_{dj}\mathbf{g}_j + \mathbf{h}, \beta^{-1}I), \tag{1}$$

where $\mathbf{h}$ is the global latent process which captures the dependence among outputs and $\boldsymbol{\ell}_d$ is the local latent process specific to the dth output which is constructed by latent processes $\{\mathbf{g}_j\}_{j=1}^{J}$ and weights $\{w_{dj}\}$. The weights $\{w_{dj}\}$ represent the local parameters which are different for D outputs. β is the inverse variance of the white Gaussian noise.

As shown in (1), the idea for constructing the output $\mathbf{y}_d$ is inspired by the COGP (Nguyen and Bonilla, 2014). The COGP models the dth output $\mathbf{y}_d$ as the weighted sum of the dth local latent process and J global latent processes, which contains $(J + D)$ GPs in total. The CGPDS uses a global latent process $\mathbf{h}$ and a local latent process $\boldsymbol{\ell}_d$ constructed by $J(J \ll D)$ latent processes $\{\mathbf{g}_j\}_{j=1}^{J}$, which includes $(J + 1)$ GPs. In a word, CGPDSs can not only capture the dependence among multiple outputs but also maintain the specific characteristics of each output with fewer parameters. Last but not the least, fewer parameters would make the model easier to learn.

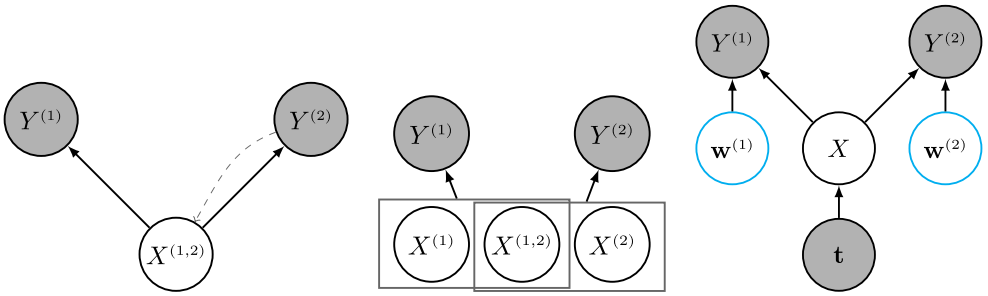
2.3 Multi-view Models Based on GPLVMs and GPDSs

In this section, we give the introduction of related multi-view models based on GPLVMs and GPDSs such as the shared GPLVM, subspace GPLVM and MRD.

The shared GPLVM assumes that all observations are generated from the same low-dimensional latent variable with additional Gaussian noise. Figure 2(a) shows the graphical model of the shared GPLVM. The dotted line represents the back-mapping from the output space, which can constrain the latent space. The assumption of sharing the same latent variable for all views is far from perfect for many datasets because this means data of all views share main generating parameters. Therefore, ideally, the shared latent variable can be used to connect all views and the private latent variables can be used to differentiate all views. The back-constraint from the second view to the latent space represents the bijective relationship between $Y^{(2)}$ and $X^{(1,2)}$. The back-constraint means that observation in the first view $Y^{(1)}$ has to be accommodated by throwing away information which does not exist in the second view $Y^{(2)}$. This model can also be considered as a feature selection model because it uses information from one view to determine what is important for the other view.

A new version of the shared GPLVM, that is, the subspace GPLVM, introduces the private latent variable for each view and a shared latent variable for all views. Figure 2(b) represents the graphical model of the subspace GPLVM. The subspace GPLVM learns a factorized latent representation within a single model. The model directly concatenates the private latent variable of each view with the shared latent variable, and then generates the data of each view. For inference, the subspace GPLVM seeks the maximum a posterior (MAP) solution for the latent space. The fact that the latent variables are not integrated out indicates that it is difficult to determine the structure of the latent space automatically. The idea of employing factorized latent space in the multi-view learning has been proposed in several works (Jia et al., 2011; Virtanen et al., 2012; Zhang et al., 2013).

The MRD can also learn a factorized latent representation and relax the previous "hard" discrete segmentation of latent space. Figure 2(c) shows the graphical model of the MRD with dynamics. A single latent variable X is used as the latent representation



(a) Shared GPLVM (b) Subspace GPLVM (c) MRD

Figure 2: Development of multi-view models based on GPLVMs and GPDSs. (a) shows the shared GPLVM where all the variances in the observations are shared in a single shared latent variable. (b) shows the subspace GPLVM which introduces private latent variables to express the variance in each view. (c) represents the MRD which uses a single latent variable and selects the shared and private latent dimensions according to the ARD weights $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ and a predetermined threshold. The shadowed nodes represent observations. The black hollow nodes represent latent variables. The cyan nodes represent parameters.

for all views where each dimension in X represents private or shared latent information. The MRD adopts variational inference with inducing points in order to integrate out the latent variable X. More precisely, the outputs of two view $Y^{(1)}$ and $Y^{(2)}$ are assumed to be independent GPs with the zero mean and an ARD covariance function, that is, $\kappa(\mathbf{x}_i, \mathbf{x}_j) = (\sigma_{ard})^2 \exp^{-\frac{1}{2}\sum_{q=1}^{Q} w_q (x_{i,q} - x_{j,q})^2}$. Two sets of ARD weights $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ in this model can be optimized in the Bayesian framework. An additional threshold δ is required to be specified manually for each dataset. By comparing ARD weights with the threshold, the MRD determines whether the dimension is private or shared and divides the latent space into three subspaces with $X = (X^{(1)}, X^s, X^{(2)})$. Here, X^s represents the shared subspace which consists of the set of dimensions $q \in [1, \cdots, Q]$ with $w_q^{(1)} > \delta$ and $w_q^{(2)} > \delta$. $X^{(1)}$ and $X^{(2)}$ are private latent subspaces of two views, respectively. $X^{(1)}$ is composed of the set of dimensions where $w_q^{(1)} > \delta$ and $w_q^{(2)} < \delta$ and analogously for $X^{(2)}$ ($w_q^{(1)} < \delta$ and $w_q^{(2)} > \delta$). There are two different versions of the MRD model, one with dynamic characteristics (with the GP prior on the latent variable) and one without dynamic characteristics.

3. Multi-view Collaborative Gaussian Process Dynamical System

In this section, we extend the CGPDS to the scenario of multi-view learning and propose the model of multi-view collaborative Gaussian process dynamical systems (McGPDSs). Figure 3 shows the graphical model of the McGPDS.

Specifically, we aim to model two views $Y^{(1)} \in \mathbb{R}^{N \times D_1}$ and $Y^{(2)} \in \mathbb{R}^{N \times D_2}$ in the same model where $\mathbf{y}_n^{(1)}$ and $\mathbf{y}_n^{(2)}$ are the observations at time $t_n \in \mathbb{R}^+$. We assume there is a shared low-dimensional latent variable $X^{(1,2)} \in \mathbb{R}^{N \times Q}$ which governs the generation of the private low-dimensional latent variables, that is, $X^{(1)} \in \mathbb{R}^{N \times Q}$ and $X^{(2)} \in \mathbb{R}^{N \times Q}$. The

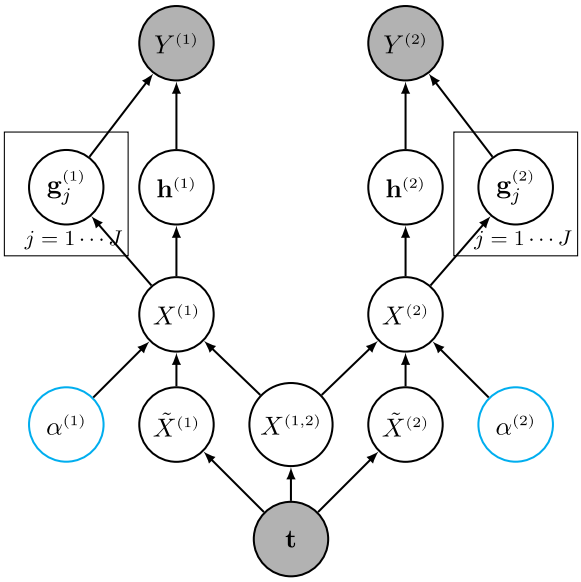


Figure 3: The graphical model for the McGPDS. The McGPDS explicitly models the dependence between private and shared latent variables and automatically learns the relevance between private and shared latent variables. The shadowed nodes represent observations. The black hollow nodes represent latent variables.

private low-dimensional latent variable for each view generates the corresponding observation. Moreover, we endow the GP prior on low-dimensional latent variables to model the dynamics. Here, N represents the number of training points. D_1 and D_2 represent the dimensions of two-view data, respectively. Q denotes the dimension of low-dimensional latent variables (with $Q \ll min(D_1, D_2)$). The superscript (1) and (2) corresponds to the first and second view, respectively. The superscript $(1, 2)$ means the shared information for two views.

Formally, the generative process is given as follows. The shared low-dimensional latent variable $X^{(1,2)}$ is assumed to be a multi-dimensional GP indexed by time t, that is

$$x_q^{(1,2)}(t) \sim \mathcal{G}P(0, \kappa_x^{(1,2)}(t, t')), q = 1, \ldots, Q, \tag{2}$$

where dimensions of the shared latent function $\mathbf{x}^{(1,2)}(t)$ are independently drawn from a GP with the covariance function $\kappa_x^{(1,2)}(t, t')$ with parameters $\boldsymbol{\theta}_x^{(1,2)}$. Since the latent variable $X^{(1,2)}$ is conditionally independent given $\mathbf{t}$, we have

$$p(X^{(1,2)}|\mathbf{t}) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{x}_q^{(1,2)}|\mathbf{0}, \mathbf{K}_{\mathbf{t},\mathbf{t}}^{(1,2)}), \tag{3}$$

where $\mathbf{K}_{\mathbf{t},\mathbf{t}}^{(1,2)}$ is the covariance matrix computed by kernel $\kappa_x^{(1,2)}(t, t')$. We also introduce two latent variables $\tilde{X}^{(1)}$ and $\tilde{X}^{(2)}$ which follow view-specific dynamic priors, i.e.,

$$p(\tilde{X}^{(1)}|\mathbf{t}) = \prod_{q=1}^{Q} \mathcal{N}(\tilde{\mathbf{x}}_q^{(1)}|\mathbf{0}, \mathbf{K}_{\mathbf{t},\mathbf{t}}^{(1)}), \tag{4}$$

$$p(\tilde{X}^{(2)}|\mathbf{t}) = \prod_{q=1}^{Q} \mathcal{N}(\tilde{\mathbf{x}}_q^{(2)}|\mathbf{0}, \mathbf{K}_{\mathbf{t},\mathbf{t}}^{(2)}), \tag{5}$$

where $\tilde{X}^{(1)}$ and $\tilde{X}^{(2)}$ are also assumed to be conditionally independent, and $\mathbf{K}_{\mathbf{t},\mathbf{t}}^{(1)}$ and $\mathbf{K}_{\mathbf{t},\mathbf{t}}^{(2)}$ are covariance matrices computed by kernels $\kappa_x^{(1)}(t,t')$ and $\kappa_x^{(2)}(t,t')$, respectively.

Let $\hat{X}^{(1)}$ be a noisy version of the shared latent variable $X^{(1,2)}$, i.e., $\hat{X}^{(1)} \sim N(\hat{X}^{(1)}|X^{(1,2)}, \epsilon^{(1)})$. The private latent variable $X^{(1)}$ is defined as a convex combination of the view-specific latent variable $\tilde{X}^{(1)}$ and $\hat{X}^{(1)}$, i.e., $X^{(1)} = (1-\alpha^{(1)})\hat{X}^{(1)} + \alpha^{(1)}\tilde{X}^{(1)}$, with the combination weight $\alpha^{(1)} \in [0,1]$ which can adjust the importance of the two combination components. The model can automatically learn the dependence between private and shared latent variables by optimizing $\alpha^{(1)}$. After integrating out $\hat{X}^{(1)}$, the conditional distribution of $X^{(1)}$ given $X^{(1,2)}$ and $\mathbf{t}$ is

$$p(X^{(1)}|X^{(1,2)}, \mathbf{t}) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{x}_q^{(1)}|(1-\alpha^{(1)})\mathbf{x}_q^{(1,2)}, (\alpha^{(1)})^2\mathbf{K}_{\mathbf{t},\mathbf{t}}^{(1)} + (1-\alpha^{(1)})^2\epsilon^{(1)}),$$

Similarly we define the private latent variable $X^{(2)}$, with

$$p(X^{(2)}|X^{(1,2)}, \mathbf{t}) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{x}_q^{(2)}|(1-\alpha^{(2)})\mathbf{x}_q^{(1,2)}, (\alpha^{(2)})^2\mathbf{K}_{\mathbf{t},\mathbf{t}}^{(2)} + (1-\alpha^{(2)})^2\epsilon^{(2)}),$$

where $\alpha^{(2)} \in [0,1]$ and $\epsilon^{(2)}$ denotes the variance of the Gaussian noise in the second view.

The setting of latent space in our model is largely different from the previous multi-view models based on GPLVMs and GPDSs, such as the shared GPLVM, subspace GPLVM and MRD. The shared GPLVM employs a single shared latent variable for all views and all variances in the observations are shared, where the private information cannot be modeled. The subspace GPLVM introduces a factorized latent space where each view is connected with an additional private latent space. The model employs MAP estimates so that the structure of the latent space cannot be automatically determined. The MRD model also employs a single latent space and determines whether a dimension is private or shared according to the weights in the ARD covariance functions and the artificially specified threshold. All the above models either use a single latent variable or do not explicitly model the relationship between private and shared latent variables (dimensions). Our model explicitly models the relevance between shared and private latent space. The relevance of the private and shared latent variables can be automatically learned by optimizing the weights $\alpha^{(1)}$ and $\alpha^{(2)}$.

