SemiCCA: Efficient semi-supervised learning of canonical correlations

Akisato Kimura[1,a)] Masashi Sugiyama[2] Takuho Nakano[3,1]
Hirokazu Kameoka[3,1] Hitoshi Sakano[1] Eisaku Maeda[1] Katsuhiko Ishiguro[1]

Abstract: Canonical correlation analysis (CCA) is a powerful tool for analyzing multi-dimensional paired data. However, CCA tends to perform poorly when the number of paired samples is limited, which is often the case in practice. To cope with this problem, we propose a semi-supervised variant of CCA named *SemiCCA* that allows us to incorporate additional unpaired samples for mitigating overfittng. The main contribution of the proposed method against previously proposed methods is its efficiency and intuitive operation: it smoothly bridges the generalized eigenvalue problems of CCA and principal component analysis (PCA), and thus its solution can be computed efficiently just by solving a single eigenvalue problem as the original CCA.

Keywords: Canonical correlation analysis, semi-supervised learning, generalized eigenproblem, principal component analysis, multi-label prediction

1. Introduction

The goal of dimensionality reduction is to obtain a low-dimensional representation of high-dimensional data samples, while preserving most of the intrinsic information contained in the original data. If dimensionality reduction is carried out appropriately, the compact representation of the data can be used for various tasks, such as visualization, noise reduction and classification.

Analyzing high-dimensional co-occurring data $(\boldsymbol{x}, \boldsymbol{y})$ is an important challenge in machine learning and pattern recognition communities, e.g., in the context of multi-view learning [1], automatic annotation of music, image and video [2], [3], [4], and sensor data mining [5], [6], [7], [8]. Canonical correlation analysis (CCA) [9] is a classical but still powerful tool for analyzing multivariate paired samples. CCA finds projection bases $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$ so that correlation between projected samples $\boldsymbol{w}_x^\top \boldsymbol{x}$ and $\boldsymbol{w}_y^\top \boldsymbol{y}$ is maximized. However, the performance of CCA tends to be degraded when the number of paired samples $(\boldsymbol{x}, \boldsymbol{y})$ is limited, while a large number of additional *unpaired* samples (i.e., $\boldsymbol{x}$-only samples and $\boldsymbol{y}$-only samples) are often available a lot in real-world applications. For example, in the case of automatic image annotation, collecting many labeled images (= paired samples $(\boldsymbol{x}, \boldsymbol{y})$) is often hard, while unlabeled images (= unpaired samples $\boldsymbol{x}$) can be easily obtained a lot. In the case of sensor data mining, data tends to be lost due to faulty devices and unstable transmissions, which produces a lot of unpaired samples.

To utilize such additional unpaired samples, Blaschko et al. [10] proposed a semi-supervised extension of kernelized CCA [11], [12] by the use of Laplacian regularization. This method enables us to find highly correlated directions that are also located on high variance directions along the data manifold. However, it is specialized to kernelized CCA, and deriving semi-supervised variants of the standard (linear) CCA is not necessarily obvious.

This paper proposes quite a simple method to extend linear CCA to semi-supervised one, that we call *SemiCCA*. The proposed method SemiCCA utilizes additional unpaired samples by smoothly bridging CCA and principal component analysis (PCA). More specifically, the generalized eigenvalue problems of CCA and PCA are combined using a trade-off parameter. Thus the solution of SemiCCA can still be obtained just by solving the combined eigenvalue problem, which is the same computational complexity as the original CCA.

2. Reviewing CCA

Consider a set of paired samples of size N, $\boldsymbol{X}_P = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)$ and $Y_P = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N)$, where each sample is a real-valued vector with dimension D_x and D_y. Without loss of generality, we assume that $\boldsymbol{X}_P$ and $\boldsymbol{Y}_P$ are both centered, that can always be achieved by subtracting the sample means from each sample. CCA is a method of finding a pair $(\boldsymbol{w}_x, \boldsymbol{w}_y) \in \mathcal{R}^{D_x} \times \mathcal{R}^{D_y}$ of basis vectors for

[1] NTT Communication Science Laboratories, NTT Corporation, 2-4 Hikaridai, Seika, Soraku, Kyoto, 619–0237 Japan.
[2] Graduate School of Information Science and Engineering, Tokyo Institute of Technology, 2-12-1 Oookayama, Meguro, Tokyo, 152–8552 Japan.
[3] Graduate School of Information Science and Technologies, the University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo, 113–8656 Japan.
[a)] akisato@ieee.org

a given set $(\boldsymbol{X}_P, \boldsymbol{Y}_P)$ of paired samples so that their normalized correlation is maximized as follows:

