Kernel Canonical Correlation Analysis and its Applications to Nonlinear Measures of Association and Test of Independence*

Su-Yun Huang,[1†] Mei-Hsien Lee[2] and Chuhsing Kate Hsiao[2]

[1]Institute of Statistical Science, Academia Sinica, Taiwan

[2]Division of Biostatistics, Institute of Epidemiology
National Taiwan University

draft, May 25, 2006

Abstract

Measures of association between two sets of random variables have long been of interest to statisticians. The classical canonical correlation analysis can characterize, but also be limited to, linear association. In this article we study nonlinear association measures using the kernel method. The introduction of kernel method from machine learning community has a great impact on statistical analysis. The kernel canonical correlation analysis (KCCA) is a method that generalizes the classical linear canonical correlation analysis to nonlinear setting. Such a generalization is nonparametric. It allows us to depict the nonlinear relation of two sets of variables and enables applications of classical multivariate data analysis originally constrained to linearity relation. Moreover, the kernel-based canonical correlation analysis no longer requires the Gaussian distributional assumption on observations, and therefore enhances greatly the applicability.

The main purpose of this article is twofold. One is to link the KCCA emerging from the machine learning community to the nonlinear canonical analysis in statistical literature, and the other is to provide the KCCA some further statistical applications including association measures, dimension reduction and test of independence without the usual Gaussian assumption. Implementation algorithms will be discussed and several examples will be illustrated.

Key words and phrases: association, canonical correlation analysis, kernel canonical correlation analysis, test of independence.

1 Introduction

The description of relation between two sets of variables has been a long interest to many researchers. Hotelling (1936) introduced the canonical correlation analysis to describe the linear relation between two sets. We use the abbreviation LCCA for linear canonical correlation analysis. The LCCA is

*Running head: Kernel canonical correlation analysis and its applications

†Corresponding author: Su-Yun Huang, Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, syhuang@stat.sinica.edu.tw.

concerned with the linear relation between two sets of variables having a joint distribution. It aims to define a new orthogonal coordinate system for each of the two sets in a way that the new pair of coordinate systems are optimal in maximizing the correlations. The new systems of coordinates are simply linear systems of the original ones. Thus, the LCCA can only be used to describe linear relation. Via such linear relation it can extract only linear subspaces for dimension reduction or features selection. Under Gaussian assumption, the LCCA can also be used for testing stochastic independence between two sets of variables. However, all these become invalid, if data are not Gaussian nor at least elliptically symmetrically distributed. Thus, one has to resort to some nonlinear and nonparametric strategy.

Motivated from the active development of statistical learning theory (Vapnik, 1998; Hastie, Tibshirani and Friedman, 2001; and references therein) and the popular and successful usage of various kernel machines (Cortes and Vapnik, 1995; Cristianini and Shawe-Taylor, 2000; Schölkopf and Smola, 2002; Herbrich, 2002), there has emerged a hybrid approach of the LCCA with a kernel machine (Akaho, 2001; Bach and Jordan, 2002), named kernel canonical correlation analysis (KCCA). The KCCA was also studied recently by Gretton, Herbrich and Smola (2003), Kuss and Graepel (2003), Hardoon, Szedmak and Shawe-Taylor (2004) and Gretton, Herbrich, Smola, Bousquet and Schölkopf (2005). All the above works are more on the computer science side. In statistical literature there has been a wide concern as well as advanced development of the nonlinear canonical correlation analysis (NLCCA), e.g., Dauxois, Romain and Viguier (1993), Dauxois and Nkiet (1997), Dauxois and Nkiet (1998), Dauxois and Nkiet (2002), Dauxois, Nkiet and Romain (2004), Ramsay and Silverman (Chapter 12, 1997), Eubank and Hsing (2005), Hsing, Liu, Brun and Dougherty (2005) among many others. The KCCA originated from the machine learning community can be regarded (from a theoretical point of view) as a special case of nonlinear canonical analysis. In this article we will bridge together the general statistical theory of nonlinear canonical analysis and the kernel algorithms of canonical analysis. We will also introduce several statistical applications.

The rest of the article is organized as follows. In Section 2 we brief the methods of LCCA and KCCA. In Section 3 we discuss some implementation issues concerning regularization and parameters selection. In Section 4 we introduce some statistical applications using the KCCA. These applications are nonlinear association measures, dimension reduction for nonlinear discriminant analysis and a test of independence. Finally, concluding remarks and discussion are given in Section 5. All relevant theoretical background, including the nonlinear canonical correlation analysis (NLCCA), its approximation, estimation and asymptotic distribution, is in the Appendix.

