Journal of Zhejiang University SCIENCE A
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Regularized canonical correlation analysis with unlabeled data*

Xi-chuan ZHOU, Hai-bin SHEN‡
(*Institute of VLSI Design, Zhejiang University, Hangzhou 310027, China*)
E-mail: zhouxc@vlsi.zju.edu.cn; shenhb@yahoo.cn

Received Mar. 25, 2008; Revision accepted June 1, 2008; Crosschecked Dec. 26, 2008

Abstract: In standard canonical correlation analysis (CCA), the data from definite datasets are used to estimate their canonical correlation. In real applications, for example in bilingual text retrieval, it may have a great portion of data that we do not know which set it belongs to. This part of data is called unlabeled data, while the rest from definite datasets is called labeled data. We propose a novel method called regularized canonical correlation analysis (RCCA), which makes use of both labeled and unlabeled samples. Specifically, we learn to approximate canonical correlation as if all data were labeled. Then, we describe a generalization of RCCA for the multi-set situation. Experiments on four real world datasets, Yeast, Cloud, Iris, and Haberman, demonstrate that, by incorporating the unlabeled data points, the accuracy of correlation coefficients can be improved by over 30%.

Key words: Canonical correlation analysis (CCA), Regularization, Unlabeled data, Generalized canonical correlation analysis (GCCA)
doi:10.1631/jzus.A0820221 **Document code:** A **CLC number:** TP301

INTRODUCTION

Canonical correlation analysis (CCA), developed by Hotelling (1936), has been a standard tool in statistical analysis. It finds two bases, which are optimal with respect to correlations; at the same time, it finds the corresponding correlations. A good introduction of CCA can be found in (Kettenring, 1971; Hardoon *et al.*, 2004). Further details and applications of CCA can be found in (Gittins, 1985; Cohen *et al.*, 2002; Kuss and Graepel, 2003; Vert and Kanehisa, 2003; Vinokourov *et al.*, 2003; Yamanishi *et al.*, 2003). Traditional CCA uses the labeled data to calculate the correlation coefficient without considering the unlabeled data points. In this paper, we propose a regularized CCA, which makes use of both labeled and unlabeled data points, to improve the accuracy of the result.

Here we briefly explain what the regularized CCA is and why it is needed. Vinokourov *et al*.(2003) used a variant of CCA, Kernel CCA, to infer a semantic representation of text from different languages. An initial sample of documents is translated by human to create a set of dual-language training documents:

$$\{\boldsymbol{x}_i^{(1)} \mid i=1, 2, \ldots, N\} = D_{\mathrm{E}},\ \{\boldsymbol{x}_i^{(2)} \mid i=1, 2, \ldots, N\} = D_{\mathrm{F}}.$$

For example, D_{E} is a set of texts labeled as English and D_{F} is labeled as French. Vinokourov *et al*. (2003) used Kernel CCA to extract semantic information from these two corpuses. However, only a small part of the total bilingual texts can be manually labeled and included in the corpus. In other words, suppose D_{total} is the total set of texts in English and French available. The languages of most documents are not explicitly labeled, which form another set defined as $D_{\mathrm{unlabeled}}$. The relation between these corpuses is $D_{\mathrm{total}}=D_{\mathrm{F}}\cup D_{\mathrm{E}}\cup D_{\mathrm{unlabeled}}$. In such a case, the original CCA or its variant Kernel CCA can calculate only canonical correlation with the labeled datasets (D_{E} and D_{F}). By incorporating regularization terms, the regularized canonical correlation analysis (RCCA) can use D_{E}, D_{F} and $D_{\mathrm{unlabeled}}$ at the same time. The abovementioned example is a typical application where RCCA is preferable. We have noted that, in order to fit in the same coordination, the two random variables should have the same dimension. This can

‡ Corresponding author
* Project (No. 5959438) supported by Microsoft (China) Co., Ltd.

be achieved by using dimension reduction or feature selection before CCA is applied.

Fig.1 illustrates the idea of this study. All data points are from class Setosa and class Virginica of the famous Iris dataset. First assume that only the data points in circles and squares are labeled, and we do not know which class the rest of the points belong to. The maximum canonical correlation calculated from these labeled data points is defined as ρ_{labeled}. On the other hand, given the class labels of all the data points, we will be able to calculate the ideal maximum canonical correlation $\rho_{\text{if-all-labeled}}$. Obviously, $\rho_{\text{labeled}} \neq \rho_{\text{if-all-labeled}}$. We define their absolute difference as the correlation error (σ_{cor}) as follows:

$$\sigma_{\text{cor}} = | \rho_{\text{if-all-labeled}} - \rho_{\text{labeled}} |. \quad (1)$$

RCCA is proposed in this study to incorporate the manifold structure illustrated by unlabeled data to reduce the correlation error defined in Eq.(1).

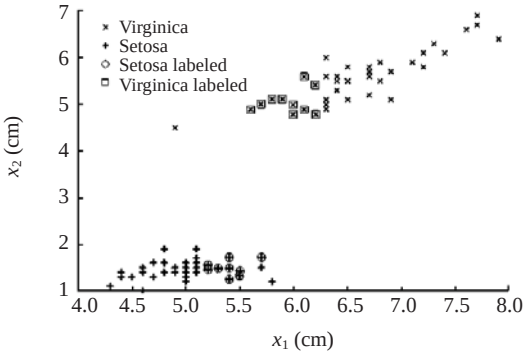

Fig.1 Illustration of regularized CCA with unlabeled data
RCCA intends to reduce the gap between CCA calculated by all data points and by labeled points (in circles and squares)

Since proposed in 1936, CCA has developed other variants (Gestel *et al*., 2001; Bach and Jordan, 2005). Earlier attempt to enhance CCA by regularization techniques has also been made. Gou and Fyfe (2004) developed a canonical correlation neural network for multi-collinearity and functional data. Inspired by their work, we introduce a different penalty term using unlabeled data to reduce the error caused by lack of labeled data. Kettenring (1971) extended the canonical correlation as a measure between two sets of variables to more than two sets, while preserving most of its properties. Hardoon *et al*. (2004) also designed a generalized CCA with optimization problem formulation. Inspired by their work, we generalize RCCA for the situation where more than two datasets are considered.

