Generalized Kernel Canonical Correlation Analysis: Criteria and Low Rank Kernel Learning

Canh Hao Nguyen, Tu Bao Ho, Jean-Michel Renders
and Nicola Cancedda

February 20, 2009
KS-RR-2009-002

Generalized Kernel Canonical Correlation Analysis: Criteria and Low Rank Kernel Learning

Canh Hao Nguyen, Tu Bao Ho, Jean-Michel Renders
and Nicola Cancedda

Email: canhhao@jaist.ac.jp
School of Knowledge Science
Japan Advanced Institute of Science and Technology

JAIST Research Report: KS-RR-2009-002

February 22, 2009

Abstract

Canonical Correlation Analysis is a classical data analysis technique for computing common correlated subspaces for two datasets. Recent advances in machine learning enable the technique to operate solely on kernel matrices, making it a kernel method with the advantages of modularity, efficiency and nonlinearity. Its performance is also improved with appropriate regularization and low-rank approximation methods, making it applicable to many practical applications.

However, the classical technique is applicable to find correlation of only two datasets. It is a practical problem that we wish to consider correlation of more than two datasets at the same time. Such problems occurs in many domains such as multilingual text processing, where we wish to find a common representation of parallel document corpora from more than two languages altogether (we call this situation *multiple view* or *multiview* for short). Generalizing CCA to more than two views face some problems, namely: finding criteria for multi-view CCA and available computational solutions for these criteria.

In this report, we analyze the criteria that have been proposed to be objective functions for multi-view CCA. We obtain that only some of them are suitable for our purpose. In these criteria, only one of them, namely MAX-VAR, has an efficient solution. We describe our algorithm for this criterion. We conduct experiments on a multi-lingual corpora. Experiment results show

that multi-view CCA brings an advantage over two view CCA when there are not too many training data are available.

We then show that low rank approximation of kernels are done independently from views. This could be a disadvantage as different views may be projected onto subspaces that may not result in correlation. We then propose a new incomplete Cholesky decomposition procedure that simultaneously decomposes all views. Experiment results show that the new ICD, by making sure the alignment of subspaces from different views, give a higher performance for multiview CCA when there are many views and a few dimensions for approximation.

1 Introduction

Introduction: Canonical Correlation Analysis (CCA) is a statistical method to find linear correlation between two (or more) multidimensional variables. It was proposed by H. Hotelling [9]. CCA can be kernelized, therefore it inherits all advantages of kernel methods [14], namely computational efficiency, modularity and nonlinearity.

The starting idea of CCA is that we may receive different views, or different data representations of inherently the same objects. In biological samples express differently in different biological experiments [17, 5]. The same documents are translated into different languages [12]. From multiple views to the same object, one wishes to find a common, latent semantic representation of the object itself. This may seem to be impossible as it is task-dependent. However, the good news is that in that common semantic representation, the object is uniquely represented. This leads to the fact that there exist transformations from different views into the common semantic space that should result in the common representation. To this end, it is the purpose of CCA to enforce the correlation of different views after being transformed into the common semantic space.

Background: CCA is the problem of finding each basic vectors sets, one for each set of (multidimensional) variables, such that the correlation between the projections of these variables onto their corresponding basics are maximized. We describe only the first basic vectors for each set for now.

Given $X = \{x_1, x_2 \cdots x_n\}$ and $Y = \{y_1, y_2 \cdots y_n\}$ where $x_i \in R^{n_1}$ and $y_i \in R^{n_2}$, the problem is to find $w_x \in R^{n_1}$ and $w_y \in R^{n_2}$ so that the projections of X onto w_x and Y onto w_y are maximized. Without loss of generality, we assume that $\sum_{i=1}^{n} x_i = 0$ and $\sum_{i=1}^{n} y_i = 0$.

Projection on w_x: $x_i \rightarrow \langle x_i, w_x \rangle$, therefore the projections of X become $\{w_x^T X\}^T = X^T w_x$ (X is considered as a column vector). Similarly, projections of Y onto w_y are $Y^T w_y$. The objective of CCA is to maximize the correlation with respect to w_x and w_y

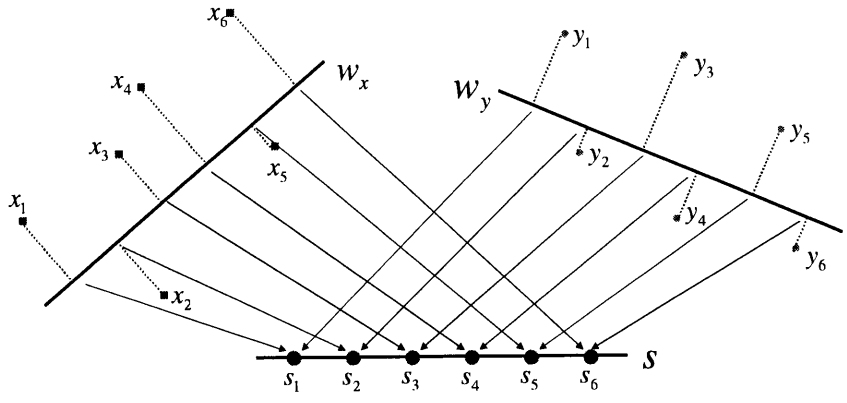

Figure 1: Canonical Correlation Analysis: projecting two sets of data onto subspaces that are maximally correlated.

$$\rho = \max_{w_x, w_y} corr(X^T w_x, Y^T w_y) = \max_{w_x, w_y} \frac{\langle X^T w_x, Y^T w_y \rangle}{||X^T w_x|| \cdot ||Y^T w_y||}. \tag{1}$$

$$\rho = \max_{w_x, w_y} \frac{w_x^T X Y^T w_y}{\sqrt{w_x^T X X^T w_x \cdot w_y^T Y Y^T w_y}}. \tag{2}$$

Denote: $C_{xx} = XX^T$, $C_{yy} = YY^T$ (these are correlation matrices), $C_{xy} = XY^T$ and $C_{yx} = YX^T$ (cross-correlation matrices).

$$\rho = \max_{w_x, w_y} \frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} w_x \cdot w_y^T C_{yy} w_y}} \tag{3}$$

Here, ρ is called *canonical correlation* and w_x, w_y are called *canonical variates* or *canonical variables.* A pictorial description of the algorithm is shown in figure 1.