2 Canonical correlation analysis

2.1 A review of linear canonical correlation analysis

Suppose the random vector X of q components has a probability distribution P on $\mathcal{X} \subset \Re^q$. We partition X into q_1 and q_2 components:

$$X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix}.$$

Let the corresponding partition of $\mathcal{X}$ be denoted by $\mathcal{X}_1 \oplus \mathcal{X}_2$. All vectors are column vectors throughout, unless transposed to row vectors. We adopt a Matlab vector and matrix convention

$$[X^{(1)}; X^{(2)}] := \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix},$$

where the semicolon denotes stacking $X^{(1)}$ on top of $X^{(2)}$ as shown in the right hand side of the equality. We are interested in finding the relation between $X^{(1)}$ and $X^{(2)}$. The LCCA describes linear

relation by reducing the correlation structure between these two sets of variables to the simplest possible form by means of linear transformations on $X^{(1)}$ and $X^{(2)}$. For the first pair of canonical variates, the LCCA seeks a pair of linear variates $\alpha' X^{(1)}$ and $\beta' X^{(2)}$ that maximize the correlation, namely, it solves the following optimization problem:

$$\rho := \max_{[\alpha;\beta]\in\Re^{q_1+q_2}} \alpha'\Sigma_{12}\beta \quad \text{subject to} \quad \alpha'\Sigma_{11}\alpha = 1 \ \text{ and } \ \beta'\Sigma_{22}\beta = 1, \tag{1}$$

where Σ_{ij}'s are the covariance matrices of $X^{(i)}$ and $X^{(j)}$, $i, j = 1, 2$. Denote the solution of (1) by $[\alpha_1; \beta_1]$, called the first pair fo canonical vectors. For the rest pairs of canonical vectors, it solves sequentially the same problem as (1) with extra constraints of iterative orthonormality: i.e., for the kth pair

$$\alpha_k'\Sigma_{11}\alpha_i = 0 \ \text{ and } \ \beta_k'\Sigma_{22}\beta_i = 0, \ \forall i = 1, \ldots, k-1. \tag{2}$$

The sets $\{\alpha_i\}$ and $\{\beta_i\}$ can be regarded as coordinate axes in a pair of new coordinate systems. The sequence of correlation coefficients $\{\rho_1, \rho_2 \ldots\}$ describes only the linear relation between $X^{(1)}$ and $X^{(2)}$. The LCCA can be characterized by pairs of canonical variates $\{(U_1, V_1), (U_2, V_2), \ldots\}$ and correlation coefficients $\{\rho_1, \rho_2 \ldots\}$, where $U_i = \alpha_i' X^{(1)}$ and $V_i = \beta_i' X^{(2)}$. The LCCA can be justified by assuming that $X^{(1)}$ and $X^{(2)}$ have a joint Gaussian distribution, and the likelihood ratio criterion for testing independent of $X^{(1)}$ and $X^{(2)}$ can be expressed entirely in terms of sample correlation coefficients based on a given data set.

Assume that we have data set $A = \{x_j\}_{j=1}^n$, where $x_j = [x_j^{(1)}; x_j^{(2)}]$ partitioned according to $\mathcal{X} = \mathcal{X}_1 \oplus \mathcal{X}_2$. Let $\mathbb{X}$ denote the data design matrix

$$\mathbb{X} := \begin{bmatrix} x_1^{(1)'} & x_1^{(2)'} \\ \vdots & \vdots \\ x_n^{(1)'} & x_n^{(2)'} \end{bmatrix}_{n\times(q_1+q_2)} := [\mathbb{X}_1 \ \mathbb{X}_2].$$

The sampling-based LCCA for $X^{(1)}$ and $X^{(2)}$ is to work on $[\mathbb{X}_1 \ \mathbb{X}_2]$. The covariance matrices in optimization problems (1) and (2) are replaced by their sample versions.

2.2 Kernel generalization of canonical correlation analysis

There are cases where linear correlations may not be adequate for describing "association" between $X^{(1)}$ and $X^{(2)}$. A natural alternative, therefore, is to explore and exploit their nonlinear relation. Several authors in the machine learning community (see Introduction) have resorted to kernel method. The so called KCCA can be viewed as a special case of NLCCA, mainly due to Dauxois *et al.* (see Introduction). Here we give a working description of the KCCA.[1]

A real-valued function $\kappa(\cdot, \cdot)$ defined on some index set $\mathcal{X} \times \mathcal{X}$ is said to be positive definite if

$$\forall n \in N, \ \forall (a_1, \ldots, a_n) \in \Re^n, \ \forall (t_1, \ldots, t_n) \in \mathcal{X}^n,$$
$$\textstyle\sum_{i,j=1}^n a_i a_j \kappa(t_i, t_j) \geq 0,$$

and κ is said to be strictly positive definite when the equality holds if and only if $a_1 = \ldots = a_n = 0$. Let $\kappa_1(\cdot, \cdot)$ and $\kappa_2(\cdot, \cdot)$ be two (strictly) positive definite kernels defined on $\mathcal{X}_1 \times \mathcal{X}_1$ and $\mathcal{X}_2 \times \mathcal{X}_2$, respectively. Then, $\kappa := \kappa_1 + \kappa_2$ is a (strictly) positive definite kernel defined on $(\mathcal{X}_1 \oplus \mathcal{X}_2) \times (\mathcal{X}_1 \oplus \mathcal{X}_2)$.

[1]The KCCA formulation here is different from that in the above mentioned works originated from the machine learning community. Our formulation is in the primal space with Aronszajn kernel map as our feature mapping; while the computer scientists' works are formulated via kernel spectrum-based feature mapping and are solved in the dual space. The two different kinds of formulation are isometrically isomorphic. Our formulation is closer to that of Dauxois *et al.* The isometric isomorphism can be found in Huang and Hwang (2006).