The mapping from $X^{(1)}$ to $Y^{(1)}$ ($X^{(1)} \mapsto Y^{(1)}$) and the mapping from $X^{(2)}$ to $Y^{(2)}$ ($X^{(2)} \mapsto Y^{(2)}$) in the McGPDS employ the same idea as the mapping from X to Y ($X \mapsto Y$) in the CGPDS (Zhao et al., 2018). Additionally, attributed to conditional independence assumption, the distributions of the outputs can be written as the product of D terms, that is,

$$p(Y^{(1)}|X^{(1)}) = \prod_{d=1}^{D} \mathcal{N}(\mathbf{y}_d^{(1)}|\sum_{j=1}^{J} w_{dj}\mathbf{g}_j^{(1)}(X^{(1)}) + \mathbf{h}^{(1)}(X^{(1)}), (\beta^{(1)})^{-1}),$$
$$p(Y^{(2)}|X^{(2)}) = \prod_{d=1}^{D} \mathcal{N}(\mathbf{y}_d^{(2)}|\sum_{j=1}^{J} w_{dj}\mathbf{g}_j^{(2)}(X^{(2)}) + \mathbf{h}^{(2)}(X^{(2)}), (\beta^{(2)})^{-1}),$$

where $\beta^{(1)}$ and $\beta^{(2)}$ are the inverse variance of the white Gaussian noise. The latent processes $\mathbf{h}^{(1)}$ and $\{\mathbf{g}_j^{(1)}\}_{j=1}^J$ are GPs indexed by input $X^{(1)}$. Similarly, latent processes $\mathbf{h}^{(2)}$ and $\{\mathbf{g}_j^{(2)}\}_{j=1}^J$ are GPs indexed by input $X^{(2)}$, and we have

$$h^{(1)}(\mathbf{x}^{(1)}) \sim \mathcal{G}P(0, \kappa_h^{(1)}(\mathbf{x}^{(1)}, \mathbf{x}^{(1)'})), \qquad h^{(2)}(\mathbf{x}^{(2)}) \sim \mathcal{G}P(0, \kappa_h^{(2)}(\mathbf{x}^{(2)}, \mathbf{x}^{(2)'})),$$

$$g_j^{(1)}(\mathbf{x}^{(1)}) \sim \mathcal{G}P(0, \kappa_g^{(1)j}(\mathbf{x}^{(1)}, \mathbf{x}^{(1)'})), \qquad g_j^{(2)}(\mathbf{x}^{(2)}) \sim \mathcal{G}P(0, \kappa_g^{(2)j}(\mathbf{x}^{(2)}, \mathbf{x}^{(2)'})),$$

where the kernels $\kappa_h^{(1)}(\mathbf{x}^{(1)}, \mathbf{x}^{(1)'})$ and $\kappa_g^{(1)j}(\mathbf{x}^{(1)}, \mathbf{x}^{(1)'})$ are parameterized by $\boldsymbol{\theta}_h^{(1)}$ and $\boldsymbol{\theta}_g^{(1)j}$, respectively. Similarly, $\boldsymbol{\theta}_h^{(2)}$ and $\boldsymbol{\theta}_g^{(2)j}$ are parameters of $\kappa_h^{(2)}(\mathbf{x}^{(2)}, \mathbf{x}^{(2)'})$ and $\kappa_g^{(2)j}(\mathbf{x}^{(2)}, \mathbf{x}^{(2)'})$.

The mappings from $X^{(1)}$ to $Y^{(1)}$ and $X^{(2)}$ to $Y^{(2}$ are different from the shared GPLVM, subspace GPLVM and MRD. The shared GPLVM, subspace GPLVM and MRD employ one GP mapping for each view to capture the common information of multiple outputs. These models can not sufficiently model the characteristics of each output, while the mappings in our model can well capture the differences and dependence among multiple outputs.

4. Inference and Learning

Given the model assumptions, we can get the joint distribution of observations and latent variables for the proposed model,

$$\begin{aligned} &p(Y^{(1)}, Y^{(2)}, H^{(1)}, H^{(2)}, G^{(1)}, G^{(2)}, X^{(1)}, X^{(1)}, X^{(1,2)}) \\ &= \prod_{\mathcal{K}\in\{(1),(2)\}} p(Y^{\mathcal{K}}|G^{\mathcal{K}}, H^{\mathcal{K}}) p(G^{\mathcal{K}}, H^{\mathcal{K}}|X^{\mathcal{K}}) p(X^{\mathcal{K}}|X^{(1,2)}, \mathbf{t}) p(X^{(1,2)}|\mathbf{t}), \end{aligned} \tag{6}$$

where the superscript $\mathcal{K} \in \{(1), (2)\}$ of a variable indicates the view the variable corresponding to and $G^{\mathcal{K}} = [(\mathbf{g}_1^{\mathcal{K}})^\top, \ldots, (\mathbf{g}_J^{\mathcal{K}})^\top]$. The marginal likelihood can be calculated by integrating out all the latent variables, which is commonly used as the goal of model learning. However, the private low-dimensional variables $X^{(1)}$ and $X^{(2)}$ cannot be integrated out because they appear nonlinearly in the inverse of the kernel matrices $\mathbf{G}_{X,X}^{(1)j}$, $\mathbf{H}_{X,X}^{(1)}$ and $\mathbf{G}_{X,X}^{(2)j}$, $\mathbf{H}_{X,X}^{(2)}$, respectively. Throughout the paper, covariance matrices are represented by bold uppercase characters with superscripts and subscripts. The corresponding GP can be inferred from the character, with $\mathbf{K}$ for x, $\mathbf{H}$ for h and $\mathbf{G}$ for g, respectively. The superscript indicates the view that the GP is from, while the subscript indicates the inputs where the covariance matrix evaluated. Following Titsias and Lawrence (2010), we make some approximations to the true posterior of the model using variational inference, thus deducing the variational lower bound of the logarithmic marginal likelihood.

4.1 Variational Lower Bound

We introduce inducing points and adopt the structured variational inference method to our model. In order to train the proposed model, we minimize the KL divergence between approximate posterior and true posterior, which is equivalent to maximizing the evidence lower bound of the logarithmic marginal likelihood.

First, we employ inducing variables to augment the model. Specifically, for each view $\mathcal{K} \in \{(1), (2)\}$ and each latent function, we introduce a set of M inducing variables. We

use $\{\mathbf{u}_j^{\mathcal{K}} \in R^M\}_{j=1}^J$ and $\mathbf{v}^{\mathcal{K}} \in R^M$ to represent the value of $g_j^{\mathcal{K}}$ at inducing inputs $Z_g^{\mathcal{K}j} \in R^{M\times Q}$ and the value of $h^{\mathcal{K}}$ at inducing points $Z_h^{\mathcal{K}} \in R^{M\times Q}$, respectively. Denote $U^{\mathcal{K}} = [(\mathbf{u}_1^{\mathcal{K}})^\top, \ldots, (\mathbf{u}_J^{\mathcal{K}})^\top]$. Attributed to the conditional independence assumption of latent variables $\{\mathbf{g}_j^{\mathcal{K}}\}_{j=1}^J$, we have $p(U^{\mathcal{K}}|\{Z_g^{\mathcal{K}j}\}_{j=1}^J) = \prod_{j=1}^J \mathcal{N}(\mathbf{u}_j^{\mathcal{K}}|0, \mathbf{G}_{Z,Z}^{\mathcal{K}j})$. $p(V^{\mathcal{K}}|X^{\mathcal{K}})$ is also assumed to be zero-mean Gaussian with covariance matrix $\mathbf{H}_{Z,Z}^{\mathcal{K}}$. The conditional Gaussian distributions are given as $p(G^{\mathcal{K}}|U^{\mathcal{K}}, X^{\mathcal{K}}) = \prod_{j=1}^J \mathcal{N}(\mathbf{g}_j^{\mathcal{K}}|\boldsymbol{\mu}_g^{\mathcal{K}j}, \widetilde{\mathbf{G}}_{X,X}^{\mathcal{K}j})$ with $\boldsymbol{\mu}_g^{\mathcal{K}j} = \mathbf{G}_{X,Z}^{\mathcal{K}j}(\mathbf{G}_{Z,Z}^{\mathcal{K}j})^{-1}\mathbf{u}_j^{\mathcal{K}}$ and $\widetilde{\mathbf{G}}_{X,X}^{\mathcal{K}j} = \mathbf{G}_{X,X}^{\mathcal{K}j} - \mathbf{G}_{X,Z}^{\mathcal{K}j}(\mathbf{G}_{Z,Z}^{\mathcal{K}j})^{-1}\mathbf{G}_{Z,X}^{\mathcal{K}j}$. Additionally, $p(H^{\mathcal{K}}|V^{\mathcal{K}}, X^{\mathcal{K}}) = \mathcal{N}(H^{\mathcal{K}}|\boldsymbol{\mu}_h^{\mathcal{K}}, \widetilde{\mathbf{H}}_{X,X}^{\mathcal{K}})$ with $\boldsymbol{\mu}_h^{\mathcal{K}} = \mathbf{H}_{X,Z}^{\mathcal{K}}(\mathbf{H}_{Z,Z}^{\mathcal{K}})^{-1}\mathbf{v}^{\mathcal{K}}$ and $\widetilde{\mathbf{H}}_{X,X}^{\mathcal{K}} = \mathbf{H}_{X,X}^{\mathcal{K}} - \mathbf{H}_{X,Z}^{\mathcal{K}}(\mathbf{H}_{Z,Z}^{\mathcal{K}})^{-1}\mathbf{H}_{Z,X}^{\mathcal{K}}$.

Then, we introduce the joint variational distribution which is assumed to be factorized as $q(\Theta^{(1)})q(\Theta^{(2)})q(X^{(1)})q(X^{(2)})q(X^{1,2})$ where $q(X^{(1)}) = \mathcal{N}(X^{(1)}|\boldsymbol{\mu}^{(1)}, S^{(1)})$, $q(X^{(2)})$=$\mathcal{N}(X^{(2)}|\boldsymbol{\mu}^{(2)}$, $S^{(2)})$ and $q(X^{(1,2)}) = \mathcal{N}(X^{(1,2)}|\boldsymbol{\mu}^{(1,2)}, S^{(1,2)})$. $q(\Theta^{(1)})$ and $q(\Theta^{(2)})$ are the variational distributions of latent variables $\{G^{(1)}, H^{(1)}, U^{(1)}, V^{(1)}\}$ and $\{G^{(2)}, H^{(2)}, U^{(2)}, V^{(2)}\}$ whose specific forms are defined as

$$q(\Theta^{\mathcal{K}}) = p(G^{\mathcal{K}}|U^{\mathcal{K}}, X^{\mathcal{K}})p(H^{\mathcal{K}}|V^{\mathcal{K}}, X^{\mathcal{K}})q(U^{\mathcal{K}})q(V^{\mathcal{K}}), \quad \mathcal{K} \in \{(1), (2)\}. \tag{7}$$

Finally, given the above assumptions, the lower bound of the logarithmic marginal likelihood can be expressed as

$$\begin{aligned}\mathcal{F}_v(q) &= \int \prod_{\mathcal{K}} q(\Theta^{\mathcal{K}})q(X^{\mathcal{K}})q(X^{(1,2)}) \log \prod_{\mathcal{K}} \frac{p(Y^{\mathcal{K}}|X^{\mathcal{K}})p(X^{\mathcal{K}}|X^{(1,2)}, \mathbf{t})p(X^{(1,2)}|\mathbf{t})}{q(\Theta^{\mathcal{K}})q(X^{\mathcal{K}})q(X^{(1,2)})} d\Theta^{\mathcal{K}} dX^{\mathcal{K}} dX^{(1,2)} \\ &= -\mathrm{KL}\big[q(X^{(1)})q(X^{(2)})q(X^{(1,2)})||p(X^{(1)}|X^{(1,2)}, \mathbf{t})p(X^{(2)}|X^{(1,2)}, \mathbf{t})p(X^{(1,2)}|\mathbf{t})\big] \\ &\quad + \sum_{\mathcal{K}} \hat{\mathcal{L}}^{\mathcal{K}}, \quad \mathcal{K} \in \{(1), (2)\}. \end{aligned} \tag{8}$$

The detailed calculation of the KL divergence is given below.

$$\begin{aligned}&\mathrm{KL}(q(X^{(1)})q(X^{(2)})q(X^{(1,2)})||p(X^{(1)}|X^{(1,2)}, \mathbf{t})p(X^{(2)}|X^{(1,2)}, \mathbf{t})p(X^{(1,2)}|\mathbf{t})) \\ &= \frac{1}{2}\sum_{q=1}^{Q}\Bigg[\big[\log|A^{(1)}| + \log|A^{(2)}| + \log|\mathbf{K}_{\mathbf{t},\mathbf{t}}^{(1,2)}| - \log|S_q^{(1,2)}| - \log|S_q^{(1)}| - \log|S_q^{(2)}|\big] \\ &\qquad + \big[(1-\alpha^{(1)})\boldsymbol{\mu}_q^{(1,2)} - \boldsymbol{\mu}_q^{(1)}\big]^\top (A^{(1)})^{-1}\big[(1-\alpha^{(1)})\boldsymbol{\mu}_q^{(1,2)} - \boldsymbol{\mu}_q^{(1)}\big] \\ &\qquad + \big[(1-\alpha^{(2)})\boldsymbol{\mu}_q^{(1,2)} - \boldsymbol{\mu}_q^{(2)}\big]^\top (A^{(2)})^{-1}\big[(1-\alpha^{(2)})\boldsymbol{\mu}_q^{(1,2)} - \boldsymbol{\mu}_q^{(2)}\big] \\ &\qquad + \mathrm{Tr}\big[\big[(1-\alpha^{(1)})^2(A^{(1)})^{-1} + (1-\alpha^{(2)})^2(A^{(2)})^{-1}\big]S_q^{(1,2)}\big] \\ &\qquad + \mathrm{Tr}\big[(\mathbf{K}_{\mathbf{t},\mathbf{t}}^{(1,2)})^{-1}\big[\boldsymbol{\mu}_q^{(1,2)}(\boldsymbol{\mu}_q^{(1,2)})^\top + S_q^{(1,2)}\big]\big] + \mathrm{Tr}\big[(A^{(1)})^{-1}S_q^{(1)} + (A^{(2)})^{-1}S_q^{(2)}\big]\Bigg], \end{aligned} \tag{9}$$

where $A^{(1)}$ and $A^{(2)}$ represent $(\alpha^{(1)})^2\mathbf{K}_{\mathbf{t},\mathbf{t}}^{(1)} + (1-\alpha^{(1)})^2\epsilon^{(1)}I$ and $(\alpha^{(2)})^2\mathbf{K}_{\mathbf{t},\mathbf{t}}^{(2)} + (1-\alpha^{(2)})^2\epsilon^{(2)}I$, respectively.

