Semi-paired Probabilistic Canonical Correlation Analysis

Bo Zhang, Jie Hao, Gang Ma, Jinpeng Yue, Zhongzhi Shi

Semi-Paired Probabilistic Canonical Correlation Analysis

Bo Zhang[1,2,4], Jie Hao[3], Gang Ma[1,2], Jinpeng Yue[1,2], Zhongzhi Shi[1]

[1] Chinese Academy of Sciences, Institute of Computing Technology, The Key Laboratory of Intelligent Information Processing, Beijing, China
{zhangb, mag, yuejp, shizz }@ics.ict.ac.cn
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Xuzhou Medical College School of Medicine Information, Xuzhou, China
haojie@xzmc.edu.cn
[4] China University of Mining and Technology, School of Computer Science and Technology, Xuzhou, China

ABSTRACT. CCA is a powerful tool for analyzing paired multi-view data. However, when facing semi-paired multi-view data which widely exist in real-world problems, CCA usually performs poorly due to its requirement of data pairing between different views in nature. To cope with this problem, we propose a semi-paired variant of CCA named SemiPCCA based on the probabilistic model for CCA. Experiments with artificially generated samples demonstrate the effectiveness of the proposed method.

KEYWORDS: canonical correlation analysis; probabilistic canonical correlation analysis; semi-paired multi-view data

1 Introduction

CCA is a data analysis and dimensionality reduction method similar to PCA. While PCA deals with only one data space, CCA is a technique for joint dimensionality reduction across two spaces that provide heterogeneous representations of the same data. In real world, we often meet with such a case that one object is represented by two or more types of features, e.g., image can be represented by color and texture features, the same person has visual and audio features. Canonical correlation analysis (CCA) is a classical but still powerful method for analyzing these paired multi-view data.

CCA requires the data be rigorously paired or one-to-one correspondence among different views due to its correlation definition. However, such requirement is usually not satisfied in real-world applications due to various reasons, e.g., (1) different sampling frequencies of sensors acquiring data or sensor faulty result in the multi-view data cannot keep one-to-one correspondence any more. (2) we are often given

adfa, p. 1, 2011.

only a few paired and a lot of unpaired multi-view data, because unpaired multi-view data are relatively easier to be collected and pairing them is difficult, time consuming, even expensive. In literature, such data is referred as semi-paired multi-view data [1], weakly-paired multi-view data [2] or partially-paired multi-view data [3]. To cope with this problem, several extensions of CCA have been proposed to utilize the meaningful prior information hidden in additional unpaired data.

In this paper, we propose a yet another semi-paired variant of CCA called SemiP-CCA, which extends the probabilistic CCA model to incorporate unpaired data into the projection. We derive an efficient EM learning algorithm for this model. Experimental results on various learning tasks show promising performance for SemiP-CCA model. It is necessary to mention that the actual meaning of "semi-" in SemiP-CCA is "semi-paired" rather than "semi-supervised" in popular semi-supervised learning literature.

This paper is organized as follows. After reviewing previous work in Section 2, we formally introduce SemiPCCA model in Section 3 and derive an EM algorithm in Section 4. Finally Section 5 illustrates experiments results and Section 6 concludes the paper.

2 Related work

In this section, we review canonical correlation analysis and some improved algorithms of CCA that can effectively deal with semi-paired multi-view data.

2.1 CCA: Canonical Correlation Analysis

Let x_1 and x_2 be two set of random variables. Consider the linear combination $u = \mathbf{W}_1^T x_1$ and $v = \mathbf{W}_2^T x_2$. The problem of canonical correlation analysis reduce to find optimal linear transformation $\mathbf{W}_1$ and $\mathbf{W}_2$, which maximizes the correlation coefficient between u and v in accordance with that between x_1 and x_2. That is:

$$\rho = \max_{\mathbf{W}_1, \mathbf{W}_2} \frac{\mathbf{W}_1^T \mathbf{\Sigma}_{12} \mathbf{W}_2}{\sqrt{\mathbf{W}_1^T \mathbf{\Sigma}_{11} \mathbf{W}_1 \cdot \mathbf{W}_2^T \mathbf{\Sigma}_{22} \mathbf{W}_2}} \tag{1}$$

