CLASS-AWARE SHARED GAUSSIAN PROCESS DYNAMIC MODEL

Ryosuke Sawata[†]*, Takahiro Ogawa*[‡] *and Miki Haseyama*[‡]

[†]Graduate School of Information Science and Technology, Hokkaido University
[‡]Faculty of Information Science and Technology, Hokkaido University
N-14, W-9, Kita-ku, Sapporo, Hokkaido, 060-0814, Japan
E-mail: {sawata, ogawa, mhaseyama}@lmd.ist.hokudai.ac.jp

ABSTRACT

A new method of Gaussian process dynamic model (GPDM), named class-aware shared GPDM (CSGPDM), is presented in this paper. One of the most difference between our CSGPDM and existing GPDM is considering class information which helps to build the class label-based latent space being effective for the following class-related tasks. In terms of representation learning, CSGPDM is optimized by considering not only a non-linear relationship but also time-series relation and discriminative information of each class label. Then CSGPDM can reflect the following three points to the estimated latent space: i) the relationship between heterogeneous input sets, ii) time-series relations lurked in each input data, and iii) class information. Therefore, when input heterogeneous sets of features have time-series relations and class information, the above CSGPDM-based latent space can be beneficial for the obtaining the new CSGPDM-based feature sets for the post classification and estimating one side of the lacking samples by bridging the input heterogeneous feature sets via the latent space. Experimental results show that the estimated CSGPDM-based latent space outperformed those of GPDM and shared GPDM (SGPDM).

Index Terms— Gaussian process, discriminant analysis, representation learning, canonical correlation analysis

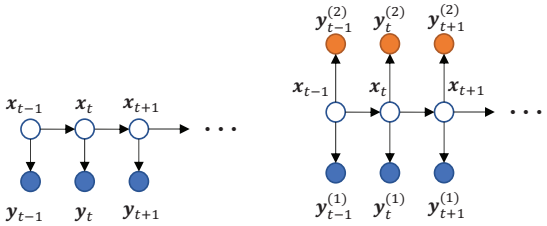
1. INTRODUCTION

Due to the development of infrastructure of information technology, we have become able to access a huge amount of and diverse datasets such as image [1, 2], speech [3–5], music [6, 7], text [8–11] and biosignal [12, 13]. In particular, the dataset consisting of more than two modalities, i.e., the multi-modal dataset, such as video [14, 15], image or sound with text captions [16, 17], etc. and the demands for accessing those have been increasing more recently. Hence, it it necessary and beneficial for us to exploit the method which can deal with multi-modal inputs effectively.

In order to handle the multi-modal inputs, canonical correlation analysis (CCA) [18] is one of the well-known and powerful methods. Specifically, CCA can estimate the low-dimensional latent space so that its space reflects the relationship between multi-modal high-dimensional inputs as the maximized correlation coefficient. It is known that the new estimated features via latent space, i.e., latent features, and the corresponding correlation which is implicitly lurked in the input multi-modal sets are beneficial for some tasks. Therefore, there are many applications utilizing CCA [19–24] and its extended versions [25–31]. However, the existing CCAs have two problems as far as we know. First, it is difficult to determine the appropriate parameters which are necessary for some terms, e.g., the kernel functions and regularization. In fact, they had to apply the grid search [32] in order to find the effective values of required parameters [20,24], and there is no guarantee that the found parameters are optimal even after applying the grid search. Second, there is no way to decide the best number of latent space's dimensions. In other words, we have to manually decide the number of the dimensions, and thus we probably lose some beneficial dimensions for the post tasks.

To resolve the above problems, we focus on the Gaussian process (GP)-based models. This is because they can simultaneously optimize the required parameters by monitoring the value of the likelihood function [33] when estimating the latent space. Furthermore, GP-based models can build a latent space so that its dimension is optimized the number decided in advance, and thus it is not necessary to omit some dimensions manually after building the latent space like CCA. We especially focus on one of the well-known GP-based method to estimate the low-dimensional latent space for the high-dimensional input, called Gaussian process latent variable model (GPLVM) [34]. Since GPLVM has been identified to be an effective probabilistic approach for dimensionality reduction, there are many types of its extended models like the aforementioned discussions regarding CCAs: shared GPLVM (SGPLVM) [35], GP dynamics model (GPDM) [36], discriminative GPLVM (DGLVM) [37], supervised GPLVM [38], etc. SGPLVM estimates the low-dimensional latent space which is shared by the heterogeneous input feature sets. In other words, SGPLVM can receive the multi-modal input sets like CCA comparing with GPLVM. Furthermore, Wang et al. [36] proposed GPDM which can reduce the dimensions of input considering the dynamics of input, i.e., time-series relation. In addition, Urtasun et al. [37] and Gao et al. [38] proposed new GPLVMs which rewrites the original cost functions so that the labels denoting the class information can be considered simultaneously. However, as far as we know, there is no GP-based method which can simultaneously consider all of the above characteristics, i.e., a) multi-modal inputs, b) time-series relation and c) class information.