Since the observations on different dimensions in each view are assumed to be conditionally independent, the term $\hat{\mathcal{L}}^{\mathcal{K}}$ can be decomposed regarding dimensions, which has the following formula.

$$\hat{\mathcal{L}}^{\mathcal{K}} = \sum_{d=1}^{D}\Bigg[\log\frac{(\beta^{\mathcal{K}})^{\frac{N}{2}}|\mathbf{H}_{Z,Z}^{\mathcal{K}}|^{\frac{1}{2}}}{(2\pi)^{\frac{N}{2}}|\beta^{\mathcal{K}}\psi_4^{\mathcal{K}} + \mathbf{H}_{Z,Z}^{\mathcal{K}}|^{\frac{1}{2}}} + \sum_{j=1}^{J}\log\frac{|\mathbf{G}_{Z,Z}^{\mathcal{K}j}|^{\frac{1}{2}}}{|\beta^{\mathcal{K}}(w_{dj}^{\mathcal{K}})^2\psi_5^{\mathcal{K}j} + \mathbf{G}_{Z,Z}^{\mathcal{K}j}|^{\frac{1}{2}}}$$

$$
\begin{aligned}
&-\frac{1}{2}(\mathbf{y}_d^{\kappa})^{\top}\Bigg(\beta^{\kappa}I-\sum_{j=1}^{J}(\beta^{\kappa})^2(w_{dj}^{\kappa})^2\psi_1^{\kappa j}(\beta^{\kappa}(w_{dj}^{\kappa})^2\psi_5^{\kappa j}+\mathbf{G}_{Z,Z}^{\kappa j})^{-1}(\psi_1^{\kappa j})^{\top}\\
&-(\beta^{\kappa})^2\psi_0^{\kappa}(\beta^{\kappa}\psi_4^{\kappa}+\mathbf{H}_{Z,Z}^{\kappa})^{-1}(\psi_0^{\kappa})^{\top}\Bigg)\mathbf{y}_d^{\kappa}-\frac{\beta^{\kappa}}{2}\psi_2^{\kappa}+\frac{\beta^{\kappa}}{2}\mathrm{Tr}(\psi_4^{\kappa}(\mathbf{H}_{Z,Z}^{\kappa})^{-1})\\
&-\frac{\beta^{\kappa}}{2}\sum_{j=1}^{J}(w_{dj}^{\kappa})^2\psi_3^{\kappa j}+\frac{\beta^{\kappa}}{2}\sum_{j=1}^{J}\mathrm{Tr}((w_{dj}^{\kappa})^2\psi_5^{\kappa j}(\mathbf{G}_{Z,Z}^{\kappa j})^{-1})\Bigg], \qquad (10)
\end{aligned}
$$

where $\psi_0^{\kappa}=\langle\mathbf{H}_{X,Z}^{\kappa}\rangle_{q(X^{\kappa})}$, $\psi_1^{\kappa j}=\langle\mathbf{G}_{X,Z}^{\kappa j}\rangle_{q(X^{\kappa})}$, $\psi_2^{\kappa}=\mathrm{Tr}(\langle\mathbf{H}_{X,X}^{\kappa}\rangle_{q(X^{\kappa})})$, $\psi_3^{\kappa j}=\mathrm{Tr}(\langle\mathbf{G}_{X,X}^{\kappa j}\rangle_{q(X^{\kappa})})$, $\psi_4^{\kappa}=\langle\mathbf{H}_{Z,X}^{\kappa}\mathbf{H}_{X,Z}^{\kappa}\rangle_{q(X^{\kappa})}$, and $\psi_5^{\kappa j}=\langle\mathbf{G}_{Z,X}^{\kappa j}\mathbf{G}_{X,Z}^{\kappa j}\rangle_{q(X^{\kappa})}$. $\langle\cdot\rangle_{q(X^{\kappa})}$ denotes expectation under the distribution $q(X^{\kappa})$. The detailed computations for the evidence lower bound and the involved statistics are given in Appendix A and B, respectively.

The computational complexity for training McGPDS is dominated by computing the inversions of the kernel matrices, and thus the computational complexity is $O(VD(J+1)M^3+(V+1)N^3)$, where V is the number of views.

4.2 Parameter Estimation

The parameters to be optimized in the proposed model include model parameters and variational parameters. The model parameters involve hyperparameters in the kernel functions of the latent variables $\{\mathbf{g}^{(1)},\mathbf{g}^{(2)},\mathbf{h}^{(1)},\mathbf{h}^{(2)},X^{(1)},X^{(2)},X^{(1,2)}\}$, e.g., σ_f^2 and α_q in the used ARD kernel $\kappa(x,x')=\sigma_f^2\exp(-\frac{1}{2}\sum_{q=1}^{Q}\alpha_q(x_q-x_q')^2)$, the inverse variance of white Gaussian noise $\{\beta^{(1)},\beta^{(2)}\}$, Gaussian noises $\{\epsilon^{(1)},\epsilon^{(2)}\}$, and weights $\{W^{(1)},W^{(2)},\alpha^{(1)},\alpha^{(2)}\}$. The variational parameters include the mean and covariance of the variational distributions, $\{\boldsymbol{\mu}^{(1)},S^{(1)},\boldsymbol{\mu}^{(2)},S^{(2)},\boldsymbol{\mu}^{(1,2)},S^{(1,2)}\}$, and the inducing inputs $\{Z^{(1)},Z_h^{(1)},Z^{(2)},Z_h^{(2)}\}$. All the parameters are jointly optimized through the gradient descent method. Here we give the update rules for variational mean and covariance matrices, in which the optimization for covariance employs the reparameterization trick inspired by Opper and Archambeau (2009). The derivation is analogous to that in Damianou et al. (2011) and Damianou et al. (2016), to which we refer the readers for more details.

The variational mean in the private latent space can be optimized by the gradient descent method and the gradient of evidence lower bound w.r.t variational mean is given by

$$
\frac{\partial\mathcal{L}}{\partial\boldsymbol{\mu}_q^{\kappa}}=\frac{\partial\hat{\mathcal{L}}^{\kappa}}{\partial\boldsymbol{\mu}_q^{\kappa}}-(A_q^{\kappa})^{-1}\big[\boldsymbol{\mu}_q^{\kappa}-(1-\alpha^{\kappa})\boldsymbol{\mu}_q^{(1,2)}\big].
$$

The private variational covariance matrix S_q^{κ} can be reparameterized as

$$
S_q^{\kappa} \;=\; ((A_q^{\kappa})^{-1}+diag(\boldsymbol{\lambda}_q^{\kappa}))^{-1},
$$

where $diag(\boldsymbol{\lambda}_q^{\kappa})=-2\frac{\partial\mathcal{F}_v(q)}{\partial S_q^{\kappa}}$ is an $N\times N$ diagonal and positive definite matrix, w.r.t which the gradient of evidence lower bound is given by

$$
\frac{\partial L}{\partial\lambda_q^{\kappa}} \;=\; -(S_q^{\kappa}\circ S_q^{\kappa})(\frac{\partial\hat{L}^{\kappa}}{\partial S_q^{\kappa}}+\frac{1}{2}\lambda_q^{\kappa}). \qquad (11)
$$

The shared variational parameters $\{\boldsymbol{\mu}^{(1,2)}, S^{(1,2)}\}$ have analytical solutions. After updating the private variational parameters, we can update the shared variational parameters by the following equations.

$$\boldsymbol{\mu}_q^{(1,2)} = S_q^{(1,2)}\big[(1-\alpha^{(1)})(A_q^{(1)})^{-1}\boldsymbol{\mu}_q^{(1)} + (1-\alpha^{(2)})(A_q^{(2)})^{-1}\boldsymbol{\mu}_q^{(2)}\big], \tag{12}$$

$$S_q^{(1,2)} = [(1-\alpha^{(1)})^2(A_q^{(1)})^{-1} + (1-\alpha^{(2)})^2(A_q^{(2)})^{-1} + (\mathbf{K_{tt}}^{(1,2)})^{-1}]^{-1}. \tag{13}$$

5. Prediction with the McGPDS

Given the trained McGPDS which can jointly model observations of two views $Y^{(1)}$ and $Y^{(2)}$ and learn the shared latent space $X^{(1,2)}$ and the private latent spaces $X^{(1)}$ and $X^{(2)}$, we aim to generate the outputs from a view given the observations from the other view. For example, generate $Y_*^{(2)} \in \mathbb{R}^{N_*\times D_2}$ using $Y_*^{(1)} \in \mathbb{R}^{N_*\times D_1}$. The McGPDS has the capability to accomplish this task by three steps, similar to MRD (Damianou et al., 2012).

In the first step, we use variational inference again to derive the posterior distributions of the latent variables $X_*^{(1)} \in \mathbb{R}^{N_*\times Q}$ and $X_*^{(1,2)} \in \mathbb{R}^{N_*\times Q}$ which are most likely to govern the generation of $Y_*^{(1)}$. We use $q(X_*^{(1)}, X_*^{(1,2)})$ to approximate $p(X_*^{(1)}, X_*^{(1,2)}|Y_*^{(1)})$. The approximate posterior distribution $q(X_*^{(1)}, X_*^{(1,2)})$ is the marginal distribution of $q(X^{(1)}, X^{(1,2)}, X_*^{(1)}, X_*^{(1,2)})$. To obtain $q(X^{(1)}, X^{(1,2)}, X_*^{(1)}, X_*^{(1,2)})$, we maximize the variational lower bound of the marginal likelihood $p(Y^{(1)}, Y_*^{(1)})$,

$$\begin{aligned}\mathcal{F}_*^{(1)} =& -\mathrm{KL}\big[q(X_*^{(1)}, X^{(1)})q(X_*^{(1,2)}, X^{(1,2)})||p(X_*^{(1)}, X^{(1)}|X_*^{(1,2)}, X^{(1,2)})p(X_*^{(1,2)}, X^{(1,2)})\big] \\ &+ \hat{\mathcal{L}}^{(1)}(Y_*^{(1)}, Y^{(1)}),\end{aligned} \tag{14}$$

where we've omitted time $\mathbf{t}$ and $\mathbf{t}_*$ for brevity. Particularly, the lower bound can be maximized using the same method as for training. The detailed calculation for $\mathcal{F}_*^{(1)}$ is given in Appendix C.

In the second step, we obtain the private latent variable which is also essential to generate data from a view. Precisely, in order to generate observations $Y_*^{(2)}$, we need to obtain the private latent variable $X_*^{(2)}$. However, just the observed test data from the first view $Y_*^{(1)}$ can hardly provide information for data in the second view $Y_*^{(2)}$ and thus it is quite difficult to obtain an exact representation of $X_*^{(2)}$. Therefore, we refer to the latent variables learned from training data, $X^{(1,2)}$ and $X^{(2)}$, and employ the nearest neighbor to obtain the private latent variable $X_*^{(2)}$. Specifically, we find the shared latent variable from training data $\bar{X}^{(1,2)}$ which is closest to $X_*^{(1,2)}$ obtained by the first step, and acquire the variational distribution of private latent variable $\bar{X}^{(2)}$ directly from training data whose indexes correspond to $\bar{X}^{(1,2)}$ to approximate the posterior of private latent variable $X_*^{(2)}$.

In the third step, we predict the output $Y_*^{(2)}$ using the marginal posterior distribution of latent variable $q(X_*^{(2)})$ obtained through the second step. Specifically, $Y_*^{(2)}$ can be calculated by

$$\begin{aligned}p(Y_*^{(2)}) = \int & p(Y_*^{(2)}|G_*^{(2)}, H_*^{(2)})p(G_*^{(2)}|X_*^{(2)}, U^{(2)})p(H_*^{(2)}|X_*^{(2)}, V^{(2)})q(U^{(2)})q(V^{(2)})q(X_*^{(2)}) \\ & dG_*^{(2)}dH_*^{(2)}dX_*^{(2)}dU^{(2)}dV^{(2)}.\end{aligned} \tag{15}$$

Note that the variational distributions $q(U^{(2)})$ and $q(V^{(2)})$ are obtained during the training phase which need not be optimized during the prediction period. Since the integration in

Algorithm 1 Prediction with the McGPDS

1: **Input:** training data for two views $Y^{(1)}$ and $Y^{(2)}$, McGPDS model trained via two-view data ($Y^{(1)}$, $Y^{(2)}$) and test data in the first view $Y_*^{(1)}$.
2: **Output**: generated observations in the second view $Y_*^{(2)}$.
3: Maximize the evidence lower bound of the marginal likelihood $p(Y_*^{(1)}, Y^{(1)})$ to obtain $q(X^{(1)}, X^{(1,2)}, X_*^{(1)}, X_*^{(1,2)})$.
4: Get the marginal distribution $q(X_*^{(1)}, X_*^{(1,2)})$ to obtain test mean $\boldsymbol{\mu}_*^{(1)}$ and $\boldsymbol{\mu}_*^{(1,2)}$ and covariance $S_*^{(1)}$ and $S_*^{(1,2)}$.
5: Find the optimal $\hat{\boldsymbol{\mu}}_*^{(2)}$ and $\hat{S}_*^{(2)}$ using the K-nearest neighbor method according to the distance between $\boldsymbol{\mu}_*^{(1,2)}$ and $\boldsymbol{\mu}^{(1,2)}$.
6: $q(X_*^{(2)}) \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_*^{(2)}, \hat{S}_*^{(2)})$.
7: Predict $Y_*^{(2)}$ using Equation (15).

(15) is analytically intractable, we follow Damianou et al. (2011) to calculate the expectation of $\mathbf{g}_*^{(2)}$ and $\mathbf{h}_*^{(2)}$ as $\mathbb{E}(\mathbf{g}_*^{(2)})$ and $\mathbb{E}(\mathbf{h}_*^{(2)})$, respectively, and estimate the expectation covariance matrices with Monte Carlo sampling. The element-wise autocovariance matrices of $\mathbf{g}_*^{(2)}$ and $\mathbf{h}_*^{(2)}$ are denoted as $\mathbb{V}(\mathbf{g}_*^{(2)})$ and $\mathbb{V}(\mathbf{h}_*^{(2)})$, respectively.

$$\mathbb{E}(\mathbf{h}_*^{(2)}) = \psi_{0*}^{(2)}\mathbf{b}_h^{(2)},$$
$$\mathbb{E}(\mathbf{g}_*^{(2)j}) = \psi_{1*}^{(2)j}\mathbf{b}_g^{(2)j},$$
$$\mathbb{V}(h_{\widetilde{n}}^{(2)j}) = \mathbf{b}_h^{(2)\top}(\psi_{4\widetilde{n}}^{(2)} - ((\psi_{0\widetilde{n}}^{(2)})^\top)\psi_{0\widetilde{n}}^{(2)})\mathbf{b}_h^{(2)} + \psi_{2*}^{(2)} - \mathrm{Tr}\big[((\mathbf{H}_{Z,Z}^{(2)})^{-1} - (\mathbf{H}_{Z,Z}^{(2)} + \beta^{(2)}\psi_4^{(2)})^{-1})\psi_{4*}^{(2)}\big],$$
$$\mathbb{V}(g_{\widetilde{n}}^{(2)j}) = \mathbf{b}_g^{(2)j\top}(\psi_{5\widetilde{n}}^{(2)j} - (\psi_{1\widetilde{n}}^{(2)j})^\top\psi_{1\widetilde{n}}^{(2)j})\mathbf{b}_g^{(2)j} + \psi_{3*}^{(2)j} - \mathrm{Tr}[((\mathbf{G}_{Z,Z}^{-1})^{(2)j} - (\mathbf{G}_{Z,Z}^{(2)j} + \beta^{(2)}w_{dj}^2\psi_5^{(2)j})^{-1})$$
$$\psi_{5*}^{(2)j}],$$