CANONICAL CORRELATION ANALYSIS

Given a sample of instance $S = ((\boldsymbol{x}_1^{(1)}, \boldsymbol{x}_1^{(2)}), (\boldsymbol{x}_2^{(1)}, \boldsymbol{x}_2^{(2)}), \ldots, (\boldsymbol{x}_N^{(1)}, \boldsymbol{x}_N^{(2)}))$, where $\boldsymbol{x}_i^{(r)} \in \mathbb{R}^d$ (r=1, 2; i=1, 2, …, N), and the superscript '(r)' of each vector indicates which set the vector belongs to. $S_x^{(1)}$ denotes the first dataset $\{\boldsymbol{x}_1^{(1)}, \boldsymbol{x}_2^{(1)}, \ldots, \boldsymbol{x}_N^{(1)}\}$ and $S_x^{(2)}$ denotes the second dataset $\{\boldsymbol{x}_1^{(2)}, \boldsymbol{x}_2^{(2)}, \ldots, \boldsymbol{x}_N^{(2)}\}$. The first stage of canonical correlation is to choose the projection vectors of $\boldsymbol{w}_x^{(1)}$ and $\boldsymbol{w}_x^{(2)}$ to maximize the correlation

$$\rho = \max_{\boldsymbol{w}_x^{(1)}, \boldsymbol{w}_x^{(2)}} \frac{\left\langle S_x^{(1)} \boldsymbol{w}_x^{(1)}, S_x^{(2)} \boldsymbol{w}_x^{(2)} \right\rangle}{\left\| S_x^{(1)} \boldsymbol{w}_x^{(1)} \right\| \left\| S_x^{(2)} \boldsymbol{w}_x^{(2)} \right\|}.$$

If $E[f(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})] = \dfrac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)})$ is used to denote the empirical expectation of $f(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})$, we can rewrite the correlation expression as

$$\max_{\boldsymbol{w}_x^{(1)}, \boldsymbol{w}_x^{(2)}} \frac{(\boldsymbol{w}_x^{(1)})^{\mathrm{T}} E_{12} \boldsymbol{w}_x^{(2)}}{\sqrt{(\boldsymbol{w}_x^{(1)})^{\mathrm{T}} E_{11} w_x^{(1)} (\boldsymbol{w}_x^{(2)})^{\mathrm{T}} E_{22} \boldsymbol{w}_x^{(2)}}},$$

where $E_{rm} = E[\boldsymbol{x}^{(r)} (\boldsymbol{x}^{(m)})^{\mathrm{T}}]$ (r, m=1, 2).

Now observe that the covariance matrix of $\boldsymbol{x}^{(1)}$, $\boldsymbol{x}^{(2)}$ is

$$\boldsymbol{C} = E[(\boldsymbol{x}^{(1)} \boldsymbol{x}^{(2)})^{\mathrm{T}} (\boldsymbol{x}^{(1)} \boldsymbol{x}^{(2)})] = \begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{pmatrix}.$$

The total covariance matrix $\boldsymbol{C}$ is a block matrix where the within-sets covariance matrices are $\boldsymbol{C}_{11}$ and $\boldsymbol{C}_{22}$, and the between-sets covariance matrices are $\boldsymbol{C}_{12} = \boldsymbol{C}_{21}^{\mathrm{T}}$. Hence, we can rewrite the correlation as

$$\rho = \max_{\boldsymbol{w}_x^{(1)}, \boldsymbol{w}_x^{(2)}} \frac{(\boldsymbol{w}_x^{(1)})^{\mathrm{T}} \boldsymbol{C}_{12} \boldsymbol{w}_x^{(2)}}{\sqrt{(\boldsymbol{w}_x^{(1)})^{\mathrm{T}} \boldsymbol{C}_{11} \boldsymbol{w}_x^{(1)} (\boldsymbol{w}_x^{(2)})^{\mathrm{T}} \boldsymbol{C}_{22} \boldsymbol{w}_x^{(2)}}}. \quad (2)$$

The CCA optimization problem formulated in Eq.(2) is equivalent to the following optimization problem:

$$\max_{\boldsymbol{w}_x^{(1)},\boldsymbol{w}_x^{(2)}} \left((\boldsymbol{w}_x^{(1)})^{\mathrm{T}}\boldsymbol{C}_{12}\boldsymbol{w}_x^{(2)}\right)$$
$$\text{s.t. } (\boldsymbol{w}_x^{(1)})^{\mathrm{T}}\boldsymbol{C}_{11}\boldsymbol{w}_x^{(1)}=1 \text{ and } (\boldsymbol{w}_x^{(2)})^{\mathrm{T}}\boldsymbol{C}_{22}\boldsymbol{w}_x^{(2)}=1. \quad (3)$$

This optimization problem is equivalent to the following generalized eigenvalue problem:

$$\boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21}\boldsymbol{w}_x^{(1)}=\lambda^2\boldsymbol{C}_{11}\boldsymbol{w}_x^{(1)}.$$

Then $\boldsymbol{w}_x^{(2)}$ can be calculated by

$$\boldsymbol{w}_x^{(2)}=\frac{1}{\lambda}\boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21}\boldsymbol{w}_x^{(1)}.$$