For a given point $x \in \mathcal{X}$, we augment x with kernel values through

$$x \mapsto \gamma_x = [\gamma_x^{(1)}; \gamma_x^{(2)}], \quad \text{where } \gamma_x^{(i)} = [\kappa_i(x^{(i)}, z_1^{(i)}); \ldots; \kappa_i(x^{(i)}, z_{m_i}^{(i)})], \quad i = 1, 2. \tag{3}$$

Then, each data point x_j is augmented by $\gamma_j = [\gamma_j^{(1)}; \gamma_j^{(2)}]$, where

$$\gamma_j^{(i)} = [\kappa_i(x_j^{(i)}, z_1^{(i)}); \ldots; \kappa_i(x_j^{(i)}, z_{m_i}^{(i)})], \quad i = 1, 2, \quad j = 1, \ldots, n.$$

By matrix notation, the augmented kernel design matrix is given by

$$\mathbb{K} := [\, \mathbb{K}_1 \; \mathbb{K}_2 \,] := \begin{bmatrix} \gamma_1^{(1)\prime} & \gamma_1^{(2)\prime} \\ \vdots & \vdots \\ \gamma_n^{(1)\prime} & \gamma_n^{(2)\prime} \end{bmatrix}_{n \times (m_1 + m_2)}. \tag{4}$$

The most common choice of $\{z_\nu^{(1)}\}$ and $\{z_\mu^{(2)}\}$ is probably $\{z_\nu^{(1)}\} = \{x_\nu^{(1)}\}_{\nu=1}^n$ and $\{z_\mu^{(2)}\} = \{x_\mu^{(2)}\}_{\mu=1}^n$, which leads to matrices with the (i, j)th entries given below:

$$\mathbb{K}_1(i, j) = \kappa_1(x_i^{(1)}, x_j^{(1)}) \;\text{ and }\; \mathbb{K}_2(i, j) = \kappa_2(x_i^{(2)}, x_j^{(2)}), \; i, j = 1, \ldots, n. \tag{5}$$

Such matrices are called full-data kernel matrices and are commonly used in variants of support vector machines. The augmented representation of x_j by $\gamma_j = [\gamma_j^{(1)}; \gamma_j^{(2)}] \in \Re^{m_1 + m_2}$ can be regarded as an alternative way of recording data measurements with high inputs. The KCCA procedure consists of two major steps:

(a) Transform data points to a kernel-augmented representation as in (3), or equivalently (4) in matrix notation. One common choice for the kernel function is the Gaussian density function. Throughout this article we use Gaussian pdf as our choice of kernel in all experiments.

(b) Carry out the LCCA on the kernel augmented data $\mathbb{K} = [\, \mathbb{K}_1 \; \mathbb{K}_2 \,]$, as contrast to working on $\mathbb{X}$ for linear canonical correlation analysis. Note that some sort of *regularization* is necessary here to solve the corresponding canonical analysis problem. It involves a spectrum decomposition (also known as singular value decomposition) of extracting the canonical variates and correlation coefficients. Detailed computational implementation is discussed in the Implementation Section.

Remark 1 *We can consider augmenting the data in $A = \{x_j\}_{j=1}^n$ by functional representation:*

$$x_j \mapsto \kappa(x_j, \cdot) \in F = \{\kappa(x_j, \cdot) = \kappa_1(x_j^{(1)}, \cdot) + \kappa_2(x_j^{(2)}, \cdot)\}_{j=1}^n.$$

That is, a data point is re-coded by a function. The KCCA is a canonical analysis for these functional data. However, these functional data are not congenital but are intentionally made so, in order to study the nonlinear relation in a canonical analysis. The kernel matrix $\mathbb{K}$ in (4) is a discretized approximation of these functional data using a specific choice of basis set. More theoretical details are in Section 2.3 and Appendix.

2.3 More on kernel-augmented data

Here we will introduce a feature mapping, which brings the original sample space $\mathcal{X}_1 \oplus \mathcal{X}_2 \subset \Re^{q_1 + q_2}$ into an infinite dimensional Hilbert space. Parallel to the classical multivariate analysis in Euclidean spaces, procedures of the same kind can be developed in Hilbert spaces for convenient nonlinear

analysis. It is often done via using a certain parametric notion of a classical method. Here, since we are studying the association measures, we adopt the notion of LCCA.

For a given positive definite kernel $\kappa(x,t)$ defined on $\mathcal{X}\times\mathcal{X}$, there exists a Hilbert space $\mathcal{H}$ consisting of a collection of functions $\{f(x):\sum_{i=1}^{m}a_i\kappa(x,t_i),\forall m\in N,\forall a_i\in\Re,\forall t_i\in\mathcal{X}\}$ and its closure, where the closure is taken with respect to the norm:

$$\|\sum_{i=1}^{m}a_i\kappa(x,t_i)\|_{\mathcal{H}}^2=\sum_{i,j=1}^{m}a_ia_j\kappa(t_i,t_j).$$

Such a Hilbert space is a reproducing kernel Hilbert space (RKHS), whose definition is given below.

Definition 1 (Reproducing kernel Hilbert space) *A Hilbert space of real-valued functions on an index set $\mathcal{X}$ satisfying the property that, all the evaluation functionals are bounded linear functionals, is called a reproducing kernel Hilbert space.*

To every RKHS of functions on $\mathcal{X}$ there corresponds a unique positive-definite kernel κ satisfying the reproducing property, $\langle f(\cdot),\kappa(x,\cdot)\rangle=f(x)$ for all f in this RKHS and all $x\in\mathcal{X}$. We say that this RKHS admits the kernel κ. The theory of RKHS is a powerful tool in statistics and probability, and in particular in stochastic processes, nonparametric inference and the recent burst of kernel learning algorithms. See Berlinet and Thomas-Agnan (2004) for a friendly clear and up-to-date account of theory, examples and computational aspects of RKHS.