Where $\mathbf{\Sigma}_{11}$ and $\mathbf{\Sigma}_{22}$ are the within-set covariance matrix and $\mathbf{\Sigma}_{12}$ is the between-sets covariance matrix. Since the solution of Eq. (1) is not affected by rescaling $\mathbf{W}_1$ and $\mathbf{W}_2$ either together or independently, the optimization of ρ is equivalent to maximizing the numerator subject to $\mathbf{W}_1^T \mathbf{\Sigma}_{11} \mathbf{W}_1 = 1$ and $\mathbf{W}_2^T \mathbf{\Sigma}_{22} \mathbf{W}_2 = 1$. Then with Lagrange multiplier method, we can get

$$\mathbf{\Sigma}_{12} \mathbf{\Sigma}_{22}^{-1} \mathbf{\Sigma}_{21} \mathbf{W}_1 = \lambda^2 \mathbf{\Sigma}_{11} \mathbf{W}_1$$

$$\mathbf{\Sigma}_{21} \mathbf{\Sigma}_{11}^{-1} \mathbf{\Sigma}_{12} \mathbf{W}_2 = \lambda^2 \mathbf{\Sigma}_{22} \mathbf{W}_2$$

Which is a generalized eigenproblem of the form $\mathbf{A}x = \lambda \mathbf{B}x$. A sequence of $\mathbf{W}_1$ and $\mathbf{W}_2$ can be obtained by eigenvectors descending ordered by the corresponding

maximal eigenvalues, which indicating the explained correlation. In some literature, CCA is often described as the following:

$$\begin{pmatrix} 0 & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{pmatrix} = \lambda \begin{pmatrix} \boldsymbol{\Sigma}_{11} & 0 \\ 0 & \boldsymbol{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{pmatrix}$$

2.2 Semi-Paired Canonical Correlation Analysis

Recently, some improved algorithms of CCA that can deal with semi-paired multi-view data have emerged. Blaschko et al. [4] proposes semi-supervised Laplacian regularization of kernel canonical correlation (SemiLRKCCA) to find a set of highly correlated directions by exploiting the intrinsic manifold geometry structure of all data (paired and unpaired). SemiCCA [5] resembles the manifold regularization [6], i.e., using the global structure of the whole training data including both paired and unpaired samples to regularize CCA. Consequently, SemiCCA seamlessly bridges CCA and principal component analysis (PCA), and inherits some characteristics of both PCA and CCA. Gu et al. [3] proposed partially paired locality correlation analysis (PPLCA), which effectively deals with the semi-paired scenario of wireless sensor network localization by virtue of the combination of the neighborhood structure information in data. Most recently, Chen et al. [1] presents a general dimensionality reduction framework for semi-paired and semi-supervised multi-view data which naturally generalizes existing related works by using different kinds of prior information. Based on the framework, they develop a novel dimensionality reduction method, termed as semi-paired and semi-supervised generalized correlation analysis (S^2GCA), which exploits a small amount of paired data to perform CCA.

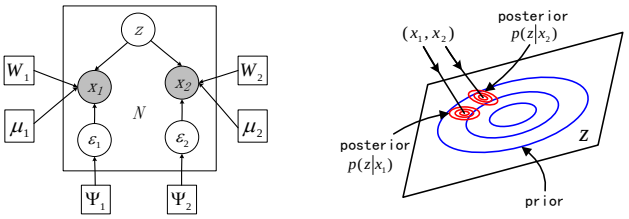
3 The SemiPCCA Model

In this section, we first review a probabilistic model for CCA in section 3.1, and then present our model.

3.1 PCCA: Probabilistic Canonical Correlation Analysis

In [7], Bach and Jordan propose a probabilistic interpretation of CCA. In this model, two random vectors $x_1 \in \mathbb{R}^{m_1}$ and $x_2 \in \mathbb{R}^{m_2}$ are considered generated by the same latent variable $z \in \mathbb{R}^d (\min(m_1, m_2) \geq d \geq 1)$ and thus the "correlated" to each other. The graphical model of the probabilistic CCA model is shown in Figure 1(a).