In this paper, we first describe the CCA algorithm in the next section, followed by its kernelization, regularization and low rank versions using Incomplete Cholesky Decomposition for computational efficiency. We then analyze its multiple views generalization criteria to see which ones are suitable and solvable for very large scale applications. We describe the formulations for the criteria and show their interpretations. Experiments are presented to show that advantage of multiview version when there are a few training data points. We when propose a new ICD algorithm that simultaneously decompose kernel matrices from different views, which is supposed to preserve more correlation so that the multiview CCA would have a higher performance.

2 Canonical Correlation Analysis

Formula 1 is equivalent to maximizing its numerator subject to the following constraints: $w_x^T C_{xx} w_x = 1$ and $w_y^T C_{yy} w_y = 1$.

Lagrangian is

$$L(\lambda_x, \lambda_y, w_x, w_y) = w_x^T C_{xy} w_y - \frac{\lambda_x}{2} \cdot (w_x^T C_{xx} w_x - 1) - \frac{\lambda_y}{2} \cdot (w_y^T C_{yy} w_y - 1). \quad (4)$$

Taking derivatives in respect to w_x and w_y give us:

$$C_{xy} w_y - \lambda_x C_{xx} w_x = 0 \quad (5)$$

$$C_{yx} w_x - \lambda_y C_{yy} w_y = 0. \quad (6)$$

Taking the constraints into account, we have $\lambda_x = \lambda_y = \lambda$.

The solution:Assuming C_{yy} is invertible.

$$w_y = \frac{C_{yy}^{-1} C_{yx} w_x}{\lambda}. \quad (7)$$

Substituting into 5 then:

$$\frac{C_{yy}^{-1} C_{yx} w_x}{\lambda} - \lambda C_{xx} w_x = 0. \quad (8)$$

$$C_{yy}^{-1} C_{yx} w_x = \lambda^2 C_{xx} w_x \quad (9)$$

Therefore, the solution to the problem can be calculated using an inversion of a matrix 7 and a generalized eigenvalue problem 9.

Note 1: Equations 5 and 6 can be rewritten as:

$$\begin{pmatrix} 0 & C_{xy} \\ C_{yx} & 0 \end{pmatrix} \begin{pmatrix} w_x \\ w_y \end{pmatrix} = \lambda \begin{pmatrix} C_{xx} & 0 \\ 0 & C_{yy} \end{pmatrix} \begin{pmatrix} w_x \\ w_y \end{pmatrix}. \quad (10)$$

$$\begin{pmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{pmatrix} \begin{pmatrix} w_x \\ w_y \end{pmatrix} = (1 + \lambda) \begin{pmatrix} C_{xx} & 0 \\ 0 & C_{yy} \end{pmatrix} \begin{pmatrix} w_x \\ w_y \end{pmatrix}. \quad (11)$$

This means that we can formulate it into a generalized eigenvalue problem [6].

2.1 Kernelization

Above is the linear version of CCA where we need to find canonical covariates that after linear projections, views are correlated. However, the class of linear functions is not able to capture nonlinear relations among views, which is the case in practice. For that reason, one needs to use a larger class of functions. It is usually the case that one settles to the space called Reproducing Kernel Hilbert Space [15], which is large enough to capture nonlinear relation and small enough not to contain many non-smooth functions. The space is defined to be any combination of *kernel function* k (being a positive semidefinite function):

$$f(\cdot) = \sum_i \alpha_i k(\cdot, x). \tag{12}$$

By the virtue of the Representer theorem [16], we know that in this particular case, canonical covariates have its representation (for both X and Y):

$$w_x = \sum_i \alpha_i k(\cdot, x_i), \; x_i \in X. \tag{13}$$

Having known that w_x lies in the span of columns of X and w_y lies in the span of columns of Y, we may rewrite the equation 3 using $w_x = X\alpha$ and $w_y = Y\beta$, where $\alpha \in R^n$ and $\beta \in R^n$:

$$\rho = \max_{\alpha,\beta} \frac{\alpha^T X^T \cdot XY^T \cdot Y\alpha}{\sqrt{(\alpha^T X^T \cdot XX^T \cdot X\alpha) \cdot (\beta^T Y^T \cdot YY^T \cdot Y\alpha)}}, \tag{14}$$

$$\rho = \max_{\alpha,\beta} \frac{\alpha^T K_x K_y \beta}{\sqrt{\alpha^T K_x^2 \alpha \cdot \beta^T K_y^2 \beta}} \tag{15}$$

given that $K_x = X^T X$ and $K_y = Y^T Y$.
Again, Lagrangian is:

$$L(\lambda_\alpha, \lambda_\beta, \alpha, \beta) = \alpha^T K_x K_y \beta - \frac{\lambda_\alpha}{2}(\alpha^T K_x^2 \alpha - 1) - \frac{\lambda_\beta}{2}(\beta^T K_y^2 \beta - 1) \tag{16}$$

In a similar fashion as before, taking derivatives in respect to α and β, we have $\lambda_\alpha = \lambda_\beta = \lambda$ and

$$K_x K_y \beta - \lambda K_x^2 \alpha = 0 \tag{17}$$

$$K_y K_x \alpha - \lambda K_y^2 \beta = 0 \tag{18}$$

The solution: When K_y is invertible then:

$$\beta = \frac{K_y^{-1} K_x \alpha}{\lambda} \tag{19}$$

and

$$K_x K_x \alpha = \lambda^2 K_x K_x \alpha \tag{20}$$

This means that perfect correlation ($\lambda = 1$) can be obtained with any α. We have the problem of overfitting here.

Note 2: Equations 17 and 18 can be rewritten as:

$$\begin{pmatrix} 0 & K_x K_y \\ K_y K_x & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \lambda \begin{pmatrix} K_x K_x & 0 \\ 0 & K_y K_y \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}. \tag{21}$$

$$\begin{pmatrix} K_x K_x & K_x K_y \\ K_y K_x & K_y K_y \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = (1+\lambda) \begin{pmatrix} K_x K_x & 0 \\ 0 & K_y K_y \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}. \tag{22}$$

We arrive at one generalized eigenvalue problem (of the $2n * 2n$ matrix).