Let κ_1 and κ_2 be two strictly positive definite kernels. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be the associated RKHS admitting κ_1 and κ_2, respectively. Consider a transformation, $\gamma:\mathcal{X}=\mathcal{X}_1\oplus\mathcal{X}_2\mapsto\mathcal{H}=\mathcal{H}_1\oplus\mathcal{H}_2$ given by

$$\gamma:x=[x^{(1)};x^{(2)}]\mapsto\kappa_1(x^{(1)},\cdot)+\kappa_2(x^{(2)},\cdot). \tag{6}$$

The original sample space $\mathcal{X}$ is then embedded into a new sample space $\mathcal{H}$ via the transformation γ, called Aronszajn kernel map, which is named after Aronszajn (1950). Each point $x\in\mathcal{X}$ is mapped to an element in $\mathcal{H}$. If we restrict γ to $\mathcal{X}_1$, it maps $\mathcal{X}_1$ into $\mathcal{H}_1$, and similarly for the restriction to $\mathcal{X}_2$. Let $\{\phi_\nu\}_{\nu=1}^{m_1}$ and $\{\psi_\mu\}_{\mu=1}^{m_2}$ be our choice of (incomplete) linear independent systems in $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively. The discretization of functional data $\{\kappa_1(x_j^{(1)},\cdot)+\kappa_2(x_j^{(2)},\cdot)\}_{j=1}^{n}$ by the systems is given by

$$[\langle\kappa_1(x_j^{(1)},\cdot),\phi_\nu(\cdot)\rangle_{\mathcal{H}_1},\langle\kappa_2(x_j^{(2)},\cdot),\psi_\mu(\cdot)\rangle_{\mathcal{H}_2}]=[\phi_\nu(x_j^{(1)}),\psi_\mu(x_j^{(2)})],$$

where $j=1,\ldots,n$, $\nu=1,\ldots,m_1$ and $\mu=1,\ldots,m_2$, which leads to an $n\times(m_1+m_2)$ kernel-augmented data matrix. If we choose $\{\kappa_1(z_\mu^{(1)},\cdot)\}_{\nu=1}^{m_1}$ and $\{\kappa_2(z_\mu^{(2)},\cdot)\}_{\mu=1}^{m_2}$ as our choice of linear independent systems for discretization, it leads to the kernel data matrix $\mathbb{K}$ in (4). That is, the KCCA is carried out by performing LCCA on discretized kernel data. The discretization depends on the choice of incomplete linear independent systems for $\mathcal{H}_1$ and $\mathcal{H}_2$.

The mapping γ transforms the problem on Euclidean space to a problem on RKSH. Nonparametric and nonlinear approach in an Euclidean space is often relatively hard to handle in theory and in computational implementation. However, the transformed problem on an RKHS becomes transparent in theory and is easier to compute, as it is a linear procedure in the RKHS.

3 Implementation

There have been various implementation algorithms for KCCA (e.g., Gretton, Herbrich and Smola, 2003; Hardoon, Szedmak and Shawe-Taylor, 2004; Kuss and Graepel, 2003; Fukumizu, Bach and Jordan, 2004; Gretton, Herbrich, Smola and Schölkopf, 2005). Our aim here is quite different from that of computer scientists. In Section 2.2 we have interpreted the working procedure of KCCA as

an LCCA functioning on the kernel augmented data $[\gamma_j^{(1)}; \gamma_j^{(2)}]$, $j = 1, \ldots, n$. The implementation will thus be different from computer scientists', too. The procedure for KCCA is already stated in Section 2.1. After bridging the kernel method and the linear procedure, the main purpose of this section is to lessen the programming burden by using code already available in common statistical/mathematical packages. The LCCA has long been a standard statistical analysis tool for measures of association, independence test, etc., and it has been implemented in various mathematical and statistical software (e.g., *canoncorr* in Matlab; *cancor* in R; *cancor* in Splus; and *proc cancorr* in SAS). These packages are ready for use for kernel-based nonlinear canonical analysis. One advantage for using a conventional LCCA code from any available package is to allow statisticians unfamiliar with kernel machines can still have an easy access to implementing KCCA. *The extra efforts required are to prepare the data in an appropriate kernel augmentation form.* In the following, we discuss two more steps, the regularization step and parameters selection step, in kernel data preparation before feeding them into the LCCA procedure. They are adopted to avoid computational instability and to enhance the computational speed and efficiency in the later LCCA procedure.

3.1 Regularization

The optimization problem to be solved is:

$$\max_{[\alpha;\beta]\in\Re^{2n}} \alpha'\Lambda_{12}\beta \quad \text{subject to} \quad \alpha'\Lambda_{11}\alpha = 1 \ \text{ and } \ \beta'\Lambda_{22}\beta = 1, \tag{7}$$

where Λ_{ik} is the sample covariance of $\{\gamma_j^{(i)}\}_{j=1}^n$ and $\{\gamma_j^{(k)}\}_{j=1}^n$. Note that often both Λ_{11} and Λ_{22} are singular or near singular, especially when $m_1 + m_2$ is large, which causes computational difficulty. The optimization problem (7) is then ill-conditioned and some sort of regularization is needed in preparing the kernel data matrices $\mathbb{K}_1$ and $\mathbb{K}_2$. There are three commonly used regularization methods for coping with ill-conditioned eigen-components extraction.

