DISTRIBUTED LABEL DEQUANTIZED GAUSSIAN PROCESS LATENT VARIABLE MODEL FOR MULTI-VIEW DATA INTEGRATION

Koshi Watanabe[†]*, Keisuke Maeda*[††]*, Takahiro Ogawa*[†††]*, Miki Haseyama*[†††]

[†]School of Engineering, Hokkaido University, Japan
[††]Office of Institutional Research, Hokkaido University, Japan
[†††]Faculty of Information Science and Technology, Hokkaido University, Japan
E-mail:{koshi,maeda,ogawa}@lmd.ist.hokudai.ac.jp, miki@ist.hokudai.ac.jp

ABSTRACT

In this paper, we present a novel method for multi-view data analysis, distributed label dequantized Gaussian process latent variable model (DLDGP). DLDGP can integrate multi-view data and class information into a common latent space. In the previous multi-view methods, the dimension of label features transformed from the class information is much smaller than those of the other modalities, which causes a dimensionality-limitation problem in the latent space. DLDGP extends the dimension of the label features by a distributed label dequantization scheme. Additionally, DLDGP calculates correlation between different classes by encoding class information into distributed features. DLDGP can correctly capture the relationship between multi-view data and obtain the latent features with high expression ability. Experimental results show the effectiveness of our method by using the open dataset.

***Index Terms*—** Label dequantization, distributed labels, Gaussian process, multi-view data.

1. INTRODUCTION

Online shopping services are becoming one of the new foundations of our daily lives. Additionally, these services provide numerous items, and recommendation systems are necessary for removing the burden on users searching for their desired items [1, 2]. In order to construct these systems, the previous approaches focused on the relationship between items in purchase history and estimated labels (e.g., rating values) assigned by users. Since these systems need to utilize multi-view data obtained from items, it is expected that the effective use of those data contributes to accurate label estimation.

Traditionally, correlation analysis-based approaches have been adopted for analyzing such multi-view data. Multi-view canonical correlation analysis (MVCCA) [3] is one of their representative methods and can derive projections mapping multi-view data to a low-dimensional common latent space by maximizing their correlation. However, the obtained projections are deterministic and linear, accuracy degradation tends to occur when the multi-view data have a complicated structure, and probabilistic interpretation approaches are desirable [4]. Multi-modal Gaussian process latent variable model (mGP-LVM) [5–7] assumes that multi-view data are generated from the common latent space based on probabilistic and non-linear projections. These projections are based on Gaussian process (GP) [8] and enable the calculation of accurate latent space for complicated data with probabilistic interpretation.

This work was partly supported by JSPS KAKENHI Grant Numbers JP21H03456 and JP20K19856.

In recommendation systems, rating values obtained from the purchase history are one of the most valuable information. These values can be regarded as class information, and various methods have been proposed for analyzing them [9–11]. Especially, discriminative GP-LVM (DGP-LVM) [9] utilizes class information as prior knowledge in the latent space. The latent representation of DGP-LVM is regularized to minimize variance within the same class and maximize that between different classes. On the other hand, in the general latent variable models, using class information as label features is known to improve their performance [12, 13]. Supervised GP-LVM [10] transforms class information into a one-hot vector and uses it as an additional modality. Although using the one-hot encoded class information as one modality is a powerful approach for multi-view data, they still have the following two problems that prevent the calculation of the effective latent space.

Problem (**i**)**:** The dimension of the latent space is limited to the minimum dimension between all features obtained from multi-view data. Generally, the dimension of the label features is much smaller than those of the other modalities. Therefore, informative correlation between different modalities may be lost because of the dimensionality limitation.

Problem (**ii**)**:** Since class information is transformed into the one-hot vector, the model hardly captures the correlation between different classes. For example, five-star rating items are expected to be closer to four-star rating items than one-star rating items. However, such between-class information is not considered due to the transformation into the one-hot vector.

In this paper, we propose a novel method, distributed label dequantized Gaussian process latent variable model (DLDGP) that can solve the above problems by the following approaches.

Approach (**i**)**:** We newly introduce a distributed label dequantization scheme into mGP-LVM. This scheme enables to obtain informative label features while increasing the number of dimensions of the original label features. Therefore, DLDGP can eliminate the dimensionality limitation in **problem** (**i**) and can calculate latent variables that better reflect the correlation between different modalities.