Note 3: Since the solution can be computed using kernel matrices K_x and K_y, nonlinear transformation of original variables can be introduced into this problem in a similar fashion as other kernel methods.

2.2 Regularization

As mention before, the cases, in which perfect correlation and any α can be a canonical covariate, would happen. Therefore, regularization is necessary to avoid nonsense solution (overfitting). It is introduced as follows:

$$\rho = \max_{\alpha,\beta} \frac{\alpha^T K_x K_y \beta}{\sqrt{(\alpha^T K_x^2 \alpha + k\|w_x\|^2) \cdot (\beta^T K_y^2 \beta + k\|w_y\|^2)}} \tag{23}$$

$$\rho = \max_{\alpha,\beta} \frac{\alpha^T K_x K_y \beta}{\sqrt{(\alpha^T K_x^2 \alpha + k\alpha^T K_x \alpha) \cdot (\beta^T K_y^2 \beta + k\beta^T K_y \beta)}} \tag{24}$$

Lagrangian is:

$$L(\lambda_\alpha, \lambda_\beta, \alpha, \beta) = \alpha^T K_x K_y \beta - \frac{\lambda_\alpha}{2}(\alpha^T K_x^2 \alpha + k\alpha^T K_x \alpha - 1) - \frac{\lambda_\beta}{2}(\beta^T K_y^2 \beta + k\beta^T K_y \beta - 1) \tag{25}$$

Taking derivatives, proving $\lambda_\alpha = \lambda_\beta = \lambda$, we arrive at:

$$K_x K_y \beta - \lambda(K_x + kI)K_x \alpha = 0, \tag{26}$$

$$K_y K_x \alpha - \lambda(K_y + kI)K_y \beta = 0. \tag{27}$$

The solution: If K_y is invertible, we have:

$$\beta = \frac{(K_y + kI)^{-1} K_x \alpha}{\lambda} \tag{28}$$

and substituting into 26 give us:

$$K_x K_y (K_y + kI)^{-1} K_x \alpha = \lambda^2 K_x (K_x + kI)^{-1} \alpha \tag{29}$$

This give us the generalized eigenvalue problem. When K_x is also invertible, we have:

$$(K_x + kI)^{-1} K_y (K_y + kI)^{-1} K_x \alpha = \lambda^2 \alpha \tag{30}$$

Note 4: Equations 26 and 27 can be rewritten as:

$$\begin{pmatrix} 0 & K_x K_y \\ K_y K_x & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \lambda \begin{pmatrix} (K_x + kI)K_x & 0 \\ 0 & (K_y + kI)K_y \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}. \tag{31}$$

$$\begin{pmatrix} (K_x + kI)K_x & K_x K_y \\ K_y K_x & (K_y + kI)K_y \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = (1+\lambda) \begin{pmatrix} (K_x + kI)K_x & 0 \\ 0 & (K_y + kI)K_y \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}. \tag{32}$$

Note 5: In Bach & Jordan, $(K_x + kI)K_x$ is approximated with $(K_x + \frac{k}{2}I)^2$, arriving at:

$$\begin{pmatrix} (K_x + \frac{k}{2}I)^2 & K_x K_y \\ K_y K_x & (K_y + \frac{k}{2}I)^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = (1+\lambda) \begin{pmatrix} (K_x + \frac{k}{2}I)^2 & 0 \\ 0 & (K_y + \frac{k}{2}I)^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}. \tag{33}$$

2.3 Efficient Computation

To reduce the computation load, it is better to approximate the kernel matrices with lower rank ones. The reasons are:

- Efficient matrix inversion and eigen-decomposition, making these methods linear in n - number of data objects, as opposed to n^3 originally.

- It is observed that many kernel matrices have low ranks. Advantages have already been taken for SVMs [4].
- It is natural to expect much linear dependency among variables in the same sets when working with CCA (if they are independent, overfitting may happen and meaningless correlations occur). Therefore, the kernel matrix is likely to be of low rank.

An efficient approximation is Incomplete Cholesky decomposition [6] or its dual implementation Gram-Schmidt orthogonalization. In the end, we will have lower rank approximations of kernel matrices as: $K_x = R_x R_x^T$ and $K_y = R_y R_y^T$, where $R_x \in R^{n*m_1}$ and $R_y \in R^{n*m_2}$. Both R_x and R_y are lower triangular matrices and $m_1, m_2 \ll n$.
In the reduced dimensional space, $\tilde{\alpha} = R_x^T \alpha \in R^{m_1}$ and $\tilde{\beta} = R_y^T \beta \in R^{m_2}$. Plugging into formulas 26 and 27, we have:

$$R_x R_x^T R_y R_y^T \beta - \lambda R_x (R_x^T R_x + kI) R_x^T \alpha = 0 \tag{34}$$

$$R_y R_y^T R_x R_x^T \alpha - \lambda R_y (R_y^T R_y + kI) R_y^T \beta = 0 \tag{35}$$

Putting R_x^T and R_y^T into the left hand side of the above equations gives us:

$$R_x^T R_x R_x^T R_y R_y^T \beta - \lambda R_x^T R_x (R_x^T R_x + kI) R_x^T \alpha = 0 \tag{36}$$

$$R_y^T R_y R_y^T R_x R_x^T \alpha - \lambda R_y^T R_y (R_y^T R_y + kI) R_y^T \beta = 0 \tag{37}$$

Let

$$\begin{aligned} Z_{xx} &= R_x^T R_x \in R^{m_1 * m_1}, \\ Z_{yy} &= R_y^T R_y \in R^{m_2 * m_2}, \\ Z_{xy} &= R_x^T R_y \in R^{m_1 * m_2}, \\ Z_{yx} &= R_y^T R_x \in R^{m_2 * m_1}. \end{aligned} \tag{38}$$