In this model, the observations of x_1 and x_2 are generated form the same latent variable z (Gaussian distribution with zero mean and unit variance) with unknown linear transformations $\mathbf{W}_1$and $\mathbf{W}_2$ by adding Gaussian noise ε_1 and ε_2, i.e. ,



(a) (b)

Fig. 1. (a) Graphical model for probabilistic CCA. The shaded nodes represent observed variables and unshaded node represents the latent variable. The box denotes a plate comprising a data set of N-independent observations. (b) Illustration of the projection of paired data onto the mean of the posterior distribution in latent space.

$$\mathrm{x}_1 = \mathbf{W}_1\mathrm{z} + \mu_1 + \varepsilon_1, \mathbf{W}_1 \in \mathbb{R}^{\mathrm{m}_1\times\mathrm{d}} \tag{2}$$

$$\mathrm{x}_2 = \mathbf{W}_2\mathrm{z} + \mu_2 + \varepsilon_2, \mathbf{W}_2 \in \mathbb{R}^{\mathrm{m}_2\times\mathrm{d}} \tag{3}$$

$$P(\mathrm{z})\sim\mathcal{N}(0,\mathbf{I}_\mathrm{d}),\ P(\varepsilon_1)\sim\mathcal{N}(0,\boldsymbol{\psi}_1),\ P(\varepsilon_2)\sim\mathcal{N}(0,\boldsymbol{\psi}_2)$$

Let $\mathrm{x} = \begin{pmatrix}\mathrm{x}_1\\ \mathrm{x}_2\end{pmatrix}$, we have

$$P(\mathrm{x})\sim\mathcal{N}(\mu,\boldsymbol{\Sigma})$$

where $\mu = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}$, $\boldsymbol{\Sigma} = \mathbf{W}\mathbf{W}^\mathrm{T} + \boldsymbol{\psi}$, $\mathbf{W} = \begin{pmatrix}\mathbf{W}_1\\ \mathbf{W}_2\end{pmatrix}$ and $\boldsymbol{\psi} = \begin{pmatrix}\boldsymbol{\psi}_1 & 0\\ 0 & \boldsymbol{\psi}_2\end{pmatrix}$. From [7], the corresponding maximum-likelihood estimations to the unknown parameters $\mathbf{W}_1$, $\mathbf{W}_2$, μ_1, μ_2, $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ are

$$\hat{\mu}_1 = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\mathrm{x}_1^\mathrm{i},\ \widehat{\mathbf{W}}_1 = \widetilde{\boldsymbol{\Sigma}}_{11}\mathbf{U}_{1\mathrm{d}}\mathbf{M}_1,\ \widehat{\boldsymbol{\psi}}_1 = \widetilde{\boldsymbol{\Sigma}}_{11} - \widehat{\mathbf{W}}_1\widehat{\mathbf{W}}_1{}^\mathrm{T}$$

$$\hat{\mu}_2 = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\mathrm{x}_2^\mathrm{i},\ \widehat{\mathbf{W}}_2 = \widetilde{\boldsymbol{\Sigma}}_{22}\mathbf{U}_{2\mathrm{d}}\mathbf{M}_2,\ \widehat{\boldsymbol{\psi}}_2 = \widetilde{\boldsymbol{\Sigma}}_{22} - \widehat{\mathbf{W}}_2\widehat{\mathbf{W}}_2{}^\mathrm{T}$$

where $\widetilde{\boldsymbol{\Sigma}}_{11}$, $\widetilde{\boldsymbol{\Sigma}}_{22}$have the same meaning of standard CCA, the columns of $\mathbf{U}_{1\mathrm{d}}$ and $\mathbf{U}_{2\mathrm{d}}$ are the first d canonical directions, $\mathbf{P}_\mathrm{d}$ is the diagonal matrix with its diagonal elements given by the first d canonical correlations and $\mathbf{M}_1$, $\mathbf{M}_2 \in \mathbb{R}^{\mathrm{d}\times\mathrm{d}}$, with spectral norms smaller the one, satisfying $\mathbf{M}_1\mathbf{M}_2{}^\mathrm{T} = \mathbf{P}_\mathrm{d}$. In our expectations, let $\mathbf{M}_1 = \mathbf{M}_2 = (\mathbf{P}_\mathrm{d})^{1/2}$. The posterior expectations of z given x_1 and x_2 are

$$\mathrm{E}(\mathrm{z}|\mathrm{x}_1) = \mathbf{M}_1{}^\mathrm{T}\mathbf{U}_{1\mathrm{d}}{}^\mathbf{T}(\mathrm{x}_1 - \hat{\mu}_1), \mathrm{E}(\mathrm{z}|\mathrm{x}_2) = \mathbf{M}_2{}^\mathrm{T}\mathbf{U}_{2\mathrm{d}}{}^\mathbf{T}(\mathrm{x}_2 - \hat{\mu}_2) \tag{4}$$

Thus, $\mathrm{E}(\mathrm{z}|\mathrm{x}_1)$ and $\mathrm{E}(\mathrm{z}|\mathrm{x}_2)$ lie in the d dimensional subspace that are identical with those of standard CCA, as illustrate in Figure 1(b).