Motivated by the aforementioned discussions, we newly propose class-aware shared GPDM (CSGPDM) in this paper. Specifically, CSGPDM let the multi-modal inputs share the low-dimensional latent space considering the time-series relation and class information lurked in each input, simultaneously. While Urtasun et al. [37] and Gao et al. [38] added class label-driven prior term to the cost function, our class-aware extension is achieved by merely rewriting the existing GPLVM-based prior term. Hence, our method which makes the target model class-aware has the following two contributions: i) it is applicable for almost all GPLVM-based models (e.g., GPDM, SGPLVM and supervised GPLVM) and ii) it does not increase the parameter to train, i.e., calculation cost, because it is achieved by merely rewriting without adding a new prior term. Since CSGPDM



(a) GPDM (b) Shared GPDM (SGPDM)

Fig. 1: Graphical models which this paper focuses on: GPDM and SGPDM. Note that our extended version of only GPDM and SGPDM are actually exploited in this paper, although our proposal can be applicable for almost all GPLVM-based methods.

not only solves the aforementioned CCAs' problems but also considers all of the above characteristics, obtaining the more effective CSGPDM-based latent features for the multi-modal inputs which have class information and time-series relation becomes feasible in our method.

2. CLASS-AWARE EXTENSION

In this section, we assume that D-dimensional observations $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_T]^{\mathrm{T}} \in \mathbb{R}^{T \times D}$ are generated by K-dimensional latent variables $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T]^{\mathrm{T}} \in \mathbb{R}^{T \times K}$, where each index $t\,(= 1, 2, \cdots, T)$ denotes the discrete-time index.

2.1. Brief Review of GPDM

As shown in Fig. 1(a), GPDM assumes the current latent variable is depending on the previous one. Specifically, GPDM considers the following Markov dynamics, i.e., time-series relation:

$$\boldsymbol{x}_t = f(\boldsymbol{x}_{t-1}; \boldsymbol{A}) + \boldsymbol{n}_{x,t}, \quad \boldsymbol{y}_t = g(\boldsymbol{x}_t; \boldsymbol{B}) + \boldsymbol{n}_{y,t}, \tag{1}$$

where $f(\bullet)$ and $g(\bullet)$ are respectively mappings with parameters $\boldsymbol{A}$ and $\boldsymbol{B}$. Furthermore, $\boldsymbol{n}_{x,t}$ and $\boldsymbol{n}_{y,t}$ are zero-mean Gaussian noise, respectively. The aforementioned means Markov process since the tth latent variable $\boldsymbol{x}_t$ only depends on the previous one $\boldsymbol{x}_{t-1}$ via $f(\bullet)$. Then each observation $\boldsymbol{y}_t$ depends on the corresponding latent variable $\boldsymbol{x}_t$ through the function $g(\bullet)$. In the original GPDM paper, the authors propose a particular nonlinear case in which $f(\bullet)$ and $g(\bullet)$ are linear combinations of the following basis functions:

$$f(\boldsymbol{x}; \boldsymbol{A}) = \sum_{k=1}^{K} \boldsymbol{a}_k \phi_k(\boldsymbol{x}), \quad g(\boldsymbol{x}; \boldsymbol{B}) = \sum_{m=1}^{M} \boldsymbol{b}_m \psi_m(\boldsymbol{x}). \tag{2}$$

By assuming Gaussian priors on the rows of $\boldsymbol{A}$ and $\boldsymbol{B}$, then we can derive the following probabilities:

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \frac{1}{(2\pi)^{TD/2}\,|\boldsymbol{K}_{\boldsymbol{Y}}|^{D/2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{K}_{\boldsymbol{Y}}^{-1}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}\right)\right), \tag{3}$$

$$p(\boldsymbol{X}) = \prod_{t=2}^{T} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \frac{p(\boldsymbol{x}_1)}{(2\pi)^{(T-1)D/2}\,|\boldsymbol{K}_{\boldsymbol{X}}|^{D/2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{K}_{\boldsymbol{X}}^{-1}\boldsymbol{X}_{2:T}\boldsymbol{X}_{2:T}^{\mathrm{T}}\right)\right), \tag{4}$$

where $\boldsymbol{K}_{\boldsymbol{Y}} \in \mathbb{R}^{T \times T}$ and $\boldsymbol{K}_{\boldsymbol{X}} \in \mathbb{R}^{(T-1)\times(T-1)}$ are respectively Gram matrices whose (i, j)th elements are calculated by using the kernel functions $k_y(\boldsymbol{x}_i, \boldsymbol{x}_j) = \psi(\boldsymbol{x}_i)\psi(\boldsymbol{x}_i)^{\mathrm{T}}$ and $k_x(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^{\mathrm{T}}$, called "kernel trick". Here, the posterior $p(\boldsymbol{X}|\boldsymbol{Y})$ can be approximated as $p(\boldsymbol{X}|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{X})p(\boldsymbol{X})$. Then, by using Eqs. (3) and (4), we can derive the log-posterior as follows:

$$\begin{aligned}
&\log p(\boldsymbol{X}|\boldsymbol{Y}) = \log p(\boldsymbol{Y}|\boldsymbol{X}) + \log p(\boldsymbol{X}) \\
&= \underbrace{-\frac{D}{2}\log|\boldsymbol{K}_{\boldsymbol{Y}}| - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{K}_{\boldsymbol{Y}}^{-1}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}\right)}_{\mathcal{L}_{\mathrm{GPLVM}}} \underbrace{-\frac{D}{2}\log|\boldsymbol{K}_{\boldsymbol{X}}| - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{K}_{\boldsymbol{X}}^{-1}\boldsymbol{X}_{2:T}\boldsymbol{X}_{2:T}^{\mathrm{T}}\right) + \log p(\boldsymbol{x}_1)}_{\mathcal{L}_{\mathrm{Markov}}}.
\end{aligned} \tag{5}$$

GPDM maximizes the above log-posterior Eq. (5), and then we obtain the parameters which are necessary for GPDM to build the latent space.

2.2. Class-aware GPDM (CGPDM)

In order for GPDM to consider class information when input multi-modalities have class labels, we newly derive the class-aware extended version of GPDM, named class-aware GPDM (CGPDM), in this section. As shown in Eq. (5), we can regard that the log-posterior of GPDM consists of two terms, i.e., $\mathcal{L}_{\mathrm{GPLVM}}$ which is the same as the log-posterior of GPLVM and $\mathcal{L}_{\mathrm{Markov}}$ aiming to consider the time-series dynamics based on the Markov process. Thus, focusing on $\mathcal{L}_{\mathrm{GPLVM}}$, it is considered that the dimensionality reduction part of GPMD depends on GPLVM. In other words, the term $\boldsymbol{K}_{\boldsymbol{Y}}^{-1}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}$ means to learn the the correlation matrix $\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}$ through the Gram matrix $\boldsymbol{K}_{\boldsymbol{Y}}$ of the latent variables based on the GPLVM's dimensionality reduction. Then we can rewrite its part by using the following block-diagonal matrix $\boldsymbol{M}$ aiming to decorrelate the pairs consisting of different class samples:

$$\begin{aligned}
&\log p(\boldsymbol{X}|\boldsymbol{Y}) \\
&:= -\frac{D}{2}\log|\boldsymbol{K}_{\boldsymbol{Y}}| - \frac{1}{2}\underbrace{\mathrm{tr}\left(\boldsymbol{K}_{\boldsymbol{Y}}^{-1}\boldsymbol{Y}\boldsymbol{M}\boldsymbol{Y}^{\mathrm{T}}\right)}_{\text{Discriminative term}} - \frac{D}{2}\log|\boldsymbol{K}_{\boldsymbol{X}}| - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{K}_{\boldsymbol{X}}^{-1}\boldsymbol{X}_{2:T}\boldsymbol{X}_{2:T}^{\mathrm{T}}\right) + \log p(\boldsymbol{x}_1),
\end{aligned} \tag{6}$$

where $\boldsymbol{M} = \mathrm{diag}\{\mathbf{1}_{n_1\times n_1}, \cdots, \mathbf{1}_{n_c\times n_c}, \cdots, \mathbf{1}_{n_C\times n_C}\}$ is a block-diagonal matrix. Note that all of the observations $\boldsymbol{y}_t$ are aligned per the class label in advance, and "$\mathbf{1}_{n\times n}$" is nth-order square matrix whose elements are all one. Furthermore, n_c $(c = 1, 2, \cdots, C)$ stands for the number of the corresponding cth class's samples. This rewriting makes the value of the correlation coefficient calculated by the pair of ith and jth observations $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$, i.e., (i, j)th elements of $\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}$, be zero if ith and jth samples respectively belong to different classes.

CGPDM maximizes the log-posterior in Eq. (6), and then the class-aware learning of dimensionality reduction considering class information and the time-relation lurked in the original input data becomes feasible. The expected results representing the differences between GPDM and CGPDM are shown in Fig. 2.

