Generalized Bayesian Canonical Correlation Analysis with Missing Modalities

Toshihiko Matsuura*[1][0000−0002−6276−2276],
Kuniaki Saito*[1][0000−0001−9446−5068],
Yoshitaka Ushiku[1], and Tatsuya Harada[1,2]

[1] The University of Tokyo, Tokyo, Japan
{matsuura, k-saito, ushiku, harada}@mi.t.u-tokyo.ac.jp
[2] RIKEN, Tokyo, Japan

Abstract. Multi-modal learning aims to build models that can relate information from multiple modalities. One challenge of multi-modal learning is the prediction of a target modality based on a set of multiple modalities. However, there are two challenges associated with the goal: Firstly, collecting a large, complete dataset containing all required modalities is difficult; some of the modalities can be missing. Secondly, the features of modalities are likely to be high dimensional and noisy. To deal with these challenges, we propose a method called Generalized Bayesian Canonical Correlation Analysis with Missing Modalities. This method can utilize the incomplete sets of modalities. By including them in the likelihood function during training, it can estimate the relationships among the non-missing modalities and the feature space in the non-missing modality accurately. In addition, this method can work well on high dimensional and noisy features of modalities. This is because, by a probabilistic model based on the prior knowledge, it is strong against outliers and can reduce the amount of data necessary for the model learning even if features of modalities are high dimensional. Experiments with artificial and real data demonstrate our method outperforms conventional methods.

Keywords: multi-modal learning· missing modalities· Bayesian inference· canonical correlation analysis

1 Introduction

In the field of machine learning, multi-modal learning, which models relationships among multiple modalities, has been studied actively. One challenge of multi-modal learning is to construct a predictive model from a set of multiple modalities to a certain modality. We call the modality to be predicted *target modality* and the modality to be used to predict a target modality *source modality*. As a model for estimating the relationship between different modalities, canonical correlation analysis (CCA) [9] is representative, and there are prior studies that actually use CCA for prediction [7]. Also, note that we will call

* Authors contributed equally.

	Source modalities		Target modality
	Modality 1	Modality 2	Modality 3
Element 1	✓	×	✓
Element 2	✓	×	×
Element 3	✓	✓	✓
Element 4	×	✓	✓
.	.	.	.
.	.	.	.
.	.	.	.

Fig. 1. An example of a real sample missing modalities. ✓means that the modality is provided and × means that it is not provided. In this figure, only Element 3 provides all modalities and the others have missing modalities in various patterns. GBCCA-M2 can utilize all elements in learning and predict a target modality from source modalities.

the set of modalities collected from an object an *element*, and we will refer to a group of elements as a *sample* respectively.

There are two challenges in building such a model: some modalities are missing for any reason. For example, in purchaser behavior prediction, some people often refuse to provide some modalities because of their privacy. As Fig. 1 shows, there are various patterns of missing modalities, which makes the problem more difficult. Further, the features of modalities likely to be high dimensional and noisy. The situation occurs when we collect a large amount of information. To deal with these challenges, we propose a method called Generalized Bayesian Canonical Correlation Analysis with Missing Modalities (GBCCA-M2). This method can learn relationships among different modalities utilizing the incomplete sets of modalities by including them in the likelihood function. This study is motivated by the previous works [13, 26] which utilized incomplete sets of modalities. These previous works were proposed to learn the relationships between two different modalities, whereas our method can deal with more than two different modalities. In addition, this method works well on high dimensional and noisy modalities thanks to the prior knowledge incorporated on the parameters of a model. The prior knowledge is introduced to control the sparsity of the weight parameters linking each latent variable to modalities, which makes the model robust to high dimensional and noisy features of modalities. The main contributions of this paper are as follows:

- We propose Generalized Bayesian Canonical Correlation Analysis with Missing Modalities (GBCCA-M2) which is a learning model that can account for elements with missing modalities in the likelihood function.
- Through an experiment using artificial data, we demonstrate that GBCCA-M2 improves prediction performance when using elements with missing modalities, and it is effective for high dimensional and noisy modalities.
- Through an experiment using real data, we demonstrate that GBCCA-M2 is more effective for predicting purchaser behavior and retrieving images from English and Japanese sentences than existing methods.

2 Related Work

Through much research on multi-modal learning, it has been determined in various tasks that performance can be improved by using multiple source modalities rather than using only one [10, 25]. CCA [9] is a method that learns relationships between two modalities. Given pairs of modalities, the model learns to project them into the latent space where they are maximally correlated. CCA has many variants, such as Generalized CCA [5], kernel extensions [15, 1], probabilistic CCA (PCCA) [3] and Bayesian CCA (BCCA) [14, 23]. Generalized CCA extends to CCA in order to capture relationships among more than two modalities. The probabilistic models such as PCCA or BCCA incorporate prior knowledge into their parameters and can learn relationships between high dimensional and noisy features of modalities. We explain the details of BCCA in Section 3. The difficulty in learning relationships between paired modalities is that it is often expensive to collect a large number of paired modalities. In reality, there are a limited number of paired modalities; however, unpaired modalities may be accessible. To overcome this problem, extensions of CCA for semi-supervised learning have been proposed (e.g., Semi CCA [13], Semi PCCA [26, 11]). Semi PCCA can deal with elements that are missing modalities by describing likelihood for use with them. However, this method can only deal with the case where only one of the two modalities is missing. Therefore, we will introduce the methods that are used for general missing data analysis below.

Statistical analysis of missing data is roughly classified as one of the following three types [17]: 1) complete case analysis (CC) [12, 20], 2) imputation of missing values, and 3) describing likelihood for use with missing data. CC is simple, but elements with missing values are not utilized. As for imputation of missing values, this includes mean imputation, regression imputation, or multiple imputation. Methods for complementing missing values by autoencoder [8] have also been developed, and the extensions, such as Cascaded Residual Autoencoder [22] attacks the cases where the modalities are missing. Since imputations of missing values mainly assume that the missing values occur randomly, they are not suitable for the case of missing modalities. As for the studies on describing likelihood for use with missing data [6, 18, 26], it is known that these methods hold looser assumptions than CC and imputation of missing values. However, these methods merely estimate the distribution of data or regress missing values. Although a regression can be performed using parameters learned by Semi PCCA, this is not suitable for the case of multi-modal learning with missing modalities. In our experiments, we use CC and mean imputation to make spuriously complete data for comparison of methods.