- Ridge-type regularization,
- Principal components approach,
- Reduced set (reduced bases) approach.

The ridge-type regularization adds a small quantity to the diagonals,

$$\alpha'(\Lambda_{11} + \lambda_1 I)\alpha = 1 \ \text{ and } \ \beta'(\Lambda_{22} + \lambda_2 I)\beta = 1,$$

to stabilize the numerical computation for solving problem (7).

The principal components approach. Let $\mathbb{K}_1$ and $\mathbb{K}_2$ be the full-data kernel matrices in (5). The PCA approach extracts leading eigen-vectors from $\mathbb{K}_1$ and $\mathbb{K}_2$, denoted by $\mathbb{U}_1$ and $\mathbb{U}_2$, respectively. Next, it projects columns of $\mathbb{K}_1$ and $\mathbb{K}_2$ onto the column spaces of $\mathbb{U}_1$ and $\mathbb{U}_2$ and obtains the reduced-rank approximation of kernel data

$$\tilde{\mathbb{K}}_1 = \mathbb{K}_1'\mathbb{U}_1 \ \text{ and } \ \tilde{\mathbb{K}}_2 = \mathbb{K}_2'\mathbb{U}_2. \tag{8}$$

The LCCA is then functioning on the reduced kernel $\tilde{\mathbb{K}} = [\tilde{\mathbb{K}}_1 \ \tilde{\mathbb{K}}_2]$ instead of the full-data kernel $\mathbb{K}$. The PCA approach corresponds to using the following linear independent systems for discretization:

$$\phi_\nu(x^{(1)}) = \sum_{j=1}^{n} \kappa_1(x^{(1)}, x_j^{(1)}) u_{j\nu}^{(1)}, \ \nu = 1, \ldots, m_1,$$

and

$$\psi_\mu(x^{(2)}) = \sum_{j=1}^{n} \kappa_2(x^{(2)}, x_j^{(2)}) u_{j\mu}^{(2)},\ \mu = 1, \ldots, m_2,$$

where $[u_{j\nu}^{(1)}] = \mathbb{U}_1$ and $[u_{j\mu}^{(2)}] = \mathbb{U}_2$. The PCA approach seeks to provide optimal linear independent systems for discretization.

The reduced set approach is also a type of reduced-rank approximation to the full-data kernel matrix. Let $A_1 = \{x_1^{(1)}, \ldots, x_n^{(1)}\}$ and $A_2 = \{x_1^{(2)}, \ldots, x_n^{(2)}\}$ denote the full sets of data on $\mathcal{X}_1$ and $\mathcal{X}_2$. The reduced set method selects a small portion of data $\tilde{A}_1$ and $\tilde{A}_2$ from the full sets to form reduced-rank kernel matrices

$$\tilde{\mathbb{K}}_1 = \mathbb{K}_1(A_1, \tilde{A}_1) \ \text{ and } \ \tilde{\mathbb{K}}_2 = \mathbb{K}_2(A_2, \tilde{A}_2), \tag{9}$$

where $\mathbb{K}_1(A_1, \tilde{A}_1) = [\kappa_1(x_i^{(1)}, x_j^{(1)})]_{x_i^{(1)} \in A_1, x_j^{(1)} \in \tilde{A}_1}$ is a thin column matrix of size $n \times m_1$ with m_1 the size of $\tilde{A}_1$, and similarly for $\tilde{\mathbb{K}}_2(A_2, \tilde{A}_2)$. The corresponding choices of linear independent systems are:

$$\phi_\nu(x^{(1)}) = \kappa_1(x^{(1)}, x_\nu^{(1)}),\ x_\nu^{(1)} \in \tilde{A}_1, \ \text{ and } \ \psi_\mu(x^{(2)}) = \kappa_2(x^{(2)}, x_\mu^{(2)}),\ x_\mu^{(2)} \in \tilde{A}_2.$$

The choice of subsets $\tilde{A}_1$ and $\tilde{A}_2$ is often made by uniform random sampling from A_1 and A_2, respectively. The approach is termed the random subset and is the most economic one among the three regularization methods in terms of computational complexity, and it leads to sparse representation with only a small number of kernel functions in the underlying model and involved in computation as well. This approach can effectively speed up the computation and cut down underlying model complexity (Lee and Mangasarian, 2001; Williams and Seeger, 2001; Lee and Huang, 2006). However, as the subset selection is purely random, it is recommended for median to large data sets.

Of course, one can always do better than pure randomness in subset selection. Snelson and Ghahramani (2005) have proposed a nice way of training the reduced set selection. For small reduced set size, their trained selection can perform much better than purely random selection. However, as they have indicated, when the reduced set size increases both their method and the purely random selection perform alike, and their method is relatively expensive to train. Another interesting article on training the reduced set selection is by Wang, Neskovic and Cooper (2005). They have proposed two different training selection methods. One is based on a statistical confidence measure and the other uses Hausdorff distance as criteria to select patterns near the decision boundary. Their experimental results show that a significant amount of kernel bases can be removed and that the random sampling performs consistently well despite its simplicity. Other sequential adaptive algorithm for bases selection is also available (Smola and Schölkopf, 2000). All these methods are more time-consuming and require some search algorithms. The uniform random subset is a simple and economic alternative.