REGULARIZED CANONICAL CORRELATION ANALYSIS WITH UNLABELED DATA

Given two sets of labeled data $\{\boldsymbol{x}_1^{(1)}, \boldsymbol{x}_2^{(1)}, \ldots, \boldsymbol{x}_l^{(1)}\}$, $\{\boldsymbol{x}_1^{(2)}, \boldsymbol{x}_2^{(2)}, \ldots, \boldsymbol{x}_l^{(2)}\}$, and a much larger unlabeled dataset $\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_u\}$, where $\boldsymbol{x}_i^{(r)}, \boldsymbol{z}_j \in \mathbb{R}^d$ (r=1, 2; i=1, 2, …, l; j=1, 2, …, u). Data matrices used later are: $\boldsymbol{X}^{(1)}=[\boldsymbol{x}_1^{(1)}, \boldsymbol{x}_2^{(1)}, \ldots, \boldsymbol{x}_l^{(1)}]$, $\boldsymbol{X}^{(2)}=[\boldsymbol{x}_1^{(2)}, \boldsymbol{x}_2^{(2)}, \ldots, \boldsymbol{x}_l^{(2)}]$, and $\boldsymbol{V}=[\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{2l+u}]$. Each data vector is a column of $\boldsymbol{V}$, satisfying

$$\boldsymbol{v}_i=\begin{cases}\boldsymbol{x}_i^{(1)}, & 1\le i\le l,\\ \boldsymbol{x}_{i-l}^{(2)}, & l+1\le i\le 2l,\\ \boldsymbol{z}_{i-2l}, & 2l+1\le i\le 2l+u.\end{cases} \quad (4)$$

Objective function

CCA aims to find a pair of projection vectors $(\boldsymbol{w}_x^{(1)}, \boldsymbol{w}_x^{(2)})$ such that the correlation coefficient ρ in Eq.(2) is maximized. When there are no sufficient training samples, over-fitting may happen. A typical way to prevent over-fitting is to impose a regularizer (Gou and Fyfe, 2004; Shawe-Taylor and Cristianini, 2004):

$$\max_{\boldsymbol{w}_x^{(1)},\boldsymbol{w}_x^{(2)}} \frac{(\boldsymbol{w}_x^{(1)})^{\mathrm{T}}\boldsymbol{C}_{12}\boldsymbol{w}_x^{(2)}}{\sqrt{F(1)F(2)}}, \quad (5)$$

where

$$F(r)=(\boldsymbol{w}_x^{(r)})^{\mathrm{T}}\boldsymbol{C}_{rr}\boldsymbol{w}_x^{(r)}+\alpha_r J(\boldsymbol{w}_x^{(r)}),\ r=1, 2.$$

The regularization term $J(\boldsymbol{w})$ controls the learning complexity of the hypothesis family, and the coefficients α_1 and α_2 control balance between the model complexity and the empirical loss. $J(\boldsymbol{w})$ provides us the flexibility to incorporate our prior knowledge into some particular applications. When a set of unlabeled examples is available, we aim to construct a $J(\boldsymbol{w})$ incorporating the manifold structure. The key to learning with unlabeled data is the prior assumption of consistency. For subspace projection, it can be interpreted as 'nearby points have similar embedding (low-dimensional representations)'. Usually this is achieved by using a graph $\boldsymbol{G}$ represented by the weight matrix $\boldsymbol{S}$, where the nodes are all the data points, both labeled and unlabeled. The edge between nodes i and j represents their similarity. We define two kinds of weight matrices in this work, a sparse p-nearest-neighbor graph $\boldsymbol{S}_1$ and an exp-weighted graph $\boldsymbol{S}_2$, as follows:

$$(S_1)_{ij}=\begin{cases}1, & \boldsymbol{v}_i\in N_p(\boldsymbol{v}_j) \text{ or } \boldsymbol{v}_j\in N_p(\boldsymbol{v}_i),\\ 0, & \text{otherwise,}\end{cases} \quad (6)$$

$$(S_2)_{ij}=\exp\left(\frac{-d^2(i, j)}{\sigma^2}\right). \quad (7)$$

$N_p(\boldsymbol{v}_i)$ denotes the set of p nearest neighbors of $\boldsymbol{v}_i$. Specifically, if two data points are linked by an edge in $\boldsymbol{S}_1$ or have a highly weighted edge in $\boldsymbol{S}_2$, they are probably from the same set. Thus, a natural regularizer can be defined as

$$J(\boldsymbol{w})=\sum_{i, j=1}^{2l+u}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v}_i-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v}_j)^2S_{ij}.$$

$J(\boldsymbol{w})$ can be written into a more compact form as

$$\begin{aligned}J(\boldsymbol{w})&=\sum_{i, j=1}^{2l+u}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v}_i-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v}_j)^2S_{ij}\\&=2\sum_i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v}_iD_{ii}\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{w}-2\sum_{i, j}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v}_iS_{ij}\boldsymbol{v}_j^{\mathrm{T}}\boldsymbol{w}\\&=2\boldsymbol{w}^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{D}-\boldsymbol{S})\boldsymbol{V}^{\mathrm{T}}\boldsymbol{w}\\&=2\boldsymbol{w}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{w}.\end{aligned} \quad (8)$$

$\boldsymbol{D}$ is a diagonal matrix, whose entries are column (or row, since $\boldsymbol{S}$ is symmetric) sum of $\boldsymbol{S}$, i.e., $D_{ii}=\sum_j S_{ij}$. $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{S}$ is the famous Laplacian matrix. With this regularizer, we can write the objective

function of our RCCA as follows:

$$\max_{\boldsymbol{w}_x^{(1)},\boldsymbol{w}_x^{(2)}} \frac{(\boldsymbol{w}_x^{(1)})^{\mathrm{T}}\boldsymbol{C}_{12}\boldsymbol{w}_x^{(2)}}{\sqrt{F'(1)F'(2)}}, \tag{9}$$

where $F'(r)=(\boldsymbol{w}_x^{(r)})^{\mathrm{T}}(\boldsymbol{C}_{rr}+\alpha_r\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}})\boldsymbol{w}_x^{(r)}$, $r=1, 2$.