Approach (**ii**)**:** We transform class information into distributed features. By this transformation, we can consider the correlation between different classes and overcome the limitation of the one-hot encoding in **problem** (**ii**).

DLDGP overcomes the dimensionality limitation and considers the correlation between different classes by the distributed expression of class information. We can simultaneously solve the above problems, and this is the main contribution in this paper. Therefore, our method archives the accuracy improvement for the label estimation task.

2. DISTRIBUTED LABEL DEQUANTIZED GP-LVM

In this section, we explain our method, DLDGP. It calculates latent variables from multi-view observed variables and class information. DLDGP can also estimate distributed label features while extending their dimensions by the distributed label dequantization scheme and derive the effective latent representation of multi-view data.

In 2.1, we describe an objective function that extends the dimension of the label features and encodes class information into distributed features. In 2.2, we explain its optimization method based on the gradient descent. In 2.3, the label estimation using the trained latent space is presented.

2.1. Objective Function of DLDGP

We define V kinds of $D^{(v)}$-dimensional observed variables as $\boldsymbol{Y}^{(v)} = [\boldsymbol{y}_1^{(v)}, \boldsymbol{y}_2^{(v)}, \ldots, \boldsymbol{y}_N^{(v)}]^\top \in \mathbb{R}^{N \times D^{(v)}}$ ($v = 1, 2, \ldots, V$; N being the number of the obtained samples) and consider that each sample can be classified into C classes. Since C is generally much smaller than the dimensions $D^{(v)}$, we newly define $D^{(L)}$-dimensional dequantized label features that extend the original class information as $\boldsymbol{Y}^{(L)} = [\boldsymbol{y}_1^{(L)}, \boldsymbol{y}_2^{(L)}, \ldots, \boldsymbol{y}_N^{(L)}]^\top \in \mathbb{R}^{N \times D^{(L)}}$ ($D^{(L)}$ being a k-fold expanded number of C). When sample n ($= 1, 2, \ldots, N$) belongs to class d ($= 1, 2, \ldots, C$), we initialize each value $\{y^{(L)}_{n,\,k(d-1)+j}\}_{j=1,2,\ldots,k}$ by a uniform distribution on $(0, 1)$. Based on mGP-LVM, we assume that all observed variables $\boldsymbol{Y}^{(v')}(v' = 1, 2, \ldots, V, L)$ are generated from q-dimensional ($q < D^{(v')}$) common latent variables $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]^\top \in \mathbb{R}^{N \times q}$ as follows:

$$\boldsymbol{y}^{(v')}_{:,i} = \boldsymbol{f}_i^{(v')}(\boldsymbol{X}) + \boldsymbol{\epsilon}^{(v')}, \tag{1}$$

where

$$\boldsymbol{f}_i^{(v')}(\boldsymbol{X}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{K}^{(v')}),$$
$$\boldsymbol{\epsilon}^{(v')} \sim \mathcal{N}(\boldsymbol{0}, (\beta^{(v')})^{-1}\boldsymbol{I}),$$

and each $\boldsymbol{f}_i^{(v')}(\cdot) \in \mathbb{R}^N (i = 1, 2, \ldots, D^{(v')})$ follows a zero-mean GP prior with a covariance matrix $\boldsymbol{K}^{(v')} \in \mathbb{R}^{N \times N}$, and $\boldsymbol{\epsilon}^{(v')}$ denotes a standard Gaussian noise with a covariance matrix $(\beta^{(v')})^{-1}\boldsymbol{I}$. Owing to the extension of the class information, we can select the dimension q of $\boldsymbol{X}$ more flexibly. By marginalizing $\boldsymbol{f}_i^{(v')}(\cdot)$, the likelihood function of $\boldsymbol{X}$ and $\boldsymbol{\theta}^{(v')}$ is given by the following equation:

$$\begin{aligned} &p(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}, \ldots, \boldsymbol{Y}^{(V)}, \boldsymbol{Y}^{(L)} | \boldsymbol{X}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(V)}, \boldsymbol{\theta}^{(L)}) \\ &= \prod_{v'} p(\boldsymbol{Y}^{(v')} | \boldsymbol{X}, \boldsymbol{\theta}^{(v')}) \\ &= \prod_{v'} \frac{1}{\sqrt{(2\pi)^{ND^{(v')}} |\boldsymbol{K}^{(v')}|^{D^{(v')}}}} \\ &\qquad \times \exp\left[-\frac{1}{2}\mathrm{tr}\left((\boldsymbol{K}^{(v')})^{-1}\boldsymbol{Y}^{(v')}(\boldsymbol{Y}^{(v')})^\top\right)\right], \end{aligned} \tag{2}$$