Both the ridge-type regularization and the principal components approach are well suited for small to median sized problems and provide better approximation than the random subset approach. As the data size n becomes extremely large, the full-data kernel matrices $\mathbb{K}_1$ and $\mathbb{K}_2$ are both of size $n \times n$, and there are problems confronted in massive data computing and training, including the memory requirement for storing the kernel data, the size of the mathematical programming problem and problems of prediction instability. In the large data case, the random subset approach is recommended. The random subset approach is also reasonably well for median size problems as well.

3.2 Parameters selection

There are three parameters involved in the entire procedure of KCCA. One is the choice of kernel, the second is the kernel window width and the last is the regularization parameter (the ridge parameter,

the number of principal components, and the reduced set size). Throughout this article we use Gaussian pdf as our choice of kernel and the window width is set to $\sqrt{10S}$ coordinatewise, where S is the one-dimensional sample variance. Such a choice is based on empirical experience, which results in good normality checks on kernel data (Huang and Hwang, 2005). The window width $\sqrt{10S}$ is a universal rule of thumb suggestion. Though it might not be optimal or even far away from an optimal window width, it gives robust satisfactory results. As for choices of random subset size and the number of principal components, we suggest users start with an adequate proportion of kernel bases and then cut down the effective dimensionality by the PCA approach depending on the availability and performance of computer facility. See individual examples in next section for further discussion. As we do not use the ridge type regularization here, we do not make any suggestion for it, but one can always resort to cross-validation, grid search, or simply by hand-pick. We refer the reader, who is interested in more advanced parameters selection scheme, to Huang, Lee, Lin and Huang (2006) and Lee, Lo and Huang (2003).

4 Statistical applications

4.1 Measures of association

The KCCA leads to the following triple: a sequence of correlation coefficients $\{\rho_1, \rho_2, \ldots\}$ arranging in decreasing order, a sequence of canonical vectors $\{[\alpha_1; \beta_1], [\alpha_2; \beta_2], \ldots\}$ and a sequence of nonlinear canonical variates $\{(f_1(X^{(1)}, g_1(X^{(2)})), (f_2(X^{(1)}, g_2(X^{(2)})), \ldots\}$. Several measures of association in the literature are constructed as functions of the correlation coefficients derived from either LCCA (see, e.g., Jensen and Mayer, 1977) or NLCCA (Dauxois and Nkiet, 1998). Here we will adopt two particular and commonly seen association measures, denoted by r. One is the maximal correlation

$$r(X^{(1)}, X^{(2)}) := \rho_1, \tag{10}$$

and the other is

$$r(X^{(1)}, X^{(2)}) := -\sum_{\nu=1}^{\infty} \log(1 - \rho_\nu^2). \tag{11}$$

Note that Bach and Jordan (2002), Gretton *et al.* (2005) and also other above-mentioned works from the machine learning community (see Introduction) all adopt (10) for association measure. However, there are many other types of association measures, as functions of correlation coefficients, discussed in statistical literature.

Below we give some examples of the usage of association measures. In Example 1 we illustrate the nonlinear association measure using (10) for two univariate random variables and compare it with the LCCA-based association measure and the rank-based Kendall's τ and Spearman's ρ. In Example 2 we demonstrate the nonlinear association between two sets of multivariate data using the first two leading correlation coefficients ρ_1 and ρ_2, and we also plot the data scatters over the first two canonical variates. In Example 3 we use the KCCA-found canonical variates for dimension reduction and use them as discriminant variates for classification problem with UCI Pendigits data set. In next subsection, we use sample estimate of (11) as the testing statistic for stochastic independence, see Example 4.

Example 1 *Synthetic data set: Association for two random variables.*
In this example, we use the first kernel canonical correlation coefficient (10) as our association measure and compare it with the linear correlation coefficient and two well-known nonparametric rank-based association measures, Kendall's τ and Spearman's ρ, on their ability of assessing nonlinear relation. In the following three cases, let $X^{(1)}$ and ϵ be two independent and identically distributed standard normals, and the models of $X^{(2)}$ considered are

I. $X^{(2)} = f(X^{(1)}) + \frac{\epsilon}{k}$ with $f(x) = \cos(\pi x)$;

II. $X^{(2)} = f(X^{(1)}) + \frac{\epsilon}{k}$ with $f(x) = \exp(-|x|)$; and

III. $X^{(2)} = X^{(1)} + \frac{\epsilon}{k}$;

where k is a constant in (0, 10]. In these models, the amount of association increases with respect to k.

In each setting, we generate 500 samples of $X^{(1)}$ and ϵ, and take account of relationship between $X^{(1)}$ and $X^{(2)}$. Kernel data using full bases are prepared and then the dimensionality is cut down by the PCA extracting 99% of data variation. For models I and II, $X^{(1)}$ and $X^{(2)}$ are nonlinearly correlated. Five quantities of association measures are considered and the corresponding curves are plotted in Figure 1(a) and 1(b). The solid curve in each plot indicates the Pearson's correlation of $X^{(2)}$ and $f(X^{(1)})$ taken as the true target association measure between $X^{(1)}$ and $X^{(2)}$. Because the models $\cos(\pi x)$ and $\exp(-|x|)$ are both symmetric about zero, the classical correlations, which can only catch linear association, for models I and II are around zero. The rank-based Kendall's τ and Spearman's ρ are also around zero. The results indicate that the linear correlation coefficients as well as the rank-based τ and ρ cannot catch the nonlinear association designed for this example. Overall speaking, kernel canonical correlation outperforms the rest. It approximates the target association well. For the linearly correlated case (model III in Figure 1(c)), the solid curve still represents the true association. The measure of Spearman's ρ is very close to the target but Kendall's τ is a bit low from the true curve. Again, kernel canonical correlation follows the true curve closely.