Regularized canonical correlation analysis

Since the solution to Eq.(9) is not affected by re-scaling $\boldsymbol{w}_x^{(1)}$ or $\boldsymbol{w}_x^{(2)}$ (Hardoon *et al*., 2004), Eq.(9) is equivalent to the following optimization problem:

$$\begin{aligned}&\max_{\boldsymbol{w}_x^{(1)},\boldsymbol{w}_x^{(2)}} \left((\boldsymbol{w}_x^{(1)})^{\mathrm{T}}\boldsymbol{C}_{12}\boldsymbol{w}_x^{(2)}\right)\\ &\text{s.t. } (\boldsymbol{w}_x^{(1)})^{\mathrm{T}}(\boldsymbol{C}_{11}+\alpha_1\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}})\boldsymbol{w}_x^{(1)}=1,\\ &\quad (\boldsymbol{w}_x^{(2)})^{\mathrm{T}}(\boldsymbol{C}_{22}+\alpha_2\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}})\boldsymbol{w}_x^{(2)}=1.\end{aligned} \tag{10}$$

Its Lagrangian function is

$$\begin{aligned}&L(\boldsymbol{w}_x^{(1)},\boldsymbol{w}_x^{(2)},\lambda_x^{(1)},\lambda_x^{(2)})\\ &=\boldsymbol{w}_x^{(1)}\boldsymbol{C}_{12}\boldsymbol{w}_x^{(2)}-\frac{\lambda_x^{(1)}}{2}[(\boldsymbol{w}_x^{(1)})^{\mathrm{T}}(\boldsymbol{C}_{11}+\alpha_1\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}})\boldsymbol{w}_x^{(1)}-1]\\ &\quad-\frac{\lambda_x^{(2)}}{2}[(\boldsymbol{w}_x^{(2)})^{\mathrm{T}}(\boldsymbol{C}_{22}+\alpha_2\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}})\boldsymbol{w}_x^{(2)}-1].\end{aligned} \tag{11}$$

Taking derivatives with respect to $\boldsymbol{w}_x^{(1)}$ and $\boldsymbol{w}_x^{(2)}$, we obtain

$$\frac{\partial L}{\partial \boldsymbol{w}_x^{(1)}}=\boldsymbol{C}_{12}\boldsymbol{w}_x^{(2)}-\lambda_x^{(1)}(\boldsymbol{C}_{11}+\alpha_1\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}})\boldsymbol{w}_x^{(1)}=\boldsymbol{0}, \tag{12a}$$

$$\frac{\partial L}{\partial \boldsymbol{w}_x^{(2)}}=\boldsymbol{C}_{21}\boldsymbol{w}_x^{(1)}-\lambda_x^{(2)}(\boldsymbol{C}_{22}+\alpha_2\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}})\boldsymbol{w}_x^{(2)}=\boldsymbol{0}. \tag{12b}$$

Considering the symmetry of $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$, it is reasonable to define $\alpha_1=\alpha_2=\alpha$. Subtracting $(\boldsymbol{w}_x^{(1)})^{\mathrm{T}}$ times Eq.(12a) from $(\boldsymbol{w}_x^{(2)})^{\mathrm{T}}$ times Eq.(12b), we have

$$\begin{aligned}&-(\boldsymbol{w}_x^{(1)})^{\mathrm{T}}\boldsymbol{C}_{12}\boldsymbol{w}_x^{(2)}+\lambda_x^{(1)}(\boldsymbol{w}_x^{(1)})^{\mathrm{T}}(\boldsymbol{C}_{11}+\alpha\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}})\boldsymbol{w}_x^{(1)}\\ &+(\boldsymbol{w}_x^{(2)})^{\mathrm{T}}\boldsymbol{C}_{21}\boldsymbol{w}_x^{(1)}-\lambda_x^{(2)}(\boldsymbol{w}_x^{(2)})^{\mathrm{T}}(\boldsymbol{C}_{22}+\alpha\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}})\boldsymbol{w}_x^{(2)}=0.\end{aligned} \tag{13}$$

Considering the constraint function in optimization problem Eq.(10) and $(\boldsymbol{w}_x^{(2)})^{\mathrm{T}}\boldsymbol{C}_{21}\boldsymbol{w}_x^{(1)}=(\boldsymbol{w}_x^{(1)})^{\mathrm{T}}\boldsymbol{C}_{12}\boldsymbol{w}_x^{(2)}$, we have $\lambda_x^{(1)}=\lambda_x^{(2)}=\lambda$. By Eq.(12) we can convert the optimization problem Eq.(10) into the following generalized eigenvalue problem:

$$\boldsymbol{C}_{12}(\boldsymbol{C}_{22}+\alpha\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}})^{-1}\boldsymbol{C}_{21}\boldsymbol{w}_x^{(1)}=\lambda^2(\boldsymbol{C}_{11}+\alpha\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}})\boldsymbol{w}_x^{(1)}. \tag{14}$$

$\boldsymbol{w}_x^{(2)}$ can be calculated by

$$\boldsymbol{w}_x^{(2)}=(\boldsymbol{C}_{22}+\alpha\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}})^{-1}\boldsymbol{C}_{21}\boldsymbol{w}_x^{(1)}/\lambda. \tag{15}$$

We are left with a generalized eigenvalue problem. In case that $\boldsymbol{C}_{11}+\alpha\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}}$ and $\boldsymbol{C}_{22}+\alpha\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}}$ are singular, we may add a small regularization term $\kappa\boldsymbol{I}$ in practice.

EFFICIENT GENERALIZATION OF RCCA

In this section we propose an efficient way of generalizing RCCA for the multi-set situation, which is based on generalized canonical correlation analysis (GCCA). GCCA intends to calculate all correlation vector pairs over R datasets. The optimization problem is given below:

$$\begin{aligned}&\min_{\boldsymbol{H},\boldsymbol{W}^{(1)},\boldsymbol{W}^{(2)},\ldots,\boldsymbol{W}^{(R)}} \frac{1}{R}\sum_{r=1}^{R}\left\|\boldsymbol{H}-\boldsymbol{X}^{(r)}\boldsymbol{W}^{(r)}\right\|_{\mathrm{F}}\\ &\text{s.t. } (\boldsymbol{w}_i^{(r)})^{\mathrm{T}}\boldsymbol{C}_{rr}\boldsymbol{w}_i^{(r)}=1,\\ &\quad (\boldsymbol{w}_i^{(r)})^{\mathrm{T}}\boldsymbol{C}_{rr}\boldsymbol{w}_j^{(r)}=0,\ \text{if } i\neq j,\\ &\quad (\boldsymbol{w}_i^{(r)})^{\mathrm{T}}\boldsymbol{C}_{rm}\boldsymbol{w}_j^{(m)}=0,\ \text{if } r\neq m \text{ and } i\neq j,\\ &\qquad r, m=1, 2, ..., R;\ i, j=1, 2, ..., k,\\ &\qquad \text{except when } r\neq m \text{ and } i=j.\end{aligned} \tag{16}$$

Eq.(16) is one form of GCCA. Hardoon *et al*. (2004) proved that the optimization problem Eq.(16) is equivalent to traditional CCA when more than two datasets are considered. It is also worth noting that, previously we only preserved the largest eigenvalue and its corresponding eigenvectors. However, we can also preserve more than one pair of projection vectors corresponding to k canonical eigenvalues.