2.3. Discriminative Shared GPDM (CSGPDM)

In order to handle the multi-modal inputs having class information and time-series relation, we focus on shared GPDM (SGPDM). As shown in Fig. 1(b), SGPDM assumes that both of the observations share the same latent variables. In other words, the latent variables $\boldsymbol{X}$ having Markov dynamics generate the multi modalities at each time step. Thus, in this subsection, we newly assume the two heterogeneous sets of multi-modal input: D_1-dimensional observations $\boldsymbol{Y}^{(1)} = [\boldsymbol{y}_1^{(1)}, \boldsymbol{y}_2^{(1)}, \cdots, \boldsymbol{y}_T^{(1)}]^{\mathrm{T}} \in \mathbb{R}^{T\times D_1}$ and D_2-dimensional observations $\boldsymbol{Y}^{(2)} = [\boldsymbol{y}_1^{(2)}, \boldsymbol{y}_2^{(2)}, \cdots, \boldsymbol{y}_T^{(2)}]^{\mathrm{T}} \in \mathbb{R}^{T\times D_2}$. Note that the latent variables $\boldsymbol{X}$ are the same as the aforementioned.

On the basis of [35, 39], in order for CGPDM to receive the above multi-modal inputs like CCAs, we additionally derive the discriminative shared GPDM (CSGPDM) in this section. Specifically, by assuming that the heterogeneous sets of input data are generated by the same latent space, we can derive the following joint condi-

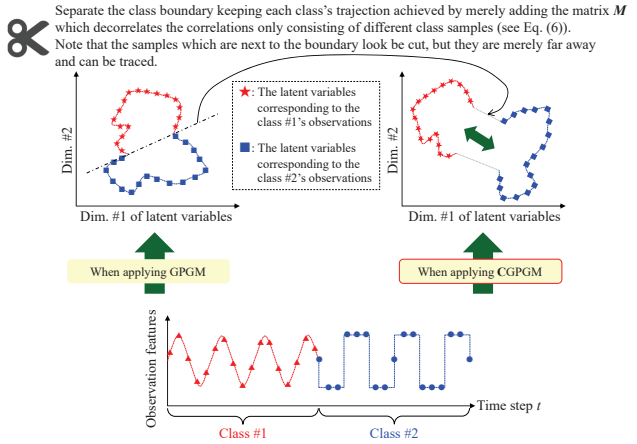


Fig. 2: An overview of the difference between GPDM and its extended version, i.e., class-aware GPDM (CGPDM), achieved by applying our proposal. Note that this example assumes there are two classes and the single observation, i.e., a modality as input, in order to simplify the explanation. In the case of the multi observations, our proposal can also be applied by adding the matrix $\boldsymbol{M}$ to all pairs of the latent variables $\boldsymbol{X}$ and each observations $\boldsymbol{Y}^{(k)}$ ($k = 1, 2, \cdots, K$; K is the total number of modalities) as explained in Eq. (8).

tional probability:

$$\begin{aligned} p(\boldsymbol{X}|\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}) &\propto p(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}|\boldsymbol{X})p(\boldsymbol{X}) \\ &= p(\boldsymbol{Y}^{(1)}|\boldsymbol{X})p(\boldsymbol{Y}^{(2)}|\boldsymbol{X})p(\boldsymbol{X}). \end{aligned} \tag{7}$$

Then we derive the following log-posterior:

$$\begin{aligned} \log p(\boldsymbol{X}|\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}) &= \log p(\boldsymbol{Y}^{(1)}|\boldsymbol{X}) + \log p(\boldsymbol{Y}^{(2)}|\boldsymbol{X}) + \log p(\boldsymbol{X}) \\ &:= -\frac{D}{2}\log\left|\boldsymbol{K}_Y^{(1)}\right| - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{K}_Y^{(1)^{-1}}\boldsymbol{Y}^{(1)}\boldsymbol{M}\boldsymbol{Y}^{(1)\mathrm{T}}\right) \\ &\quad - \frac{D}{2}\log\left|\boldsymbol{K}_Y^{(2)}\right| - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{K}_Y^{(2)^{-1}}\boldsymbol{Y}^{(2)}\boldsymbol{M}\boldsymbol{Y}^{(2)\mathrm{T}}\right) \\ &\quad - \frac{D}{2}\log|\boldsymbol{K}_X| - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{K}_X^{-1}\boldsymbol{X}_{2:T}\boldsymbol{X}_{2:T}^{\mathrm{T}}\right) + \log p(\boldsymbol{x}_1), \end{aligned} \tag{8}$$

where the Gram matrices $\boldsymbol{K}_Y^{(1)}$ and $\boldsymbol{K}_Y^{(1)}$ are respectively calculated by using the corresponding right upper indices' observations, i.e., $\boldsymbol{Y}^{(1)}$ and $\boldsymbol{Y}^{(2)}$.

In common with the discussion regarding CGPDM, by inserting $\boldsymbol{M}$ in the correlation matrices, estimation of CSGPDM-based latent space considering class information and the time-relation lurked in the original input data becomes feasible. Moreover, our modification can be applied to not only GPDM and SGPDM but also any other GPLVM-based method by similarly using the matrix $\boldsymbol{M}$ for all pairs each of which connects the latent variables with the arbitrary modal's input like the explanation of Fig. 2.