We have finally:

$$Z_{xx} Z_{xy} \tilde{\beta} - \lambda Z_{xx} (Z_{xx} + kI) \tilde{\alpha} = 0 \tag{39}$$

$$Z_{yy} Z_{yx} \tilde{\alpha} - \lambda Z_{yy} (Z_{yy} + kI) \tilde{\beta} = 0 \tag{40}$$

Assuming (again) that Z_{xx} and Z_{yy} are invertible, then:

$$Z_{xy} \tilde{\beta} - \lambda (Z_{xx} + kI) \tilde{\alpha} = 0 \tag{41}$$

$$Z_{yx} \tilde{\alpha} - \lambda (Z_{yy} + kI) \tilde{\beta} = 0 \tag{42}$$

The solution:

$$\tilde{\beta} = \frac{(Z_{yy} + kI)^{-1} Z_{yx} \tilde{\alpha}}{\lambda} \tag{43}$$

$$Z_{yx}(Z_{yy} + kI)^{-1} Z_{yx} \tilde{\alpha} = \lambda^2 (Z_{xx} + kI)\tilde{\alpha} \tag{44}$$

To compute $(Z_{xx} + kI)^{-1}$ and $(Z_{yy} + kI)^{-1}$ efficiently, do a complete Cholesky decomposition, $(Z_{xx} + kI) = S_x S_x^T$ and $(Z_{yy} + kI) = S_y S_y^T$ where $S_x \in R^{m_1 * m_1}$ and $S_y \in R^{m_2 * m_2}$ are lower triangular matrices.

$$S_x^{-1} Z_{xy} (S_y^T)^{-1} S_y^{-1} Z_{yx} (S_x^T)^{-1} \tilde{\alpha} = \lambda^2 \tilde{\alpha}. \tag{45}$$

Note 6: The formulas 41 and 42 can be rewritten in a matrix form:

$$\begin{pmatrix} 0 & Z_{xy} \\ Z_{yx} & 0 \end{pmatrix} \begin{pmatrix} \tilde{\alpha} \\ \tilde{\beta} \end{pmatrix} = \lambda \begin{pmatrix} (Z_{xx} + kI) & 0 \\ 0 & (Z_{yy} + kI) \end{pmatrix} \begin{pmatrix} \tilde{\alpha} \\ \tilde{\beta} \end{pmatrix}. \tag{46}$$

$$\begin{pmatrix} (Z_{xx} + kI) & Z_{xy} \\ Z_{yx} & (Z_{yy} + kI) \end{pmatrix} \begin{pmatrix} \tilde{\alpha} \\ \tilde{\beta} \end{pmatrix} = (1 + \lambda) \begin{pmatrix} (Z_{xx} + kI) & 0 \\ 0 & (Z_{yy} + kI) \end{pmatrix} \begin{pmatrix} \tilde{\alpha} \\ \tilde{\beta} \end{pmatrix}. \tag{47}$$

We then arrive at a generalized eigenvalue problem of a matrix with size $m_1 + m_2$. Once $\tilde{\alpha}$ and $\tilde{\beta}$ are computed, original solution can be recovered as:

$$\begin{aligned} \alpha &= (R_x R_x^T)^{-1} R_x \tilde{\alpha}, \\ w_x &= X\alpha. \end{aligned} \tag{48}$$

3 Generalization to m views

So far, we are dealing with correlation of two sets of variables. This section describe how to deal with m sets of variabels, called m views. here we first analyze criteria to generalize CCA to m views. We then describe MAXVAR, the criterion with known efficient solutions.

In practice, it is anticipated that sometimes data express in more than two views. It is the case of documents are translated in many languages. An example is Acquis, the European constitutional body. Each document are translated into 23 languages. Of course, taking into account two views would be a solution. However, having only two views may bring more canonical covariates than it should, spurious correlations may happen. Having additional views would give a more reliable estimations of correlation among view, making the set of canonical covariates smaller. By having a smaller set of canonical covariates, with the assumption that all semantic representations are

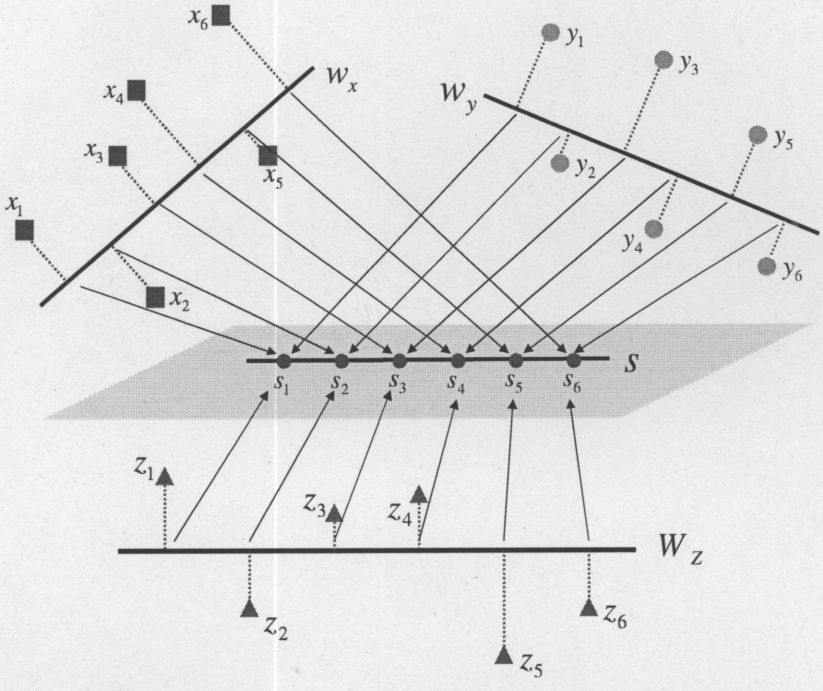

Figure 2: Generalizing CCA into multiview: projecting data from different views into a common subspace S, where semantically equivalent data points are projected into an unique point.

captured by those covariates, we expect to have a more concise (smaller and more reliable) representation of semantics. A pictorial description of multiview CCA is shown in figure 3

The central problem for generalizing CCA into multiview is the objective function of the multiview *correlation*. As in the classical two-view case, correlation is the objective function. However, correlation is defined for two random variables only. When generalizing into more than two views, there is nothing such as *correlation*. Here come the problem of defining what is a *correlation* for multiview version. There are attempts to define *correlation* for multiview, but one must bear in mind that there are several constraints. First, it must carry the meaning of correlation for our purpose of capturing common semantics of different views. Second, for a practical reason, the computational mean must be available for such objective functions. Here, we mean that there must be a solution to optimize the objective functions that scales well with large or massive data sets.