Let $\boldsymbol{X}^{(1)}$, $\boldsymbol{X}^{(2)}$, …, $\boldsymbol{X}^{(R)}$ be R matrices of data corresponding to definite sets with sizes $l\times d_1$, $l\times d_2$, …, $l\times d_R$, respectively. Each row of $\boldsymbol{X}^{(r)}$ (r=1, 2, …, R) is a data vector. Note that this definition is different from that in previous sections. Let $\boldsymbol{H}$ be an unknown matrix with size $l\times k$, where k is the number of canonical eigenvectors reserved. The columns of the matrices

$\boldsymbol{X}^{(1)}$, $\boldsymbol{X}^{(2)}$, …, $\boldsymbol{X}^{(R)}$ are centralized. We assume that the columns of every matrix $\boldsymbol{X}^{(r)}$ (r=1, 2, …, R) are linearly independent. A notation to simplify the formulae is introduced as $\boldsymbol{C}_{rm}=(\boldsymbol{X}^{(r)})^{\mathrm{T}}\boldsymbol{X}^{(m)}$.

We consider arbitrary linear combinations of the columns of these matrices in the form of $\boldsymbol{X}^{(r)}\boldsymbol{w}_i^{(r)}$ (i=1, 2, …, k; r=1, 2,…, R). Let $\boldsymbol{W}^{(r)}=[\boldsymbol{w}_1^{(r)}, \boldsymbol{w}_2^{(r)}, \ldots, \boldsymbol{w}_k^{(r)}]$ (r=1, 2, …, R) be matrices comprising the vectors of the linear combinations of columns. We are looking for linear combinations of the columns of the known matrices and a corresponding $\boldsymbol{H}$ to be the optimal solution to the optimization problem Eq.(16) (Hardoon *et al*., 2004).

It is not easy to see that RCCA incorporates geometric structure of the data by adding special regularization terms, e.g., $\alpha\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}}$, into the optimization constraints of CCA as in Eq.(10). Inspired by this idea, in order to generalize RCCA, we add the regularization terms into the constraint functions of Eq.(16). So the general form of RCCA can be described by the following optimization problem:

$$\begin{aligned}
&\min_{\boldsymbol{H},\boldsymbol{W}^{(1)},\boldsymbol{W}^{(2)},\ldots,\boldsymbol{W}^{(R)}} \frac{1}{R}\sum_{r=1}^{R}\left\|\boldsymbol{H}-\boldsymbol{X}^{(r)}\boldsymbol{W}^{(r)}\right\|_{\mathrm{F}}\\
&\text{s.t. } (\boldsymbol{w}_i^{(r)})^{\mathrm{T}}(\boldsymbol{C}_{rr}+\alpha\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}})\boldsymbol{w}_i^{(r)}=1,\\
&\quad (\boldsymbol{w}_i^{(r)})^{\mathrm{T}}(\boldsymbol{C}_{rr}+\alpha\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}})\boldsymbol{w}_j^{(r)}=0,\ \text{if } i\neq j,\\
&\quad (\boldsymbol{w}_i^{(r)})^{\mathrm{T}}(\boldsymbol{C}_{rm}+\alpha\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}})\boldsymbol{w}_j^{(m)}=0, \text{if } r\neq m \text{ and } i\neq j,\\
&\quad r,m=1,2,\ldots,R;\ i,j=1,2,\ldots,k,\\
&\quad \text{except when } r\neq m \text{ and } i=j.
\end{aligned} \tag{17}$$

Apply substitutions

$$\boldsymbol{w}_i^{(r)}=(\boldsymbol{C}_{rr}+\alpha\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}})^{-1/2}\boldsymbol{y}_i^{(r)}. \tag{18}$$

We assume that $\boldsymbol{C}_{rr}+\alpha\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}}$ has full rank. This assumption usually holds, because $\boldsymbol{C}_{rr}$ has full rank when the columns of the matrix $\boldsymbol{X}^{(r)}$ are independent. We can always decrease α to achieve this assumption. This optimization problem can be transformed into a simpler form. First, we modify the set of constraints. To make this modification readable the following notation is introduced:

$$\begin{aligned}
\boldsymbol{D}_{rm}=&(\boldsymbol{C}_{rr}+\alpha\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}})^{-1/2}(\boldsymbol{C}_{rm}+\alpha\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}})\\
&\cdot(\boldsymbol{C}_{mm}+\alpha\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}})^{-1/2},\\
&r,m=1,2,\ldots,R.
\end{aligned} \tag{19}$$

So the constraint functions of Eq.(17) can be rewritten as

$$\begin{aligned}
&(\boldsymbol{y}_i^{(r)})^{\mathrm{T}}\boldsymbol{y}_i^{(r)}=1,\\
&(\boldsymbol{y}_i^{(r)})^{\mathrm{T}}\boldsymbol{y}_i^{(r)}=0,\ \text{if } i\neq j,\\
&(\boldsymbol{y}_i^{(r)})^{\mathrm{T}}\boldsymbol{D}_{rm}\boldsymbol{y}_i^{(m)}=0,\ \text{if } i\neq j \text{ and } r\neq m,\\
&r=1,2,\ldots,R;\ i,j=1,2,\ldots,k.\\
&\text{except when } r\neq m \text{ and } i=j.
\end{aligned} \tag{20}$$