From the above discussions, our proposal enables GPDM and SGPDM to estimate the class label-driven latent spaces resulting in CGPDM and CSGPDM. Since our modification is merely adding the matrix $\boldsymbol{M}$ aiming to mask the elements consisting of different classes' samples as shown in Eqs. (6) and (8), our method has the following two contributions: i) it is applicable for not only GPDM and SGPDM but also almost all GPLVM-based models and ii) it does not increase the parameter to train, i.e., calculation cost.

3. EXPERIMENTS

We examined the validity of our method by visualizing the built latent spaces. Furthermore, we conducted tracking experiments which projected the points for inference into the predicted latent space and tracked the future points one by one based on the dynamic models' Markov process.

3.1. Dataset

To apply our method, we prepared the artificial datasets. At first, we defined the latent variables $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T]^{\mathrm{T}} \in \mathbb{R}^{T\times 2}$ each of which is two-dimensional vector $\boldsymbol{x}_t = [x_{t,(1)}, x_{t,(2)}]^{\mathrm{T}}$ ($t = 1, 2, \cdots, T$). By using the parameter θ ($0 \le \theta \le 2\pi$), we specifically defined two elements of the vector $\boldsymbol{x}_t$ as follows:

$$x_{t,(1)} = \begin{cases} \cos^3(\theta) & \text{if } 0 \le \theta \le \pi \\ \cos(\theta) & \text{else} \end{cases}, \quad x_{t,(2)} = \begin{cases} \sin^3(\theta) & \text{if } 0 \le \theta \le \pi \\ \sin(\theta) & \text{else} \end{cases}, \tag{9}$$
$$\text{s.t. } \theta = \left[\frac{2\pi}{T}, \frac{4\pi}{T}, \cdots, 2\pi\right].$$

In our experiments, we employed $T = 100$, and thus the latent variables consisted of two parametric half curves, each half of which was respectively known as asteroid (class #1) and circle (class #2), as shown in Fig. 3(a). Next, by using the latent variables $\boldsymbol{X}$, we also defined the two observations $\boldsymbol{Y}^{(1)} = [\boldsymbol{y}_1^{(1)}, \boldsymbol{y}_2^{(1)}, \cdots, \boldsymbol{y}_T^{(1)}]^{\mathrm{T}} \in \mathbb{R}^{T\times D_1}$ and $\boldsymbol{Y}^{(2)} = [\boldsymbol{y}_1^{(2)}, \boldsymbol{y}_2^{(2)}, \cdots, \boldsymbol{y}_T^{(2)}]^{\mathrm{T}} \in \mathbb{R}^{T\times D_2}$. In our experiments, we experimentally defined $D_1 = 5$ and $D_2 = 5$. Specifically, they were generated by using the latent variables as follows:

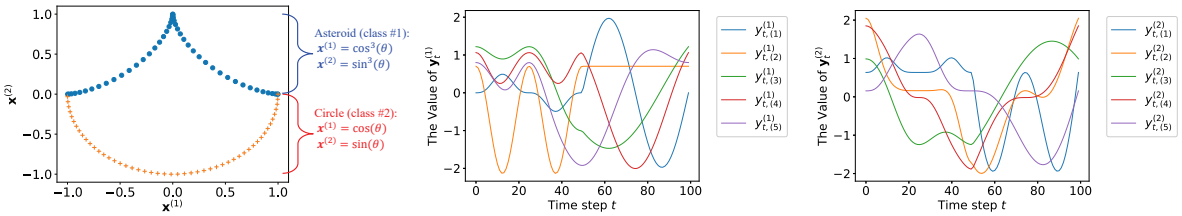
$$\boldsymbol{y}_t^{(1)} = \begin{bmatrix} y_{t,(1)}^{(1)} \\ y_{t,(2)}^{(1)} \\ y_{t,(3)}^{(1)} \\ y_{t,(4)}^{(1)} \\ y_{t,(5)}^{(1)} \end{bmatrix} = \begin{bmatrix} x_{t,(1)}x_{t,(2)} \\ (x_{t,(1)})^2 + (x_{t,(2)})^2 \\ x_{t,(1)} + x_{t,(2)} \\ (x_{t,(1)})^2 - x_{t,(2)} \\ x_{t,(1)} + (x_{t,(2)})^2 \end{bmatrix}, \quad \boldsymbol{y}_t^{(2)} = \begin{bmatrix} y_{t,(1)}^{(2)} \\ y_{t,(2)}^{(2)} \\ y_{t,(3)}^{(2)} \\ y_{t,(4)}^{(2)} \\ y_{t,(5)}^{(2)} \end{bmatrix} = \begin{bmatrix} (x_{t,(1)})^2 x_{t,(2)} \\ (x_{t,(1)})^2 + (x_{t,(1)} + x_{t,(2)})^2 \\ x_{t,(1)} - x_{t,(2)} \\ x_{t,(1)}x_{t,(2)} + x_{t,(1)} \\ x_{t,(1)}x_{t,(2)} + x_{t,(2)} \end{bmatrix}. \tag{10}$$

The visualizations of $\boldsymbol{X}$, $\boldsymbol{Y}^{(1)}$ and $\boldsymbol{Y}^{(2)}$ are summarized in Fig. 3.