$$\begin{aligned}
&\rho(\boldsymbol{X}_P, \boldsymbol{Y}_P) \\
&= \max_{(\boldsymbol{w}_x, \boldsymbol{w}_y) \in \mathcal{R}^{D_x} \times \mathcal{R}^{D_y}} \frac{\langle \boldsymbol{X}_P^\top \boldsymbol{w}_x, \boldsymbol{Y}_P^\top \boldsymbol{w}_y \rangle}{\left\| \boldsymbol{X}_P^\top \boldsymbol{w}_x \right\|_F \cdot \left\| \boldsymbol{Y}_P^\top \boldsymbol{w}_y \right\|_F} \\
&= \max_{(\boldsymbol{w}_x, \boldsymbol{w}_y) \in \in \mathcal{R}^{D_x} \times \mathcal{R}^{D_y}} \frac{\boldsymbol{w}_x^\top \boldsymbol{S}_{Pxy} \boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x^\top \boldsymbol{S}_{Pxx} \boldsymbol{w}_x} \sqrt{\boldsymbol{w}_y^\top \boldsymbol{S}_{Pyy} \boldsymbol{w}_y}},
\end{aligned}$$

where $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ is the inner product of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, $\|\boldsymbol{X}\|_F$ is the Frobenius norm of a matrix $\boldsymbol{X}$, $\boldsymbol{X}^\top$ is a transpose of a matrix $\boldsymbol{X}$, $\boldsymbol{S}_{Pxx}$, $\boldsymbol{S}_{Pyy}$ and $\boldsymbol{S}_{Pxy}$ are sample covariance matrices of paired samples

$$\begin{aligned}
&\boldsymbol{S}_{Pxx} = \boldsymbol{X}_P \boldsymbol{X}_P^\top / N, \qquad &&\boldsymbol{S}_{Pyy} = \boldsymbol{Y}_P \boldsymbol{Y}_P^\top / N, \\
&\boldsymbol{S}_{Pxy} = \boldsymbol{X}_P \boldsymbol{Y}_P^\top / N. &&
\end{aligned}$$

The maximum of the function $\rho(\boldsymbol{X}_P, \boldsymbol{Y}_P)$ is not affected by re-scaling $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$ either together or independently. Therefore, the maximization of $\rho(\boldsymbol{X}_P, \boldsymbol{Y}_P)$ is equivalent to the maximizing the numerator $\boldsymbol{w}_x^\top \boldsymbol{S}_{Pxy} \boldsymbol{w}_y$ of $\rho(\boldsymbol{X}_P, \boldsymbol{Y}_P)$ subject to

$$\boldsymbol{w}_x^\top \boldsymbol{S}_{Pxx} \boldsymbol{w}_x = \boldsymbol{w}_y^\top \boldsymbol{S}_{Pyy} \boldsymbol{w}_y = 1.$$

Taking derivatives of the corresponding Lagrangian with respect to $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$, we obtain

$$\begin{aligned}
&\boldsymbol{S}_{Pxy} \boldsymbol{w}_y - \lambda \boldsymbol{S}_{Pxx} \boldsymbol{w}_x = \boldsymbol{0}, \\
&\boldsymbol{S}_{Pxy}^\top \boldsymbol{w}_x - \lambda \boldsymbol{S}_{Pyy} \boldsymbol{w}_y = \boldsymbol{0}.
\end{aligned}$$

Therefore, the solution $(\boldsymbol{w}_x, \boldsymbol{w}_y)$ is given as the solution of the following generalized eigenvalue problem:

$$\begin{pmatrix} \boldsymbol{0} & \boldsymbol{S}_{Pxy} \\ \boldsymbol{S}_{Pxy}^\top & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{pmatrix} = \lambda \begin{pmatrix} \boldsymbol{S}_{Pxx} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}_{Pyy} \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{pmatrix}. \tag{1}$$

Picking up the top D_z (should be $D_z \leq \min(D_x, D_y)$) generalized eigenvectors as row vectors, we can obtain D_z-dimensional mappings $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$.

3. Proposed method: SemiCCA

3.1 Semi-supervised Setup

As described in the previous section, CCA is only feasible to paired samples. CCA cannot directly deal with unpaired samples. Meanwhile, if only a small number of paired samples are available, CCA tends to overfit them. The main objective of the proposed method *SemiCCA* is to give a new insight that can overcome this weakness.

CCA can be regarded as a method for supervised dimensionality reduction under the task of multi-label prediction, where a companion $\boldsymbol{y}$ of a pair is expressed by a binary class vector $\boldsymbol{y} \in \{0,1\}^{D_y}$. In this set up, SemiCCA is to extend CCA as a supervised method to a semi-supervised one. Namely, preserving the global structure of all the samples in an unsupervised manner can be better than only relying too much on information provided by a small number of labeled samples to avoid overfitting. Not only that, SemiCCA

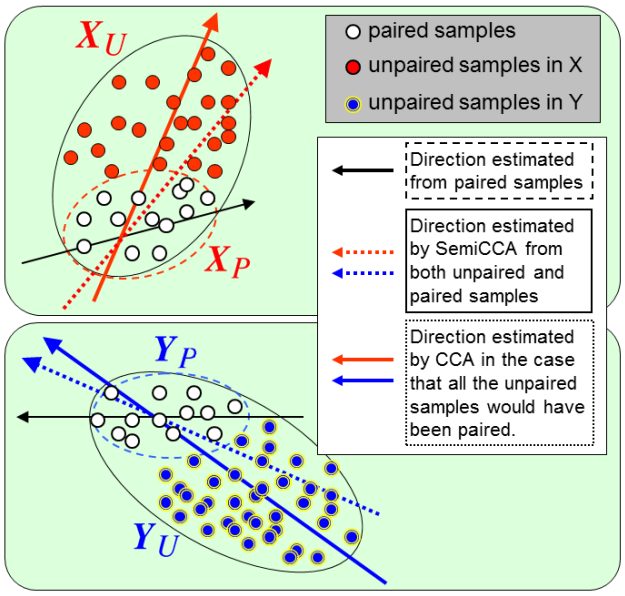

Fig. 1 Effects of unpaired samples in SemiCCA

can be also applied to any types of co-occurring real-valued sample pairs, where several pairs are forced to be unpaired.