where $\mathbb{V}(h_{\widetilde{n}}^{(2)j})$ denotes the $\widetilde{n}$th entry of $\mathbb{V}(\mathbf{h}_*^{(2)})$, and $\mathbb{V}(g_{\widetilde{n}}^{(2)j})$ denotes the $(\widetilde{n} \times j)$th entry of $\mathbb{V}(\mathbf{g}_*^{(2)})$. $\mathbf{b}_h^{(2)} = \beta^{(2)}(\mathbf{H}_{Z,Z}^{(2)} + \beta^{(2)}\psi_4^{(2)})^{-1}(\psi_0^{(2)})^\top\mathbf{y}^{(2)}, \mathbf{b}_g^{(2)j} = \beta^{(2)}(\mathbf{G}_{Z,Z}^{(2)j} + \beta^{(2)}\psi_5^{(2)j})^{-1}(\psi_1^{(2)j})^\top\mathbf{y}^{(2)}$, $\psi_{0*}^{(2)} = \langle\mathbf{H}_{X_*,Z}^{(2)}\rangle_{q(X_*^{(2)})}$, $\psi_{1*}^{(2)j} = \langle\mathbf{G}_{X_*,Z}^{(2)j}\rangle_{q(X_*^{(2)})}$, $\psi_{2*}^{(2)} = \mathrm{Tr}(\langle\mathbf{H}_{X_*,X_*}^{(2)}\rangle_{q(X_*^{(2)})}), \psi_{3*}^{(2)j} = \mathrm{Tr}(\langle\mathbf{G}_{X_*,X_*}^{(2)j}\rangle_{q(X_*^{(2)})})$, $\psi_{4*}^{(2)} = \langle\mathbf{H}_{Z,X_*}^{(2)}\mathbf{H}_{X_*,Z}^{(2)}\rangle_{q(X_*^{(2)})}$, $\psi_{5*}^{(2)j} = \langle\mathbf{G}_{Z,X_*}^{(2)j}\mathbf{G}_{X_*,Z}^{(2)j}\rangle_{q(X_*^{(2)})}$, $\psi_{0\widetilde{n}}^{(2)} = \langle\mathbf{H}_{X_{\widetilde{n}},Z}^{(2)}\rangle_{q(X_{\widetilde{n}}^{(2)})}, \psi_{1\widetilde{n}}^{(2)j} = \langle\mathbf{G}_{X_{\widetilde{n}},\mathbf{u}}^{(2)j}\rangle_{q(X_{\widetilde{n}}^{(2)})}$, $\psi_{4\widetilde{n}}^{(2)} = \langle\mathbf{H}_{Z,X_{\widetilde{n}}}^{(2)}\mathbf{K}_{\mathbf{h}_{\widetilde{n}},Z}^{(2)}\rangle_{q(X_{\widetilde{n}}^{(2)})}, \psi_{5\widetilde{n}}^{(2)j} = \langle\mathbf{G}_{Z,X_{\widetilde{n}}}^{(2)j}\mathbf{G}_{X_{\widetilde{n}},Z}^{(2)j}\rangle_{q(X_{\widetilde{n}}^{(2)})}$, $\widetilde{n} = 1, \ldots, N_*$, $d = 1, \ldots, D$ and $j = 1, \ldots, J$. Since $Y_{*d}^{(2)} = \sum_{j=1}^J w_{dj}^{(2)}\mathbf{g}_*^{(2)j} + \mathbf{h}_*^{(2)}, d \in [1 \ldots D]$, the expectation and covariance of $Y_{*d}^{(2)}$ are $\mathbb{E}(Y_{*d}^{(2)}) = \sum_{j=1}^J w_{dj}^{(2)}\mathbb{E}(\mathbf{g}_*^{(2)j}) + E(\mathbf{h}_*^{(2)})$ and $\mathbb{V}(Y_{*d}^{(2)}) = \sum_{j=1}^J (w_{dj}^{(2)})^2\mathbb{V}(\mathbf{g}_*^{(2)j}) + \mathbb{V}(\mathbf{h}_*^{(2)}) + (\beta^{(2)})^{-1}I$, where $(\mathbf{y}_*^{(2)})^\top = [(\mathbf{y}_{*1}^{(2)})^\top, \ldots, (\mathbf{y}_{*D}^{(2)})^\top]$. The whole prediction process is shown in Algorithm 1.

6. Experiments

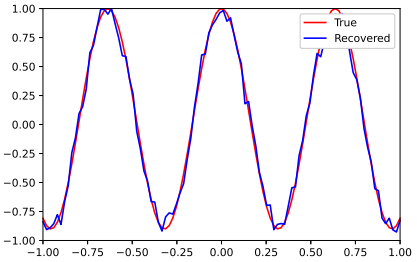
In order to validate the effectiveness of the proposed McGPDS, we conduct experiments on five multi-view datasets including two synthetic datasets and three real world datasets[1]. We

1. For an implementation of McGPDS in Matlab, see https://github.com/mcgpds/mcgpds.

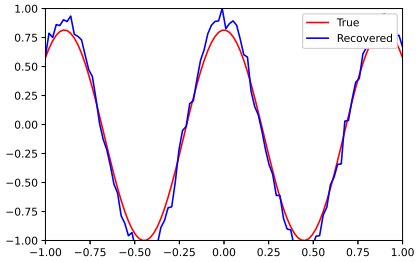
evaluate our model in two different kinds of tasks. The first is recovering the structures of the latent variables when the correlation between the shared and private latent variables is strong. The second is generating data from one view given data from the other view.

For comparison, all models are trained with the same initializations and we set $J = 1$ in the proposed model. For the toy data experiments, we use linear kernel without inducing points and the dimension of each view's private latent variable is set to 1. For the real-world data experiments, we use RBF kernel with the variance initialized to 1. We use 100 inducing points and the dimension of each view's private latent variable is set to 5 unless otherwise stated. For all the experiments, alpha is initialized to 0.5 for each view and the mixture weights in the output layer are independently initialized from a Gaussian distribution with 0 mean and 0.01 variance. For the K-nearest neighbor method, we set $K = 1$. In the experiments, the shared GPLVM refers to the new version of the shared GPLVM, namely, the subspace GPLVM. For MRD, we follow the setting in Damianou et al. (2012). All experiments are repeated five times, and the average results are reported as the final results. The root mean square error (RMSE) and mean standardized log loss (MSLL) are used as the performance measures. MSLL is the mean negative log probability of all the test data, where the predictive density is given by (15). The lower the RMSE and MSLL are, the better the performance is.

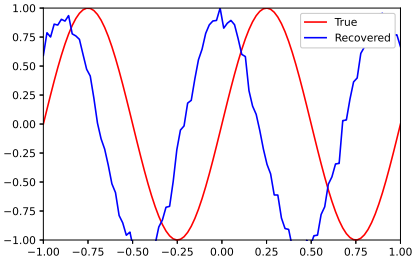
6.1 Toy Data



(a) Private Signal ($cos(\pi^2 t)$)

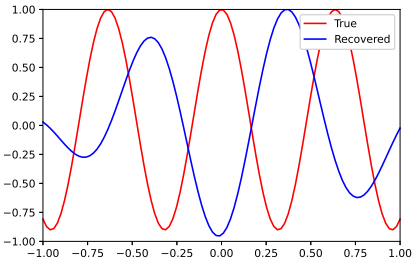


(b) Private Signal ($cos(\sqrt{5}\pi t)$)

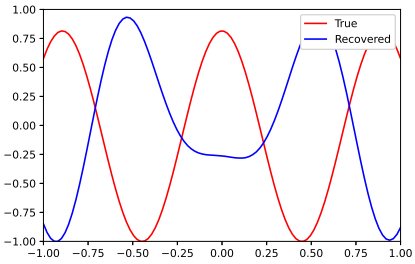


(c) Shared Signal ($sin(2\pi t)$)

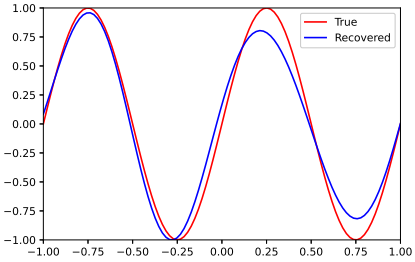
Figure 4: The results of McGPDSs on the toy dataset. Red lines represent true signals, and blue lines represent recovered signals.



(a) Private Signal ($cos(\pi^2 t)$)



(b) Private Signal ($cos(\sqrt{5}\pi t)$)



(c) Shared Signal ($sin(2\pi t)$)

Figure 5: The results of MRD on the toy dataset. Red lines represent true signals, and blue lines represent recovered signals.

First, we conduct the experiment on a synthetic dataset which is similar to the one used by Salzmann et al. (2010) and Jia et al. (2010). We first generate three one-dimensional latent variables using three signals: $cos(\pi^2 t)$ and $cos(\sqrt{5}\pi t)$ which generate the private latent variables, and $sin(2\pi t)$ which generates the shared latent variable. Then, we use the randomly generated projection matrices to map the one-dimensional private latent variables to the ten-dimensional space and the one-dimensional shared latent variable to the five-dimensional space. The two-view sequential data $Y^{(1)}$ and $Y^{(2)}$ are constructed by concatenating the ten-dimensional private variable of each view with the five-dimensional shared variable. Therefore, both the generated sequences $Y^{(1)}$ and $Y^{(2)}$ are in 15 dimensions in total, that is, $\mathbf{y}_i^{(1)}, \mathbf{y}_i^{(2)} \in \mathcal{R}^{15}$.

The proposed model is capable of learning the latent variables corresponding to the observed sequential data. We use the McGPDS with a linear kernel function to recover the latent signals: the private signals ($cos(\pi^2 t)$ and $cos(\sqrt{5}\pi t)$) and the shared signal ($sin(2\pi t)$). We compare our model with the state-of-the-art GP-based multi-view dynamical system, i.e., MRD with dynamics.

Figure 4 shows the recovery results of the latent signals by our model. Specifically, Figure 4(a), (b) and (c) show the true signals as well as the recovered signals by McGPDS for $cos(\pi^2 t)$, $cos(\sqrt{5}\pi t)$ and $sin(2\pi t)$, respectively. As shown in Figure 4, the recovered signals almost exactly match the true signals (up to a translation), which demonstrates that our model has the ability to learn an effective latent representation even when private latent variables are orthogonal to shared latent variables. As a comparison, Figure 5 shows the

results of the MRD with dynamics on this toy dataset. Figure 5(a) and (b) shows that the recovered private signals by the MRD deviates significantly from the true signals in both view. The only recovered signal that matches the true signal is the shared signal, as shown in Figure 5(d).

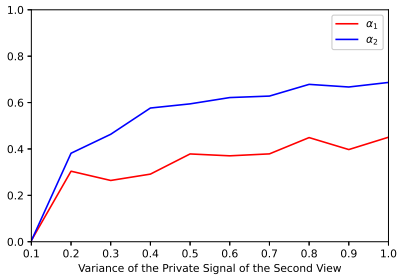


Figure 6: The learned $\alpha^{(1)}$ and $\alpha^{(2)}$ of McGPDS on the toy dataset with different variance of private latent signal of view 2.

Next, we test the interpretability of the learned combination weights, $\alpha^{(1)}$ and $\alpha^{(2)}$, on another synthetic dataset. The shared latent signal is generated from $sin(2\pi t)$ and the private latent signals are generated from pure Gaussian noise. The variance of the Gaussian noise of the first view is fixed to 0.1, while that of the second view varies from 0.1 to 1. The observations from each view is constructed by concatenating its private signal and the shared signal. Under this construction, the first view almost only contains the shared signal, while the ratio of the private and the shared signals in the second view increases with the variance of the former. We plot the learned $\alpha^{(1)}$ and $\alpha^{(2)}$ against the variance of the private latent variable of view 2 in Figure 6. As expected, the learned $\alpha^{(2)}$ increases with the variance of the private signal of the second view, which coincides with the change of the significant of the private signal. The learned $\alpha^{(1)}$ also increases but at a slower rate, since large noise in the second view adds difficulty in recovering the shared signal and the view-specific dynamic has to complement.

6.2 Human Motion Data

In this experiment, we use the human motion data which contain a set of 3D human poses and their corresponding silhouettes. The data are collected by Agarwal and Triggs (Ankur and Bill, 2006). We use 566 frames for training which contain 5 sequences corresponding to walking motions in different directions. The test data is a separate walking sequence of 158 frames. The pose data are 63-dimensional joint location vectors, and the silhouette data are 100-dimensional histogram of oriented gradients (HOG) vectors. We consider the task of generating data from a view given the other view, that is, we generate the corresponding 3D human poses given the silhouette data. We use the RBF kernel for all GPs and 100 inducing points for McGPDS. Dimensions of the shared and both private latent variables are set to 5 for all the models.

As described in the previous section, given test data in the first view $Y^{(1)}_{test}$, McGPDS optimises the private latent variables in the first view $X^{(1)}_{test}$ and the shared latent points $X^{(1,2)}_{test}$. Then, the training latent variables $X^{(2)}$ in the second view are selected as the test private

Table 1: The RMSE and MSLL on the human motion dataset.

	RMSE	MSLL
NNYspace	2.65±0.00	-
NNXspace (X learned by MRD)	3.19±0.03	-
NNXspace (X learned by McGPDS)	2.40±0.03	-
Shared GPLVM	5.15±0.01	3.41±0.17
MRD without dynamics	5.03±0.01	3.37±0.03
MRD with dynamics	2.65±0.01	3.01±0.25
Independent CGPDS	2.69±0.13	3.22±0.23
McGPDS	**2.37± 0.03**	**2.60±0.05**
McGPDS+GPLVM	2.62± 0.04	3.78±0.25
McGPDS+Linear	2.81± 0.15	-

latent variables $X_{test}^{(2)}$ according to the similarity of $X^{(1,2)}$ and $X_{test}^{(1,2)}$. Finally, McGPDS generates a set of novel poses $Y_{test}^{(2)}$ based on these selected training latent points $X^{(2)}$.

In this experiment, we compare our model with seven different methods, the nearest neighbor (NN) in silhouette space (NNYspace), the NN method in the X space (X learned by MRD), the NN method in the X space (X learned by McGPDS), the shared Gaussian process latent variable model (GPLVM), the MRD without dynamics, the MRD with dynamics and the independent CGPDS model. NNYspace finds the predicted 3D pose from training data whose silhouette is the closest to the corresponding test silhouette. Similarly, NNXspace finds the predicted 3D pose from training data whose shared latent information is the closest to the corresponding shared information of test data. The independent CGPDS model use one CGPDS on each view independently. To demonstrate the usefulness of the two key components in McGPDS, i.e., modelling the private latent variables using GPS with the mixture mean and covariance, and modelling the map from the private latent variables to observations with CGPDS, we conduct ablation studies for them. More specifically, we run two methods, McGPDS+GPLVM, which is McGPDS with the prior of the private latent variables replaced by that of GPLVM, and McGPDS+Linear, which is McGPDS with the output coupling layer replaced by a linear map, on the human motion dataset with the other setting unchanged.

Table 1 shows the RMSE and MSLL on the human motion dataset. As shown in Table 1, our model (McGPDS) obtains the lowest RMSE 2.37 ± 0.03 and the lowest MSLL 2.60 ± 0.05, which means that our model outperforms the state-of-the-art model (MRD with dynamics). Both McGPDS and MRD with dynamics outperform the independent CGPDS model, which confirms the usage of shared latent space structures. In addition, NNXspace (X learned by McGPDS) performs better than NNXspace (X learned by MRD). The ablation studies also confirms the usage of the two key components. Figure 7 demonstrates the results visually. As shown in Figure 7, the 3D poses generated by our model are closest to the true poses.