——— Put Figure 1 here ———

Example 2 *Synthetic data set: Association for two sets of variables.*
Let X_1 and X_2 be independent and identically distributed random variables having uniform distribution over the interval (-2,2). Let

$$Y_1 = X_1^2 + 0.1\epsilon_1 \quad \text{and} \quad Y_2 = \cos(\pi X_2) + 0.1\epsilon_2, \tag{12}$$

where ϵ's are standard normal random noises. Let $X^{(1)} = [X_1; X_2]$ and $X^{(2)} = [Y_1; Y_2]$. In each simulation, we sample 1000 pairs of $[X^{(1)}; X^{(2)}]$ from the described model and calculate two leading pairs of sample canonical correlation coefficients and variates using both the linear and kernel CCA. In contrast to the PCA approach in Example 1, here we use the uniform random subset of size 200 as our choice of regularization. Of course, one can use the PCA approach with the full set kernel bases of size 1000. The PCA with full set bases is better accurate but computationally more intensive. For demonstration purpose, we choose the computationally-economic random subset approach.

The mutual correlation carried in the two leading pairs of canonical variates found by LCCA and KCCA are listed in Table 1 for 30 replicate runs. Reported are the average correlation coefficients and their standard errors. It is evident that kernel canonical variates capture more association between two groups of data. We randomly choose a copy of the 30 replicate runs and plot these 1000 data points along the two leading pairs of sample canonical variates found by LCCA and KCCA in Figure 2. Figure 2(a) is the data scatter of the first pair of linear canonical variates and Figure 2(b) is the data scatter for the second pair of linear canonical variates. There is indication of some strong association left unexplained. Figure 2(c) and Figure 2(d) are data scatters of the first and the second pair of kernel canonical variates, respectively. They show strong correlation within each pair of kernel canonical variates. The correlations are 0.993 and 0.965, respectively. After the kernel transformation, the distribution of pairs of canonical variates is indeed well depicted as elliptically symmetrically distributed. This phenomenon implies the applicability of many multivariate data analysis tools, which originally are constrained by Gaussian assumption, now can extend to work on kernel data. Note that if the noise level in (12) is larger, then the ellipses in Figures 2(c) and 2(d) will be thicker and of rounder shape. But the basic elliptical scatters will remain.

——— Put Table 1 and Figure 2 here ———

Example 3 *Dimension reduction for nonlinear discriminant using pen-based recognition of handwritten digits.*
In this example we utilize the kernel canonical variates as discriminant variates for multiple classification. Take, for instance, a k-classes problem, we first construct a k-dimensional class indicator variable "$x^{(2)} = (c_1, \ldots, c_k)$" as follows:

$$c_i = \begin{cases} 1, & \text{if an instance with input } x^{(1)} \text{ is from the } i\text{th class,} \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

The pendigits data set is taken from the UCI machine learning data bases. The number of training instances is 7494 and the number of testing instances is 3498. For each instance there are 16 input measurements (i.e., $x^{(1)}$ is 16-dimensional) and a corresponding class label from $\{0, 1, 2, \ldots, 9\}$. A Gaussian kernel with the covariance matrix $\text{diag}(10S_1, \ldots, 10S_{16})$ is used, and a random subset of training inputs of size 300 (stratified over 10 classes with 30 instances each) is adopted to produce the kernel data for input measurements. We keep the class label variable $x^{(2)}$ as in (13) without subject to kernel transformation. Then, the LCCA is acting on kernel training inputs and their associated class labels, and it leads to $(k-1)$ canonical variates as discriminant directions. The KCCA in this example first extends the data to nonlinear high inputs. The kernel trick augments the original input measurements in $\Re^{16}$ by high input measurements in $\Re^{300}$ using 300 kernel bases. Then it cuts down the dimensionality from 300 to 9 for discriminant purpose. Scatter plots of data along kernel canonical variates and along linear canonical variates are given below. To avoid excessive ink we only sample 20 test instances per digit to produce the data scatter plots. Different classes are labeled with distinct symbols. Figure 3 are data scatters along kernel canonical variates. Digits 0, 4, 6, and 8 can be easily separated when plotting the first versus the second kernel canonical variates, and thus they are removed from the next plot. Digits 2 and 3 are identified in the second versus the third kernel canonical variates, digits 5 and 9 are in 3rd-and-4th, and digits 1 and 7 are identified in the 4th-and-5th variates. In fact, the first three kernel canonical variates can pretty much classify a majority of the ten digits already; while in Figure 4, it is till difficult to discriminate the ten digits even with 5 leading linear canonical variates. The 9-dimensional subspace spanned by kernel canonical variates based on training inputs is used as the designated discriminant subspace. Test instances are projected onto this subspace and classification is made by Fisher's linear discriminant analysis in this KCCA-found discriminant subspace, i.e., a test instance is assigned to the class with the closest class center (according to Mahalanobis distance). The accuracy of the kernel classification on test data can achieve the rate 97.24% (with standard error 0.056% over 10 replicate samplings of random subsets in the regularization step).