where $\boldsymbol{\theta}^{(v')}$ are parameters of the kernel functions $k^{(v')}(\cdot, \cdot)$. Although we can derive the latent variables $\boldsymbol{X}$ by maximizing Eq. (2), we cannot calculate those corresponding to new observed variables in the test phase [14]. To solve this problem, DLDGP introduces the back constraints technique [15] and derives an optimized projection instead of deriving the latent variables directly. The latent variables $\boldsymbol{X}$ calculated by the projection $g(\cdot)$ are given by the following equation:

$$\boldsymbol{X} = g(\boldsymbol{Y}^{(v)}; \boldsymbol{W}), \tag{3}$$

where $\boldsymbol{W}$ is a set of parameters of the projection $g(\cdot)$, and we adopt a multi-layer perceptron [16] with three layers as the projection in our method. By this procedure, we can obtain the latent variables corresponding to the new observed variables based on the mapping by the projection $g(\cdot)$. In DLDGP, the parameters are updated by solving the following log-likelihood maximization problem:

$$\{\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{Y}}^{(L)}\} = \arg\max_{\boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{Y}^{(L)}} \sum_{v'} \log p(\boldsymbol{Y}^{(v')} | \boldsymbol{X}, \boldsymbol{\theta}^{(v')}). \tag{4}$$

Note that, we define $\boldsymbol{\Theta} = \{\boldsymbol{\theta}^{(v')}\}_{v'=1,2,\ldots,V,L}$ to simplify the notation.

In the proposed method, we provide the class information and distributed expression for the dequantized label features $\boldsymbol{Y}^{(L)}$ by two constraint functions. Firstly, we let the values of $\boldsymbol{Y}^{(L)}$ be finite by a norm constraint $c(\cdot)^{(\mathrm{L2})}$ as follows:

$$\begin{aligned} c^{(\mathrm{L2})}(\boldsymbol{y}_n^{(L)}) &= (\boldsymbol{y}_n^{(L)})^\top \boldsymbol{y}_n^{(L)} - 1 \\ &= 0. \end{aligned} \tag{5}$$

To reflect the class information in $\boldsymbol{Y}^{(L)}$, we define a mask matrix $\boldsymbol{M}_n \in \mathbb{R}^{D^{(L)} \times D^{(L)}}$ for each sample whose diagonal elements from $k(d-1)+1$ to kd are 1, and the other elements are 0 when sample n belongs to class d. By the mask matrix $\boldsymbol{M}_n$, a class constraint $c^{(Q)}(\cdot)$ is given as follows:

$$\begin{aligned} c^{(Q)}(\boldsymbol{y}_n^{(L)}) &= (\boldsymbol{y}_n^{(L)})^\top \boldsymbol{M}_n \boldsymbol{y}_n^{(L)} - \alpha \\ &= 0 \qquad (0 \leqq \alpha \leqq 1), \end{aligned} \tag{6}$$

where α is a precision parameter to control the distribution of the values of the dequantized label features. If α is close to 1, the values $\{y^{(L)}_{n,\,k(d-1)+j}\}_{j=1,2,\ldots,k}$ tend to become larger, and the other values in $\boldsymbol{y}_n^{(L)}$ tend to become smaller. By adjusting α appropriately, we can control the influence on each value in $\boldsymbol{y}_n^{(L)}$ and can consider the correlation between different classes. Furthermore, by optimizing $\boldsymbol{Y}^{(L)}$ to satisfy Eqs. (5) and (6), DLDGP can calculate the latent variables with improved representation. From the above, we define the optimization problem of DLDGP as follows:

$$\begin{aligned} \{\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{Y}}^{(L)}\} = \arg\max_{\boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{Y}^{(L)}} &\sum_{v'} \log p(\boldsymbol{Y}^{(v')} | \boldsymbol{W}, \boldsymbol{\theta}^{(v')}) \\ \text{s.t. } c^{(\mathrm{L2})}(\boldsymbol{y}_n^{(L)}) = 0,\ &c^{(Q)}(\boldsymbol{y}_n^{(L)}) = 0. \end{aligned} \tag{7}$$