To better understand the impact of dimensionality and number of inducing points in McGPDS, we plot the RMSE and MSLL against total dimension of private latent variables in Figure 8(a) and the RMSE, MSLL and training time against number of inducing points in Figure 8(b). Figure 8(a) shows that the RMSE of McGPDS decreases as the total dimension

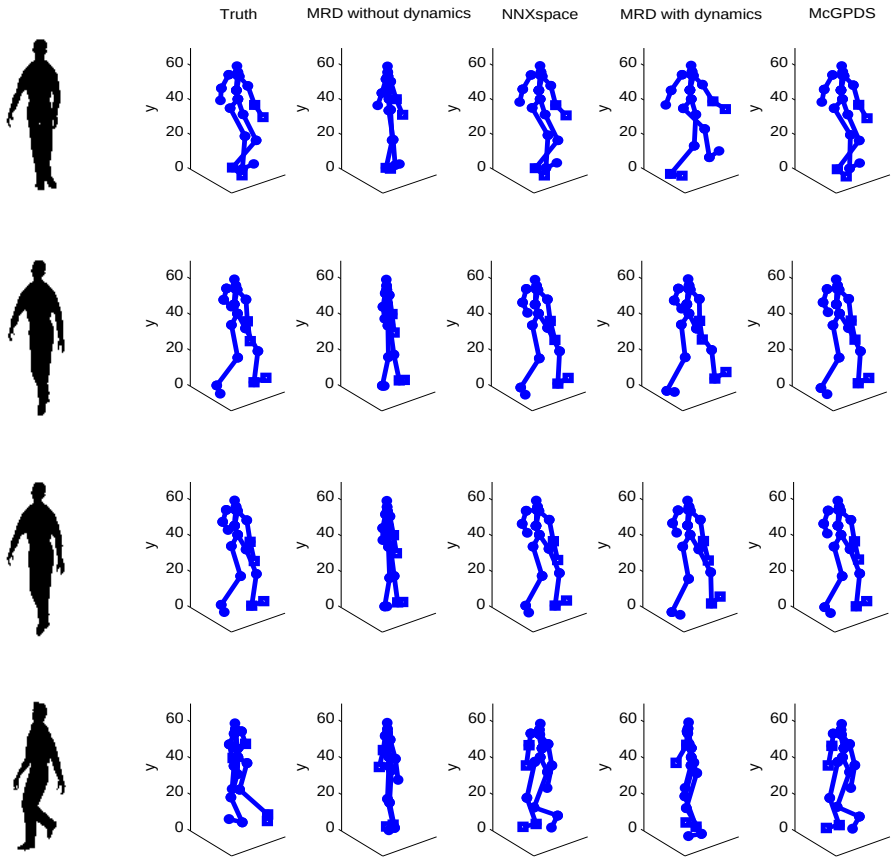
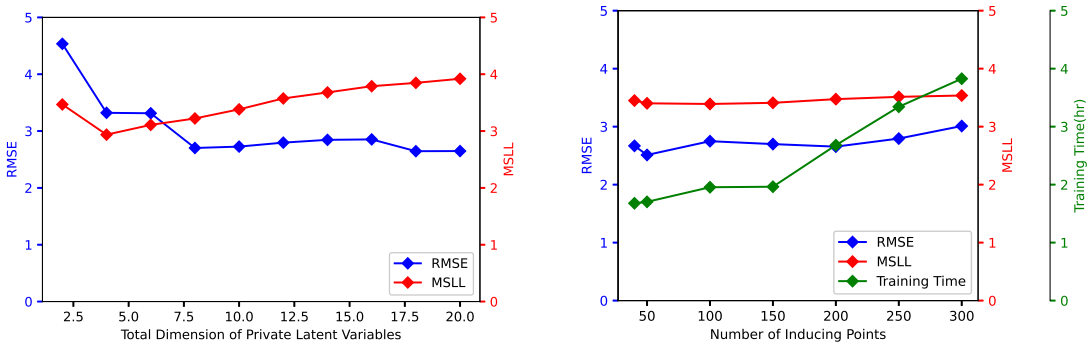


Figure 7: The results of generating 3D poses given silhouettes. The left-most side of each line represents the test silhouette. The remaining parts, from left to right, are the true poses, poses generated by MRD without dynamics, poses generated by NNXspace (X learned by the McGPDS), poses generated by MRD with dynamics, and poses generated by McGPDS, respectively.

of private latent variables increases, implying that larger latent space provides McGPDS more capability to capture multiview dynamics. The increase of MSLL is possibly due to the increase of number of variables, which encourages the model to upweight the KL divergence term in the ELBO, leading to an increase in the variance of the likelihood and thus in the MSLL. Figure 8(b) shows how the training time increases with the number of inducing points, while the impact of the latter on RMSE and MSLL is moderate.

6.3 CUAVE Data

In this experiment, we employ the CUAVE data which are composed of the videos showing a person speaking Arabic numerals and the corresponding Mel frequency cepstral coefficients (mfcc) features of the audio signals. Each video is represented by a 3750-dimensional vector and each mfcc feature is represented by a 13-dimensional vector. We use 194 frames of videos and mfcc features as training data and 51 frames of videos for testing. Our task is to



(a) Impact of Dimensionality (b) Impact of Num. of Inducing Points

Figure 8: (a) RMSE and MSLL of McGPDS with different total dimension of private latent variables on the human motion dataset. (b) RMSE, MSLL and training time(hr) of McGPDS with different number of inducing points on the human motion dataset.

Table 2: The RMSE and MSLL on the CUVAE dataset.

	RMSE	MSLL
NNYspace	1.31±0.00	-
NNXspace (X learned by MRD)	1.70±0.10	-
NNXspace (X learned by McGPDS)	1.38±0.15	-
Shared GPLVM	1.61±0.01	4.70±0.27
MRD without dynamics	1.29±0.01	4.34±0.13
MRD with dynamics	1.24±0.03	3.45±0.20
McGPDS	**1.19±0.03**	**1.94±0.07**

generate mfcc features given the frames of the videos. We use the RBF kernel for all GPs and 100 inducing points for McGPDS. Dimensions of the shared and both private latent variables are set to 5 for all the models.

From Table 2, we can see that our model obtains the best performance (with the lowest RMSE 1.19 ± 0.03 and lowest MSLL 1.94 ± 0.07) on the CUAVE dataset. The method NNXspace (X learned by McGPDS) is also better than NNXspace (X learned by MRD) in the CUAVE dataset. These results show that our model can obtain more reasonable latent representation, and thus generate observations closer to the truth.

6.4 Classification

In the final experiment, we examine McGPDS on a classification task. We use the Oil dataset, which contains 1000 12-dimensional examples from 3 classes. The observations constitute the first view, while the corresponding labels are taken as the second view in the form of one-hot encoding. Following the setting of Damianou et al. (2012), we select 10 random subsets of the data with increasing number of training points and compare to the NN method in the data space. Figure 9 shows that the accuracy of McGPDS is worse than

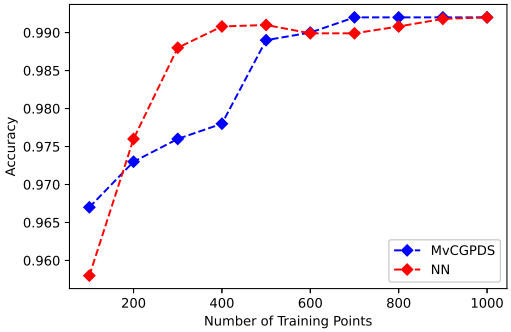


Figure 9: Accuracy of McGPDS and NN on the Oil Dataset.

NN when the training set is small and is comparable to NN as the number of training points increases. There are two possible reasons for the mediocre performance of McGPDS on small-size non-dynamic data. First, McGPDS uses three GPs to model the time dynamics, while the time stamps of non-dynamic data provide little, if not misleading, information on the observations. Second, McGPDS uses a mixture of GPs to model the observations, while the observations from view 2 of the used dataset is just the one-hot representation of labels. Both of these could potentially make McGPDS perform not so well on non-dynamic small data. We leave the application of McGPDS to classification for future work.

7. Conclusion

In this paper, we have proposed the McGPDS, which extends the CGPDS into the scenario of multi-view learning with flexible and general modeling in the latent space. As a novel hierarchical multi-view framework, the McGPDS takes full use of the characteristics of the multi-view data and the advantages of the CGPDS. The setting on the latent space is elastic and reasonable, where the relationship between private and shared latent variables can be learned adaptively via optimizing weights. We introduce inducing points and employ variational inference to integrate out the latent variables. The proposed model is trained through maximizing the evidence lower bound.

The effectiveness of our model for multi-view learning has been empirically validated on synthetic and real-world two-view datasets. For future work, we will extend our model beyond the current two views. The methodology can be similar to the current scenario, but deriving the ELBO for more-than-two-view cases is non-trivial and the applications of generating one or multiple views from other views will be more challenging.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Projects 62076096 and 62006078, Shanghai Municipal Project 20511100900, Chenguang Program of the Shanghai Education Development Foundation and the Shanghai Municipal Education Commission under Grant 19CG25, the Open Research Fund of KLATASDS-MOE, and the

Fundamental Research Funds for the Central Universities. Corresponding Author: Shiliang Sun.

Appendix A. Derivation of the Evidence Lower Bound for Training

In this section, we give the detailed derivation of the evidence lower bound for training data. Given two views of data, $Y^{(1)}$ and $Y^{(2)}$, the joint probability distribution for the proposed model is given by

$$p(Y^{(1)}, Y^{(2)}, H^{(1)}, H^{(2)}, G^{(1)}, G^{(2)}, X^{(1)}, X^{(2)}, X^{(1,2)}) = p(Y^{(1)}|G^{(1)}, H^{(1)})p(Y^{(2)}|G^{(2)}, H^{(2)})$$
$$p(G^{(1)}, H^{(1)}|X^{(1)})p(G^{(2)}, H^{(2)}|X^{(2)})p(X^{(1)}|X^{(1,2)}, \mathbf{t})p(X^{(2)}|X^{(1,2)}, \mathbf{t})p(X^{(1,2)}|\mathbf{t}). \quad (16)$$

We can get the logarithm of the marginal likelihood by integrating the latent variables,

$$p(Y^{(1)}, Y^{(2)}) = \int p(Y^{(1)}|G^{(1)}, H^{(1)})p(Y^{(2)}|G^{(2)}, H^{(2)})p(G^{(1)}, H^{(1)}|X_{(1)})p(G^{(2)}, H^{(2)}|X^{(2)})$$
$$p(X^{(1)}|X^{(1,2)}, \mathbf{t})p(X^{(2)}|X^{(1,2)}, \mathbf{t})p(X^{(1,2)}|\mathbf{t})dH^{(1)}dH^{(2)}dG^{(1)}dG^{(2)}dX^{(1)}dX^{(2)}dX^{(1,2)}. \quad (17)$$

Note that the integration w.r.t $X^{(1)}$ and $X^{(2)}$ is intractable, because $X^{(1)}$ appears nonlinearly in the inverse of the matrices $\mathbf{G}^{(1)}_{X,X}$ and $\mathbf{H}^{(1)}_{X,X}$ and $X^{(2)}$ appears nonlinearly in the inverse of the matrices $\mathbf{G}^{(2)}_{X,X}$ and $\mathbf{H}^{(2)}_{X,X}$. Therefore, we introduce inducing variables U and V to augment the model and compute the lower bound of its logarithmic marginal likelihood. The augmented joint probability density takes the form as

$$\begin{aligned} p(&Y^{(1)},Y^{(2)}, H^{(1)}, H^{(2)}, G^{(1)}, G^{(2)}, U^{(1)}, U^{(2)}, V^{(1)}, V^{(2)}, X^{(1)}, X^{(2)}, X^{(1,2)}) \\ =&p(\mathbf{Y}^{(1)}|G^{(1)}, H^{(1)})p(G^{(1)}|U^{(1)}, X^{(1)})p(H^{(1)}|V^{(1)}, X^{(1)})p(U^{(1)}|X^{(1)})p(V^{(1)}|X^{(1)}) \\ &p(Y^{(2)}|G^{(2)}, H^{(2)})p(G^{(2)}|U^{(2)}, X^{(2)})p(H^{(2)}|V^{(2)}, X^{(2)})p(U^{(2)}|X^{(2)})p(V^{(2)}|X^{(2)}) \\ &p(X^{(1)}|X^{(1,2)}, \mathbf{t})p(X^{(2)}|X^{(1,2)}, \mathbf{t})p(X^{(1,2)}|\mathbf{t}). \end{aligned} \quad (18)$$

In the above formula, $p(U^{(1)}|X^{(1)})$ and $p(U^{(2)}|U^{(2)})$ are zero-mean Gaussian with covariance matrices $\mathbf{G}^{(1)}_{Z,Z}$ and $\mathbf{G}^{(2)}_{Z,Z}$ and $p(V^{(1)}|X^{(1)})$ and $p(V^{(2)}|X^{(2)})$ are zero-mean Gaussian with covariance matrices $\mathbf{H}^{(1)}_{Z,Z}$ and $\mathbf{H}^{(2)}_{Z,Z}$. Precisely, they are expressed as

$$p(U^{(1)}|X^{(1)}) = \prod_{j=1}^{J} \mathcal{N}(\mathbf{u}^{(1)}_j; 0, \mathbf{G}^{(1)j}_{Z,Z}), \quad (19)$$
$$p(U^{(2)}|X^{(2)}) = \prod_{j=1}^{J} \mathcal{N}(\mathbf{u}^{(2)}_j; 0, \mathbf{G}^{(2)j}_{Z,Z}), \quad (20)$$
$$p(V^{(1)}|X^{(1)}) = \mathcal{N}(V^{(1)}; 0, \mathbf{H}^{(1)}_{Z,Z}), \quad (21)$$
$$p(V^{(2)}|X^{(2)}) = \mathcal{N}(V^{(2)}; 0, \mathbf{H}^{(2)}_{Z,Z}). \quad (22)$$

The conditional distributions for latent variables G and H given the inducing variables U and V are Gaussian, which have the following forms.

$$p(G^{\mathcal{K}}|U^{\mathcal{K}}, X^{\mathcal{K}}) = \prod_{j=1}^{J} \mathcal{N}(\mathbf{g}^{\mathcal{K}}_j; \boldsymbol{\mu}^{\mathcal{K}j}_g, \widetilde{\mathbf{K}}^{\mathcal{K}j}_g), \quad (23)$$
$$p(H^{\mathcal{K}}|V^{\mathcal{K}}, X^{\mathcal{K}}) = \mathcal{N}(H^{\mathcal{K}}; \boldsymbol{\mu}^{\mathcal{K}}_h, \widetilde{\mathbf{K}}^{\mathcal{K}}_h), \quad (24)$$

where $\mathcal{K} \in \{(1),(2)\}$. The specific expressions for the related statistics are $\boldsymbol{\mu}_g^{\mathcal{K}j} = \mathbf{G}_{X,Z}^{\mathcal{K}j}(\mathbf{G}_{Z,Z}^{\mathcal{K}j})^{-1}\mathbf{u}_j^{\mathcal{K}}$, $\widetilde{\mathbf{K}}_g^{\mathcal{K}j} = \mathbf{G}_{X,X}^{\mathcal{K}j} - \mathbf{G}_{X,Z}^{\mathcal{K}j}(\mathbf{G}_{Z,Z}^{\mathcal{K}j})^{-1}\mathbf{G}_{Z,X}^{\mathcal{K}j}$, $\boldsymbol{\mu}_h^{\mathcal{K}} = \mathbf{H}_{X,Z}^{\mathcal{K}}(\mathbf{H}_{Z,Z}^{\mathcal{K}})^{-1}\mathbf{v}^{\mathcal{K}}$ and $\widetilde{\mathbf{K}}_h^{\mathcal{K}} = \mathbf{H}_{X,X}^{\mathcal{K}} - \mathbf{H}_{X,Z}^{\mathcal{K}}(\mathbf{H}_{Z,Z}^{\mathcal{K}})^{-1}\mathbf{H}_{Z,X}^{\mathcal{K}}$.