3.2 SemiPCCA: Semi-paired Probabilistic Canonical Correlation Analysis

Consider a set of paired samples of size N_p, $\mathbf{X}_1^{(P)} = \{(x_1^i)\}_{i=1}^{N_p}$ and $\mathbf{X}_2^{(P)} = \{(x_2^i)\}_{i=1}^{N_p}$, where each sample x_1^i (resp. x_2^i) is represented as a vector with dimension of m_1 (resp. m_2). When the number of paired of samples is small, CCA tends to overfit the given paired samples. Here, let us consider the situation where unpaired samples $\mathbf{X}_1^{(U)} = \{(x_1^j)\}_{j=N_p+1}^{N_1}$ and/or[1] $\mathbf{X}_2^{(U)} = \{(x_2^k)\}_{k=N_p+1}^{N_2}$ are additional provided, where $\mathbf{X}_1^{(U)}$ and $\mathbf{X}_2^{(U)}$ might be independently generated. Since the original CCA and PCCA cannot directly incorporate such unpaired samples, we proposed a novel method named Semi-paired PCCA (SemiPCCA) that can avoid overfitting by utilizing the additional unpaired samples. See Figure 2 for an illustration of the graphical model of the SemiPCCA model.

The whole observation is now $D = \{(x_1^i, x_2^i)\}_{i=1}^{N_p} \cup \{(x_1^j)\}_{j=N_p+1}^{N_1} \cup \{(x_2^k)\}_{k=N_p+1}^{N_2}$. The likelihood, with the independent assumption of all the data points, is calculated as

$$\mathcal{L}(\theta) = \prod_{i=1}^{N_p} P(x_1^i, x_2^i; \theta) \prod_{j=N_p+1}^{N_1} P(x_1^j; \theta) \prod_{k=N_p+1}^{N_2} P(x_2^k; \theta) \tag{5}$$

In SemiPCCA model, for paired samples $\{(x_1^i, x_2^i)\}_{i=1}^{N_p}$, x_1^i and x_2^i are considered generated by the same latent variable z^i and $P(x_1^i, x_2^i; \theta)$ is calculated as in PCCA model, i.e.

$$P(x_1^i, x_2^i; \theta) \sim \mathcal{N}\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\mathbf{W}_1\mathbf{W}_1^T + \boldsymbol{\psi}_1 & \mathbf{W}_1\mathbf{W}_2^T\\ \mathbf{W}_2\mathbf{W}_1^T & \mathbf{W}_2\mathbf{W}_2^T + \boldsymbol{\psi}_2\end{pmatrix}\right)$$

Whereas for unpaired observations $\mathbf{X}_1^{(U)} = \{(x_1^j)\}_{j=N_p+1}^{N_1}$ and $\mathbf{X}_2^{(U)} = \{(x_2^k)\}_{k=N_p+1}^{N_2}$, x_1^j and x_2^k are separately generated from the latent variable z_1^j and z_2^k with linear transformations $\mathbf{W}_1$ and $\mathbf{W}_2$ by adding Gaussian noise ε_1 and ε_2. From Eq. (2) and Eq. (3),

$$P(x_1^j; \theta) = \int P(x_1^j | z_1^j) P(z_1^j)\, dz_1^j \sim \mathcal{N}(\mu_1, \mathbf{W}_1\mathbf{W}_1^T + \boldsymbol{\psi}_1)$$

$$P(x_2^k; \theta) = \int P(x_2^k | z_2^k) P(z_2^k)\, dz_2^k \sim \mathcal{N}(\mu_2, \mathbf{W}_2\mathbf{W}_2^T + \boldsymbol{\psi}_2)$$

1 In the context of automatic image annotation, $\mathbf{X}_1^{(U)}$ only exists, whereas $\mathbf{X}_2^{(U)}$ is empty.

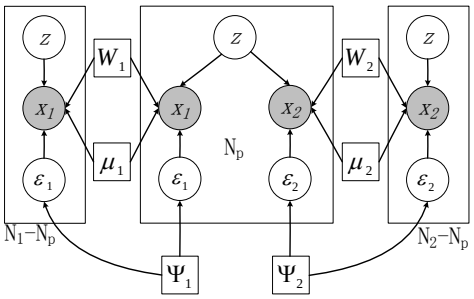


Fig. 2. Graphical model for probabilistic CCA. The box denotes a plate comprising a data set of N_p paired observations, and additional unpaired samples.