3 Generalized Bayesian Canonical Correlation Analysis with Missing Modalities

Since our proposed method is motivated by Bayesian CCA (BCCA) [14, 23] and Semi CCA [13], we will first review these two methods separately.

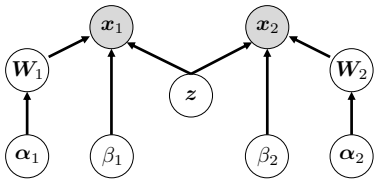

Fig. 2. Graphical illustration of the BCCA model as a plate diagram. The shaded nodes indicate the two observed variables, and the other nodes indicate the model parameters to be estimated. The latent variable $\boldsymbol{z}$ captures the correlation between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

3.1 Bayesian Canonical Correlation Analysis

BCCA [14, 23] is a method that adapts the hierarchical Bayesian model to CCA [9]. Fujiwara et al. [7] proposed a new BCCA model to reconstruct images from human brain information. As shown in Fig. 2, the new model captures the relationships between the two modalities. In the model, modalities $\boldsymbol{x}_i \in \mathbb{R}^{d_i}$ $(i = 1, 2)$ are generated by common latent variables $\boldsymbol{z} \in \mathbb{R}^{d_z}, d_z \leq \min(d_i)$ and weight matrices $\boldsymbol{W}_i \in \mathbb{R}^{d_i \times d_z}$, where d_i and d_z represent the dimension of modalities and latent variables, respectively. In addition, weight matrices are controlled by parameters $\boldsymbol{\alpha}_i \in \mathbb{R}^{d_i \times d_z}$. The likelihood of the modalities is

$$P(\boldsymbol{x}_i|\boldsymbol{W}_i, \boldsymbol{z}) \propto \exp\left(-\frac{1}{2}\sum_{n=1}^{N}\left(\boldsymbol{x}_i(n) - \boldsymbol{W}_i\boldsymbol{z}(n)\right)^{\mathrm{T}}\beta_i\left(\boldsymbol{x}_i(n) - \boldsymbol{W}_i\boldsymbol{z}(n)\right)\right), \quad (1)$$

where $\beta_i \mathrm{I}_{d_z} (\beta_i \in \mathbb{R}^1)$ represents covariance of the Gaussian distribution, I_d represents a $d \times d$ identity matrix, and, N represents the sample size. The prior distribution of latent variables is

$$P_0(\boldsymbol{z}) \propto \exp\left(-\frac{1}{2}\sum_{n=1}^{N}\|\boldsymbol{z}(n)\|^2\right). \quad (2)$$

Latent variables are generated from the Gaussian distribution whose mean is $\mathbf{0}$ and whose covariance is I. The prior distribution of weight matrices is

$$P_0(\boldsymbol{W}_i|\boldsymbol{\alpha}_i) \propto \exp\left(-\frac{1}{2}\sum_{s=1}^{d_i}\sum_{t=1}^{d_z}\alpha_{i_{(s,t)}}W_{i_{(s,t)}}^2\right). \quad (3)$$

The (s, t) element of weight matrices is generated from the Gaussian distribution whose mean is $W_{i_{(s,t)}}$ and whose covariance is $\alpha_{i(s,t)}$. Weight matrices are controlled by hyper-parameters $\boldsymbol{\alpha}_i$, whose hyper-prior distribution is

$$P_0(\boldsymbol{\alpha}_i) = \prod_{s=1}^{d_i}\prod_{t=1}^{d_z}\mathcal{G}(\alpha_{i_{(s,t)}}|\overline{\alpha}_{i_{(s,t)}}, \gamma_{i_{(s,t)}}), \quad (4)$$

where $\mathcal{G}(\alpha|\overline{\alpha}, \gamma)$ is the Gamma distribution whose mean is $\overline{\alpha}$ and whose confidence parameter is γ. This probability model (Eq. (1), (4)) is known as automatic

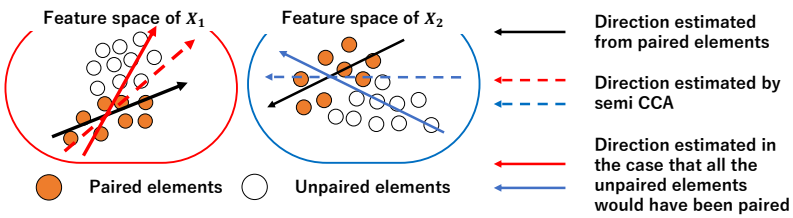

Fig. 3. An example of spatial estimation by Semi CCA. By using unpaired elements, we can estimate a direction closer to that estimated in the case that all the unpaired elements would have been paired, than by using only paired elements.

relevance determination (ARD) [19], which drives unnecessary components to zero. The prior distribution of observation noise β_i is

$$P_0(\beta_i) = \frac{1}{\beta_i}, \tag{5}$$

which is called non-informative priors. Parameters are estimated by variational Bayesian inference [2], and the predictive distribution of the target modality is driven using these estimated parameters.

3.2 Semi Canonical Correlation Analysis

Semi CCA [13] is a method that extends CCA to a semi-supervised one by combining CCA and principal component analysis (PCA). We denote the group of elements whose modalities are paired as P, the ones whose are not paired as U, and the sample covariance matrices as $\boldsymbol{\Sigma}$s. The solution of Semi CCA can be obtained by solving the following general eigenvalue problem.

$$B \begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{pmatrix} = \lambda C \begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{pmatrix}, \tag{6}$$

$$B = \beta \begin{pmatrix} \mathbf{0} & \boldsymbol{\Sigma}_{12}^{(P)} \\ \boldsymbol{\Sigma}_{21}^{(P)} & \mathbf{0} \end{pmatrix} + (1-\beta) \begin{pmatrix} \boldsymbol{\Sigma}_{11}^{(P+U)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22}^{(P+U)} \end{pmatrix}, \tag{7}$$

$$C = \beta \begin{pmatrix} \boldsymbol{\Sigma}_{12}^{(P)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{21}^{(P)} \end{pmatrix} + (1-\beta) \begin{pmatrix} \mathrm{I}_{D_1} & \mathbf{0} \\ \mathbf{0} & \mathrm{I}_{D_2} \end{pmatrix}. \tag{8}$$

β represents the contribution ratio of CCA to PCA. Similar to this, we introduce contribution rates of elements missing modalities to GBCCA-M2. Semi CCA has an application in probabilistic models, such as Probabilistic Semi CCA [26, 11]. However, they are not suitable for high dimensional and noisy features of modalities because the premise that weight matrices become sparse in learning is not assumed, which causes overfitting.