The next subsection will discuss the proposed criteria and we will analyze to see if they satisfy those constraints.

Criteria	Usage in ML	Suitability	Has solutions
SUMCOR	Min-sum-of-distance	Yes	No
MAXVAR	MKCCA	Yes	Yes
SSQCOR	No	Yes	No
MINVAR	KerICA	No	Yes
GENVAR	KerICA	No	Yes

Table 1: Analysis of CCA generalization criteria.

3.1 Analysis of criteria

For the two view problem as before, correlation is the only objective function has been taken into account so far. However, when having more than two views, there are many ways to define correlation among a set of projections from different views. The objectives for multiview CCA can be analyzed from the correlation matrix of all projected vectors. Here, we discuss only the objective functions for solving the first correlation variates. The following correlation variates can be computed in the same way, with an additional constraint of orthogonality of projections. Fortunately, this constraint is automatically satisfied by the solutions.

First is some notations. Suppose that having m views, X_1, X_2 ... X_m, which are row-wise centered. The target is to find m projections into w_1, w_2 ... w_m such that $X_i^T w_i$ are *maximally correlated* (the meanings of *maximally correlated* are defined later). Denote: $C_{ij} = X_i X_j^T$, $l_i = \sqrt{(w_i^T C_{ii} w_i)}$ be the length of the projected vector, $e_i = \frac{X_i^T w_i}{l_i}$ would be the normalized projected vector. Denote: C is the full covariance matrix comprising of C_{ij} and D is the block-diagonal matrix of C_{ii}. We now have: $\|e_i\| = 1$ and $1^T e_i = 0$. The correlation matrix of e_i is a $m * m$ matrix $\Phi = \{\phi_{ij}\}_{ij=1}^m$ satisfying:

$$\phi_{ij} =< e_i, e_j > . \tag{49}$$

There are some works in Statistics that propose criteria for generalizing CCA into more than m views [13], [10], and [7]. All the criteria basically measure how well projections of views are grouped together in the feature space. Hence, most of them are based on the matrix Φ. The criteria proposed in [10] are milestones. Other works just modify them to incorporate other information such as covariance into the criteria. The list of criteria is in the first column of table 3.1. It is noteworthy that all these criteria are equivalent to each other when $m = 2$.

The second column shows how these criteria are known and used in the Machine Learning community. We analyze these criteria to see if which ones would be suitable for our purpose, named Suitability in the table 3.1. The last column shows if using

these criteria may give a efficient enough algorithm for large scale application. SUMCOR is known to be the sum of square distance measure. It is the most intuitive criterion without any known efficient solution. In fact, it is equivalent to a multivariate eigenvalue problem where there exist no theories to shed a light on its solutions. It is known to have a very large number of solutions in [3]. MAXVAR is known to be a generalized eigenvalue problem and has been proposed and used in [1, 13, 17]. SSQCOR has not been used due to the fact that it is not known to be formulated to any easily solvable problem. The two criteria MINVAR and GENVAR are not considered to be suitable for generalized CCA for the following reason. As long as there is any linear dependency among projections of all the views, MINVAR and GENVAR would get the maximum value, indicating a perfect correlations among all projections. In fact, linear dependency does not guarantee perfect correlations for more than two random variables. These criteria are, therefore, used to measure independency in ICA [1].

3.2 MAXVAR

We start with another objective function:

$$\rho_3 = \max_{w} \sum_{i=1}^{m} \sum_{j=1}^{m} w_i^T C_{ij} w_j, \tag{50}$$

subject to: $\sum w_i^T C_{ii} w_i = 1$.

Theorem 1: ρ_3 is the maximum generalized eigenvalue (λ_{max}) of the following generalized eigenvalue problem.

$$\begin{pmatrix} C_{11} & C_{12} & \cdots & C_{1m} \\ C_{21} & C_{22} & \cdots & C_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ C_{m1} & C_{m2} & \cdots & C_{mm} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{pmatrix} = \lambda \cdot Diag(C_{11}, C_{22}, \cdots C_{mm}) \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{pmatrix} \tag{51}$$

Proof: Lagrangian of the objective function:

$$L(w, \lambda) = \sum_{i=1}^{m} \sum_{j=1}^{m} w_i^T C_{ij} w_j - \lambda(\sum w_i^T C_{ii} w_i - 1). \tag{52}$$

Taking derivatives with respect to w_i for all i:

$$\sum_{j=1}^{m} C_{ij} w_j = \lambda C_{ii} w_i. \tag{53}$$

Multiplying w_i^T to the left hand side of each part of 53, we have:

$$\sum_{j=1}^{m} w_i^T C_{ij} w_j = \lambda w_i^T C_{ii} w_i. \tag{54}$$

Summing over i from 54 then:

$$\sum_{i=1}^{m}\sum_{j=1}^{m} w_i^T C_{ij} w_j = \lambda. \tag{55}$$

We have ρ_3 is an eigenvalue of 51. Now we need to prove it is the largest one.

If $\rho_3 = \lambda < \lambda_{max}$ then the exists a (normalized) w_{max} is the generalized eigenvector corresponding to the eigenvalue λ_{max}. Then:

$$\lambda_{max} = \sum_{i=1}^{m}\sum_{j=1}^{m} w_{i-max}^T C_{ij} w_{j-max} > \lambda = \rho_3. \tag{56}$$

This contradicts the optimality of ρ_3. Hence, $\rho_3 = \lambda_{max}$.

Note 7: One can conclude that maximizing ρ_3 is equivalent to finding the maximum eigenvalue of the generalized eigenvalue problem.This is different from the problem in **??** only at having the same λ instead of different λ_i. The difference between ρ_1 and ρ_3 is the weighting of each (normalized) projected vector e_i with its length l_i. ρ_3 is, in fact, a weighted version of ρ_1.