3.3 Projections in SemiPCCA model

Analogous to the PCCA model, in SemiPCCA model the projection of paired observation $\left(\mathrm{x}_1^{\mathrm{i}}, \mathrm{x}_2^{\mathrm{i}}\right)$ is directly given by Eq. (4).

Although this result looks similar as that in PCCA model, the learning of $\mathbf{W}_1$ and $\mathbf{W}_2$ are influenced by those unpaired samples. Unpaired samples reveal the global structure of whole the samples in each domain. 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.

4 EM learning for SemiPCCA

The log likelihood of the observations in $\mathcal{L}(\theta)$ is a sum of three parts. Therefore in E-step we need to deal with them differently. For each paired observation i in $\left\{\left(\mathrm{x}_1^{\mathrm{i}}, \mathrm{x}_2^{\mathrm{i}}\right)\right\}_{\mathrm{i}=1}^{\mathrm{N_p}}$, we estimate the posterior distribution of z^{i} given $\left(\mathrm{x}_1^{\mathrm{i}}, \mathrm{x}_2^{\mathrm{i}}\right)$. This is done using

$$P\left(\mathrm{z}^{\mathrm{i}} \middle| \mathrm{x}_1^{\mathrm{i}}, \mathrm{x}_2^{\mathrm{i}}; \theta\right) \sim \mathcal{N}\left(\mathbf{W}^{\mathbf{T}}(\mathbf{W}\mathbf{W}^{\mathbf{T}} + \boldsymbol{\psi})^{-\mathbf{1}}\left(\begin{pmatrix}\mathrm{x}_1^{\mathrm{i}} \\ \mathrm{x}_2^{\mathrm{i}}\end{pmatrix} - \mu\right), \mathbf{I} - \mathbf{W}^{\mathbf{T}}(\mathbf{W}\mathbf{W}^{\mathbf{T}} + \boldsymbol{\psi})^{-\mathbf{1}}\mathbf{W}\right)$$

and we calculate the expectation with respect to the posterior distribution $P\left(\mathrm{z}^{\mathrm{i}} \middle| \mathrm{x}_1^{\mathrm{i}}, \mathrm{x}_2^{\mathrm{i}}; \theta\right)$ as

$$\langle \mathrm{z}^{\mathrm{i}} \rangle = \mathbf{W}^{\mathbf{T}}(\mathbf{W}\mathbf{W}^{\mathbf{T}} + \boldsymbol{\psi})^{-\mathbf{1}}\left(\begin{pmatrix}\mathrm{x}_1^{\mathrm{i}} \\ \mathrm{x}_2^{\mathrm{i}}\end{pmatrix} - \mu\right) \tag{6}$$

$$\langle \mathrm{z}^{\mathrm{i}} \mathrm{z}^{\mathrm{i}^{\mathrm{T}}} \rangle = \langle \mathrm{z}^{\mathrm{i}} \rangle \langle \mathrm{z}^{\mathrm{i}} \rangle^{\mathrm{T}} + \mathbf{I} - \mathbf{W}^{\mathbf{T}}(\mathbf{W}\mathbf{W}^{\mathbf{T}} + \boldsymbol{\psi})^{-\mathbf{1}}\mathbf{W} \tag{7}$$

For unpaired points $\left\{\left(\mathrm{x}_1^{\mathrm{j}}\right)\right\}_{\mathrm{j}=\mathrm{N_p}+1}^{\mathrm{N_1}}$, latent variable $\mathrm{z}_1^{\mathrm{j}}$ is only conditioned on $\mathrm{x}_1^{\mathrm{j}}$,

which can be calculated posterior distribution via

$$P\left(z_1^j \middle| x_1^j; \theta\right) \sim \mathcal{N}\left(\mathbf{W_1}^{\mathbf{T}}\left(\mathbf{W_1}\mathbf{W_1}^{\mathbf{T}} + \boldsymbol{\psi}_1\right)^{-1}\left(x_1^j - \mu_1\right), \mathbf{I} - \mathbf{W_1}^{\mathbf{T}}\left(\mathbf{W_1}\mathbf{W_1}^{\mathbf{T}} + \boldsymbol{\psi}_1\right)^{-1}\mathbf{W_1}\right)$$

and we calculate the expectation with respect to the posterior distribution $P\left(z_1^j \middle| x_1^j; \theta\right)$ as

$$\langle z_1^j \rangle = \mathbf{W_1}^{\mathbf{T}}\left(\mathbf{W_1}\mathbf{W_1}^{\mathbf{T}} + \boldsymbol{\psi}_1\right)^{-1}\left(x_1^j - \mu_1\right) \tag{8}$$

$$\langle z_1^j z_1^{j\,T} \rangle = \langle z_1^j \rangle \langle z_1^j \rangle^T + \mathbf{I} - \mathbf{W_1}^{\mathbf{T}}\left(\mathbf{W_1}\mathbf{W_1}^{\mathbf{T}} + \boldsymbol{\psi}_1\right)^{-1}\mathbf{W_1} \tag{9}$$