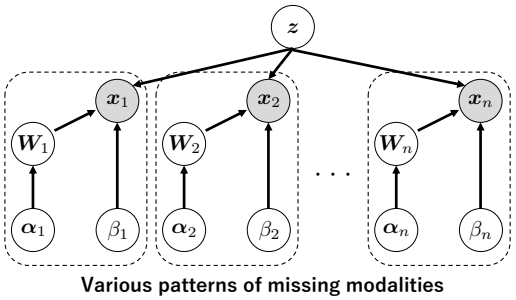

Fig. 4. Graphical illustration of the GBCCA-M2 model as a plate diagram. Each element used in learning has two or more modalities and various missing patterns. The shaded nodes $\boldsymbol{x}_i$ indicate the observed variables. The latent variable z captures the correlations among the $\boldsymbol{x}_i$s.

3.3 Generalized Bayesian Canonical Correlation Analysis with Missing Modalities

As mentioned in Section 1, some of the modalities are missing and the features of them are high dimensional and noisy. Considering these characteristics, the following functions are necessary for our method: F1) dealing with various patterns of missing modalities, F2) dealing with more than two different modalities, F3) highly accurate prediction for high dimensional and noisy features of modalities. BCCA meets F3, so we extend it so as to meet F1 and F2 through the proposed GBCCA-M2. We construct the model of GBCCA-M2 while considering the following: 1) the number of modalities should be increased more than two, and all modalities are generated from common latent variables and 2) the contribution rates to the likelihood are changed according to how many modalities are missing. The graphical model of GBCCA-M2 is shown in Fig. 4. Now, we introduce the likelihood and prior distribution of GBCCA-M2, parameter estimation by a variational Bayesian inference, and the prediction of target modality using source modalities and estimated parameters.

The likelihood and prior distribution: The likelihood of modalities is

$$P\left(\boldsymbol{x}_i|\boldsymbol{W}_i,\boldsymbol{z}\right)=\prod_{m=1}^{M}P\left(\boldsymbol{x}_i^{(m)}|\boldsymbol{W}_i,z^{(m)}\right)^{\eta_m} \tag{9}$$

$$P\left(\boldsymbol{x}_i^{(m)}|\boldsymbol{W}_i,\boldsymbol{z}^{(m)}\right)\propto\exp\left(-\frac{1}{2}\sum_{n=1}^{N_i^{(m)}}\left(\boldsymbol{x}_i^{(m)}(n)-\boldsymbol{W}_i\boldsymbol{z}^{(m)}(n)\right)^{\mathrm{T}}\beta_i\left(\boldsymbol{x}_i^{(m)}(n)-\boldsymbol{W}_i\boldsymbol{z}^{(m)}(n)\right)\right), \tag{10}$$

where $\boldsymbol{x}_i^{(m)}$ represents the i-th modality of an element that has m sets of modalities, M represents the number of modalities, and $N_i^{(m)}$ represents the number of elements which have m sets of modalities and the i-th modality of them is not missing. Moreover, we introduce contribution rates η_m of elements missing

modalities to the likelihood function and change them according to the degree of missing modalities. Especially, the more modalities are missing, the smaller the contribution rates should be ($\eta_1 < \eta_2 < \eta_3 < \cdots$), and the more elements missing modalities are, the smaller contribution rates should be, which is reflected in Fig. 5. Owing to them, we can properly utilize elements missing modalities. As with BCCA, prior distributions and the hyper-prior distribution of each parameter are as follows:

$$P_0\left(\boldsymbol{z}^{(m)}\right) \propto \exp\left(-\frac{1}{2}\sum_{n=1}^{N^{(m)}} \|\boldsymbol{z}^{(m)}(n)\|^2\right), \tag{11}$$

$$P_0\left(\boldsymbol{W}_i|\boldsymbol{\alpha}_i\right) \propto \exp\left(-\frac{1}{2}\sum_{s=1}^{d_i}\sum_{t=1}^{d_z} \alpha_{i_{(s,t)}} W_{i_{(s,t)}}^2\right), \tag{12}$$

$$P_0(\boldsymbol{\alpha}_i) = \prod_{s=1}^{d_i}\prod_{t=1}^{d_z} \mathcal{G}\left(\alpha_{i_{(s,t)}}|\overline{\alpha}_{i_{(s,t)}}, \gamma_{i_{(s,t)}}\right), \tag{13}$$

$$P_0\left(\beta_i\right) = \frac{1}{\beta_i}. \tag{14}$$

Parameter estimation by variational Bayesian inference: Given the likelihood (Eq. (9), (10)); the prior distribution (Eq. (11), (12), (14)); and the hyper-prior distribution (Eq. (13)), weight matrices are estimated as the posterior distribution $P(\boldsymbol{W}_1, \cdots, \boldsymbol{W}_M|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_M)$. This posterior distribution is obtained by marginalizing the joint posterior distributions with respect to latent variables and variance parameters $\boldsymbol{\alpha}_i, \beta_i$ as follows:

$$\begin{aligned} P\left(\boldsymbol{W}_1, \cdots, \boldsymbol{W}_M|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_M\right) = \int \mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{\alpha}_1 \cdots \mathrm{d}\boldsymbol{\alpha}_M \mathrm{d}\beta_1 \cdots \mathrm{d}\beta_M \\ P(\boldsymbol{W}_1, \cdots, \boldsymbol{W}_M, \boldsymbol{z}, \boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_M, \beta_1, \cdots, \beta_M|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_M). \end{aligned} \tag{15}$$