DLDGP can eliminate the dimensionality limitation in **problem (i)** by defining the dequantized label features with the constraints in Eq. (7) and the lack of the class correlation in **problem (ii)** by Eq. (6). Furthermore, DLDGP can find the latent structure in the new sample by the projection in Eq. (3), and this leads to the accurate estimation of its label features. The details of the label estimation are presented in subsection 2.3, and we next explain the optimization method of our model in the following subsection.

2.2. Update of Parameters of DLDGP

In this subsection, we explain how to solve the optimization problem shown in Eq. (7). Since the dequantized label features $\boldsymbol{Y}^{(L)}$ are

constrained by Eqs. (5) and (6), we solve the optimization problem with respect to $\boldsymbol{W}$ and $\boldsymbol{\Theta}$ and the problem with respect to $\boldsymbol{Y}^{(L)}$, alternatively.

2.2.1. Update of $\boldsymbol{W}$ and $\boldsymbol{\Theta}$

The optimization problem of $\boldsymbol{W}$ and $\boldsymbol{\Theta}$ is given as follows:

$$\{\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{\Theta}}\} = \arg\max_{\boldsymbol{W},\boldsymbol{\Theta}} \sum_{v'} \log p(\boldsymbol{Y}^{(v')}|\boldsymbol{W}, \boldsymbol{\theta}^{(v')}). \tag{8}$$

Since Eq. (8) is equivalent to the optimization problem of mGP-LVM, we solve this problem by the scaled conjugate gradient method [17]. By solving Eq. (8), we can obtain the latent features which reflect the correlation between multi-view data with the projection $g(\cdot)$.

2.2.2. Update of $\boldsymbol{Y}^{(L)}$

The optimization problem of $\boldsymbol{Y}^{(L)}$ is given as follows:

$$\begin{aligned}\widehat{\boldsymbol{Y}}^{(L)} &= \arg\max_{\boldsymbol{Y}^{(L)}} \log p(\boldsymbol{Y}^{(L)}|\boldsymbol{W}, \boldsymbol{\theta}^{(v')}) \\ &= \arg\min_{\boldsymbol{Y}^{(L)}} \frac{1}{2}\mathrm{tr}\left((\boldsymbol{K}^{(L)})^{-1}\boldsymbol{Y}^{(L)}(\boldsymbol{Y}^{(L)})^\top\right) \\ &\quad \text{s.t. } c^{(\mathrm{L2})}(\boldsymbol{y}_n^{(L)}) = 0,\ c^{(\mathrm{Q})}(\boldsymbol{y}_n^{(L)}) = 0. \end{aligned} \tag{9}$$

This equation can be solved by the augmented Lagrangian method (ALM) [18]. The augmented Lagrangian function $\Phi(\cdot)$ of Eq. (9) is given by the following equation:

$$\begin{aligned}&\Phi(\boldsymbol{Y}^{(L)}, \rho, \Lambda) \\ &= \frac{1}{2}\mathrm{tr}\left((\boldsymbol{K}^{(L)})^{-1}\boldsymbol{Y}^{(L)}(\boldsymbol{Y}^{(L)})^\top\right) \\ &\quad + \frac{\rho}{2}\sum_{n=1}^{N}\left[\left\{c^{(\mathrm{L2})}(\boldsymbol{y}_n^{(L)})\right\}^2 + \left\{c^{(Q)}(\boldsymbol{y}_n^{(L)})\right\}^2\right] \\ &\quad + \sum_{n=1}^{N}\left\{\lambda_n^{(\mathrm{L2})}c^{(\mathrm{L2})}(\boldsymbol{y}_n^{(L)}) + \lambda_n^{(Q)}c^{(Q)}(\boldsymbol{y}_n^{(L)})\right\}, \end{aligned} \tag{10}$$

where ρ is a penalty parameter, and $\Lambda = \{\lambda_n^{(\mathrm{L2})}, \lambda_n^{(Q)}\}_{n=1,2,\ldots,N}$ are Lagrange multipliers. Equation (10) can be minimized by the simple gradient method [19], and differentiation of Eq. (10) is given as follows:

$$\begin{aligned}\frac{\partial\Phi}{\partial\boldsymbol{Y}^{(L)}} &= (\boldsymbol{K}^{(L)})^{-1}\boldsymbol{Y}^{(L)} + 2\boldsymbol{P}^{(\mathrm{L2})}\boldsymbol{D}^{(L2)} + 2\boldsymbol{P}^{(Q)}\boldsymbol{D}^{(Q)} \\ &\quad + 2\boldsymbol{R}^{(\mathrm{L2})}\boldsymbol{D}^{(\mathrm{L2})} + 2\boldsymbol{R}^{(Q)}\boldsymbol{D}^{(Q)}, \end{aligned} \tag{11}$$

where

$$\boldsymbol{P}^{(t)} = \mathrm{diag}(\rho\, c^{(t)}(\boldsymbol{y}_1^{(L)}),\ \rho\, c^{(t)}(\boldsymbol{y}_2^{(L)}), \ldots,\ \rho\, c^{(t)}(\boldsymbol{y}_N^{(L)})), \tag{12}$$

$$\boldsymbol{R}^{(t)} = \mathrm{diag}(\lambda_1^{(t)},\ \lambda_2^{(t)}, \ldots,\ \lambda_N^{(t)}) \quad (t \in \{\mathrm{L2}, Q\}), \tag{13}$$

and $\boldsymbol{D}^{(\mathrm{L2})}$ and $\boldsymbol{D}^{(Q)}$ are respectively differentiation of Eqs. (5) and (6) and denoted as follows:

$$\boldsymbol{D}^{(\mathrm{L2})} = \boldsymbol{Y}^{(L)}, \tag{14}$$

$$\boldsymbol{D}^{(Q)} = \left[\boldsymbol{M}_1\boldsymbol{y}_1^{(L)}, \boldsymbol{M}_2\boldsymbol{y}_2^{(L)}, \ldots, \boldsymbol{M}_N\boldsymbol{y}_N^{(L)}\right]^\top. \tag{15}$$

On the basis of ALM, we update $\boldsymbol{Y}^{(L)}$, ρ and Λ in each iteration as follows:

$$\boldsymbol{Y}_{\tau+1}^{(L)} \leftarrow \boldsymbol{Y}_{\tau}^{(L)} - \eta\frac{\partial\Phi}{\partial\boldsymbol{Y}^{(L)}}\Big|_{\boldsymbol{Y}^{(L)}=\boldsymbol{Y}_\tau^{(L)}}, \tag{16}$$

$$\lambda_{n,\tau+1}^{(t)} \leftarrow \lambda_{n,\tau}^{(t)} + \mu_{n,\tau}^{(t)}\, c^{(t)}(\boldsymbol{y}_n^{(L)}), \tag{17}$$

$$\rho_{\tau+1} \leftarrow \gamma\rho_\tau \quad (\gamma > 1), \tag{18}$$

where η is a learning rate, γ is a constant to increase the number of the penalty parameter ρ, and τ indicates a step of the optimization procedure. By the above calculation, DLDGP infers the dequantized labels while considering the correlation between different classes.

2.3. Estimation of Label Features Corresponding to New Observed Variables

In this subsection, we explain the estimation of the label features from a new sample. By using the projection $g(\cdot)$ in 2.1, we can calculate the latent variables corresponding to the new sample $\boldsymbol{y}_*$ as follows:

$$\boldsymbol{x}_* = g(\boldsymbol{y}_*; \widehat{\boldsymbol{W}}). \tag{19}$$

If the latent variables $\boldsymbol{x}_*$ are obtained, we can estimate label features $\boldsymbol{y}_*^{(L)}$ corresponding to $\boldsymbol{x}_*$ based on the mean prediction of GP as follows:

$$\begin{aligned}\boldsymbol{y}_*^{(L)} &= (\boldsymbol{k}_{:,*}^{(L)})^\top(\widehat{\boldsymbol{K}}^{(L)})^{-1}\boldsymbol{Y}^{(L)}, \\ \boldsymbol{k}_{:,*}^{(L)} &= [k^{(L)}(\boldsymbol{x}_1, \boldsymbol{x}_*), k^{(L)}(\boldsymbol{x}_2, \boldsymbol{x}_*), \ldots, k^{(L)}(\boldsymbol{x}_N, \boldsymbol{x}_*)]^\top. \end{aligned} \tag{20}$$

In this way, we can estimate the labels corresponding to the new sample with the calculations of Eqs. (19) and (20).