For unpaired points $\left\{\left(x_2^k\right)\right\}_{k=N_p+1}^{N_2}$, latent variable z_2^k is only conditioned on x_2^k, which can be calculated posterior distribution via

$$P\left(z_2^k \middle| x_2^k; \theta\right) \sim \mathcal{N}\left(\mathbf{W_2}^{\mathbf{T}}\left(\mathbf{W_2}\mathbf{W_2}^{\mathbf{T}} + \boldsymbol{\psi}_2\right)^{-1}\left(x_2^k - \mu_2\right), \mathbf{I} - \mathbf{W_2}^{\mathbf{T}}\left(\mathbf{W_2}\mathbf{W_2}^{\mathbf{T}} + \boldsymbol{\psi}_2\right)^{-1}\mathbf{W_2}\right)$$

and we calculate the expectation with respect to the posterior distribution $P\left(z_2^k \middle| x_2^k; \theta\right)$ as

$$\langle z_2^k \rangle = \mathbf{W_2}^{\mathbf{T}}\left(\mathbf{W_2}\mathbf{W_2}^{\mathbf{T}} + \boldsymbol{\psi}_2\right)^{-1}\left(x_2^j - \mu_2\right) \tag{10}$$

$$\langle z_2^k z_2^{k\,T} \rangle = \langle z_2^k \rangle \langle z_2^k \rangle^T + \mathbf{I} - \mathbf{W_2}^{\mathbf{T}}\left(\mathbf{W_2}\mathbf{W_2}^{\mathbf{T}} + \boldsymbol{\psi}_2\right)^{-1}\mathbf{W_2} \tag{11}$$

In the M-step, we maximize the complete log-likelihood $\mathcal{L}(\theta)$ by setting the partial derivatives of the complete log likelihood with respect to each parameter to zero, holding $P\left(z^i \middle| x_1^i, x_2^i; \theta\right)$, $P\left(z_1^j \middle| x_1^j; \theta\right)$ and $P\left(z_2^k \middle| x_2^k; \theta\right)$ fixed from the E-step.

For means of x_1 and x_2 we have

$$\hat{\mu}_1 = \tilde{\mu}_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} x_1^i, \hat{\mu}_2 = \tilde{\mu}_2 = \frac{1}{N_2}\sum_{i=1}^{N_2} x_2^i \tag{12}$$

Which are just the sample means. Since they are always the same in all EM iterations, we can centre the data $\mathbf{X}_1^{(P)} \cup \mathbf{X}_1^{(U)}$, $\mathbf{X}_2^{(P)} \cup \mathbf{X}_2^{(U)}$ by subtracting these means in the beginning and ignore these parameters in the learning process. So for simplicity we change the notation x_1^i,x_2^i, x_1^j and x_2^k to be the centred vectors in the following.

For the two mapping matrices, we have the updates

$$\widehat{\mathbf{W_1}} = \left[\sum_{i=1}^{N_p} x_1^i \langle z^i \rangle^T + \sum_{j=N_p+1}^{N_1} x_1^j \langle z_1^j \rangle^T\right]\left[\sum_{i=1}^{N_p} \langle z^i z^{i\,T} \rangle + \sum_{j=N_p+1}^{N_1} \langle z_1^j z_1^{j\,T} \rangle\right]^{-1} \tag{13}$$

$$\widehat{\mathbf{W}}_2 = \left[\sum_{i=1}^{N_p} x_2^i \langle z^i \rangle^T + \sum_{k=N_p+1}^{N_2} x_2^k \langle z_2^k \rangle^T\right] \left[\sum_{i=1}^{N_p} \langle z^i z^{iT} \rangle + \sum_{k=N_p+1}^{N_2} \langle z_2^k z_2^{kT} \rangle\right]^{-1} \quad (14)$$