We now adopt the variational inference method to approximately compute the integral. Specifically, we introduce a joint variational distribution $q(\Omega)$ over all the latent variables denoted by Ω, which has the factorized form as

$$q(\Omega) = q(\Theta^{(1)})q(\Theta^{(2)})q(X^{(1)})q(X^{(2)})q(X^{(1,2)}), \tag{25}$$

where

$$\begin{aligned}
q(X^{(1)}) &= \mathcal{N}(X^{(1)}|\boldsymbol{\mu}^{(1)}, S^{(1)}),\\
q(X^{(2)}) &= \mathcal{N}(X^{(2)}|\boldsymbol{\mu}^{(2)}, S^{(2)}),\\
q(X^{(1,2)}) &= \mathcal{N}(X^{(1,2)}|\boldsymbol{\mu}^{(1,2)}, S^{(1,2)}),\\
q(\Theta^{(1)}) &= p(G^{(1)}|U^{(1)}, X^{(1)})p(H^{(1)}|V^{(1)}, X^{(1)})q(U^{(1)})q(V^{(1)}),\\
q(\Theta^{(2)}) &= p(G^{(2)}|U^{(2)}, X^{(2)})p(H^{(2)}|V^{(2)}, X^{(2)})q(U^{(2)})q(V^{(2)}).
\end{aligned}$$

The evidence lower bound of the logarithmic marginal likelihood $\log p(Y^{(1)}, Y^{(2)})$ is

$$\begin{aligned}
\mathcal{F}_v(q,\theta) &= \int q(\Theta^{(1)})q(X^{(1)}) \log \frac{p(Y^{(1)}|G^{(1)}, H^{(1)})p(G^{(1)}|X^{(1)})p(H^{(1)}|X^{(1)})}{q(\Theta^{(1)})} dG^{(1)}dH^{(1)}dX^{(1)}\\
&+ \int q(\Theta^{(2)})q(X^{(2)}) \log \frac{p(Y^{(2)}|G^{(2)}, H^{(2)})p(G^{(2)}|X^{(2)})p(H^{(2)}|X^{(2)})}{q(\Theta^{(2)})} dG^{(2)}dH^{(2)}dX^{(2)}\\
&- \int q(X^{(1)})q(X^{(2)})q(X^{(1,2)}) \log \frac{q(X^{(1)})q(X^{(2)})q(X^{(1,2)})}{p(X^{(1)}|X^{(1,2)}, \mathbf{t})p(X^{(2)}|X^{(1,2)}, \mathbf{t})p(X^{(1,2)}|\mathbf{t})} dX^{(1,2)}dX^{(1)}dX^{(2)}\\
&= \hat{\mathcal{L}}^{(1)} + \hat{\mathcal{L}}^{(2)} - \mathrm{KL}(q(X^{(1)})q(X^{(2)})q(X^{(1,2)})||p(X^{(1)}|X^{(1,2)}, \mathbf{t})p(X^{(2)}|X^{(1,2)}, \mathbf{t})p(X^{(1,2)}|\mathbf{t})).
\end{aligned} \tag{26}$$

The detailed computation of the first term $\hat{\mathcal{L}}^{(1)}$ in Equation (26) is given by

$$\hat{\mathcal{L}}^{(1)} = \int q(U^{(1)}, V^{(1)})q(X^{(1)}) \log \frac{p(Y^{(1)}|U^{(1)}, V^{(1)}, X^{(1)})p(U^{(1)}, V^{(1)})}{q(U^{(1)}, V^{(1)})} dU^{(1)}dV^{(1)}dX^{(1)}, \tag{27}$$

where $\log p(Y^{(1)}|U^{(1)}, V^{(1)}, X^{(1)})$ in the lower bound can be approximated by

$$\begin{aligned}
\log p(Y^{(1)}|U^{(1)}, V^{(1)}, X^{(1)}) &\geq \langle \log p(Y^{(1)}|G^{(1)}, H^{(1)})\rangle_{p(G^{(1)}, H^{(1)}|U^{(1)}, V^{(1)})}\\
&= \sum_{d=1}^{D} \langle \log p(Y_d^{(1)}|G^{(1)}, H^{(1)})\rangle_{p(G^{(1)}|U^{(1)})p(H^{(1)}|V^{(1)})}\\
&= \sum_{d=1}^{D} \Bigg[\log \mathcal{N}(Y_d^{(1)}|\sum_{j=1}^{J} w_{dj}^{(1)}\boldsymbol{\mu}_g^{(1)j} + \boldsymbol{\mu}_h^{(1)}, (\beta^{(1)})^{-1}I) - \frac{\beta^{(1)}}{2}\mathrm{Tr}(\widetilde{\mathbf{K}}_h^{(1)})\\
&\qquad - \frac{\beta^{(1)}}{2}\mathrm{Tr}(\sum_{j=1}^{J}(w_{dj}^{(1)})^2(\widetilde{\mathbf{K}}_g^{(1)j})) \Bigg].
\end{aligned} \tag{28}$$

As the outputs $Y^{(1)}$ are conditionally independent, the lower bound can be written as a sum of D terms, that is, $\hat{\mathcal{L}}^{(1)} = \sum_{d=1}^{D} \hat{\mathcal{L}}_d^{(1)}$, where $\hat{\mathcal{L}}_d^{(1)}$ is given by

$$\hat{\mathcal{L}}_d^{(1)} = \int q(\mathbf{u}^{(1)}, \mathbf{v}^{(1)}) q(X^{(1)}) \log \frac{\mathcal{N}(\mathbf{y}_d^{(1)} | \sum_{j=1}^{J} w_{dj}^{(1)} \boldsymbol{\mu}_g^{(1)j} + \boldsymbol{\mu}_h^{(1)}, (\beta^{(1)})^{-1} I) p(\mathbf{u}^{(1)}, \mathbf{v}^{(1)})}{q(\mathbf{u}^{(1)}, \mathbf{v}^{(1)})}$$
$$d\mathbf{u}^{(1)} d\mathbf{v}^{(1)} dX^{(1)} - \int \frac{\beta^{(1)}}{2} \mathrm{Tr}(\widetilde{\mathbf{K}}_h^{(1)}) q(X^{(1)}) dX^{(1)} - \int \frac{\beta^{(1)}}{2} \mathrm{Tr}(\sum_{j=1}^{J} (w_{dj}^{(1)})^2 \widetilde{\mathbf{K}}_g^{(1)j}) q(X^{(1)}) dX^{(1)}.$$

By changing the integration order, we get

$$\hat{\mathcal{L}}_d^{(1)} = \int q(\mathbf{u}^{(1)}, \mathbf{v}^{(1)}) \log \frac{e^{\langle \log N(\mathbf{y}_d^{(1)}; \sum_{j=1}^{J} w_{dj}^{(1)} \boldsymbol{\mu}_g^{(1)j} + \boldsymbol{\mu}_h^{(1)}, (\beta^{(1)})^{-1} I) \rangle_{q(X^{(1)})}} p(\mathbf{u}^{(1)}, \mathbf{v}^{(1)})}{q(\mathbf{u}^{(1)}, \mathbf{v}^{(1)})} d\mathbf{u}^{(1)} d\mathbf{v}^{(1)}$$
$$- \frac{\beta^{(1)}}{2} \mathrm{Tr}(\langle \widetilde{\mathbf{K}}_h^{(1)} \rangle_{q(X^{(1)})}) - \frac{\beta^{(1)}}{2} \mathrm{Tr}(\langle \sum_{j=1}^{J} (w_{dj}^{(1)})^2 \widetilde{\mathbf{K}}_g^{(1)j} \rangle_{q(X^{(1)})}), \tag{29}$$

where the optimal variational distribution $q(\mathbf{u}^{(1)}, \mathbf{v}^{(1)})$ for the dth output that gives rise to this lower bound is

$$q(\mathbf{u}^{(1)}, \mathbf{v}^{(1)}) \propto e^{\langle \log \mathcal{N}(\mathbf{y}_d^{(1)}; \sum_{j=1}^{J} w_{dj}^{(1)} \boldsymbol{\mu}_g^{(1)j} + \mu_h^{(1)}, (\beta^{(1)})^{-1} I) \rangle_{q(X^{(1)})}} p(\mathbf{u}^{(1)}, \mathbf{v}^{(1)}). \tag{30}$$

The optimal variational distribution is analytically Gaussian,

$$\begin{aligned} q(\mathbf{u}^{(1)}, \mathbf{v}^{(1)}) =& \mathcal{N}\left(\mathbf{v}^{(1)}; \mathbf{H}_{Z,Z}^{(1)} (\beta^{(1)} \psi_4^{(1)} + \mathbf{H}_{Z,Z}^{(1)})^{-1} (\psi_0^{(1)})^\top \beta^{(1)} \mathbf{y}_d^{(1)}, \mathbf{H}_{Z,Z}^{(1)} (\beta^{(1)} \psi_4^{(1)} + \mathbf{H}_{Z,Z}^{(1)})^{-1} \mathbf{H}_{Z,Z}^{(1)}\right) \\ & \cdot \prod_{j=1}^{J} \mathcal{N}\left(\mathbf{u}_j^{(1)}; \mathbf{G}_{Z,Z}^{(1)j} ((\beta^{(1)} (w_{dj}^{(1)})^2 \psi_5^{(1)j} + \mathbf{G}_{Z,Z}^{(1)j})^{-1} (\psi_1^{(1)j})^\top w_{dj}^{(1)} \beta^{(1)} \mathbf{y}_d^{(1)}, \right. \\ & \left. \mathbf{G}_{Z,Z}^{(1)j} ((\beta^{(1)} (w_{dj}^{(1)})^2 \psi_5^{(1)j} + \mathbf{G}_{Z,Z}^{(1)j})^{-1} \mathbf{G}_{Z,Z}^{(1)j}\right), \end{aligned} \tag{31}$$

where $\psi_0^{(1)} = \langle \mathbf{H}_{X,Z}^{(1)} \rangle_{q(X^{(1)})}$, $\psi_1^{(1)j} = \langle \mathbf{G}_{X,Z}^{(1)j} \rangle_{q(X^{(1)})}$, $\psi_2^{(1)} = \mathrm{Tr}(\langle \mathbf{H}_{X,X}^{(1)} \rangle_{q(X^{(1)})})$, $\psi_3^{(1)j} = \mathrm{Tr}(\langle \mathbf{G}_{X,X}^{(1)j} \rangle_{q(X^{(1)})})$, $\psi_4^{(1)} = \langle \mathbf{H}_{Z,X}^{(1)} \mathbf{H}_{X,Z}^{(1)} \rangle_{q(X^{(1)})}$ and $\psi_5^{(1)j} = \langle \mathbf{G}_{Z,X}^{(1)j} \mathbf{G}_{X,Z}^{(1)j} \rangle_{q(X^{(1)})}$.

Furthermore, the optimal lower bound can be obtained using Jensen's inequality,

$$\begin{aligned} \hat{\mathcal{L}}_d^{(1)} \leq & \log \int e^{\langle \log \mathcal{N}(\mathbf{y}_d^{(1)}; \sum_{j=1}^{J} w_{dj}^{(1)} \boldsymbol{\mu}_g^{(1)j} + \boldsymbol{\mu}_h^{(1)}, (\beta^{(1)})^{-1} I) \rangle_{q(X^{(1)})}} p(\mathbf{u}^{(1)}, \mathbf{v}^{(1)}) d\mathbf{u}^{(1)} d\mathbf{v}^{(1)} \\ & - \frac{\beta^{(1)}}{2} \mathrm{Tr}(\langle \widetilde{\mathbf{K}}_h^{(1)} \rangle_{q(X^{(1)})}) - \frac{\beta^{(1)}}{2} \mathrm{Tr}(\langle \sum_{j=1}^{J} (w_{dj}^{(1)})^2 \widetilde{\mathbf{K}}_g^{(1)j} \rangle_{q(X^{(1)})}) \\ = & \log \left[\frac{(\beta^{(1)})^{\frac{N}{2}} |\mathbf{G}_{Z,Z}^{(1)}|^{\frac{1}{2}} |\mathbf{H}_{Z,Z}^{(1)}|^{\frac{1}{2}}}{(2\pi)^{\frac{N}{2}} |\beta^{(1)} (w_{dj}^{(1)})^2 \psi_5^{(1)} + \mathbf{G}_{Z,Z}^{(1)}|^{\frac{1}{2}} |\beta^{(1)} \psi_4^{(1)} + \mathbf{H}_{Z,Z}^{(1)}|^{\frac{1}{2}}} \exp\{-\frac{1}{2} (\mathbf{y}_d^{(1)})^\top F_d^{(1)} \mathbf{y}_d^{(1)}\} \right] \\ & - \frac{\beta^{(1)}}{2} \mathrm{Tr}(\langle \widetilde{\mathbf{K}}_h^{(1)} \rangle_{q(X^{(1)})}) - \frac{\beta^{(1)}}{2} \mathrm{Tr}(\langle \sum_{j=1}^{J} (w_{dj}^{(1)})^2 \widetilde{\mathbf{K}}_g^{(1)j} \rangle_{q(X^{(1)})}), \end{aligned} \tag{32}$$

where $F_d^{(1)} = \beta^{(1)} I - (\beta^{(1)})^2 (w_{dj}^{(1)})^2 \psi_1^{(1)} (\beta^{(1)} (w_{dj}^{(1)})^2 \psi_5^{(1)} + \mathbf{G}_{Z,Z}^{(1)})^{-1} (\psi_1^{(1)})^\top - (\beta^{(1)})^2 \psi_0^{(1)} (\beta^{(1)} \psi_4^{(1)} + \mathbf{H}_{Z,Z}^{(1)})^{-1} (\psi_0^{(1)})^\top$.