This joint posterior distribution cannot be calculated analytically, so it is approximated by using a trial distribution with the following factorization based on variational Bayes inference.

$$\begin{aligned} &Q\left(\boldsymbol{W}_1, \cdots, \boldsymbol{W}_M, \boldsymbol{z}, \boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_M, \beta_1, \cdots, \beta_M\right) \\ =&Q_W(\boldsymbol{W}_1) \cdots Q_W(\boldsymbol{W}_M) Q_z(\boldsymbol{z}) Q_\alpha(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_M, \beta_1, \cdots, \beta_M). \end{aligned} \tag{16}$$

The trial distribution of weight matrices $Q_W\left(\boldsymbol{W}_i\right)$ is

$$Q_W\left(\boldsymbol{W}_i\right) = \prod_{s=1}^{d_i}\prod_{t=1}^{d_z} \mathcal{N}\left(W_{i_{(s,t)}}|\overline{W}_{i_{(s,t)}}, \sigma_{i_{(s,t)}}^{-1}\right), \tag{17}$$

$$\overline{W}_{i_{(s,t)}} = \overline{\beta}_i \sigma_{i_{(s,t)}}^{-1} \sum_{m=1}^{M}\left(\eta_m \cdot \sum_{n=1}^{N_i^{(m)}} x_{i_s}^{(m)}(n) z_t^{(m)}(n)\right), \tag{18}$$

$$\sigma_{i_{(s,t)}}^{-1} = \overline{\beta}_i \sum_{m=1}^{M}\left(\eta_m \cdot \sum_{n=1}^{N_i^{(m)}} {z_t^{(m)}}^2(n) + N_i^{(m)} \Sigma_{z^{(m)}(t,t)}^{-1}\right) + \overline{\alpha}_{i_{(s,t)}}. \tag{19}$$

The trial distribution of latent variable $Q_z\left(\boldsymbol{z}^{(m)}\right)$ is

$$Q_z\left(\boldsymbol{z}^{(m)}\right) = \prod_{n=1}^{N^{(m)}} \mathcal{N}\left(\boldsymbol{z}^{(m)}(n)|\overline{\boldsymbol{z}}^{(m)}(n), \boldsymbol{\Sigma}_{z^{(m)}}^{-1}\right), \tag{20}$$

$$\overline{\boldsymbol{z}}^{(m)}(n) = \boldsymbol{\Sigma}_{z^{(m)}}^{-1} \sum_{i=1}^{M} \eta_m \overline{\beta}_i \overline{\boldsymbol{W}}_i^{\mathrm{T}} \boldsymbol{x}_i^{(m)}(n), \tag{21}$$

$$\boldsymbol{\Sigma}_{z^{(m)}} = \sum_{i=1}^{M} \left[\eta_m \overline{\beta}_i \left(\overline{\boldsymbol{W}}_i^{\mathrm{T}} \overline{\boldsymbol{W}}_i + \boldsymbol{\Sigma}_{W_i}^{-1}\right)\right] + \mathrm{I}, \tag{22}$$

$$\boldsymbol{\Sigma}_{W_i} = \mathrm{diag}\left(\left[\sum_{s=1}^{d_i} \sigma_{i_{(s,1)}}, \cdots, \sum_{s=1}^{d_i} \sigma_{i_{(s,d_z)}}\right]\right). \tag{23}$$

Finally, the trial distribution of the inverse variances $Q_\alpha(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_M, \beta_1, \cdots, \beta_M)$ is further factorized to $Q_\alpha(\boldsymbol{\alpha}_1) \cdots Q_\alpha(\boldsymbol{\alpha}_M) Q_\alpha(\beta_1) \cdots Q_\alpha(\beta_M)$. The expected values of $\boldsymbol{\alpha}_i$ and β_i are

$$\overline{\alpha}_{i_{(s,t)}} = \left(\frac{1}{2} + \gamma_{i0_{(s,t)}}\right) \left(\frac{1}{2}\overline{W}_{i_{(s,t)}}^2 + \frac{1}{2}\sigma_{i_{(s,t)}}^{-1} + \gamma_{i0_{(s,t)}} \alpha_{i0_{(s,t)}}^{-1}\right)^{-1}, \tag{24}$$

$$\begin{aligned} \overline{\beta}_i = d_i N_i^{(M)} \Bigg\{ & \sum_{n=1}^{N_i^{(M)}} \|\boldsymbol{x}_i(n) - \overline{\boldsymbol{W}}_i \overline{\boldsymbol{z}}(n)\|^2 \\ & + \mathrm{Tr}\left[\boldsymbol{\Sigma}_{W_i}^{-1} \left(\sum_{n=1}^{N_i^{(M)}} \boldsymbol{z}(n)\boldsymbol{z}^{\mathrm{T}}(n) + N_i^{(M)} \boldsymbol{\Sigma}_z^{-1}\right) + N_i^{(M)} \boldsymbol{\Sigma}_z^{-1} \overline{\boldsymbol{W}}_i^{\mathrm{T}} \overline{\boldsymbol{W}}_i\right] \Bigg\}^{-1}, \end{aligned} \tag{25}$$

where $\gamma_{i0_{(s,t)}}, \alpha_{i0_{(s,t)}}$ are constant values (zero in our study). For estimating β_i, only elements having all modalities are used. By calculating $Q_W(\boldsymbol{W}_i)$, $Q_z(\boldsymbol{z})$, and $Q_\alpha(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_M, \beta_1, \cdots, \beta_M)$ successively, the parameter are estimated.