Eq.(20) is equivalent to singular decomposition problems of the matrices $\boldsymbol{D}_{rm}$. If we consider the matrix $\boldsymbol{D}_{rm}$ for a fixed pair of the indexes r, m and apply the singular decomposition, we have

$$\boldsymbol{D}_{rm}=(\boldsymbol{Y}^{(r)})^{\mathrm{T}}\boldsymbol{\Lambda}_{rm}\boldsymbol{Y}^{(m)}. \tag{21}$$

Columns of $\boldsymbol{Y}^{(r)}$ and $\boldsymbol{Y}^{(m)}$ equal $\boldsymbol{y}_i^{(r)}$ and $\boldsymbol{y}_i^{(m)}$ respectively satisfying $(\boldsymbol{Y}^{(r)})^{\mathrm{T}}\boldsymbol{Y}^{(r)}=\boldsymbol{I}$, $(\boldsymbol{Y}^{(m)})^{\mathrm{T}}\boldsymbol{Y}^{(m)}=\boldsymbol{I}$. The singular decomposition $\boldsymbol{\Lambda}_{rm}$ is a diagonal matrix. Constraint functions of Eq.(17) do not contain the items having indexes with the properties $r\neq m$ and $i=j$. The singular values of $\boldsymbol{D}_{rm}$ are given as

$$(\boldsymbol{y}_i^{(r)})^{\mathrm{T}}\boldsymbol{D}_{rm}\boldsymbol{y}_i^{(m)}=\boldsymbol{\Lambda}_{ii}. \tag{22}$$

The consequence of the singular decomposition is that the set of the feasible solutions to the optimization problem with constraints Eq.(20) is equal to the set of the singular vectors of the matrices $\boldsymbol{D}_{rm}$. To express the objective function of Eq.(17), we use the notations

$$\boldsymbol{U}_r=\boldsymbol{X}^{(r)}(\boldsymbol{C}_{rr}+\alpha\boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}})^{-1/2}. \tag{23}$$

We can derive another statement about the optimal solution. Exploiting the definition of the Frobenius norm, the objective function of Eq.(17) can be rewritten as a sum of the Euclidean norm of the column vectors, where $\boldsymbol{h}_i$ denotes the ith column of the matrix $\boldsymbol{H}$:

$$\begin{aligned}
\frac{1}{R}\sum_{r=1}^{R}\left\|\boldsymbol{H}-\boldsymbol{X}^{(r)}\boldsymbol{W}^{(r)}\right\|_{\mathrm{F}}&=\frac{1}{R}\sum_{r=1}^{R}\sum_{i=1}^{k}\left\|\boldsymbol{h}_i-\boldsymbol{X}^{(r)}\boldsymbol{w}_i^{(r)}\right\|_2^2\\
&=\frac{1}{R}\sum_{r=1}^{R}\sum_{i=1}^{k}\left\|\boldsymbol{h}_i-\boldsymbol{U}_r\boldsymbol{y}_i^{(r)}\right\|_2^2\\
&=\frac{1}{R}\sum_{r=1}^{R}\sum_{i=1}^{k}\left\langle\boldsymbol{h}_i-\boldsymbol{U}_r\boldsymbol{y}_i^{(r)},\ \boldsymbol{h}_i-\boldsymbol{U}_r\boldsymbol{y}_i^{(r)}\right\rangle.
\end{aligned} \tag{24}$$

Combined with the constraint functions rewritten in Eq.(20), the Lagrangian function of the optimization is

$$L=\frac{1}{R}\sum_{r=1}^{R}\sum_{i=1}^{k}\left\langle \boldsymbol{h}_i-\boldsymbol{U}_r\boldsymbol{y}_i^{(r)},\ \boldsymbol{h}_i-\boldsymbol{U}_r\boldsymbol{y}_i^{(r)}\right\rangle$$
$$+\sum_{r=1}^{R}\sum_{i=1}^{k}\lambda_{r,ii}(1-(\boldsymbol{y}_i^{(r)})^{\mathrm{T}}\boldsymbol{y}_i^{(r)})+\sum_{r=1}^{R}\sum_{i=1,j=1,i\neq j}^{k}\lambda_{r,ij}(-(\boldsymbol{y}_i^{(r)})^{\mathrm{T}}\boldsymbol{y}_j^{(r)})$$
$$+\sum_{r=1,m=1,r\neq m}^{R}\sum_{i=1,j=1,i\neq j}^{k}\lambda_{rm,ij}(-(\boldsymbol{y}_i^{(r)})^{\mathrm{T}}\boldsymbol{y}_j^{(r)}). \quad (25)$$

We disregard the constant $1/R$ from the objective function Eq.(17). After computing the partial derivatives, where $\boldsymbol{h}_i$ signs the ith column of the matrix $\boldsymbol{H}$, we obtain

$$\frac{\partial L}{\partial \boldsymbol{h}_i}=\sum_{r=1}^{R}(2\boldsymbol{h}_i-2\boldsymbol{U}_r\boldsymbol{y}_i^{(r)})=\boldsymbol{0},\ i=1,2,...,k, \quad (26)$$

$$\frac{\partial L}{\partial \boldsymbol{y}_i^{(r)}}=2\boldsymbol{U}_r^{\mathrm{T}}\boldsymbol{U}_r\boldsymbol{y}_i^{(r)}-2\boldsymbol{U}_r^{\mathrm{T}}\boldsymbol{h}_i-2\lambda_{r,ij}\sum_{j}^{k}\boldsymbol{y}_j^{(r)}$$
$$-2\sum_{r=1,m=1,r\neq m}^{R}\sum_{i=1,j=1,i\neq j}^{k}\lambda_{rm,ij}\boldsymbol{D}_{rm}\boldsymbol{y}_j^{(m)}=\boldsymbol{0}, \quad (27)$$
$$r=1,2,...,R;\ i=1,2,...,k.$$

From Eq.(26) $\boldsymbol{h}_i$ can be expressed as

$$\boldsymbol{h}_i=\frac{1}{R}\sum_{r=1}^{R}\boldsymbol{U}_r\boldsymbol{y}_i^{(r)},\quad i=1,2,...,k. \quad (28)$$

Thus, we can replace the variable $\boldsymbol{H}$ in Eq.(17) by an expression of the other variables without changing the optimum value or the optimal solution.