Let us explain the idea of SemiCCA using an illustrative two-dimensional data set depicted in Fig. 1, where paired (resp. unpaired) samples are plotted with white (resp. red and blue). When only the paired samples $(\boldsymbol{X}_P, \boldsymbol{Y}_P)$ are used, poor projection bases may be obtained by CCA due to overfittng, as shown at the black arrows in Fig. 1. In contrast, unpaired samples

$$\begin{aligned}
\boldsymbol{X}_U &= (\boldsymbol{x}_{N+1}, \boldsymbol{x}_{N+2}, \ldots, \boldsymbol{x}_{N_x}) \\
&= (\boldsymbol{x}_{U,1}, \boldsymbol{x}_{U,2}, \ldots, \boldsymbol{x}_{U,N_x-N}), \\
\boldsymbol{Y}_U &= (\boldsymbol{y}_{N+1}, \boldsymbol{y}_{N+2}, \ldots, \boldsymbol{y}_{N_x}) \\
&= (\boldsymbol{y}_{U,1}, \boldsymbol{y}_{U,2}, \ldots, \boldsymbol{y}_{U,N_y-N})
\end{aligned}$$

can be used for revealing the global structure in each domain, as shown at the colored arrows in Fig. 1. Note once a basis in one sample space is rectified, the corresponding bases in the other sample space is also rectified so that correlations between two bases are maximized (cf. the dotted arrows in Fig. 1).

3.2 Algorithm

Motivated by the above illustration, we develop a novel method for effectively incorporating unpaired samples into the original CCA. The proposed method, SemiCCA, combines CCA with only the paired samples and principal component analysis (PCA) with all the samples including paired and unpaired samples. More specifically, we integrate the eigenvalue problems of CCA and PCA since this allows us to compute the combined solution efficiently. The solution of SemiCCA is given by the leading generalized eigenvectors of the following generalized eigenvalue problem:

$$\overline{\boldsymbol{C}} \begin{pmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{pmatrix} = \lambda \underline{\boldsymbol{C}} \begin{pmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{pmatrix}, \tag{2}$$

where

$$\overline{\boldsymbol{C}} = \beta \begin{pmatrix} \mathbf{0} & \boldsymbol{S}_{Pxy} \\ \boldsymbol{S}_{Pxy}^\top & \mathbf{0} \end{pmatrix} + (1-\beta) \begin{pmatrix} \boldsymbol{S}_{xx} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_{yy} \end{pmatrix},$$
$$\underline{\boldsymbol{C}} = \beta \begin{pmatrix} \boldsymbol{S}_{Pxx} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_{Pyy} \end{pmatrix} + (1-\beta) \begin{pmatrix} \boldsymbol{I}_{D_x} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{D_y} \end{pmatrix},$$

S_{xx} and S_{yy} are sample covariance matrices of all the pairs

$$\boldsymbol{S}_{xx} = \left(\boldsymbol{X}_P \boldsymbol{X}_P^\top + \boldsymbol{X}_U \boldsymbol{X}_U^\top\right) \Big/ N_x,$$
$$\boldsymbol{S}_{yy} = \left(\boldsymbol{Y}_P \boldsymbol{Y}_P^\top + \boldsymbol{Y}_U \boldsymbol{Y}_U^\top\right) \Big/ N_y,$$

and β is a constant named *a trade-off parameter* taking a value in $[0, 1]$. The parameter β controls the trade-off between CCA and PCA. Namely, when $\beta = 1$, Eq. (2) is reduced to the CCA eigenvalue problem Eq. (1), while when $\beta = 0$ Eq. (2) is reduced to the PCA eigenvalue problem, under the assumption that $\boldsymbol{X} = (\boldsymbol{X}_P, \boldsymbol{X}_U)$ and $\boldsymbol{Y} = (\boldsymbol{Y}_P, \boldsymbol{Y}_U)$ are uncorrelated. In general, SemiCCA with a trade-off parameter $0 < \beta < 1$ inherits the properties of both CCA and PCA so that the global structure in each domain and the co-occurrence information of paired samples are smoothly controlled.

One may use different trade-off parameters in $\overline{\boldsymbol{C}}$ and $\underline{\boldsymbol{C}}$ to increase the flexibility. However, this in turn makes the trade-off parameter choice laborious. For this reason, we focus on using the single shared trade-off parameter β for both $\overline{\boldsymbol{C}}$ and $\underline{\boldsymbol{C}}$, as the first step.