Geometrical interpretation of the weights: The weights l_i are introduced from w_i, which are, in turn, introduced by the generalized eigenvalue problem. However, we can get some ideas from equation 54.

Having $l_i e_i = X_i^T w_i$, the right hand side of the equation 54 is:

$$\lambda w_i^T C_{ii} w_i = \lambda \cdot l_i^2. \tag{57}$$

The left hand side is

$$\sum_{j=1}^{m} w_i^T C_{ij} w_j = w_i^T X_i \sum_{j=1}^{m} W_j^T x_j =< l_i e_i, \sum_{j=1}^{m} l_j e_j > . \tag{58}$$

Denote $le = \sum_{j=1}^{m} l_j e_j$ be the sum of all un-normalized projected vectors, then $\lambda \cdot l_i^2 =< l_i e_i, le >$.

$$l_i = \frac{l}{\lambda} < e_i, e > . \tag{59}$$

$$< e_i, le > = \lambda l_i. \tag{60}$$

Hence, the weight of a projected vector e_i in the objective function is the dot product between it and the weighted mean.

In light of the correlation matrix Φ, we can see that:

$$\begin{pmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1m} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{m1} & \phi_{m2} & \cdots & \phi_{mm} \end{pmatrix} \begin{pmatrix} l_1 \\ l_2 \\ \vdots \\ l_m \end{pmatrix} = \begin{pmatrix} \sum_j < e_1, l_j e_j > \\ \sum_j < e_2, l_j e_j > \\ \vdots \\ \sum_j < e_m, l_j e_j > \end{pmatrix} = \begin{pmatrix} < e_1, le > \\ < e_2, le > \\ \vdots \\ < e_m, le > \end{pmatrix} = \lambda \begin{pmatrix} l_1 \\ l_2 \\ \vdots \\ l_m \end{pmatrix} \tag{61}$$

Refer to 60 for the last equation.

This means that the weights l_i form a eigenvector for the correlation matrix. This also means that the objective function ρ_3, equivalently λ (see 55), is actually *maximizing an eigenvalue* of the correlation matrix Φ.

Theorem 2: λ_{max} is the maximum eigenvalue of all the correlation matrices obtained from projections of X_i into w_i.

Proof: Denote $l_i = \|X_i^T w_i\|$, $L = \{l_1, l_2 \cdots l_m\}^T$, $e_i = \frac{X_i^T w_i}{l_i}$ as usual.

$$\begin{aligned} \lambda_{max} &= \max_W \frac{W^T C W}{W^T D W} \\ &= \max_W \frac{\sum_{i=1}^m \sum_{j=1}^m w_i^T C_{ij} w_j}{\sum_{i=1}^m w_i^T C_{ii} w_i} \\ &= \max_W \frac{\sum_{i=1}^m \sum_{j=1}^m < l_i e_i, l_j e_j >}{\sum_{i=1}^m l_i^2} \\ &= \max_W \frac{\sum_{i=1}^m \sum_{j=1}^m l_i l_j < e_i, e_j >}{\sum_{i=1}^m l_i^2} \\ &= \max_W \frac{L^T \Phi L}{L^T L}. \end{aligned} \tag{62}$$

One can observe that we can scale w_i independently of other w_j, therefore l_i can receive any value independently of other l_j. As W runs all over its space, L also runs all over its space.

Therefore,

$$\begin{aligned} \lambda_{max} &= \max_W \frac{L^T \Phi L}{L^T L} \\ &= \max_L \frac{L^T \Phi L}{L^T L}. \end{aligned} \tag{63}$$

Then, λ_{max} is also the maximum eigenvalue of all eigenvalues of all correlation matrices. This means that ρ_3 is the MAXVAR criterion in [10].

3.2.1 Kernelization and Regularization

In a similar fashion as [1], we can come to the kernelized form of the objective as finding the maximal eigenvalue of:

$$\begin{pmatrix} K_1K_1 & K_1K_2 & \cdots & K_1K_m \\ K_2K_1 & K_2K_2 & \cdots & K_2K_m \\ \vdots & \vdots & \ddots & \vdots \\ K_mK_1 & K_mK_2 & \cdots & K_mK_m \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix} = \lambda \cdot Diag(K_1K_1, K_2K_2, \cdots K_mK_m) \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix}. \tag{64}$$

If we regularize like PLS w^Tw, then the above problem becomes:

$$\begin{pmatrix} 0 & K_1K_2 & \cdots & K_1K_m \\ K_2K_1 & 0 & \cdots & K_2K_m \\ \vdots & \vdots & \ddots & \vdots \\ K_mK_1 & K_mK_2 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix}$$

$$= \lambda \cdot Diag(K_1K_1 + kK_1, K_2K_2 + kK_2, \cdots K_mK_m + kK_m) \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix}. \tag{65}$$

If we want to regularize futher $\alpha^T\alpha$, the it becomes:

$$\begin{pmatrix} 0 & K_1K_2 & \cdots & K_1K_m \\ K_2K_1 & 0 & \cdots & K_2K_m \\ \vdots & \vdots & \ddots & \vdots \\ K_mK_1 & K_mK_2 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix}$$

$$= \lambda \cdot Diag(K_1K_1 + kI, K_2K_2 + kI, \cdots K_mK_m + kI) \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix}. \tag{66}$$

3.2.2 Efficient Computation

For the PLS regularization, proceeding similarly as in section 5 or previous subsection, we arrive at the reduced dimensional form:

$$\begin{pmatrix} 0 & Z_{12} & \cdots & Z_{1_m} \\ Z_{21} & 0 & \cdots & Z_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{m1} & Z_{m2} & \cdots & 0 \end{pmatrix} \begin{pmatrix} \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \\ \vdots \\ \tilde{\alpha}_m \end{pmatrix} = \tilde{\lambda} \cdot Diag(Z_{11}+kI, Z_{22}+kI, \cdots Z_{mm}+kI) \begin{pmatrix} \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \\ \vdots \\ \tilde{\alpha}_m \end{pmatrix}. \tag{67}$$