By using the dequantized and distributed label features for the feature integration, DLDGP can consider the correlation between the multi-view data and that between different classes accurately. From the above, the proposed method contributes to the improvement of the estimation accuracy.

3. EXPERIMENTS

3.1. Dataset

To verify the effectiveness of our method, we used Amazon[1] review database [20, 21]. This database provides 142.8 million review data that contain review text, rating values on a scale of one to five and metadata about products. We randomly selected 50 users who have 1,000 review data and collected product data including images, titles, description and prices of the products. To utilize these review data and product data, we performed the feature extraction by pre-trained models. Specifically, we extracted 768-dimensional features from the review text by sentence-BERT [22] as review features ($v = 1, D^{(1)} = 768$). In the product data, we extracted 4,096-dimensional features from the product images by CaffeNet [23] pre-trained on ImageNet [24] and text features from the description and titles by sentence-BERT. As the entire features, we concatenated the image features, the two kinds of text features and the one-dimensional price data. We regarded these 5,693-dimensional features as product features ($v = 2, D^{(2)} = 5,693$). We used the rating values as the class information ($C = 5$).

[1] https://www.amazon.co.jp/

Table 1: Average of MAE and MZE of DLDGP and comparative methods with their standard deviations.

	MAE	MZE
DLDGP	$\mathbf{0.584 \pm 0.289}$	$\mathbf{0.407 \pm 0.165}$
mGP-LVM [5]	0.906 ± 0.315	0.588 ± 0.148
mLDGP [14]	0.754 ± 0.321	0.511 ± 0.163
sMVCCA [25]	1.82 ± 0.589	0.807 ± 0.068
Deep CCA [26]	0.878 ± 0.326	0.607 ± 0.163
Deep FM [27]	1.47 ± 0.329	0.865 ± 0.103
GPC [28]	0.884 ± 0.278	0.577 ± 0.081

3.2. Experimental Conditions

In this experiment, we divided the dataset into nine to one and regarded them as the training and test sets, respectively. Furthermore, we conducted ten-fold cross-validation for each user. We trained the projection $g(\cdot)$ in 2.1 and 2.3, which maps the product features to the latent space by using the training data in the above cross-validation. In the test phase, we calculated the latent variables corresponding to the product features of a test sample by Eq. (19) and estimated its rating value by Eq. (20). We set the parameters as $\alpha = 0.8$ and $k = 5$, i.e., we extended the rating values on a scale of 1 to 25 with the distributed expression ($D^{(L)} = 25$). We set the dimension of the latent space as $q = 20$, and this was achieved by the above extension of the rating information.

We compared our method with the following methods: mGP-LVM [5], multi-modal label dequantized Gaussian process latent variable model (mLDGP) [14], supervised MVCCA (sMVCCA) [25], deep canonical correlation analysis (deep CCA) [26], deep factorization machine (deep FM) [27] and Gaussian process classifier (GPC) [28]. mGP-LVM does not consider the distributed label dequantization, and mLDGP does not consider the distributed expression of class information. Therefore, these two methods were used to verify the effectiveness of our novel approaches. sMVCCA is a representative correlation analysis-based method for multi-view data using label features as one modality. Deep CCA is a CCA-based method that introduces deep learning which is getting much attention in the field of machine learning. Deep FM is a remarkable method in the field of recommendation systems and is applicable to the rating estimation task. Since GPC is the regression model based on GP, we compared DLDGP with a non-embedding approach. In mLDGP and mGP-LVM, we estimated the label features in the same manner as that of our method. In sMVCCA and deep CCA, we calculated latent variables corresponding to the test features by the projection and estimated its rating value based on the distances between the latent variables of the test sample and those of the training samples.

We evaluated DLDGP and the comparative methods by mean absolute error (MAE) and mean zero-one error (MZE) of the estimated rating values. Since these values of DLDGP and mLDGP were on the scale of 1-25, we converted the values of 1-5, 6-10, 11-15, 16-20 and 21-25 into 1, 2, 3, 4 and 5, respectively.