3.3 Some extensions

We have focused on the case where two sets of samples are given so far. However, the proposed method SemiCCA can be easily extended to multiple data sets by considering correlations over all pairs of samples [13]. For example, we can formulate SemiCCA for a triad $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ of sample sets, as follows:

$$\overline{\boldsymbol{C}} \begin{pmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \\ \boldsymbol{w}_z \end{pmatrix} = \lambda \underline{\boldsymbol{C}} \begin{pmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \\ \boldsymbol{w}_z \end{pmatrix},$$

where

$$\overline{\boldsymbol{C}} = \beta \begin{pmatrix} \mathbf{0} & \boldsymbol{S}_{(P)xy} & \boldsymbol{S}_{(P)xz} \\ \boldsymbol{S}_{(P)xy}^\top & \mathbf{0} & \boldsymbol{S}_{(P)yz} \\ \boldsymbol{S}_{(P)xz}^\top & \boldsymbol{S}_{(P)yz}^\top & \mathbf{0} \end{pmatrix} + (1-\beta) \begin{pmatrix} \boldsymbol{S}_{xx} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_{yy} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{S}_{zz} \end{pmatrix},$$
$$\underline{\boldsymbol{C}} = \beta \begin{pmatrix} \boldsymbol{S}_{(P)xx} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_{(P)yy} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{S}_{(P)zz} \end{pmatrix} + (1-\beta) \begin{pmatrix} \boldsymbol{I}_{D_x} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{D_y} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{I}_{D_z} \end{pmatrix}.$$

Of cource, the above discussion can be applied to more than 3 sample set in the same way.

We can also obtain a kernelized variant of SemiCCA by using the standard kernel trick and the technique of pairwise expression [14]. A covariance matrix $\boldsymbol{S}_{xy}$ can be converted to the following pairwise expression (see [14] for details):

$$\boldsymbol{S}_{xy} = \boldsymbol{X}(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{X}^\top = \boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^\top,$$

where $\boldsymbol{W}$ is a matrix so that all the elements are 1, $\boldsymbol{D}$ is a diagonal matrix so that the n-th diagonal element is $D_{n,n} = \sum_{m=1}^{N} W_{n,m}$, and $\boldsymbol{L}$ is called a graph Laplacian matrix $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$. In the same way, the identity matrix can be expressed with the following pairwise form:

$$\boldsymbol{I}_{D_x} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^\dagger \boldsymbol{X}^\top,$$

where $\boldsymbol{X}^\dagger$ denotes the Moore-Penrose generalized inverse of a matrix $\boldsymbol{X}$. Therefore, we can express the eigenvalue problem (Eq. (2)) solved in SemiCCA as

$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix} \overline{\boldsymbol{L}} \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{pmatrix} = \lambda \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix} \underline{\boldsymbol{L}} \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{pmatrix}, \tag{3}$$

where

$$\overline{\boldsymbol{L}} = \beta \begin{pmatrix} \mathbf{0} & \boldsymbol{L}_{(P)xy} \\ \boldsymbol{L}_{(P)xy}^\top & \mathbf{0} \end{pmatrix} + (1-\beta) \begin{pmatrix} \boldsymbol{L}_{xx} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{L}_{yy} \end{pmatrix},$$
$$\underline{\boldsymbol{L}} = \begin{pmatrix} \boldsymbol{L}_{(P)xx} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{L}_{(P)yy} \end{pmatrix} + (1-\beta) \begin{pmatrix} (\boldsymbol{X}^\top \boldsymbol{X})^\dagger & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{Y}^\top \boldsymbol{Y})^\dagger \end{pmatrix}.$$

Here, we introduce the following expressions with appropriate vectors $\boldsymbol{\alpha}_X$, $\boldsymbol{\alpha}_Y \in \mathcal{R}^N$ as follows:

$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{pmatrix} = \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{X}\boldsymbol{\alpha}_x \\ \boldsymbol{Y}\boldsymbol{\alpha}_y \end{pmatrix} = \begin{pmatrix} \boldsymbol{K}_x\boldsymbol{\alpha}_x \\ \boldsymbol{K}_y\boldsymbol{\alpha}_y \end{pmatrix}$$

where $\boldsymbol{K}_x = \{K_{x(i,j)}\}_{i,j=1}^N$ and $\boldsymbol{K}_y = \{K_{y(i,j)}\}_{i,j=1}^N$ are $N \times N$ matrices with

$$K_{x(i,j)} = \boldsymbol{x}_i^\top \boldsymbol{x}_j, \qquad K_{y(i,j)} = \boldsymbol{y}_i^\top \boldsymbol{y}_j.$$

Then, multiplying Eq. (3) by $(\boldsymbol{X}^\top, \boldsymbol{Y}^\top)$ from the left-hand side yields

$$\begin{pmatrix} \boldsymbol{K}_x \\ \boldsymbol{K}_y \end{pmatrix} \overline{\boldsymbol{L}} \begin{pmatrix} \boldsymbol{K}_x \\ \boldsymbol{K}_y \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{\alpha}_x \\ \boldsymbol{\alpha}_y \end{pmatrix} = \lambda \begin{pmatrix} \boldsymbol{K}_x \\ \boldsymbol{K}_y \end{pmatrix} \underline{\boldsymbol{L}} \begin{pmatrix} \boldsymbol{K}_x \\ \boldsymbol{K}_y \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{\alpha}_x \\ \boldsymbol{\alpha}_y \end{pmatrix}.$$