If the $\alpha^T\alpha$ regularization then:

$$\begin{pmatrix} 0 & Z_{12} & \cdots & Z_{1_m} \\ Z_{21} & 0 & \cdots & Z_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{m1} & Z_{m2} & \cdots & 0 \end{pmatrix} \begin{pmatrix} \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \\ \vdots \\ \tilde{\alpha}_m \end{pmatrix} = \lambda \cdot Diag(Z_{11}+kZ_{11}^{-1}, Z_{22}+kZ_{22}^{-1}, \cdots Z_{mm}+kZ_{mm}^{-1}) \begin{pmatrix} \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \\ \vdots \\ \tilde{\alpha}_m \end{pmatrix}. \tag{68}$$

Original solution can be recovered by:

$$\begin{aligned} \alpha_i &= K_{ii}^{-1} R_i \tilde{\alpha}_i, \\ R_i &= U_i \Lambda_i V_i, \\ K_{ii}^{-1} &= U_i \Lambda_i^{-2} U_i, \\ w_i &= X_i U_i \Lambda_i^{-1} V_i \tilde{\alpha}_i. \end{aligned} \tag{69}$$

4 Algorithm and Experiments

4.1 Algorithm

The algorithm we implemented is using MAXVAR criterion, which is known to be the generalized eigenvalue problem. There are two part of the algorithms: learning, i.e. determining the canonical covariate and testing, i.e. projecting test data into the common subspace. Learning part follows strictly the description in section 2.3 and 3.2. The the generalized eigenvalue problem, we use the BLOPEX package [11].

There are two ways for the testing part. One is to go from $\tilde{\alpha}_i$ to α_i and w_i, then testing data is projected into these canonical covariate directly. Another way, as described in [8], is to project the data into the low dimensional space and perform projections there. We found that the second way is not only much more efficient, but also give much more stable results. Unstability and inefficiency come from the fact that in the former way, one needs to pseudo-inverse the reduced dimensional kernel matrices. Matrix pseudo-inversion is very expensive and unstable.

We has tried and found that the implementation is scalable so far for 30000 training data for 4 languages (totalling 120000 training data points).

4.2 Experiments

We would like to evaluate this algorithm to see whether having more than two views can be beneficial for some cross language information retrieval tasks. We choose the Mate Retrieval task in our experiments. The data is the alignments of the JRC-Acquis Corpus [1]. It contains alignment of sentences from pair languages. As we need a multiple view corpus, we merged from different pairs together to create on multiple alignment corpus. About merging alignment pairs, about 90% are retained after merging. Any conflict in merging is discarded.

The experiments were setup as follows. The baseline is the two language case, en and fr, and multiple view version is four languages: de, en, fr, and es. We trained models for the two and four view settings. In the testing face, we query in en and expect the answer in fr in both cases. Euclidean distance is used to retrieve mates. We used two measures for the result, top 1% retrieval and average mate rank. Training data were randomly sampled from the corpus. The sizes of training data were 50, 100, 200, 500, 1000, 2000 and 5000. For each size, we sampled 5 training data sets. For testing, we sampled 10 testing sets of size 1000.

Parameters of the experiments were set as follows. We used linear kernel. The maximum dimension of the low dimensional approximation was 300. Regularization with k were set as recommended in [1]. For the generalized eigenvalue problem, we run the 4000 iterations. We extracted 30 eigenvalue-eigenvector pairs. Results are depicted in the figure 4.2. On the horizontal axes are the sizes of training sets. The vertical axes are the measures with respect to different training sizes. Any point in the graphs is the average of 50 measures from 5 models on 10 data sets.

The conclusion we derived from this graph is that having four views does not improve the retrieval rates when the training sizes are 500 or more. The two view version is more tailored to the language pair. The multi-view version is beneficial when there are few training data. The reason is that in that case, two view version is not relible enough and additional data helps. When there are more data, the two view version is more tailored to the task.

5 Low Rank Kernel Learning

Suppose that views lie in spaces R_1 R_2, $\cdots$ R_m. We wish to find for each view a subspace $R_i^* \subset R_i$ so that after projecting data onto these subspaces, data from

[1] http://www.jrc.it/langtech

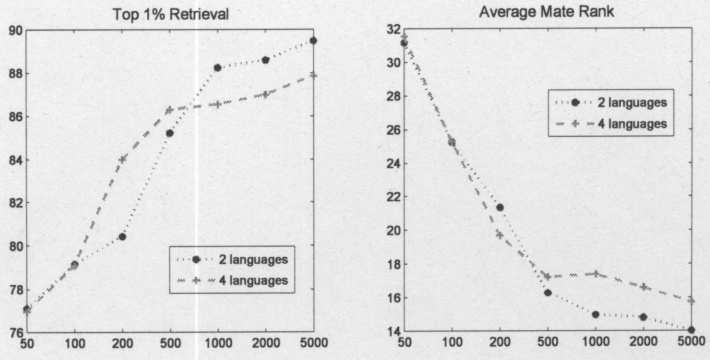

Figure 3: Retrieval results of different sizes of the set.

different views become correlated.

However, using original spaces R_i to discover R_i^* is, being a generalized eigenvalue problem with MAXVAR as analyzed in the previous section, is computationally expensive for moderate sizes. One usually use some efficient approximation. As for mKCCA, in the kernel setting, the usual technique is using low rank approximation of kernel matrices to have efficient training.

Customizing Incomplete Cholesky Decomposition for CCA. The problem is that not the same semantic space from different views are attracted as ICD is carried out independently from each view. We found that in the 5000 sizes data sets, for the two view version, about 80% of the data objects are in common when ICD picks them for the spanning set of the low dimensional space. For four-view version, the number drops to about 64%. In order to ensure that the same semantic space are extracted from different languages, one needs to have a customized ICD for this purpose. This resembles the work in [2], and indeed is mentioned as a future work.

5.1 Incomplete Cholesky Decomposition with side information

Our idea is to use ICD that makes sure that 100% of data picked are in common. This can be done by simultaneously decomposing all data matrices at the same time. The data object picked by by ICD each time is common among all views; the one which overall reduces the most the trace of the projected data matrices is picked.

5.2 Experiments

Experiments are carried out on the same data set JRC-Acquis as the previous section.