EXPERIMENT

We designed two sets of experiments to test RCCA and performed comparative experiments over four real world datasets, i.e., Yeast, Cloud, Iris, and Haberman.

All the four real world datasets consist of several subcategories. We chose two categories from each dataset and analyzed the canonical correlation between them. The detailed information of the datasets used is listed in Table 1.

Table 1 Datasets used in the experiments

Dataset	Classes	Labeled	Unlabeled
Haberman	Both two	16	130
Yeast	Cytosolic, Nuclear	85	698
Cloud	Both two	204	1640
Iris	Setosa, Virginica	10	80

The first experiment (Figs.2 and 3) assumes that 20% of the data points in each dataset are labeled. We compared different parameter settings of RCCA using the p-nearest-neighbor Laplacian matrix. The parameters we selected are shown in Table 2. Next we compared the performance of RCCA with two different Laplacian matrices defined in Eqs.(6) and (7).

The second experiment (Fig.4) compares RCCA with CCA when the share of labeled data varies. In this experiment, RCCA uses a dense Laplacian matrix defined in Eq.(7) (σ=1).

Table 2 Correlation error with different parameter settings over the four datasets

α	p	Correlation error				
		Cloud	Haberman	Iris	Yeast	Average
0.10	2	0.37	0.22	0.28	0.60	0.35
0.10	3	0.54	0.02	0.06	0.71	0.33
0.01	5	0.33	0.32	0.41	0.57	0.41
0.01	10	0.53	0.00	0.00	0.72	0.31

Fixed number of labeled data

The comparison work is twofold, illustrated in Figs.2 and 3 respectively. First we compare CCA trained by labeled data with RCCA trained by both labeled data and unlabeled data (Fig.2). Next we compared the effectiveness of sparse and dense Laplacian matrices defined in Eqs.(6) and (7) (Fig.3).

The aim of this study is to reduce the correlation error caused by lack of labeled data (defined in Eq.(1)). The results of this experiment are remarkable, showing that RCCA can reduce the correlation error into near zero.

In Fig.2, RCCA uses a sparse p-nearest-neighbor Laplacian matrix (Eq.(10)). Besides p, the other parameters are $\alpha_1=\alpha_2=\alpha$ in Eq.(9), which control the effectiveness of the regularization term. Note that if α=0, RCCA becomes CCA. The correlation error first goes down as p increases, and then begins to rise after its minimum point. However, through the whole path, RCCA outperforms CCA. The speed of RCCA to

reach the minimum point is controlled by parameter α. The correlation error usually reaches its minimum value with a small p when α=0.01.

In Fig.3, we compare the RCCA quipped with different Laplacian matrices. The solid line (α=0) corresponds to RCCA with a dense radial basis function (RBF) Laplacian matrix defined in Eq.(7) (σ=1). We compared it with different settings of p-nearest-neighbor Laplacian matrices. Generally speaking, the correlation error reaches its minimum point with a relatively large α and begins to rise if we continue to reduce α. The choice of the Laplacian matrix is less important than the choice of p. RCCA with different Laplacian matrices has close minimum values. From the experiments, we know that α satisfying $0.001 \leq \alpha \leq 0.1$ is preferred.

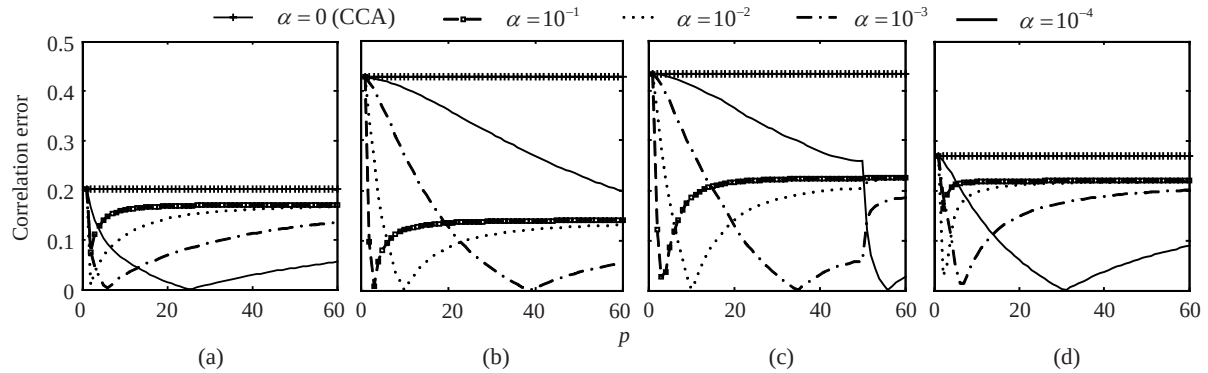

Fig.2 RCCA with a *p*-nearest-neighbor Laplacian matrix
(a) Cloud; (b) Haberman; (c) Iris; (d) Yeast

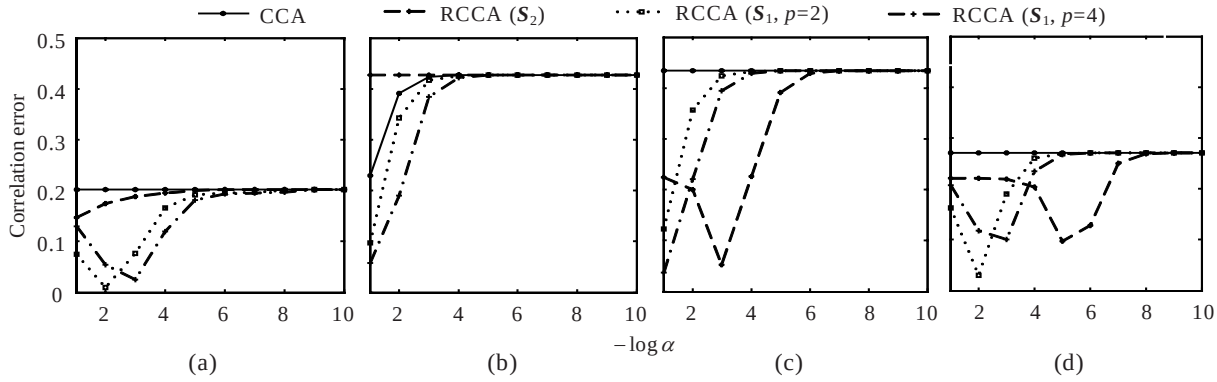

Fig.3 RCCA with different Laplacian matrices
(a) Cloud; (b) Haberman; (c) Iris; (d) Yeast