The above equation implies that the samples appear only via their inner products, which means that $\boldsymbol{K}_x = \{K_{x(i,j)}\}_{i,j=1}^N$ and $\boldsymbol{K}_y = \{K_{y(i,j)}\}_{i,j=1}^N$ can be replaced by Gram matrices, each of whose component can be decomposed with a pair (ϕ_x, ϕ_y) of functions as

$$K_{x(i,j)} = K_x(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi_x(\boldsymbol{x}_i)^\top \phi_x(\boldsymbol{x}_j),$$
$$K_{y(i,j)} = K_y(\boldsymbol{y}_i, \boldsymbol{y}_j) = \phi_y(\boldsymbol{y}_i)^\top \phi_y(\boldsymbol{y}_j).$$

The kernelized version of SemiCCA can be integrated into the work by Blaschko et al [10] with the introduction of Laplacian regularization to inhibit overfitting.

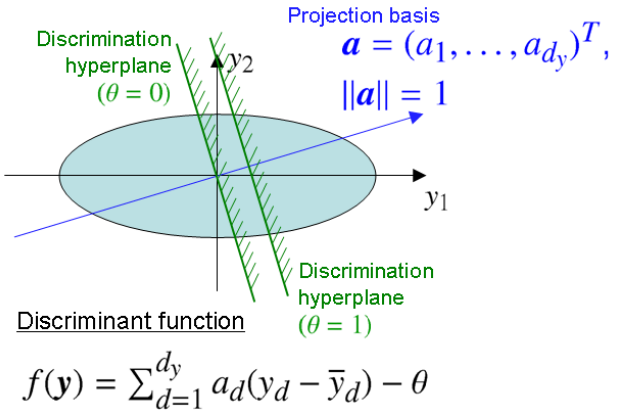

Fig. 2 How to generate artificial data

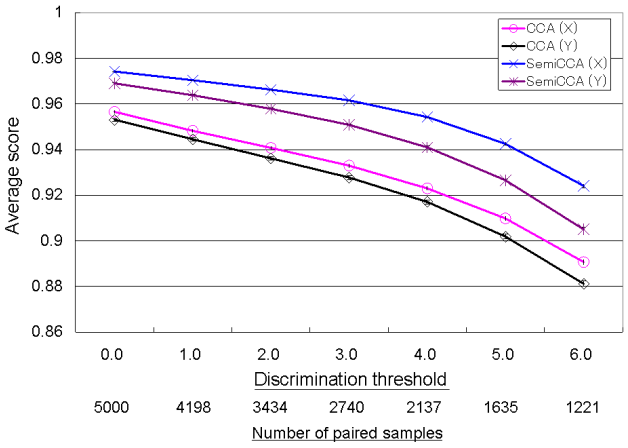

Fig. 3 Average evaluation score for artificial data

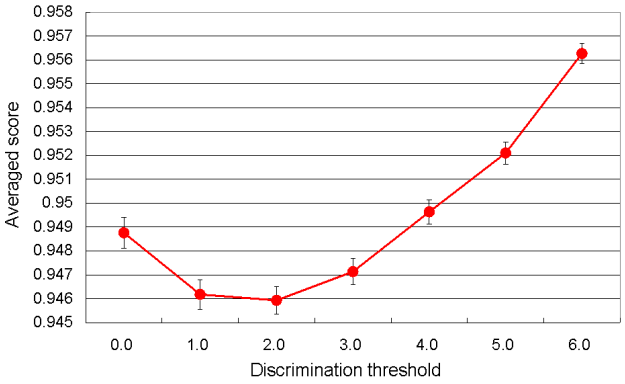

Fig. 4 Average trade-off parameter taking the highest score

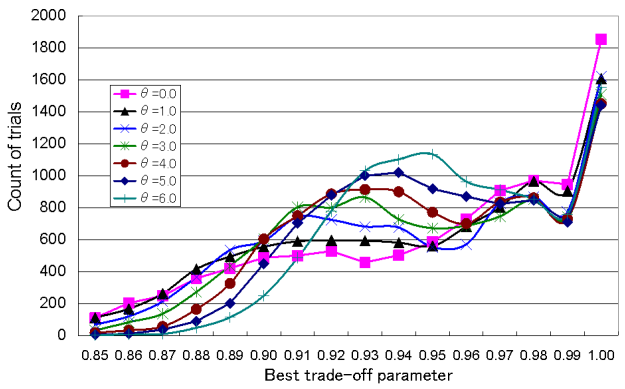

Fig. 5 Histogram of trade-off parameters taking the highest score.