3.3. Results and Discussion

Table 1 shows the average of MAE and MZE of each method with their standard deviation. By comparing DLDGP with mGP-LVM, we can confirm that the distributed label dequantization scheme contributes to the accuracy improvement for the rating estimation. It is considered that the dimensional flexibility by the dequantization enables to obtain the correlation between multi-view data well. Furthermore, DLDGP achieves higher accuracy than mLDGP, and from this result, we can confirm the effectiveness of the distributed label expression. Thus, we can reasonably argue that considering the class correlation can improve the estimation accuracy. The estimation of DLDGP, mGP-LVM and mLDGP is more accurate than that of sMVCCA, and we can give an insight that the probabilistic approach is beneficial to calculate the common latent space of multi-view data. In addition, DLDGP achieves higher accuracy than deep CCA and deep FM. This implies that DLDGP is effective compared to approaches in different areas. From the comparison with GPC, the embedding approach is suitable for the multi-view data analysis compared to the simple regression model. From the above discussion, the effectiveness of our proposed method has been confirmed.

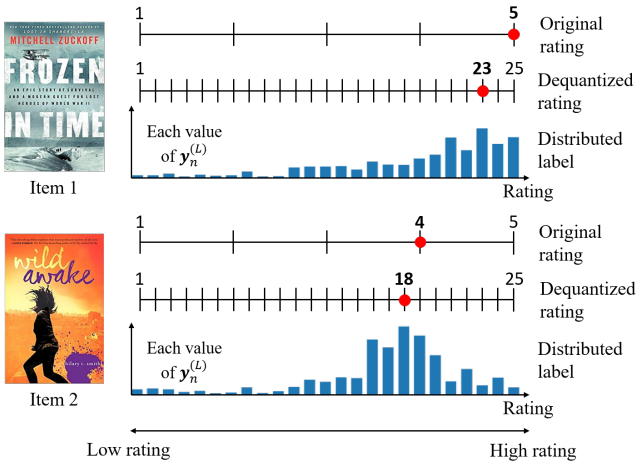

Fig. 1: Examples of the estimated labels by DLDGP.

Figure 1 shows examples of the optimized labels and their corresponding rating values during the training of DLDGP. DLDGP succeeds in extending class information while holding the original rating information, and this result shows the effectiveness of the constraints in Eq. (7) and the optimization explained in 2.2. Furthermore, the values of the label features are distributed with peaks, thus, it allows us to consider the correlation between different classes. Consequently, DLDGP can capture the between-class information by solving the one-hot encoding problem and achieve the improvement of the estimation accuracy.

4. CONCLUSION

In this paper, we have presented a novel method, DLDGP, which can integrate multi-view data into the common latent space. Furthermore, our method can overcome the dimensionality limitation of the label features by the distributed label dequantization scheme and consider the class correlation while optimizing the label features with constraint functions. With these contributions, DLDGP can integrate multi-view data and express class information more accurately. The experimental results have shown the effectiveness of our method.

5. REFERENCES

[1] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proc. the 10th int'l conf. World Wide Web*, 2001, pp. 285–295.

[2] Yehuda Koren, Robert Bell, and Chris Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[3] Jon R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[4] Guoli Song, Shuhui Wang, Qingming Huang, and Qi Tian, "Similarity Gaussian process latent variable model for multimodal data analysis," in *Proc. the IEEE Int'l Conf. Computer Vision*, 2015, pp. 4050–4058.

[5] Aaron P. Shon, Keith Grochow, Aaron Hertzmann, and Rajesh P. N. Rao, "Learning shared latent structure for image synthesis and robotic imitation," in *Proc. Advances in Neural Information Processing Systems*, 2005, pp. 1233–1240.

[6] Carl Henrik Ek, Philip H. S. Torr, and Neil D. Lawrence, "Gaussian process latent variable models for human pose estimation," in *Int'l Workshop on Machine Learning for Multimodal Interaction*, 2007, pp. 132–143.

[7] Stefanos Eleftheriadis, Ognjen Rudovic, and Maja Pantic, "Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition," *IEEE Trans. Image Processing*, vol. 24, no. 1, pp. 189–204, 2015.

[8] Carl Edward Rasmussen, "Gaussian processes in machine learning," *Advanced Lectures on Machine Learning*, vol. 1, no. 1, pp. 63–71, 2003.