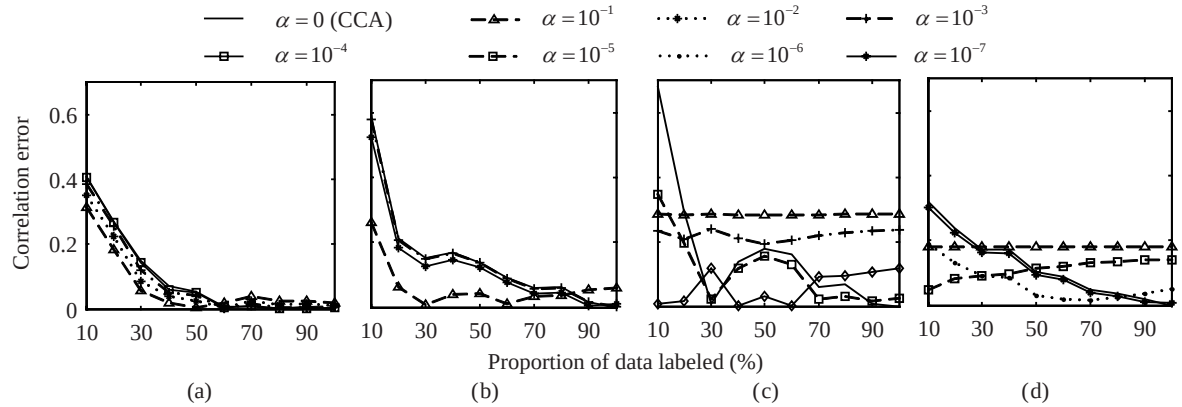

Fig.4 RCCA with various shares of labeled data
(a) Cloud; (b) Haberman; (c) Iris; (d) Yeast

Various shares of labeled data

We also designed an experiment to compare RCCA with CCA trained by various shares of labeled data. In this set of experiments, RCCA used an RBF Laplacian matrix with σ=1.

The results of this experiment are described in Fig.4. Obviously, the correlation error of CCA goes down when more data points are labeled. The correlation error of RCCA is more complex when the number of labeled data points increases. The performance of RCCA depends on the distribution of the labeled data and its manifold property. Generally speaking, RCCA with larger α gives a better performance with fewer labeled data. On the other hand, if more labeled data are given, a smaller α is preferred. In most applications, the share of labeled data does not change much, so α can still be modeled as a constant. In order to choose α automatically, we can use cross validation based methods to find the best α for a certain application.

CONCLUSION

In this paper, we propose a new variant of the canonical correlation analysis (CCA) algorithm, called regularized CCA (RCCA), which can efficiently use both labeled and unlabeled data points. The labeled data points are used to maximize canonical correlation, while the unlabeled data points are used as the regularization term. We have two contributions. First, we proposed RCCA and tested it with four real world datasets including Yeast, Cloud, Iris, and Haberman. Then we proposed an efficient method to generalize RCCA for the multi-set situation. Experimental results demonstrate the effectiveness of our algorithm.

ACKNOWLEDGEMENT

Thank Siming Wei, Rui He and Linxiao Zhao very much for pre-reading this paper and their insightful suggestions.

References

Bach, F.R., Jordan, M.I., 2005. A Probabilistic Interpretation of Canonical Correlation Analysis. Technical Report. Department of Statistics, University of California, Berkeley, CA.

Cohen, J., West, S.G., Cohen, P., Aiken, L., 2002. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey, USA.

Gestel, T.V., Suykens, J., Brabanter, J.D., 2001. Kernel Canonical Correlation Analysis and Least Squares Support Vector Machines. Int. Conf. on Artificial Neural Networks, **2130**:384-389. [doi:10.1007/3-540-44668-0_54]

Gittins, R., 1985. Canonical analysis: a review with applications in ecology. *Psychometrika*, **51**(3):495-497. [doi:10.1007/BF02294071]

Gou, Z.K., Fyfe, C., 2004. A canonical correlation neural network for multicollinearity and functional data. *Neural Networks*, **17**(2):285-293. [doi:10.1016/j.neunet.2003.07.002]

Hardoon, D., Szedmak, S., Shawe-Taylor, J., 2004. Canonical correlation analysis: an overview with application to learning methods. *Neural Comput.*, **16**(12):2639-2664. [doi:10.1162/0899766042321814]

Hotelling, H., 1936. Relations between two sets of variants. *Biometrika*, **28**(3-4):321-377. [doi:10.1093/biomet/28.3-4.321]

Kettenring, J., 1971. Canonical analysis of several sets of variables. *Biometrika*, **58**(3):433-451. [doi:10.1093/biomet/58.3.433]

Kuss, M., Graepel, T., 2003. The Geometry of Kernel Canonical Correlation Analysis. Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

Shawe-Taylor, J., Cristianini, N., 2004. Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge, UK.

Vert, J., Kanehisa, M., 2003. Graph-driven Features Extraction from Micro-array Data Using Diffusion Kernels and Kernel CCA. Advances in Neural Information Processing Systems. MIT Press, Cambridge, MA.

Vinokourov, A., Shawe-Taylor, J., Cristianini, N., 2003. Inferring a Semantic Representation of Text via Cross-language Correlation Analysis. Advances in Neural Information Processing Systems. MIT Press, Cambridge, MA.

Yamanishi, Y., Vert, J., Nakaya, A., Kanehisa, M., 2003. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical. *Bioinformatics*, **19**(Suppl. 1):323-330. [doi:10.1093/bioinformatics/btg1045]