4. Fundamental evaluations

We first evaluated the performance of the proposed method using the artificial data set created as follows: Consider a simple Gaussian latent model, where the latent random variable (corresponding to a canonical variable in the framework of CCA) is denoted by Z and the observable random variables are X and Y. We drew samples $\{\boldsymbol{z}_i\}_{i=1}^{N_z}$ from a standard normal distribution independently, $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{D_z})$, where $D_z = 10$ is the dimension of the random variable Z. The number N_z of samples was set to $N_z = 10000$. The means and covariance matrices of the conditional (Gaussian) densities $p(X|Z)$ and $p(Y|Z)$ were determined randomly. More specifically, we randomly generated each component of transformation matrices $\boldsymbol{T}_x$ and $\boldsymbol{T}_y$ and means $\overline{\boldsymbol{x}}$ and $\overline{\boldsymbol{y}}$ from $\mathcal{N}(0, 1)$. Then complete paired samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N_z}$ were created as

$$\boldsymbol{x}_i = \boldsymbol{T}_x \boldsymbol{z}_i + \overline{\boldsymbol{x}} + \boldsymbol{\delta}_{x,i}, \qquad \boldsymbol{\delta}_{x,i} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{X|Z}),$$
$$\boldsymbol{y}_i = \boldsymbol{T}_y \boldsymbol{z}_i + \overline{\boldsymbol{y}} + \boldsymbol{\delta}_{y,i}, \qquad \boldsymbol{\delta}_{y,i} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{Y|Z}),$$

where each component of $\boldsymbol{\Sigma}_{X|Z}$ and $\boldsymbol{\Sigma}_{Y|Z}$ was generated from the folded standard normal distribution. The dimensions of the samples were set to $D_x = 15$ and $D_y = 20$.

Then, we removed several samples from $\{\boldsymbol{y}_i\}_{i=1}^{N_z}$ to artificially generate unpaired samples, as depicted in Fig. 2. Here, we used the following linear discriminant function $f(\cdot)$ to remove samples:

$$f(\boldsymbol{y}) = \sum_{d=1}^{d_y} a_d(y_d - \overline{y}_d) - \theta, \tag{4}$$

where $\boldsymbol{a} = (a_1, \ldots, a_{d_y})^\top$ is a coefficient vector satisfying $\|\boldsymbol{a}\| = 1$, and θ is a threshold value such that the larger θ is, the more samples are removed. A sample $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ was kept paired if $f(\boldsymbol{y}_i) > 0$, and $\boldsymbol{y}_i$ was removed otherwise.

We compare the proposed SemiCCA with the standard CCA. We evaluated the performance of (Semi)CCA by the weighted sum of cosine distances defined as follows:

$$\sum_{i=1}^{r} \lambda_i^* \frac{\boldsymbol{w}_{x,i}^\top \boldsymbol{w}_{x,i}^*}{\|\boldsymbol{w}_{x,i}\| \cdot \|\boldsymbol{w}_{x,i}^*\|},$$

where $\boldsymbol{w}_{x,i}^*$ and λ_i^* are the "true" eigenvectors and eigenvalues. We took an oracle setting for selecting the trade-off parameter β. Namely, we adopted the trade-off parameter β marking the highest score for each trial.

Fig. 3 shows the evaluation scores averaged over 10000 independent trials for several discrimination thresholds θ, and also shows the average number of paired samples for each discrimination threshold. The results indicate that SemiCCA outperforms the standard CCA; it is note worthy that even when the number of unpaired samples is not so large, SemiCCA performs better than the original CCA.

Fig. 4 shows the trade-off parameter taking the highest score averaged over all the trials, and Fig. 5 depicts the

histogram of the best trade-off parameters. The results imply that the best trade-off parameters have a concave profile with respect to the number of paired samples. Since standard errors of the best trade-off parameters were relatively small, we expect to obtain similar results not only for oracle settings but also for cross validation scenarios. The results also indicate that the best trade-off parameters were usually close to 1, i.e., the effect of PCA is only mildly incorporated. Nevertheless, the performance is much improved, as shown in Fig. 3.

5. Applications to multi-label prediction

5.1 Method

We applied the proposed method SemiCCA to multi-label prediction, and evaluated its performance with automatic audio annotation. The baseline was proposed by Nakayama et al [15] and Harada et al [16], which is based on a simple latent model with the same structure as *probabilistic Latent Semantic Analysis* (pLSA) [17], [18].

Feature vectors were extracted from audios $\boldsymbol{G} = \{\boldsymbol{g}_n\}_{n=1}^{N_x}$ and associated text labels $\boldsymbol{W} = \{\boldsymbol{w}_n\}_{n=1}^{N}$, where N is the number of labeled samples and N_x is the total number of samples including labeled and unlabeled samples (should be $N \leq N_x$ and in most cases $N \ll N_x$). Each text label $\boldsymbol{w}_n$ was composed of text words selected from a word set given in advance. We utilized *Bag of Features* (BoF) with dimension $D_x = 1024$ as audio features $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^{N_x}$. We adopted word existence vectors $\boldsymbol{Y} = \{\boldsymbol{y}_n\}_{n=1}^{N}$ as text features, where each element represents an existence or absence of a specific word and thus the dimension D_y of text features was equal to the number of classes.