Finally the noise levels are updated as

$$\widehat{\boldsymbol{\psi}}_1 = \frac{1}{N_1}\left\{\sum_{i=1}^{N_p}\left(x_1^i - \widehat{\mathbf{W}}_1\langle z^i\rangle\right)\left(x_1^i - \widehat{\mathbf{W}}_1\langle z^i\rangle\right)^T + \sum_{j=N_p+1}^{N_1}\left(x_1^j - \widehat{\mathbf{W}}_1\langle z_1^j\rangle\right)\left(x_1^j - \widehat{\mathbf{W}}_1\langle z_1^j\rangle\right)^T\right\} \quad (15)$$

$$\widehat{\boldsymbol{\psi}}_2 = \frac{1}{N_2}\left\{\sum_{i=1}^{N_p}\left(x_2^i - \widehat{\mathbf{W}}_2\langle z^i\rangle\right)\left(x_2^i - \widehat{\mathbf{W}}_2\langle z^i\rangle\right)^T + \sum_{k=N_p+1}^{N_2}\left(x_2^k - \widehat{\mathbf{W}}_2\langle z_2^k\rangle\right)\left(x_2^k - \widehat{\mathbf{W}}_2\langle z_2^k\rangle\right)^T\right\} \quad (16)$$

The whole algorithm is summarized in Table 1.

Table 1. Algorithm of learning in SemiPCCA Model

Input: Paired observations $\left\{\left(x_1^i, x_2^i\right)\right\}_{i=1}^{N_p}$. Unpaired observations $\left\{\left(x_1^j\right)\right\}_{j=N_p+1}^{N_1}$ and $\left\{\left(x_2^k\right)\right\}_{k=N_p+1}^{N_2}$. A desired dimension d.

1: Initialize model parameters$\theta = \{\mathbf{W}_1, \mathbf{W}_2, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2\}$.

2: Calculate the sample means (12) and centre the data $\mathbf{X}_1^{(P)} \cup \mathbf{X}_1^{(U)}$, $\mathbf{X}_2^{(P)} \cup \mathbf{X}_2^{(U)}$.

3: **repeat**

{E-step}

4: **for** i = 1 to N_p **do**

5: Calculate Eq. (6) and Eq. (7) for paired data $\left(x_1^i, x_2^i\right)$;

6: **end for**

7: **for** j = N_p+1 to N_1 **do**

8: Calculate Eq. (8) and Eq. (9) for unpaired data $\left(x_1^j\right)$;

9: **end for**

10: **for** k = N_p+1 to N_2 **do**

11: Calculate Eq. (10) and Eq. (11) for unpaired data $\left(x_2^k\right)$;

12: **end for**

{M-step}

13: Update $\mathbf{W}_1$ and $\mathbf{W}_2$ via Eq. (13) and Eq. (14);

14: Update $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ via Eq. (15) and Eq. (16);

15: **until** the change of θ is smaller than a threshold.

Output: θ and projection vectors $\langle z^i \rangle (i = 1 \ldots N_p)$ which are obtained from E-step.

5 Experiments

The performance of the proposed method is evaluated using the artificial data set created as follows: we drew samples $\left\{z^i\right\}_{i=1}^{N}$ from $\mathcal{N}(0, \mathbf{I}_d)$ of dimension $d = 2$ and number of samples $N = 300$. Then the complete paired samples $\left\{\left(x_1^i, x_2^i\right)\right\}_{i=1}^{N}$ were created as

$$x_1 = \mathbf{T}_1 z + \varepsilon_1, \mathbf{T}_1 \in \mathbb{R}^{m_1 \times d}$$

$$x_2 = \mathbf{T}_2 z + \varepsilon_2, \mathbf{T}_2 \in \mathbb{R}^{m_2 \times d}$$

Where $P(\varepsilon_1) \sim \mathcal{N}\left(0, \begin{bmatrix} 0.75 & 0.5 \\ 0.5 & 0.75 \end{bmatrix}\right)$, $P(\varepsilon_2) \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\right)$, $\mathbf{T}_1 = \begin{bmatrix} 0.6 & -1/\sqrt{2} \\ 0.8 & -1/\sqrt{2} \end{bmatrix}$, $\mathbf{T}_2 = \begin{bmatrix} 0.3 & -0.7 \\ 0.4 & 0.7 \end{bmatrix}$. The dimension of samples are set to $m_1 = 2$ and $m_2 = 2$.