As before, we train on 5 data samples of different sizes (from 50 to 5000). The models are then used to test on 10 samples of size 1000. The figure shows an average of 50 results, which are the success rates of retrieving the right mate within the top 1%. mKCCA setting is as the previous section, with the low rank approximation of kernel matrices of 50 (set fixed on these experiments).

It shows that for two-view version, the new ICD algorithm gives a some performance improvement. However, on the four-view version, the improvements are seen more clearly. The reason could be the fact that the four-view version has less data points in common.

6 Conclusion and Future Works

The contributions of this paper are: an overall analysis of criteria for generalizing CCA into multiple views and a new ICD algorithm specially for generalized CCA. We showed that the MAXVAR, as proposed and used by others are the only criterion with known efficient solution. We provided an implementation. On the dataset we evaluated on, the multiple view versions show to improve performances of Mate Retrieval tasks when there are few training data. The new ICD algorithm we proposed simultaneously decompose kernels from different views to make sure that they retain more correlation on the projected subspaces. Experiments show that new ICD algorithm gives generalized CCA a higher performance.

References

[1] Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2003.

[2] Francis R. Bach and Michael I. Jordan. Predictive low-rank decomposition for kernel methods. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 33–40, New York, NY, USA, 2005. ACM Press.

[3] Moody T. Chu and J. Loren Watterson. On a multivariate eigenvalue problem, part i: algebraic theory and a power method. *SIAM Journal on Scientific Computing*, 14(5):1089–1106, 1993.

[4] Shai Fine and Katya Scheinberg. Efficient svm training using low-rank kernel representations. *Journal Machine Learning Research*, 2:243–264, 2002.

[5] Bernd Fischer, Volker Roth, and Joachim M. Buhmann. Time-series alignment by non-negative multiple generalized canonical correlation analysis. *BMC Bioinformatics*, to appear.

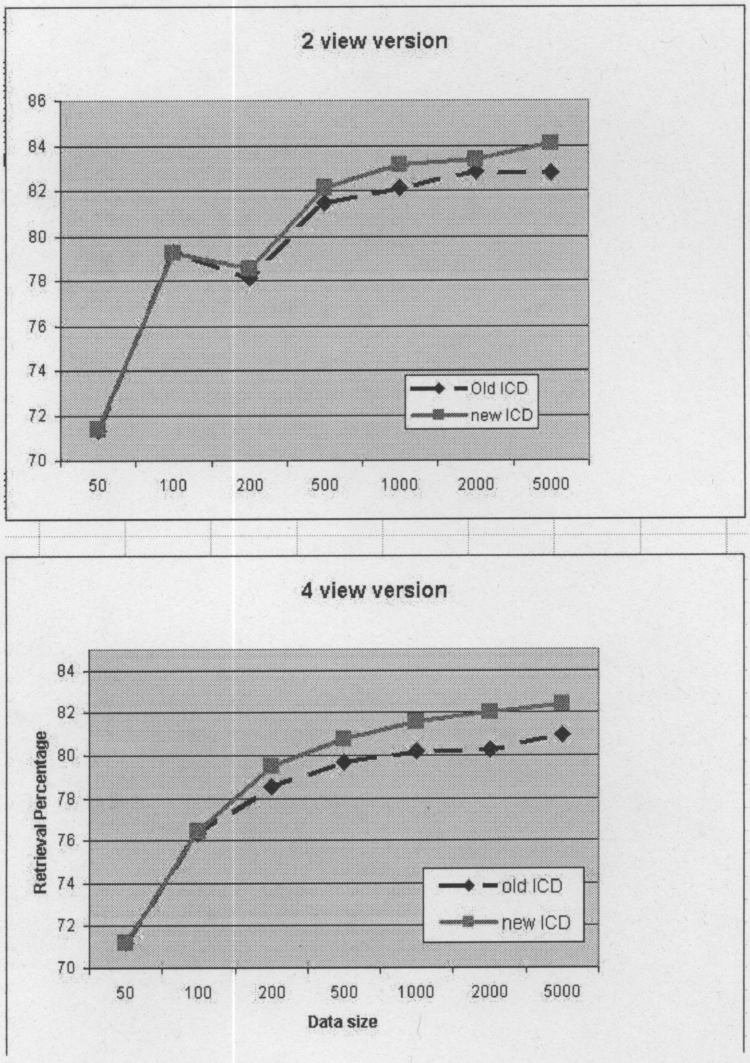

Figure 4: Comparing the original and new ICD algorithms for mKCCA. New ICD algorithm gives mKCCA a higher performance, especially when there are more views and smaller rank is used to approximate kernels.

[6] Gene H. Golub and Charles F. Van Loan. *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences)*. The Johns Hopkins University Press, October 1996.

[7] Mohamed Hanafi and Henk A. L. Kiers. Analysis of k sets of data, with differential emphasis on agreement between and within sets. *Computational Statstistics and Data Analysis*, 51(3):1491–1508, 2006.

[8] David R. Hardoon, Sandor R. Szedmak, and John R. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[9] Harold Hotelling. Relations between two sets of variables. *Biometrika*, 28:321–372, 1936.

[10] Jon R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58:433–451, 1971.

[11] A. V. Knyazev, M. E. Argentati, I. Lashuk, and E. E. Ovtchinnikov. Block locally optimal preconditioned eigenvalue xolvers (blopex) in hypre and petsc. *IAM Journal on Scientific Computing (SISC)*, to appear.

[12] Yaoyong Li and John Shawe-Taylor. Using kcca for japanese-english cross-language information retrieval and document classification. *Journal of Intelligent Information Systems*, 27(2):117–133, 2006.

[13] George Michailidis and Jan de Leeuw. The gifi system for descriptive multivariate analysis. *Statistical Science*, 1998.

[14] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004.

[15] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, September 1998.

[16] Grace Wahba. *Spline models for observational data*. SIAM [Society for Industrial and Applied Mathematics], 1990.

[17] Yoshihiro Yamanishi, Jean-Philippe Vert, A. Nakaya, and Minoru Kanehisa. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *ISMB (Supplement of Bioinformatics)*, pages 323–330, 2003.