[9] Raquel Urtasun and Trevor Darrell, "Discriminative Gaussian process latent variable model for classification," in *Proc. the 24th Int'l Conf. Machine Learning*, 2007, pp. 927–934.

[10] Xinbo Gao, Xiumei Wang, Dacheng Tao, and Xuelong Li, "Supervised Gaussian process latent variable model for dimensionality reduction," *IEEE Trans. Systems, Man, and Cybernetics, Part B*, vol. 2, no. 41, pp. 425–434, 2010.

[11] Yuta Kawachi, Yuma Koizumi, and Noboru Harada, "Complementary set variational autoencoder for supervised anomaly detection," in *Proc. the IEEE Int'l Conf. Acoustics Speech and Signal Processing*, 2018, pp. 2366–2370.

[12] Shipeng Yu, Kai Yu, Volker Tresp, Hans-Peter Kriegel, and Mingrui Wu, "Supervised probabilistic principle component analysis," in *Proc. the 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, 2006, pp. 464–473.

[13] Masanao Matsumoto, Keisuke Maeda, Naoki Saito, Takahiro Ogawa, and Miki Haseyama, "Supervised fractional-order embedding multiview canonical correlation analysis via ordinal label dequantization for image interest estimation," *IEEE Access*, vol. 9, no. 1, pp. 21810–21822, 2021.

[14] Masanao Matsumoto, Keisuke Maeda, Naoki Saito, Takahiro Ogawa, and Miki Haseyama, "Multi-modal label dequantized Gaussian process latent variable model for ordinal label estimation," in *Proc. the IEEE Int'l Conf. Acoustics Speech and Signal Processing*, 2021, pp. 3985–3989.

[15] Neil D. Lawrence and Joaquin Quiñonero-Candela, "Local distance preservation in the GP-LVM through back constraints," in *Proc. the 23rd Int'l Conf. Machine Learning*, 2006, pp. 513–520.

[16] Dennis W. Ruck, Steven K. Rogers, Matthew Kabrisky, Mark E. Oxley, and Bruce W. Suter, "The multilayer perceptron as an approximation to a Bayes optimal discriminant function," *IEEE Trans. Neural Networks*, vol. 1, no. 4, pp. 296–298, 1990.

[17] Martin F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, no. 4, pp. 525–533, 1993.

[18] Andrew R. Conn, Nicholas I.M. Gould, and Philippe Toint, "A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds," *SIAM Journal on Numerical Analysis*, vol. 28, no. 2, pp. 545–572, 1991.

[19] Roberto Battiti, "First- and second-order methods for learning: Between steepest descent and newton's method," *Neural Computation*, vol. 4, no. 2, pp. 141–166, 1992.

[20] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel, "Image-based recommendations on styles and substitutes," in *Proc. the 38th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, 2015, pp. 43–52.

[21] Ruining He and Julian McAuley, "Ups and Downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *Proc. the 25th Int'l Conf. World Wide Web*, 2016, pp. 507–517.

[22] Nils Reimers and Iryna Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proc. the 2019 Conf. Empirical Methods in Natural Language Processing*, 2019, pp. 3982–3992.

[23] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. the 22nd ACM int'l conf. on Multimedia*, 2014, pp. 675–678.

[24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. the IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[25] George Lee, Asha Singanamalli, Haibo Wang, Michael D. Feldman, Stephen R. Master, Natalie N. C. Shih, Elaine Spangler, Timothy Rebbeck, John E. Tomaszewski, and Anant Madabhushi, "Supervised multi-view canonical correlation analysis (sMVCCA): Integrating histologic and proteomic features for predicting recurrent prostate cancer," *IEEE Trans. Medical Imaging*, vol. 34, no. 1, pp. 284–297, 2015.

[26] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *Proc. the 30th Int'l Conf. on Machine Learning*, 2013, pp. 1247–1255.

[27] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He, "DeepFM: A factorization-machine based neural network for CTR prediction," in *Proc. the 26th Int'l Joint Conf. Artificial Intelligence*, 2017, p. 1725–1731.

[28] David Barber and Christopher Williams, "Gaussian processes for Bayesian classification via hybrid Monte Carlo," in *Proc. Advances in Neural Information Processing Systems*, 1997, pp. 340–346.