3.2. Results

To initialize latent variables, we adopted the five methods summarized in Table 1. All of the visualization results are depicted in Fig. 4. Note that the blue and orange plots are the estimated latent variables each of which corresponds to classes #1 and #2 and is connected in the order of time series. Meanwhile, the green and red plots are the tracking results which tracked the testing sample by inputting it to the SGPDM and CSGPDM recursively. Specifically, we adopted the two sample of ground truth latent variables, i.e., samples #1 (=$\boldsymbol{x}_1$) and #51 (=$\boldsymbol{x}_{51}$), as starting points and track their following 49 points based on the Markov process. Since $\boldsymbol{x}_1$ and $\boldsymbol{x}_{51}$ are respectively starting points of classes #1 and #2 as shown in Fig. 3(a), their following 49 tracking results should be ideally over-plotted on the predicted asteroid and circle if the model is trained successfully. As shown in Fig. 4, it is confirmed that all estimated latent variables, i.e., blue and orange trajectories, of CSGPDM were separated successfully compared to those of SGPDM as we expected in Fig. 2. Therefore, it is worth noting that our extended version of SGPDM, i.e., CSGPDM, can newly build class-aware latent space compared to SGPDM. However, focusing on the green and red tracking results, the results except Figs. 4(i) and (j), i.e., the tracking results without DLPCCA-based initialization, tended to be failure to grasp the correct latent variables since their green and red lines did not trace the corresponding blue and orange ones. On the other hand, as shown in Fig. 4(j), CSGPDM with DLPCCA-based initialization was able to not only build the latent space correctly but also project the testing samples so that they roughly succeed to trace the corresponding blue and orange lines.

To evaluate the aforementioned in the numerical way, we calculated the mean Euclidean distance (MED) (= $\sum_i \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|/T$) between the latent variable and the corresponding tracked result. All results are summarized in the right column of Table 1. As shown in this ta-



(a) Latent variables $\boldsymbol{X}$ (b) Observation $\boldsymbol{Y}^{(1)}$ (see Eq. (10)) (c) Observation $\boldsymbol{Y}^{(2)}$ (see Eq. (10))

Fig. 3: The visualizations of the artificial datasets used in our experiments. Note that the both of $\boldsymbol{Y}^{(1)}$ and $\boldsymbol{Y}^{(2)}$ were normalized so that each dimension of them has zero mean and one standard deviation in advance.

Table 1: Experimental results and the summary of each method's characteristic. Note that all latent spaces were normalized in order for them to be under the fair condition.

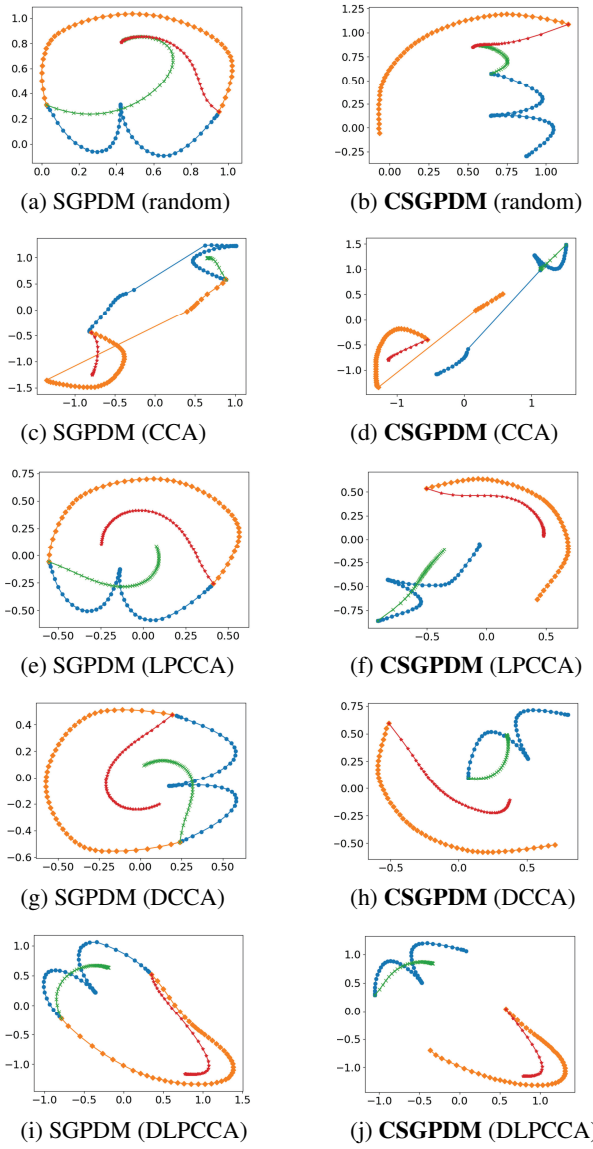
Method		Class information	Local structure-based non-linear correlation	MED		
Model	Init.			class #1	class #2	ALL
SGPDM	Random	-	-	1.80	**1.13**	**1.46**
CSGPDM				**1.53**	1.60	1.57
SGPDM	CCA [18]	×	×	**0.85**	0.77	0.81
CSGPDM				0.93	**0.63**	**0.78**
SGPDM	LPCCA [27]	×	✓	0.97	1.23	1.10
CSGPDM				**0.66**	**1.08**	**0.87**
SGPDM	DCCA [26]	✓	×	0.91	**1.35**	1.13
CSGPDM				**0.70**	1.38	**1.04**
SGPDM	DLPCCA [29]	✓	✓	0.59	0.62	0.61
CSGPDM				**0.50**	**0.51**	**0.51**