Next, a latent model was estimated from feature vectors $(\boldsymbol{X}, \boldsymbol{Y})$ with the help of (Semi)CCA. The first step was to generate latent variables $\boldsymbol{Z} = \{\boldsymbol{z}_n\}_{n=1}^{N_x}$ with (Semi)CCA. More specifically, a function $f_x : \mathcal{R}^{D_x} \rightarrow \mathcal{R}^{D_z}$ was derived from $(\boldsymbol{X}, \boldsymbol{Y})$ as training samples with SemiCCA, and latent variables $\boldsymbol{Z}$ are generated from $(\boldsymbol{X}, \boldsymbol{Y})$ with f_x. The dimension D_z of latent variables was experimentally determined. We set the function f_x as $f_x(\boldsymbol{x}) = \Lambda^{1/2}\boldsymbol{W}_x\boldsymbol{x}$. The second step was to set up a latent model. The latent model was described by the following equations:

$$
\begin{aligned}
p(\boldsymbol{x}, \boldsymbol{y}) &= \frac{1}{N_x}\sum_{n=1}^{N_x} p(\boldsymbol{x}|\boldsymbol{z}_n)p(\boldsymbol{y}|\boldsymbol{z}_n), \\
p(\boldsymbol{x}|\boldsymbol{z}_n) &\propto \exp\left(-\frac{\|f_x(\boldsymbol{x}) - \boldsymbol{z}_n\|}{2\gamma^2}\right), \\
p(\boldsymbol{y}|\boldsymbol{z}_n) &= \prod_{d=1}^{d_y} p(y_d|\boldsymbol{z}_n), \\
p(y_d = 1|\boldsymbol{z}_n) &= \mu\delta(1 - y_{n,d}) + (1 - \mu)N_d/N, \\
p(y_d = 0|\boldsymbol{z}_n) &= 1 - p(y_d = 1|\boldsymbol{z}_n),
\end{aligned}
$$

where $y_{n,d}$ is the d-th element of $\boldsymbol{y}_n$, N_d is the number of samples containing the d-th word in labeled samples, μ is a parameter representing how reliable a given label is, δ is the Kronecker delta, an operator $\propto$ stands for proportion, and γ is a positive constant. According to the preceding study [15], we set $\mu = 0.99$ and $\gamma = 1.0$.

Once the model estimation has been finished, we can execute annotation within the same framework through *maximum a posteriori* (MAP) estimation. More specifically, the text feature $\widehat{\boldsymbol{y}}$ of the most probable text label $\widehat{\boldsymbol{w}}$ can be derived by using a feature $\boldsymbol{x}^{(g)}$ extracted from a given audio $\boldsymbol{g}^{(g)}$, as follows:

$$
\begin{aligned}
\widehat{\boldsymbol{y}} &= \operatorname*{argmax}_{\boldsymbol{y}\in[0,1]^{D_y}} p(\boldsymbol{y}|\boldsymbol{x}^{(g)}) \\
&= \operatorname*{argmax}_{\boldsymbol{y}\in[0,1]^{D_y}} \frac{\sum_{n=1}^{N_x} p(\boldsymbol{x}^{(g)}|\boldsymbol{z}_n)p(\boldsymbol{y}|\boldsymbol{z}_n)}{\sum_{n=1}^{N_x} p(\boldsymbol{x}^{(g)}|\boldsymbol{z}_n)}.
\end{aligned}
$$

Since a conditional density $p(\boldsymbol{y}|\boldsymbol{z}_n)$ for a text feature $\boldsymbol{y}$ is modeled as an element-wise independent distribution

$$
p(\boldsymbol{y}|\boldsymbol{z}_n) = \prod_{d=1}^{D_y} p(y_d|\boldsymbol{z}_n),
$$

the annotation problem can be rewritten to the following simple form:

$$
\widehat{y}_d = \frac{\sum_{n=1}^{N_x} p(\boldsymbol{x}^{(g)}|\boldsymbol{z}_n)p(y_d = 1|\boldsymbol{z}_n)}{\sum_{n=1}^{N_x} p(\boldsymbol{x}^{(g)}|\boldsymbol{z}_n)}.
$$

When $\widehat{y}_d$ exceeds a pre-defined threshold θ_d, the text word of index d is provided to the given image $\boldsymbol{g}^{(g)}$.

5.2 Automatic audio annotation

For experiments of automatic audio annotation, we use the data collected from a audio sharing service called Freesound [*1], which consists of various audio files annotated with word tags such as "people", "noisy", and "restaurant". The goal is to predict the existence of each tag for a new audio clip. We downloaded 2012 audio clips from among all files containing any of pre-defined 230 text labels, 3-60 seconds in length and with a sampling rate 44.1kHz. We then randomly selected 1000 clips as labeled training samples, 912 clips as unlabeled training samples and the rest (= 100 samples) as samples for evaluation. As a fundamental feature comprising a BoF, *Mel-frequency Cepstral Coefficients* (MFCCs), the first and second instantaneous derivatives (Δ- and $\Delta\Delta$-MFCC) were used for audio frame features. We adopted the precision rate and recall rate as the evaluation measures.

Fig. 6 shows the experimental results, where the horizontal axis stands for the recall rate, and the vertical axis represents the precision rate. We compared the proposed method SemiCCA utilizing both labeled and unlabeled samples with the standard CCA utilizing only labeled samples and the standard CCA in the case that all the unlabeled samples would have been labeled [*2] Fig. 6 indicates that latent space extraction based on SemiCCA was effective for automatic audio annotation.

[*1] `http://www.freesound.org`

[*2] All the audio clips in the dataset used in this experiment have text labels, and a part of them were considered as unlabeled to build a semi-supervised setup.