Therefore, the closed-form of the first term $\hat{\mathcal{L}}^{(1)}$ in the lower bound of the approximated logarithmic marginal log-likelihood in Equation (26) is given by

$$\begin{aligned}
\hat{\mathcal{L}}^{(1)} = \sum_{d=1}^{D} \Bigg[& \log \frac{(\beta^{(1)})^{\frac{N}{2}} |\mathbf{H}_{Z,Z}^{(1)}|^{\frac{1}{2}}}{(2\pi)^{\frac{N}{2}} |\beta^{(1)} \psi_4^{(1)} + \mathbf{H}_{Z,Z}^{(1)}|^{\frac{1}{2}}} + \sum_{j=1}^{J} \log \frac{|\mathbf{G}_{Z,Z}^{(1)j}|^{\frac{1}{2}}}{|\beta^{(1)} (w_{dj}^{(1)})^2 \psi_5^{(1)j} + \mathbf{G}_{Z,Z}^{(1)j}|^{\frac{1}{2}}} \\
& - \frac{1}{2} (\mathbf{y}_d^{(1)})^\top \Bigg(\beta^{(1)} I - \sum_{j=1}^{J} (\beta^{(1)})^2 (w_{dj}^{(1)})^2 \psi_1^{(1)j} (\beta^{(1)} (w_{dj}^{(1)})^2 \psi_5^{(1)j} + \mathbf{G}_{Z,Z}^{(1)j})^{-1} (\psi_1^{(1)j})^\top - (\beta^{(1)})^2 \psi_0^{(1)} \\
& (\beta^{(1)} \psi_4^{(1)} + \mathbf{H}_{Z,Z}^{(1)})^{-1} (\psi_0^{(1)})^\top \Bigg) \mathbf{y}_d^{(1)} - \frac{\beta^{(1)}}{2} \psi_2^{(1)} + \frac{\beta^{(1)}}{2} \mathrm{Tr}(\psi_4^{(1)} (\mathbf{H}_{Z,Z}^{(1)})^{-1}) - \frac{\beta^{(1)}}{2} \sum_{j=1}^{J} (w_{dj}^{(1)})^2 \psi_3^{(1)j} \\
& + \frac{\beta^{(1)}}{2} \sum_{j=1}^{J} \mathrm{Tr}((w_{dj}^{(1)})^2 \psi_5^{(1)j} (\mathbf{G}_{Z,Z}^{(1)j})^{-1}) \Bigg],
\end{aligned} \tag{33}$$

and similarly for $\hat{\mathcal{L}}^{(2)}$,

$$\begin{aligned}
\hat{\mathcal{L}}^{(2)} = \sum_{d=1}^{D} \Bigg[& \log \frac{(\beta^{(2)})^{\frac{N}{2}} |\mathbf{H}_{Z,Z}^{(2)}|^{\frac{1}{2}}}{(2\pi)^{\frac{N}{2}} |\beta^{(2)} \psi_4^{(2)} + \mathbf{H}_{Z,Z}^{(2)}|^{\frac{1}{2}}} + \sum_{j=1}^{J} \log \frac{|\mathbf{G}_{Z,Z}^{(2)j}|^{\frac{1}{2}}}{|\beta^{(2)} (w_{dj}^{(2)})^2 \psi_5^{(2)j} + \mathbf{G}_{Z,Z}^{(2)j}|^{\frac{1}{2}}} \\
& - \frac{1}{2} (\mathbf{y}_d^{(2)})^\top \Bigg(\beta^{(2)} I - \sum_{j=1}^{J} (\beta^{(2)})^2 (w_{dj}^{(2)})^2 \psi_1^{(2)j} (\beta^{(2)} (w_{dj}^{(2)})^2 \psi_5^{(2)j} + \mathbf{G}_{Z,Z}^{(2)j})^{-1} (\psi_1^{(2)j})^\top - (\beta^{(2)})^2 \psi_0^{(2)} \\
& (\beta^{(2)} \psi_4^{(2)} + \mathbf{H}_{Z,Z}^{(2)})^{-1} (\psi_0^{(2)})^\top \Bigg) \mathbf{y}_d^{(2)} - \frac{\beta^{(2)}}{2} \psi_2^{(2)} + \frac{\beta^{(2)}}{2} \mathrm{Tr}(\psi_4^{(2)} (\mathbf{H}_{Z,Z}^{(2)})^{-1}) - \frac{\beta^{(2)}}{2} \sum_{j=1}^{J} (w_{dj}^{(2)})^2 \psi_3^{(2)j} \\
& + \frac{\beta^{(2)}}{2} \sum_{j=1}^{J} \mathrm{Tr}((w_{dj}^{(2)})^2 \psi_5^{(2)j} (\mathbf{G}_{Z,Z}^{(2)j})^{-1}) \Bigg].
\end{aligned} \tag{34}$$

For the calculation of KL divergence, for simplification, we employ $A_q^{(1)}$ and $A_q^{(2)}$ to represent $(\alpha^{(1)})^2 \mathbf{K}_{\mathbf{t},\mathbf{t}}^{(1)} + (1-\alpha^{(1)})^2 \epsilon^{(1)} I$ and $(\alpha^{(2)})^2 \mathbf{K}_{\mathbf{t},\mathbf{t}}^{(2)} + (1-\alpha^{(2)})^2 \epsilon^{(2)} I$, respectively. Then the specific calculation is given below.

$$\begin{aligned}
& \mathrm{KL}(q(X^{(1)}) q(X^{(2)}) q(X^{(1,2)}) || p(X^{(1)} | X^{(1,2)}, \mathbf{t}) p(X^{(2)} | X^{(1,2)}, \mathbf{t}) p(X^{(1,2)} | \mathbf{t})) \\
= & \frac{1}{2} \sum_{q=1}^{Q} \Bigg[\left[\log |A_q^{(1)}| + \log |A_q^{(2)}| + \log |\mathbf{K}_{\mathbf{t},\mathbf{t}}^{(1,2)}| - \log |S_q^{(1,2)}| - \log |S_q^{(1)}| - \log |S_q^{(2)}| \right] \\
& + \left[(1-\alpha^{(1)}) \boldsymbol{\mu}_q^{(1,2)} - \boldsymbol{\mu}_q^{(1)} \right]^\top (A_q^{(1)})^{-1} \left[(1-\alpha^{(1)}) \boldsymbol{\mu}_q^{(1,2)} - \boldsymbol{\mu}_q^{(1)} \right] \\
& + \left[(1-\alpha^{(2)}) \boldsymbol{\mu}_q^{(1,2)} - \boldsymbol{\mu}_q^{(2)} \right]^\top (A_q^{(2)})^{-1} \left[(1-\alpha^{(2)}) \boldsymbol{\mu}_q^{(1,2)} - \boldsymbol{\mu}_q^{(2)} \right] \\
& + \mathrm{Tr}\left[\left[(1-\alpha^{(1)})^2 (A_q^{(1)})^{-1} + (1-\alpha^{(2)})^2 (A_q^{(2)})^{-1} \right] S_q^{(1,2)} \right] \\
& + \mathrm{Tr}\left[(\mathbf{K}_{\mathbf{t},\mathbf{t}}^{(1,2)})^{-1} \left[\boldsymbol{\mu}_q^{(1,2)} (\boldsymbol{\mu}_q^{(1,2)})^\top + S_q^{(1,2)} \right] \right] + \mathrm{Tr}\left[(A_q^{(1)})^{-1} S_q^{(1)} + (A_q^{(2)})^{-1} S_q^{(2)} \right] \Bigg].
\end{aligned} \tag{35}$$

Appendix B. Computation of Statistics $\psi_0, \psi_1, \psi_2, \psi_3, \psi_4, \psi_5$

$\psi_0^{(1)}$, $\psi_1^{(1)}$, $\psi_0^{(2)}$ and $\psi_1^{(2)}$ are $N \times M$ matrices. $\psi_2^{(1)}$, $\psi_3^{(1)}$, $\psi_2^{(2)}$ and $\psi_3^{(2)}$ are scalars. $\psi_4^{(1)}$, $\psi_5^{(1)}$, $\psi_4^{(2)}$, $\psi_5^{(2)}$ are $(J \times M) \times (J \times M)$ matrices. we use the ARD kernel $\kappa_{ARD}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp(-\frac{1}{2}\sum_{q=1}^{Q} \alpha_p(\mathbf{x}_q - \mathbf{x}'_q)^2)$, and obtain

$$
\begin{aligned}
(\psi_0^{(1)})_{n,m} &= (\langle \mathbf{H}_{X,Z}^{(1)} \rangle_{q(X^{(1)})})_{n,m} = \int \kappa^{(1)^h}(\mathbf{x}_n^{(1)}, \mathbf{z}_m^{(1)^h}) \mathcal{N}(\mathbf{x}_n^{(1)} | \boldsymbol{\mu}_n^{(1)}, \mathbf{S}_n^{(1)}) d\mathbf{x}_n^{(1)} \\
&= \frac{(\sigma_f^2)^{(1)^h}}{\prod_{q=1}^{Q} (S_{nq}^{(1)} \alpha_q^{(1)^h} + 1)^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \sum_{q=1}^{Q} \frac{(\mathrm{z}_{mq}^{(1)^h} - \mu_{nq}^{(1)})^2 \alpha_q^{(1)^h}}{S_{nq}^{(1)} \alpha_q^{(1)^h} + 1} \right), \qquad (36) \\
(\psi_1^{(1)^j})_{n,m} &= (\langle \mathbf{G}_{X,Z}^{(1)^j} \rangle_{q(X^{(1)})})_{n,m} = \int \kappa_j^{(1)}(\mathbf{x}_n^{(1)}, \mathbf{z}_m^{(1)}) \mathcal{N}(\mathbf{x}_n^{(1)} | \boldsymbol{\mu}_n^{(1)}, \mathbf{S}_n^{(1)}) d\mathbf{x}_n^{(1)} \\
&= \frac{(\sigma_f^2)_j^{(1)}}{\prod_{q=1}^{Q} (S_{nq}^{(1)} \alpha_{jq}^{(1)} + 1)^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \sum_{q=1}^{Q} \frac{(z_{mq}^{(1)} - \mu_{nq}^{(1)})^2 \alpha_{jq}^{(1)}}{S_{nq}^{(1)} \alpha_{jq}^{(1)} + 1} \right), \qquad (37) \\
\psi_2^{(1)} &= \mathrm{Tr}(\langle \mathbf{H}_{X,X}^{(1)} \rangle_{q(X^{(1)})}) = N(\sigma_f^2)^{(1)^h}, \qquad (38) \\
\psi_3^{(1)^j} &= \mathrm{Tr}(\langle \mathbf{G}_{X,X}^{(1)^j} \rangle_{q(X^{(1)})}) = N(\sigma_f^2)_j^{(1)}, \qquad (39) \\
(\psi_4^{(1)})_{m,m'} &= (\langle \mathbf{H}_{Z,X}^{(1)} \mathbf{H}_{X,Z}^{(1)} \rangle_{q(X^{(1)})})_{m,m'} \\
&= \sum_{n=1}^{N} \int k^{(1)^h}(\mathbf{x}_n^{(1)}, \mathbf{z}_m^{(1)^h}) k^{(1)^h}(\mathbf{x}_n^{(1)}, \mathbf{z}_{m'}^{(1)^h}) \mathcal{N}(\mathbf{x}_n^{(1)} | \boldsymbol{\mu}_n^{(1)}, \mathbf{S}_n^{(1)}) d\mathbf{x}_n^{(1)} \\
&= (\sigma_f^4)^{(1)^h} \sum_{n=1}^{N} \prod_{q=1}^{Q} \exp \left(\frac{-\frac{\alpha_q^{(1)^h} (\mathbf{z}_{mq}^{(1)^h} - \mathbf{z}_{m'q}^{(1)^h})^2}{4} - \frac{\alpha_q^{(1)^h} (\boldsymbol{\mu}_{np}^{(1)} - \frac{\mathbf{z}_{mq}^{(1)^h}}{2} - \frac{\mathbf{z}_{m'q}^{(1)^h}}{2})^2}{2\alpha_q^{(1)^h} S_{nq}^{(1)} + 1}}{(2\alpha_q^{(1)^h} S_{nq}^{(1)} + 1)^{\frac{1}{2}}} \right), \qquad (40) \\
(\psi_5^{(1)^j})_{m,m'} &= (\langle \mathbf{G}_{Z,X}^{(1)^j} \mathbf{G}_{X,Z}^{(1)^j} \rangle_{q(X^{(1)})})_{m,m'} \\
&= \sum_{n=1}^{N} \int k_j^{(1)}(\mathbf{x}_n^{(1)}, \mathbf{z}_m^{(1)}) k_j^{(1)}(\mathbf{x}_n^{(1)}, \mathbf{z}_{m'}^{(1)}) \mathcal{N}(\mathbf{x}_n^{(1)} | \boldsymbol{\mu}_n^{(1)}, \mathbf{S}_n^{(1)}) d\mathbf{x}_n^{(1)} \\
&= (\sigma_f^4)_j^{(1)} \sum_{n=1}^{N} \prod_{q=1}^{Q} \exp \left(\frac{-\frac{\alpha_{jq}^{(1)} (\mathbf{z}_{mq}^{(1)} - \mathbf{z}_{m'q}^{(1)})^2}{4} - \frac{\alpha_{jq}^{(1)} (\boldsymbol{\mu}_{nq}^{(1)} - \frac{\mathbf{z}_{mq}^{(1)}}{2} - \frac{\mathbf{z}_{m'q}^{(1)}}{2})^2}{2\alpha_{jq}^{(1)} S_{nq}^{(1)} + 1}}{(2\alpha_{jq}^{(1)} S_{nq}^{(1)} + 1)^{\frac{1}{2}}} \right). \qquad (41)
\end{aligned}
$$

The statistics $\psi_0^{(2)}, \psi_1^{(2)}, \psi_2^{(2)}, \psi_3^{(2)}, \psi_4^{(2)}, \psi_5^{(2)}$ in the second view have the similar formulas.