Predictive distribution: When the new set of source modalities $\boldsymbol{X}_{\mathrm{new}} \in \mathfrak{P}(\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{M-1}\})$, where $\mathfrak{P}$ represents a power set (a set of all subsets), is obtained, the predictive distribution of the target modality $\boldsymbol{x}_{M_{\mathrm{new}}}$ is

$$P(\boldsymbol{x}_{M\mathrm{new}}|\boldsymbol{X}_{\mathrm{new}}) = \int \mathrm{d}\boldsymbol{W}_M \mathrm{d}\boldsymbol{z}_{\mathrm{new}} P(\boldsymbol{x}_{M\mathrm{new}}|\boldsymbol{W}_M, \boldsymbol{z}_{\mathrm{new}}) Q(\boldsymbol{W}_M) P(\boldsymbol{z}_{\mathrm{new}}|\boldsymbol{X}_{\mathrm{new}}). \tag{26}$$

When the random variable $\boldsymbol{W}_M$ is replaced with the estimated $\overline{\boldsymbol{W}}_M$, the predictive distribution is

$$P(\boldsymbol{x}_{M\mathrm{new}}|\boldsymbol{X}_{\mathrm{new}}) \simeq \int \mathrm{d}\boldsymbol{z}_{\mathrm{new}} P(\boldsymbol{x}_{M\mathrm{new}}|\boldsymbol{z}_{\mathrm{new}}) P(\boldsymbol{z}_{\mathrm{new}}|\boldsymbol{X}_{\mathrm{new}}), \tag{27}$$

$$P(\boldsymbol{x}_{M\mathrm{new}}|\boldsymbol{z}_{\mathrm{new}}) \propto \exp\left[-\frac{1}{2}\overline{\beta}_M \|\boldsymbol{x}_{M\mathrm{new}} - \overline{\boldsymbol{W}}_M \boldsymbol{z}_{\mathrm{new}}\|^2\right]. \tag{28}$$

Since the distribution $P(\boldsymbol{z}_{\mathrm{new}}|\boldsymbol{X}_{\mathrm{new}})$ is an unknown distribution, it is approximated based on the test distribution $Q_z(\boldsymbol{z})$ (Eq. (20)). The approximate distri-

bution is obtained by using only the term related to $\boldsymbol{x}_{i\mathrm{new}}$ included in $\boldsymbol{X}_{\mathrm{new}}$.

$$\tilde{Q}_z(\boldsymbol{z}_{\mathrm{new}}) = \mathcal{N}\left(\boldsymbol{z}|\overline{\boldsymbol{z}}_{\mathrm{new}}, \boldsymbol{\Sigma}_{z\mathrm{new}}^{-1}\right), \tag{29}$$

$$\overline{\boldsymbol{z}}_{\mathrm{new}} = \sum_{i=1}^{M-1} \overline{\beta}_i \boldsymbol{\Sigma}_{z\mathrm{new}}^{-1} \overline{\boldsymbol{W}}_i^{\mathrm{T}} \boldsymbol{x}_{i\mathrm{new}}, \tag{30}$$

$$\boldsymbol{\Sigma}_{z\mathrm{new}} = \sum_{i=1}^{M-1} \left(\overline{\beta}_i \left(\overline{\boldsymbol{W}}_i^{\mathrm{T}} \overline{\boldsymbol{W}}_i + \boldsymbol{\Sigma}_{W_i}^{-1}\right)\right) + \mathrm{I}. \tag{31}$$

Finally, the prediction distribution $P(\boldsymbol{x}_{M\mathrm{new}}|\boldsymbol{X}_{\mathrm{new}})$ is

$$\begin{aligned} P(\boldsymbol{x}_{M\mathrm{new}}|\boldsymbol{X}_{\mathrm{new}}) &\simeq \int \mathrm{d}\boldsymbol{z}_{\mathrm{new}} P\left(\boldsymbol{x}_{M\mathrm{new}}|\boldsymbol{z}_{\mathrm{new}}\right) \tilde{Q}_z\left(\boldsymbol{z}_{\mathrm{new}}\right) \\ &= \mathcal{N}\left(\boldsymbol{x}_{M\mathrm{new}}|\overline{\boldsymbol{x}}_{M\mathrm{new}}, \boldsymbol{\Sigma}_{M\mathrm{new}}^{-1}\right), \end{aligned} \tag{32}$$

$$\overline{\boldsymbol{x}}_{M\mathrm{new}} = \overline{\boldsymbol{W}}_M \boldsymbol{\Sigma}_{z\mathrm{new}}^{-1} \sum_{i=1}^{M-1} \overline{\beta}_i \overline{\boldsymbol{W}}_i^{\mathrm{T}} \boldsymbol{x}_{i\mathrm{new}}, \tag{33}$$

$$\boldsymbol{\Sigma}_{M\mathrm{new}} = \overline{\boldsymbol{W}}_M \boldsymbol{\Sigma}_{z\mathrm{new}}^{-1} \overline{\boldsymbol{W}}_M^{\mathrm{T}} + \overline{\beta}_M^{-1} \mathrm{I}. \tag{34}$$

4 Preliminary Investigation

We conducted three experiments to investigate the basic characteristics of GBCCA-M2 using artificially generated data. In this section, we firstly describe the common experimental setup and then explain each experiment.

4.1 Common experimental setup

As a method for generating artificial data, we used a simple Gaussian latent model. The latent variables are denoted by $\boldsymbol{Z}_{\mathrm{gen}} = \left\{\boldsymbol{z}_{\mathrm{gen}}(n)\right\}_{n=1}^{N} \in \mathbb{R}^{d_{z_{\mathrm{gen}}}}$ and observed modalities are denoted by $\boldsymbol{X}_i = \left\{\boldsymbol{x}_i(n)\right\}_{n=1}^{N} \in \mathbb{R}^{d_i}$. In this section, we considered the case of three observed modalities. $d_{z_{\mathrm{gen}}}$ and d_i represent the dimension of the latent variables and modalities respectively, and N represents the sample size. Latent variables were extracted independently from $\mathcal{N}\left(\mathbf{0}, \mathrm{I}_{d_z}\right)$. $\boldsymbol{x}_i(n)$ were generated as follows: $\boldsymbol{x}_i(n) = \boldsymbol{W}_i \boldsymbol{z}_{\mathrm{gen}}(n) + \boldsymbol{\mu}_i + \boldsymbol{\delta}_i(n)$, where each row of $\boldsymbol{W}_i$ was extracted from $\mathcal{N}(\mathbf{0}, \mathrm{I}_{d_{z_{\mathrm{gen}}}})$, mean $\boldsymbol{\mu}_i$ was extracted from $\mathcal{N}(\mathbf{0}, \mathrm{I}_{d_i})$, and covariance of noise $\boldsymbol{\delta}_i(n)$ was determined as follows:

$$\boldsymbol{\delta}_i(n) = \alpha \left(\mathrm{I}_{d_i} + \sum_{j=1}^{\frac{d_{z_{\mathrm{gen}}}}{2}} \boldsymbol{u}_j(n) \boldsymbol{u}_j(n)^{\mathrm{T}}\right). \tag{35}$$