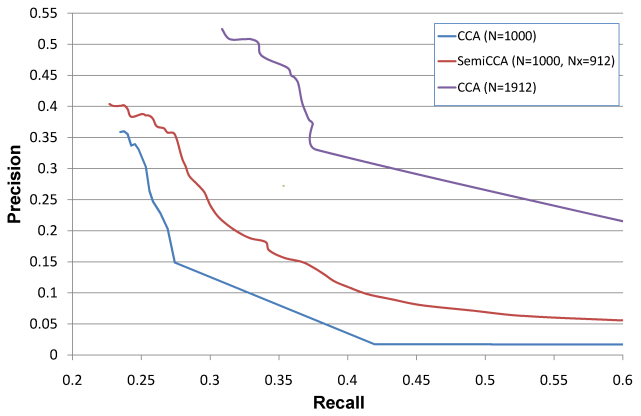

Fig. 6 Precision-recall curve for automatic audio annotation with Freesound dataset.

6. Concluding Remarks

In this paper, we proposed a semi-supervised extension of CCA that we call *SemiCCA*. Our formulation is quite simple and also intuitively understandable. Namely, SemiCCA smoothly bridges CCA with paired samples and PCA with paired and unpaired samples by a trade-off parameter. We evaluated its experimental performance, and revealed the effectiveness of SemiCCA against the original CCA.

In our future work, we will clarify some relationships between the proposed method SemiCCA and Bayesian modeling [19], [20], [21], and apply SemiCCA to other challenging real-world problems such as multi-modal event correlation analysis for audio-video synchronization, audio-visual speech recognition and sensor data mining.

References

[1] D.R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," Neural Computation, vol.16, no.12, pp.2639–2664, 2004.

[2] J.S. Downie, "The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research," Acoustical Science and Technology, vol.29, no.4, pp.247–255, 2008.

[3] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results." `http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html`, 2012.

[4] A.F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," Proc. ACM International Workshop on Multimedia Information Retrieval (MIR), pp.321–330, 2006.

[5] N. Correa, T. Adali, Y.O. Li, and V. Calhoun, "Canonical correlation analysis for data fusion and group inferences," IEEE Signal Processing Magazine, vol.27, no.4, pp.39 –50, 2010.

[6] A. Pezeshki, M.R. Azimi-Sadjadi, and L.L. Scharf, "Undersea target classification using canonical correlation analysis," IEEE Journal of Oceanic Engineering, vol.32, no.4, pp.948–955, 2007.

[7] A.A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multi-temporal remote sensing data," IEEE Transactions on Image Processing, vol.11, no.3, pp.293–305, 2002.

[8] I. Schizas, G. Giannakis, and Z.Q. Luo, "Distributed estimation using reduced-dimensionality sensor observations," IEEE Transactions on Signal Processing, vol.55, no.8, pp.4284 –4299, aug. 2007.

[9] H. Hotelling, "Analysis of a complex of statistical variables into principal components," Journal of Educational Psychology, vol.24, 1933.

[10] M.B. Blaschko, C.H. Lampert, and A. Gretton, "Semi-supervised Laplacian regularization of kernel canonical correlation analysis," Proc. European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD), pp.133–145, 2008.

[11] P.L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," Proc. International Joint Conference on Neural Networks (IJCNN), Los Alamitos, CA, USA, p.4614, 2000.

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

[13] H. Yanai and S. Puntanen, "Partial canonical correlation associated with the inverse and some generalized inverse of a partitioned dispersion matrix," Proc. Pacific Area Statistical Conference on Statistical Sciences and Data Analysis, pp.253–264, 1993.

[14] M. Sugiyama, T. Idé, S. Nakajima, and J. Sese, "Semi-supervised local Fisher discriminant analysis for dimensionality reduction," Machine Learning, vol.78, no.1, pp.35–61, 2010.

[15] H. Nayayama, T. Harada, Y. Kuniyoshi, and N. Otsu, "High-performance image annotation and retrieval for weakly labeled images," Proc. Pasific-Rim Conference on Multimedia (PCM), pp.601–610, 2008.

[16] T. Harada, H. Nakayama, and Y. Kuniyoshi, "Image annotation retrieval based on efficient learning of contextual latent space," Proc. IEEE International Conference on Multimedia and Expo (ICME), pp.858–861, 2009.

[17] T. Hofmann, "Probabilistic latent semantic indexing," Proc. International ACM SIGIR Conference on Research and development in Information Retrieval (SIGIR), pp.50–57, 1999.

[18] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," Machine Learning, vol.42, no.1, pp.177–196, 2001.

[19] F.R. Bach and M.I. Jordan, "A probabilistic interpretation of canonical correlation analysis," Tech. Rep. 688, Department of Statistics, University of California, Berkeley, 2005.

[20] C. Wang, "Variational bayesian approach to canonical correlation analysis," IEEE Transactions on Neural Networks, vol.18, no.3, pp.905 –910, 2007.

[21] S. Virtanen, A. Klami, and S. Kaski, "Bayesian cca via group sparsity," Proc. IEEE International Conference on Machine Learning (ICML), pp.457–464, 2011.