Appendix C. Derivation of Varitional Lower Bound for Testing

Given test data in the first view $Y_*^{(1)}$, we maximize a variational lower bound on the logarithmic marginal likelihood $\log p(Y^{(1)}, Y_*^{(1)})$ which can be expressed as follows. For

brevity, we've omitted time $\mathbf{t}$ and $\mathbf{t}_*$.

$$\begin{aligned}
\mathcal{F}_*^{(1)} &= \log \int p(Y_*^{(1)}, Y^{(1)}|X_*^{(1)}, X^{(1)}) p(X_*^{(1)}, X^{(1)}|X_*^{(1,2)}, X^{(1,2)}) p(X_*^{(1,2)}, X^{(1,2)}) \\
&\quad dX_*^{(1,2)} dX_*^{(1)} dX^{(1)} dX^{(1,2)} \\
&\geq \int q(X_*^{(1)}, X^{(1)}) q(X_*^{(1,2)}, X^{(1,2)}) q(G^{(1)}) q(H^{(1)}) \\
&\quad \log \frac{p(Y_*^{(1)}, Y^{(1)}|X_*^{(1)}, X^{(1)}) p(X_*^{(1)}, X^{(1)}|X_*^{(1,2)}, X^{(1,2)}) p(X_*^{(1,2)}, X^{(1,2)})}{q(X_*^{(1)}, X^{(1)}) q(X_*^{(1,2)}, X^{(1,2)}) q(G^{(1)}) q(H^{(1)})} \\
&\quad dX_*^{(1,2)} dX_*^{(1)} dX^{(1)} dX^{(1,2)} dG^{(1)} dH^{(1)} \\
&= \int q(G^{(1)}) q(H^{(1)}) q(X_*^{(1)}, X^{(1)}) \log \frac{p(Y_*^{(1)}, Y_*^{(1)}|X_*^{(1)}, X^{(1)})}{q(G^{(1)}) q(H^{(1)})} dX_*^{(1)} dX^{(1)} dG^{(1)} dH^{(1)} \\
&+ \int q(X_*^{(1)}, X^{(1)}) q(X_*^{(1,2)}, X^{(1,2)}) \log \frac{p(X_*^{(1)}, X^{(1)}|X_*^{(1,2)}, X^{(1,2)}) p(X_*^{(1,2)}, X^{(1,2)})}{q(X_*^{(1)}, X^{(1)}) q(X_*^{(1,2)}, X^{(1,2)})} \\
&\quad dX_*^{(1)} dX_*^{(1,2)} dX^{(1)} dX^{(1,2)} \\
&= \widetilde{\mathcal{L}}^{(1)}(Y_*^{(1)}, Y^{(1)}) - \mathrm{KL}\Big[q(X_*^{(1)}, X^{(1)}) q(X_*^{(1,2)}, X^{(1,2)}) || p(X_*^{(1)}, X^{(1)}|X_*^{(1,2)}, X^{(1,2)}) \\
&\quad p(X_*^{(1,2)}, X^{(1,2)}) \Big], \qquad (42)
\end{aligned}$$

The quantity $\mathcal{F}_*^{(1)}$ can be maximized using the same method as for training. In addition, parameters of the new variational distribution $q(X^{(1)}, X_*^{(1)})$ are jointly optimized because $X^{(1)}$ and $X_*^{(1)}$ are coupled in $q(X^{(1)}, X_*^{(1)})$, and so are $q(X^{(1,2)}, X_*^{(1,2)})$. Specially, the quantity $\widetilde{\mathcal{L}}^{(1)}(Y_*^{(1)}, Y^{(1)})$ can be expressed as

$$\begin{aligned}
&\widetilde{\mathcal{L}}^{(1)}(Y_*^{(1)}, Y^{(1)}) = \sum_{d=1}^{D} \Bigg[\log \frac{{\beta^{(1)}}^{\frac{N+N_*}{2}} |\mathbf{H}_{Z,Z}^{(1)}|^{\frac{1}{2}}}{(2\pi)^{\frac{N+N_*}{2}} |\beta^{(1)} \tilde{\psi_4}^{(1)} + \mathbf{H}_{Z,Z}^{(1)}|^{\frac{1}{2}}} + \sum_{j=1}^{J} \log \frac{|\mathbf{G}_{Z,Z}^{(1)j}|^{\frac{1}{2}}}{|\beta^{(1)} (w_{dj}^{(1)})^2 \tilde{\psi_5}^{(1)j} + \mathbf{G}_{Z,Z}^{(1)j}|^{\frac{1}{2}}} \\
&- \frac{1}{2} (\tilde{\mathbf{y}}_{\mathbf{d}}^{(1)})^\top \Bigg(\beta^{(1)} I - \sum_{j=1}^{J} {\beta^{(1)}}^2 (w_{dj}^{(1)})^2 \tilde{\psi_1}^{(1)j} (\beta^{(1)} (w_{dj}^{(1)})^2 \tilde{\psi_5}^{(1)j} + \mathbf{G}_{Z,Z}^{(1)j})^{-1} (\tilde{\psi_1}^{(1)j})^\top - {\beta^{(1)}}^2 \tilde{\psi_0} \\
&(\beta^{(1)} \tilde{\psi_4}^{(1)} + \mathbf{H}_{Z,Z}^{(1)})^{-1} (\tilde{\psi_0})^\top \Bigg) \tilde{\mathbf{y}}_{\mathbf{d}}^{(1)} - \frac{\beta^{(1)}}{2} \tilde{\psi_2} + \frac{\beta^{(1)}}{2} \mathrm{Tr}(\tilde{\psi_4}^{(1)} (\mathbf{H}_{Z,Z}^{(1)})^{-1}) - \frac{\beta^{(1)}}{2} \sum_{j=1}^{J} \tilde{\psi_3}^{(1)j} \\
&+ \frac{\beta^{(1)}}{2} \sum_{j=1}^{J} \mathrm{Tr}((w_{dj}^{(1)})^2 \tilde{\psi_5}^{(1)j} (\mathbf{G}_{Z,Z}^{(1)j})^{-1}) \Bigg], \qquad (43)
\end{aligned}$$

and the KL divergence can be expressed as

$$\begin{aligned}
&\mathrm{KL}\Big[q(X^{(1)}, X_*^{(1)}) q(X^{(1,2)}, X_*^{(1,2)}) || p(X^{(1)}, X_*^{(1)}|X^{(1,2)}, X_*^{(1,2)}) p(X^{(1,2)}, X_*^{(1,2)}) \Big] \\
&= \frac{1}{2} \sum_{q=1}^{Q} \Big[\log |\tilde{A}_q^{(1)}| + \log |\tilde{\mathbf{K}}_{\mathbf{t},\mathbf{t}}^{(1,2)}| - \log |\tilde{S}_q^{(1,2)}| - \log |\tilde{S}_q^{(1)}| + \mathrm{Tr}\big[(\tilde{A}_q^{(1)})^{-1} \tilde{S}_q^{(1)} \big]
\end{aligned}$$

$$+ \quad \left[(1-\alpha^{(1)})\tilde{\boldsymbol{\mu}}_q^{(1,2)} - \tilde{\boldsymbol{\mu}}_q^{(1)}\right]^\top (\tilde{A}_q^{(1)})^{-1}\left[(1-\alpha^{(1)})\tilde{\boldsymbol{\mu}}_q^{(1,2)} - \tilde{\boldsymbol{\mu}}_q^{(1)}\right]$$
$$+ \quad \mathrm{Tr}\left[(1-\alpha^{(1)})^2(\tilde{A}_q^{(1)})^{-1}\tilde{S}_q^{(1,2)}\right] + \mathrm{Tr}\left[(\tilde{\mathbf{K}}_{\mathbf{x},\mathbf{x}}^{(1,2)})^{-1}\left[\tilde{\boldsymbol{\mu}}_q^{(1,2)}(\tilde{\boldsymbol{\mu}}_q^{(1,2)})^\top + \tilde{S}_q^{(1,2)}\right]\right]\Bigg]. \tag{44}$$

where $\tilde{\psi_0}^{(1)} = \langle \mathbf{H}_{X,Z}^{(1)} \rangle_{q(X^{(1)},X_*^{(1)})}$, $\tilde{\psi}_1^{(1)j} = \langle \mathbf{G}_{X,Z}^{(1)j} \rangle_{q(X^{(1)},X_*^{(1)})}$, $\tilde{\psi}_2^{(1)} = \mathrm{Tr}(\langle \mathbf{H}_{X,X}^{(1)} \rangle_{q(X^{(1)},X_*^{(1)})})$, $\tilde{\psi}_3^{(1)j} = \mathrm{Tr}(\langle \mathbf{G}_{X,X}^{(1)j} \rangle)_{q(X^{(1)},X_*^{(1)})}$, $\tilde{\psi_4}^{(1)} = \langle \mathbf{H}_{Z,X}^{(1)} \mathbf{H}_{X,Z}^{(1)} \rangle_{q(X^{(1)},X_*^{(1)})}$ and $\tilde{\psi_5}^{(1)j} = \langle \mathbf{G}_{Z,X}^{(1)j} \mathbf{G}_{X,Z}^{(1)j} \rangle_{q(X^{(1)},X_*^{(1)})}$.

References

M. R. Amini and C. Goutte. A co-classification approach to learning from multilingual corpora. *Machine Learning*, 79:105–121, 2010.

K. Andreas and G. Carlos. Nonmyopic active learning of Gaussian processes: an exploration-exploitation approach. In *Proceedings of the 24th International Conference on Machine Learning*, pages 449–456, 2007.

A. Ankur and T. Bill. Recovering 3D human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:1–8, 2006.

K. Barnard, P. Duygulu, D. Forsyth, D. Nando N. Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

D. M. Blei and M. I. Jordan. Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 127–134, 2003.

T. Cao, V. Jojic, S. Modla, D. Powell, K. Czymmek, and M. Niethammer. Robust multimodal dictionary learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 259–266, 2013.

N. Chen, J. Zhu, and E. P. Xing. Predictive subspace learning for multi-view data: a large margin approach. *Advances in Neural Information Processing Systems*, 23:361–369, 2010.

D. A. Cohn and T. Hofmann. The missing link – A probabilistic model of document content and hypertext connectivity. *Advances in Neural Information Processing Systems*, 14: 430–436, 2001.

A. C. Damianou, M. K. Titsias, and N. D. Lawrence. Variational Gaussian process dynamical systems. *Advances in Neural Information Processing Systems*, 24:2510–2518, 2011.

A. C. Damianou, C. H. Ek, M. K. Titsias, and N. D. Lawrence. Manifold relevance determination. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1–8, 2012.

A. C. Damianou, M. K. Titsias, and N. D. Lawrence. Variational inference for latent variables and uncertain inputs in Gaussian processes. *Journal of Machine Learning Research*, 17: 1425–1486, 2016.

Z. Ding, M. Shao, and Y. Fu. Robust multi-view representation: A unified perspective from multi-view learning to domain adaption. In *Proceedings of the 27th International Joint Conferences on Artificial Intelligence*, pages 5434–5440, 2018.

J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634, 2015.

C. H. Ek and N. D. Lawrence. *Shared Gaussian process latent variable models*. PhD thesis, Oxford Brookes University, 2009.

F. Feng, X. Wang, and R. Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd International Conference on Multimedia*, pages 7–16, 2014.

M. Feurer, B. Letham, and E. Bakshy. Scalable meta-learning for Bayesian optimization using ranking-weighted Gaussian process ensembles. In *Proceedings of the 36th Automatic Machine Learning Workshop at International Conference on Machine Learning*, pages 1–15, 2018.

J. Hu, J. Lu, and Y. Tan. Sharable and individual multi-view metric learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:2281–2288, 2018.

Y. Jia, M. Salzmann, and T. Darrell. Factorized latent spaces with structured sparsity. *Advances in Neural Information Processing Systems*, 23:982–990, 2010.

Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In *Proceedings of the 13th IEEE International Conference on Computer Vision*, pages 2407–2414, 2011.

P. Jing, Y. Su, L. Nie, X. Bai, J. Liu, and M. Wang. Low-rank multi-view embedding learning for micro-video popularity prediction. *IEEE Transactions on Knowledge and Data Engineering*, 30:1519–1532, 2018.

A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

N. D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in Neural Information Processing Systems*, 17:329–336, 2004.

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

N. D. Lawrence and I. M. Jordan. Semi-supervised learning via Gaussian processes. *Advances in Neural Information Processing Systems*, 18:753–760, 2005.

Y. Li, M. Yang, and Z. M. Zhang. A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 10:1–20, 2018.

W. Liu, D. Tao, J. Cheng, and Y. Tang. Multiview Hessian discriminative sparse coding for image annotation. *Computer Vision and Image Understanding*, 118:50–60, 2014.

M. Lüthi, T. Gerig, C. Jud, and T. Vetter. Gaussian process morphable models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:1860–1873, 2018.

J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks. *arXiv preprint , arXiv:1412.6632*, pages 1–17, 2014.

J. R. Medina, H. Borner, S. Endo, and S. Hirche. Impedance-based Gaussian processes for modeling human motor behavior in physical and non-physical interaction. *IEEE Transactions on Biomedical Engineering*, 63:1–12, 2019.

I. Muslea, S. Minton, and C. A. Knoblock. Active + semi-supervised learning = robust multi-view learning. In *Proceedings of the 19th International Conference on Machine Learning*, pages 435–442, 2002.

I. Muslea, S. Minton, and C. A. Knoblock. Active learning with multiple views. *Journal of Artificial Intelligence Research*, 27:203–233, 2006.

J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 689–696, 2011.

V. T. Nguyen and E. Bonilla. Collaborative multi-output Gaussian processes. In *Proceedings of the 30th Uncertainty in Artificial Intelligence*, pages 643–652, 2014.

M. Opper and C. Archambeau. The variational Gaussian approximation revisited. *Neural Computation*, 21:786–792, 2009.

E. Puyol, B. Ruijsink, B. Gerber, M. Amzulescu, H. Langet, M. De Craene, J. A. Schnabel, P. Piro, and A. P. King. Regional multi-view learning for cardiac motion analysis: Application to identification of dilated cardiomyopathy patients. *IEEE Transactions on Biomedical Engineering*, 65:1–9, 2018.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Process for Machine Learning*. MIT Press, 2nd edition, 2006.

M. Salzmann, C. H. Ek, R. Urtasun, and T. Darrell. Factorized orthogonal latent spaces. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pages 701–708, 2010.

A. Shon, K. Grochow, A. Hertzmann, and R. P. Rao. Learning shared latent structure for image synthesis and robotic imitation. *Advances in Neural Information Processing Systems*, 18:1233–1240, 2006.

N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep Boltzmann machines. *Advances in Neural Information Processing Systems*, 25:2222–2230, 2012.

S. Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23: 2031–2038, 2013.

S. Sun, L. Mao, Z. Dong, and L. Wu. *Multiview Machine Learning*. Springer, 1st edition, 2019.

M. K. Titsias and N. D. Lawrence. Bayesian Gaussian process latent variable model. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pages 844–851, 2010.

S. Tulsiani, A. Efros, and J. Malik. Multi-view consistency as supervisory signal for learning shape and pose prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2018.

S. Virtanen, Y. Jia, A. Klami, and T. Darrell. Factorized multi-modal topic model. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pages 1–9, 2012.

X. Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 235–243, 2009.

J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models. *Advances in Neural Information Processing Systems*, 19:1441–1448, 2006.

W. Wang, R. Arora, K. Livescu, and J. Bilmes. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1083–1092, 2015.

H. Wei, P. Zhu, M. Liu, J. P. How, and S. Ferrari. Automatic pan–tilt camera control for learning Dirichlet process Gaussian process mixture models of multiple moving targets. *IEEE Transactions on Automatic Control*, 64:159–173, 2019.

X. Wei, H. Huang, L. Nie, F. Feng, R. Hong, and T. Chua. Quality matters: Assessing cQA pair quality via transductive multi-view learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4482–4488, 2018.

E. P. Xing, R. Yan, and A. G. Hauptmann. Mining associated text and images with dual-wing harmoniums. *arXiv preprint , arXiv:1207.1423*, pages 1–9, 2012.

C. Xu, D. Tao, and C. Xu. A survey on multi-view learning. *arXiv preprint , arXiv:1304.5634*, pages 1–59, 2013.

C. Zhang, C. H. Ek, A. Damianou, and H. Kjellstrom. Factorized topic models. In *Proceedings of the 1st International Conference on Learning Representations*, pages 1–9, 2013.

J. Zhao and S. Sun. Variational dependent multi-output Gaussian process dynamical systems. *Journal of Machine Learning Research*, 17:1–36, 2016.

J. Zhao, J. Fei, and S. Sun. A variant of Gaussian process dynamical systems. Technical report, East China Normal University, 2018.