$\boldsymbol{u}_j(n)$ were extracted independently from $\mathcal{N}\left(\mathbf{0}, \mathrm{I}_{d_i}\right)$. The magnitude of the noise is controlled by α, which was changed in the experiment evaluating robustness

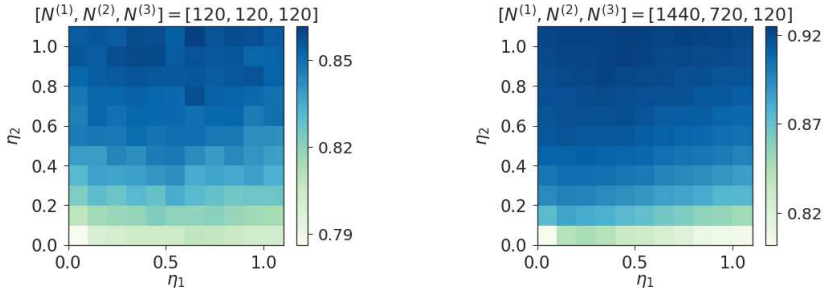

Fig. 5. Prediction performance when the contribution rates were changed.

against noise, and fixed in the other experiments. The number of elements in the test data was set to 500. $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ were set to the source modalities, and $\boldsymbol{X}_3$ the target modality. The evaluation was performed by calculating the cosine similarities between the predicted modality and that of the test data.

4.2 Contribution rates of elements missing modalities

In GBCCA-M2, in order to utilize elements with various missing patterns efficiently, we introduced the contribution rates of elements missing modalities to the likelihood as shown in Eq. (9). In this experiment, we investigated the change in prediction performance when contribution rates were changed.

The dimension of each modality was set as $[d_1, d_2, d_3, d_{z_{\text{gen}}}]$=$[250, 250, 250, 50]$. When the number of modalities was three, the patterns of missing modalities were divided into three categories of elements with one, two, and three modalities, respectively. We defined the number of elements with m sets of modalities $N^{(j)}$ and set them as $[N^{(1)}, N^{(2)}, N^{(3)}] = [120, 120, 120]$ and $[N^{(1)}, N^{(2)}, N^{(3)}] = [1440, 720, 120]$ (refer to Fig. 5). Moreover, the modality an element was missing was made uniform in each pattern. This was the same in all experiments. In Eq. (9), we fixed η_3 at 1.0 and varied η_1 and η_2 by increments of 0.1 in the range 0 to 1.0. Also, the dimension of latent variable $\boldsymbol{z}$ used in the proposed method was set to 150. Experiments were repeated ten times for each set of (η_1, η_2), and the average of cosine similarity was calculated.

The experimental results are shown in Fig. 5. Since the cosine similarity became maximal when η_2 was in the range 0.9 to 1.0, η_2 should be set to a value close to 1.0. This is because even if one modality is missing, it is possible to estimate parameters with the remaining two modalities. On the other hand, since the cosine similarity became maximal when η_1 was in the range 0.4 to 0.6, η_1 should be set to be smaller than η_2. This is because the element with one modality seems to be useful for estimating the distribution in the feature space of each modality, but it seems to deteriorate the estimation of the relationships between modalities. Moreover, since η_1 and η_2, which maximized cosine similarity when $[N^{(1)}, N^{(2)}, N^{(3)}] = [1440, 720, 120]$, were lower than when

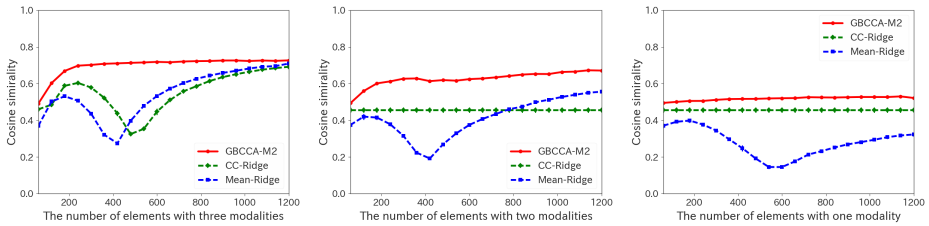

Fig. 6. Predict performance when the number of elements was changed.

$[N^{(1)}, N^{(2)}, N^{(3)}] = [120, 120, 120]$, the contribution rates should be decreased as the number of elements missing modalities increases.

4.3 The number of elements in training

GBCCA-M2 utilizes elements missing modalities by including them in the likelihood function. In this experiment, we changed the number of elements in training according to the degree of missing modalities and investigated whether GBCCA-M2 can utilize elements missing modalities effectively.

Among the three kinds of missing patterns, the number of elements of any two patterns was fixed, and the number of elements of the remaining one pattern was changed. The number of elements to be fixed was set to 60 and the number of elements to be changed was set to $60, 120, \cdots, 1200$. The dimension of each modality and the contribution rates were set as follows: $[d_1,\ d_2,\ d_3,\ d_{z_{\mathrm{gen}}}] = [250,\ 250,\ 250,\ 50]$, $[\eta_1, \eta_2, \eta_3] = [0.4, 0.9, 1.0]$. Also, the dimension of latent variable $\boldsymbol{z}$ used in the GBCCA-M2 was set to 150. We used the following two methods for comparison: 1) CC and ridge regression (CC-Ridge) and 2) mean imputation and ridge regression (Mean-Ridge). CC-Ridge removes elements with missing modalities and performs ridge regression using the remaining elements. Ridge regression is a learning method that adds a square of the weight to the loss function in the linear least squares method and obtains a weight that minimizes it. Mean-Ridge substitutes the mean value of elements in the missing modalities and performs ridge regression.