ble, we can confirm the performances of CSGPDM initialized with any one CCA tended to be superior to the corresponding SGPDM when they were initialized in the same way. Therefore, first of all, it is important to consider the correlation between both input modalities by using any CCAs, and then we argue that CSGPDM successfully projects the unknown testing samples into the latent space rather than SGPDM. In particular, the best performances regarding classes #1, #2 and all samples were respectively obtained from DSGPCM with DLPCCA-based initialization (see the highlighted yellow cells in Table 1). Thus, the validity of CSGPDM with DLPCCA-based initialization was also confirmed in terms of not only visualization but also numerical evaluation.

From the aforementioned, we argue that our extension aiming to make SGPDM class-aware is valid for multi-modal inputtings in order to estimate the latent space having class information. Especially, using class information and local structure-based non-linear correlation between the input multiple modalities simultaneously, i.e., applying DLPCCA, for initialization can build the following class-aware latent space, successfully.

4. CONCLUSION

In this paper, we have proposed the new scheme making the GP-based dimensionality reduction method class-aware from the perspective of the representation learning. Specifically, merely rewriting the GPLVM-based prior term enables GP-based dimensionality reduction models to be class-aware one. Since almost all GP-based dimensionality reduction models have the GPLVM-based prior term, our proposal is applicable without increasing additional parameter to learn. In particular, our extended version of shared Gaussian process dynamic model (SGPDM), named class-aware SGPDM (CSGPDM) can receive multi-modal inputs and estimate low-dimensional latent space considering class information and time relations lurked in the original multi-modal them. Experimental results showed that the estimated CSGPDM-based latent space outperformed those of SGPDM in terms of utilization of class information.



(a) SGPDM (random) (b) **CSGPDM** (random)
(c) SGPDM (CCA) (d) **CSGPDM** (CCA)
(e) SGPDM (LPCCA) (f) **CSGPDM** (LPCCA)
(g) SGPDM (DCCA) (h) **CSGPDM** (DCCA)
(i) SGPDM (DLPCCA) (j) **CSGPDM** (DLPCCA)

Fig. 4: The visualization results of SGPDM and CSGPDM. The blue and orange lines denote the trajectories of models' latent variables, i.e., the predicted asteroid and circle. By utilizing $\boldsymbol{x}_1$ and $\boldsymbol{x}_{51}$ as starting points and input them to the models recursively, the adjacent points $(\boldsymbol{x}_1, \hat{\boldsymbol{x}}_2, \cdots, \hat{\boldsymbol{x}}_{50})$ and $(\boldsymbol{x}_{51}, \hat{\boldsymbol{x}}_{52}, \cdots, \hat{\boldsymbol{x}}_{100})$ were predicted, and they were over-plotted and connected using green and red lines to visualize the trajectories of the estimated asteroid and circle.

5. REFERENCES

[1] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Tech. Rep., University of Toronto, 2009.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. of IEEE conference on computer vision and pattern recognition (CVPR)*, 2009, pp. 248–255.

[3] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multi-channel environmental noise recordings," in *Proc. of 21st International Congress on Acoustics*, Montreal, Canada, June 2013, Acoustical Society of America.

[4] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[5] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[6] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Dec. 2017.

[7] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv*, vol. abs/1510.08484, 2015.

[8] G. A. Miller, "WordNet: A lexical database for english," *Communications of the ACM*, vol. 38, no. 1, pp. 39–41, 1995.

[9] I. Katakis, G. Tsoumakas, and I. Vlahavas, "Multilabel text classification for automated tag suggestion," in *Proc. of ECML PKDD Discovery Challenge, Antwerp, Belgium*, 2008, pp. 75–83.

[10] J. Shetty and J. Adibi, "Enron email dataset," Tech. Rep., 2004.