We removed several samples from $\{x_2^i\}_{i=1}^N$ by a simple linear discrimination. As a discriminant function, we used $f(x_2) = a^T x_2 - \theta$ where $a = (a_1, \ldots, a_{m_2})^T$, and θ is the discrimination threshold such that the larger θ we set, the more samples removed. A sample (x_1^i, x_2^i) was kept paired if $f(x_2^i) > 0$, and (x_1^i, x_2^i) was removed otherwise. Then, we compare the proposed SemiPCCA with the original CCA and PCCA. We evaluated the performance of (Semi)CCA by the weighted sum of cosine distances defined as follows:

$$C(W_x, W_x^*, \Lambda^*) = \sum_{i=1}^{d} \lambda_i^* \frac{w_{x,i}^T w_{x,i}^*}{\|w_{x,i}\| \cdot \|w_{x,i}^*\|}$$

Where $W_x^* = (w_{x,1}^*, w_{x,2}^*, \cdots, w_{x,d}^*)^T$ and $\Lambda^* = diag(\lambda_1^*, \lambda_2^*, , \lambda_d^*)$ are the "true" first d canonical correlation directions and coefficients of fully paired samples. [5]

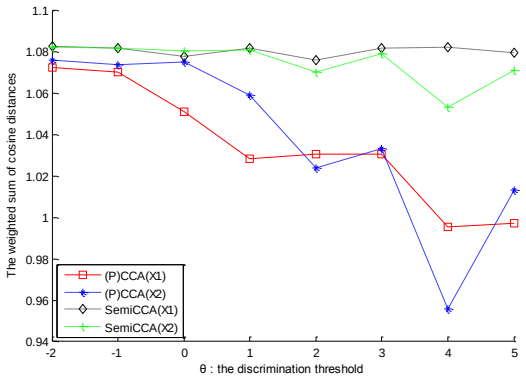


Fig. 3. Average cosine distances for artificial data

Figure 3 shows the weighted sum of cosine distances averaged over 1000 independent trials for different discrimination thresholds θ from -2 to 5. The results indicate that SemiPCCA tends to outperform the ordinary (P)CCA; it is noteworthy that even when the number of unpaired samples is not so large, SemiPCCA performs better than the original (P)CCA.

6 Conclusions

In this paper, we proposed a new semi-paired variant of CCA that we named SemiPCCA. Unlike the previous semi-paired CCA, our model is based on the probabilistic model for CCA and also intuitively comprehensive. We evaluated its performance by using artificially generated samples, and SemiPCCA performs better than the original (P)CCA.

Our future work includes some comparison of SemiPCCA with other semi-paired variants of CCA, evaluated its performance by using other data, such as PASCAL Visual Object Challenge (VOC) data sets.

References

1. Chen, X., Chen, S., Xue, H., and Zhou, X.: A unified dimensionality reduction framework for semi-paired and semi-supervised multi-view data. Pattern Recognition, 45, (5), 2005-2018(2012)
2. Lampert, C.H., Kroemer, O.: Weakly-paired Maximum Covariance Analysis for Multimodal Dimensionality Reduction and Transfer Learning. In: 11th European Conference on Computer Vision, pp. 566-579. Springer, Crete, Greece(2010)
3. Gu, J., Chen, S., and Sun, T.: Localization with Incompletely Paired Data in Complex Wireless Sensor Network. IEEE Transactions on Wireless Communications. 10, (9), 2841-2849(2011)
4. Blaschko, M., Lampert, C., and Gretton, A.: Semi-supervised Laplacian Regularization of Kernel Canonical Correlation Analysis. In: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), pp. 133-145. Springer, Antwerp, Belgium(2008)
5. Kimura, A., Kameoka, H., Sugiyama, M., and Nakano, T.: SemiCCA: Efficient Semi-supervised Learning of Canonical Correlations. In: 20th International Conference on Pattern Recognition (ICPR), pp. 2933-2936. IEEE Press, Istanbul, Turkey (2010)
6. Belkin, M., Niyogi, P., and Sindhwani, V.: Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. The Journal of Machine Learning, 7, 2399-2434(2006)
7. Bach, F.R., and Jordan, M.I.: A Probability Interpretation of Canonical Correlation Analysis. Technical Report 688, Department of Statistics, Universityof California, Berkeley(2005)

Acknowledgements. This work is supported by the National Program on Key Basic Research Project (973 Program) (No. 2013CB329502), National Natural Science Foundation of China (No.61035003, No.61202212, No.61072085, No.60933004), National High-tech R&D Program of China (863 Program) (No.2012AA011003), National Science and Technology Support Program （2012BA107B02）.