Fig. 6 shows the experimental results. When the number of elements with two modalities was increased in GBCCA-M2, the prediction performance approximately monotonically increased. This may be because elements with two modalities have a positive effect on the relationship estimation between the non-missing modalities and the estimation of the feature amount space in the non-missing modality. On the other hand, when the number of elements with one modality was increased, the prediction performance improved only in the range where the number of elements was small. This may be because when contribution rates are fixed, as the number of elements with one modality is increased, the negative effect on relationship estimation between the non-missing modality and the missing modality increases. Therefore, if η_1 is set appropriately, it should be possible to use elements with one modality effectively for learning.

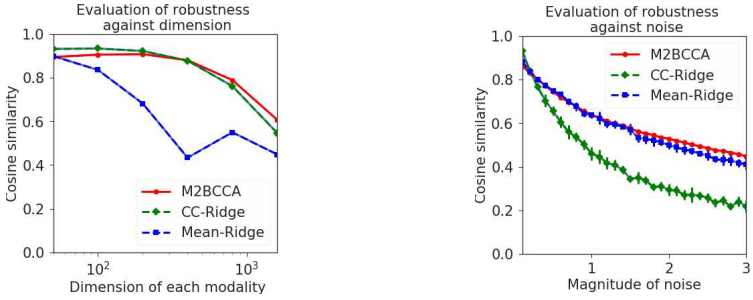

Fig. 7. Prediction performances when the dimension of modalities was changed (left) and when the noise of modalities was changed (right).

4.4 Evaluating robustness against dimension and noise

We described that the features of modalities are likely to be high dimensional and noisy. In order to show the effectiveness of GBCCA-M2 for such modalities, we conducted experiments to evaluate robustness against dimension and noise.

In the experiments evaluating robustness against dimension, we changed the parameter β, which represents the size of the dimension. The dimension of each modality was set to $[d_1, d_2, d_3, d_{z_{\text{gen}}}] = [50\beta, 50\beta, 50\beta, 10\beta]$, and the dimension of the latent variable z used in GBCCA-M2 was set to $30\,\beta$. We set β to 1, 2, 4, 8, 16, or 32. In the experiment evaluating robustness against noise, we changed α ,which controlled the magnitude of the noise (Eq.(35)), by increments of 0.1 in the range 0.1 to 3.0. The dimension of each modality was set to $[d_1, d_2, d_3, d_{z_{\text{gen}}}] = [250, 250, 250, 50]$ and the dimension of latent variable $\boldsymbol{z}$ used in GBCCA-M2 was set to 150. In both experiments, the numbers of elements in training were set to 120 for all missing patterns. Also, the contribution rates were set as follows: $[\eta_1, \eta_2, \eta_3] = [0.4, 0.9, 1.0]$. As the comparison method, we used the same two methods as in the Experiment in Section 4.3.

Fig. 7 shows the experimental results. When the dimension or noise of modality increased, GBCCA-M2 achieved higher prediction performance than the comparison methods. This may be because GBCCA-M2 is based on BCCA, which is effective for high dimensional and noisy features of modalities. Experimental results show that GBCCA-M2 is also effective for such cases.

5 Experiment with Real Data

5.1 Purchaser Behavior Prediction

We conducted an experiment to show the effectiveness of GBCCA-M2 using real purchaser dataset, in which modalities are actually missing. For the purchaser dataset, we used the INTAGE Single Source Panel (i-SSP) dataset from INTAGE Inc. This dataset includes attributes, purchase histories, and television program viewing information for the half year from January 1st, 2016 to June 30th, 2016.

Attribute	Purchase	TV	The number of elements
✓	×	×	2683
✓	✓	×	893
✓	×	✓	2297
✓	✓	✓	809

Table 1. The number of elements by missing patterns in purchaser's data.

Method	Cosine similarity	MAE	RMSE
GBCCA-M2	**0.408**	**104.8**	**383.2**
CC-Ridge	0.278	153.6	564.4
Mean-Ridge	0.397	109.3	397.7
CC-BCCA	0.404	113.5	390.2
Mean-BCCA	0.407	105.6	390.9
Semi CCA	0.402	105.6	389.4

Table 2. Comparison of each method in the actual purchaser's data.

In the attribute data, we converted the nominal scales such as occupation and residence to one-hot expression and used the proportional scales as they were. Purchasing information includes purchase data of beer, chocolate, and shampoo. We used the total number of purchases for each manufacturer as one modality. For the television program viewing information, we used the average television viewing time for each television program only if it was 20 hours or more. As a result of the above operation, the dimension of attribute information was 89, that of purchase situation was 67, and that of TV program viewing information was 226. Table 1 indicates the number of elements for each missing pattern. We extracted 100 elements with three modalities randomly as test data and used the remaining elements as learning data. We set the contribution rates and the dimension of the latent variable in GBCCA-M2 as follows: $[\eta_1, \eta_2, \eta_3, d_z] = [0.3, 0.8, 1.0, 30]$. In addition to CC-Ridge and Mean-Ridge, we used CC and BCCA (CC-BCCA), mean imputation and BCCA (Mean-BCCA), and Semi CCA for comparison. Television program viewing information was predicted from source modalities (i.e., attribute and purchase history). As the evaluation index, we calculated the following indexes using the predicted vector and the actual vector: 1) cosine similarity, 2) mean absolute error (MAE), and 3) root mean square error (RMSE). We did this 30 times and calculated the average.

Table 2 shows the experimental results. As for all evaluation index, GBCCA-M2 achieved best. This may be because GBCCA-M2 is effective for purchaser data in which features of modalities are high dimensional and noisy and there are many elements missing modalities. From the above findings, the effectiveness of GBCCA-M2 for a real purchaser dataset can be seen clearly.