[11] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov. 2019, pp. 188–197, Association for Computational Linguistics.

[12] Y. Weibo, "EEG data of simple and compound limb motor imagery," 2014.

[13] C. J. Markiewicz, K. J. Gorgolewski, F. Feingold, R. Blair, Y. O. Halchenko, E. Miller, N. Hardcastle, J. Wexler, O. Esteban, M. Goncavles, A. Jwa, and R. Poldrack, "The OpenNeuro resource for sharing of neuroscience data," *eLife*, vol. 10, pp. e71774, oct 2021.

[14] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," *arXiv*, vol. abs/1705.06950, 2017.

[15] S. Abu-El-Haija, N. Kothari, J. Lee, A. P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "YouTube-8M: A Large-Scale Video Classification Benchmark," *arXiv*, vol. abs/1609.08675, 2016.

[16] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

[17] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2020.

[18] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3, pp. 321–377, 1936.

[19] J. Blitzer, D. Foster, and S. Kakade, "Domain adaptation with coupled subspaces," in *Proc. of the Conference on Artificial Intelligence and Statistics*, 2011.

[20] H. Ohkushi, T. Ogawa, and M. Haseyama, "Music recommendation according to human motion based on kernel CCA-based relationship," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, pp. 121, 2011.

[21] W. Li, L. Duan, D. Xu, and I. W. Tsang, "Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1134–1148, June 2014.

[22] T. Ogawa and M. Haseyama, "2D semi-supervised CCA-based inpainting including new priority estimation," in *Proc. of IEEE International Conference on the Image Processing (ICIP)*, 2014, pp. 1837–1841.

[23] Y. Yeh, C. Huang, and Y. F. Wang, "Heterogeneous domain adaptation and classification by exploiting the correlation subspace," *IEEE Trans. on Image Processing*, vol. 23, no. 5, pp. 2009–2018, 2014.

[24] R. Sawata, T. Ogawa, and M. Haseyama, "Novel audio feature projection using KDLPCCA-based correlation with EEG features for favorite music classification," *IEEE Trans. on Affective Computing*, vol. 10, no. 3, pp. 430–444, 2019.

[25] X. Chang, T. Xiang, and T. M. Hospedales, "Scalable and effective deep CCA via soft decorrelation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1488–1497.

[26] T.-K. Sun, S.-C. Chen, Z. Jin, and J.-Y. Yang, "Kernelized discriminative canonical correlation analysis," in *Proc. of International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*, Nov 2007, vol. 3, pp. 1283–1287.

[27] Y. Peng, D. Zhang, and J. Zhang, "A new canonical correlation analysis algorithm with local discrimination.," *Neural Processing Letters*, vol. 31, no. 1, pp. 1–15, 2010.

[28] T. Sun and S. Chen, "Locality preserving CCA with applications to data visualization and pose estimation," *Image and Vision Computing*, vol. 25, no. 5, pp. 531 – 543, 2007.

[29] X. Zhang, N. Guan, Z. Luo, and L. Lan, "Discriminative locality preserving canonical correlation analysis," in *Pattern Recognition*, vol. 321 of *Communications in Computer and Information Science*, pp. 341–349. Springer Berlin Heidelberg, 2012.

[30] S. Akaho, "A kernel method for canonical correlation analysis," in *Proc. of the International Meeting of the Psychometric Society (IMPS)*, 2001.

[31] R. Sawata, T. Ogawa, and M. Haseyama, "The extraction of individual music preference based on deep time-series CCA," in *Proc. of IEEE 8th Global Conference on Consumer Electronics (GCCE)*, 2019, pp. 15–16.

[32] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Tech. Rep., Department of Computer Science, 2003.

[33] M. Titsias and N. D. Lawrence, "Bayesian gaussian process latent variable model," in *Proc. of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 13–15 May 2010, vol. 9, pp. 844–851, PMLR.

[34] N. D. Lawrence, "Gaussian process latent variable models for visualisation of high dimensional data," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, 2003, pp. 329–336.

[35] C. H. Ek, *Shared Gaussian Process Latent Variable Models*, Ph.D. thesis, Oxford Brookes University, 2009.

[36] J. Wang, A. Hertzmann, and D. J. Fleet, "Gaussian process dynamical models," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, 2005, vol. 18.

[37] R. Urtasun and T. Darrell, "Discriminative gaussian process latent variable model for classification," in *Pro. of the 24th International Conference on Machine Learning (ICML)*, 2007, p. 927–934.

[38] X. Gao, X. Wang, D. Tao, and X. Li, "Supervised gaussian process latent variable model for dimensionality reduction," *IEEE Trans. on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 2, pp. 425–434, 2011.

[39] J. Chen, M. Kim, Y. Wang, and Q. Ji, "Switching gaussian process dynamic models for simultaneous composite motion tracking and recognition," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 2655–2662.