5.2 Image Retrieval from English and Japanese Sentences

In this section, we report results on image retrieval from English and Japanese sentences learned with the dataset in which we made some modalities missing intentionally. In addition to MSCOCO [16] dataset, we used STAIR Captions [24], which is a Japanese image caption dataset based on images from MSCOCO. As the feature of images, we extracted the 4096-dimensional activations from 19-layer VGG model [21], and as the feature of sentences, we used

Method	R@1	R@5	R@10
GBCCA-M2	**0.092**	**0.292**	**0.439**
CC-Ridge	0.074	0.263	0.411
Mean-Ridge	0.080	0.265	0.412
CC-BCCA	0.084	0.268	0.409
Mean-BCCA	0.082	0.262	0.382
Semi CCA	0.071	0.244	0.371

Table 3. Comparison of each method in the sentence-to-image retrieval.

tf-idf-weighted bag-of-words vectors. For English, we pre-processed all the sentences with WordNet's lemmatizer [4] and removed stop words. For Japanese, we removed stop words and all parts of speech other than nouns, verbs, adjectives, and adjectival verbs. The final dimensions of English and Japanese sentences were 6245 and 7278, respectively. In training, we used 9000 elements (i.e, images and their corresponding English and Japanese sentences), made 50 percents modalities missing randomly, and reduced the dimension of each modality to 1000 by PCA. For the evaluation, we used 1000 elements. We retrieved images form English and Japanese sentences and calculated Recall@K ($K = 1, 5, 10$). We set the contribution rates and the dimension of the latent variable as follows: $[\eta_1, \eta_2, \eta_3, d_z] = [0.3, 0.8, 1.0, 750]$ and used same methods in Section 5.1 as comparison methods. Table 3 shows the experimental results. We can see that GBCCA-M2 gives best results in all methods. By using GBCCA-M2, we can retrieve images more accurately by utilizing elements missing modalities.

6 Conclusion

In this study, we considered the two challenges associated with multi-modal learningand proposed GBCCA-M2, which utilizes elements missing modalities and can work well on high dimensional and noisy features of modalities. Moreover, we conducted experiments using artificially generated data as well as real data. The findings obtained in this study are as follows: 1) in order to utilize the elements missing modalities, it is effective to change the contribution rates to likelihood according to the degree of missing modalities, 2) GBCCA-M2, which uses a hierarchical Bayesian model, is effective for high dimensional and noisy features of modalities, and 3) because GBCCA-M2 is suited to the case that there are many elements missing modalities, and the features of modalities are high dimensional and noisy, it is effectively used for such multi-modal applications.

Acknowledgments

This work was partially funded by ImPACT Program of Council for Science, Technology and Innovation (Cabinet Office, Government of Japan) and INTAGE HOLDINGS Inc.

References

1. Akaho, S.: A kernel method for canonical correlation analysis. In: Proceedings of the International Meeting of the Psychometric Society. Springer-Verlag (2001)
2. Attial, H.: Inferring parameters and structure of latent variable models by variational bayes. In: Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. pp. 21–30 (1999)
3. Bach, F.R., Jordan, M.I.: A probabilistic interpretation of canonical correlation analysis (2005), tR 688
4. Bird, S., Loper, E.: Nltk: The natural language toolkit. In: Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions (2014)
5. Carroll, J.: Generalization of canonical correlation analysis to three or more sets of variables. In: Proceedings of the American Psychological Association. vol. 3, pp. 227–228 (1968)
6. Enders, C.K., Bandalos, D.L.: The relative performance of full information maximum likelihood estimation for missing data in structural equation models. Structural Equation Modeling: A Multidisciplinary Journal **8**(3), 430–457 (2001)
7. Fujiwara, Y., Miyawaki, Y., Kamitani, Y.: Estimating image bases for visual image reconstruction from human brain activity. In: Advances in Neural Information Processing Systems 22, pp. 576–584. Curran Associates, Inc. (2009)
8. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**, 504–507 (2006)
9. Hotelling, H.: Relations between two sets of variates. Biometrika **28**(3), 321–377 (1936)
10. Huiskes, M.J., Thomee, B., Lew, M.S.: New trends and ideas in visual concept detection: the mir flickr retrieval evaluation initiative. In: Proceedings of the international conference on Multimedia information retrieval. pp. 527–536 (2010)
11. Kamada, C., Kanezaki, A., Harada, T.: Probabilistic semi-canonical correlation analysis. In: Proceedings of the 23rd ACM International Conference on Multimedia. pp. 1131–1134 (2015)
12. Kim, J.O., Curry, J.: The treatment of missing data in multivariate analysis. Sociological Methods & Research **6**(2), 215–240 (1977)
13. Kimura, A., Kameoka, H., Sugiyama, M., Nakano, T., Maeda, E., Sakano, H., Ishiguro, K.: Semicca: Efficient semi-supervised learning of canonical correlations. Information and Media Technologies **8**(2), 311–318 (2013)
14. Klami, A., Kaski, S.: Local dependent components. In: Proceedings of the 24th International Conference on Machine Learning. pp. 425–432 (2007)
15. Lai, P.L., Fyfe, C.: Kernel and nonlinear canonical correlation analysis. International Journal of Neural Systems **10**, 365–377 (2000)
16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision. pp. 740–755 (2014)
17. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. John Wiley & Sons (2002)
18. Loh, P.L., Wainwright, M.J.: High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In: Advances in Neural Information Processing Systems. pp. 2726–2734. Curran Associates, Inc. (2011)
19. Neal, R.M.: Bayesian learning for Neural Networks. Springer-Verlag NewYork (1996)

20. Roth, P.L.: Missing data: A concdptual review for applied psychologists. Personnel Psychology **47**(3), 537–560 (1994)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: arXiv:1409.1556 (2014)
22. Tran, L., Liu, X., Zhou, J., Jin, R.: Missing modalities imputation via cascaded residual autoencoder. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4971–4980 (2017)
23. Wang, C.: Variational bayesian approach to canonical correlation analysis. IEEE Transactions on Neural Networks **18**(3), 905–910 (2007)
24. Yoshikawa, Y., Shigeto, Y., Takeuchi, A.: Stair captions: Constructing a large-scale japanese image caption dataset. In: In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (vol. 2: Short Papers),. vol. 2: Short Papers, pp. 417–421 (2017)
25. You, Q., Luo, J., Jin, H., Yang, J.: Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. pp. 13–22 (2016)
26. Zhang, B., Hao, J., Ma, G., Yue, J., Shi, Z.: Semi-paired probabilistic canonical correlation analysis. In: International Conference on Intelligent Information Processing. pp. 1–10. Springer (2014)