——— Put Figures 3 & 4 here ———

4.2 Test of independence between two sets of variables

In this section we generalize the use of Bartlett's test of independence (Bartlett, 1947a; 1947b) to kernel augmented data. The theoretical foundation of this generalization is based on Dauxois and Nkiet (1998), while the computational algorithm is based on the formulation of kernel machine. The idea is to approximate the KCCA by a suitable LCCA on augmented data through certain linear independent systems of functions, and then this approximation by LCCA is estimated based on empirical data. The following statistic is used for testing independence. Bartlett's modification is adopted.

$$\left(n - \frac{1}{2}(m_1 + m_2 + 1)\right) r_n \sim \chi_{(m_1-1)(m_2-1)},$$

where r_n is the sample estimate of (11) based on reduced kernel data $\tilde{\mathbb{K}}_1$ and $\tilde{\mathbb{K}}_2$ by PCA reduction (8), and m_1 and m_2 are the column ranks of $\tilde{\mathbb{K}}_1$ and $\tilde{\mathbb{K}}_2$, respectively. We leave the technical details to the Appendix. This test can be easily done with standard statistical package. Here we use the Matlab m-file *canoncorr* for Example 4. It has the implementation of Bartlett's independent test. Tests of two sets of variables for several distributions are carried out on reduced-rank kernel data. The reduction is done by PCA extracting 99% of data variation. The first 5 distributions in this study are taken from Dauxois and Nkiet (1998) and the last one is similar to Example 2 in Section 4.1 without the additive noise.

Example 4 *Independence tests.*
In the first five cases, we are interested in testing the independence between X and Y that are both one-dimensional. In the last case VI, the independence between two vectors X and Y are tested, where both X and Y are two-dimensional.

I. $X \sim N(0,1)$ and $Y = X^2$;

II. $[X;Y] \sim$ uniform distribution on the unit disk;

III. $[X;Y] \sim$ bivariate standard normal with correlation ρ;

IV. $[X;Y] \sim$ a mixture distribution: $P_\theta = \theta Q_1 + (1-\theta)Q_2$ with $\theta = 0.5$, where Q_1 is the distribution of $[X; X^2]$ with X a univariate standard normal and Q_2 is the bivariate standard normal with correlation $\rho = 0.25$;

V. a mixture distribution: $P_\theta = \theta Q_1 + (1-\theta)Q_2$ with $\theta = 0.75$;

VI. $X = [X_1; X_2]$ has a bivariate uniform distribution over $(-2,2)\times(-2,2)$ and $Y = [X_1^2; \cos(\pi X_2)]$.

Five hundreds sample points are drawn from each of the above distributions. The total number of replicate runs is 100. In each run, the kernel transformation is first performed on X and Y separately, and then the PCA is conducted to extract leading variates accounting for 99% of variation. The power estimates (or type-I error estimate for case III-1, $\rho = 0$) and their standard errors for independence test are reported in Table 2. It is seen that the KCCA-based test has power of 1 in all cases except for case III-2 of $\rho = 0.2$; while the LCCA catches the relation only when linearity is present (cases III-2,3,4). The pattern remains so even when the relation between X and Y are complex (cases IV and V). For the type I error (case III-1, $\rho = 0$), the KCCA-based test has a bit underestimated the typ-I error, while and LCCA-based test has a bit overestimated the type-I error, but the difference is not prominent. In general, the KCCA performs about the same as the order 2 and order 3 spline approximations, but much better than the LCCA-based tests. The kernel approximation provides an alternative choice besides splines. However, our computation is a lot easier than the spline counterpart.

——— Put Table 2 here ———

5 Conclusion

In this paper, we discuss the use of kernel method for studying relation of two sets of variables via canonical analysis. We describe the procedure of KCCA and discuss its implementation technique. There are several points worth mentioning. We link the NLCCA by Dauxois *et al.* to the currently popular KCCA emerging from the machine learning community. The linkage provides NLCCA an easy-to-compute implementation through kernel trick. We also demonstrate a couple of statistical applications.

The LCCA provides a way for association study by extracting leading canonical variates. The KCCA can be used for the same purpose in a similar manner, while the association study is in the kernel associated RKHS instead of an Euclidean space. The LCCA can be used for dimension reduction, too. It finds linear dimension reduction subspaces in $\mathcal{X}_1$ and $\mathcal{X}_2$, again, by extracting leading canonical variates; while the KCCA first maps the data into a very high dimensional space via a nonlinear transformation and then finds dimension reduction subspaces therein. In this way the KCCA allows nonlinear dimension reduction directions, which are actually functional directions. The KCCA can be used for nonlinear discriminant as well. Only the explanatory variables are subject to kernel augmentation and the class labels remain unchanged. In the pen digit data the KCCA is used to select the leading discriminant variates and classification is done in the subspace spanned by these functional discriminant variates. A test instance is assigned to the nearest group center according to Mahalanobis distance. Traditional tests of independence apply mostly the statistics that take into account only linear relationship and rely on the Gaussian assumption. With the KCCA-based independence test, the distributional assumption can be avoided. As seen in the simulations, it outperforms the LCCA-based test and is compatible with the nonparametric approach using order 2 or 3 splines. Note that, due to the reproducing property of the kernel associated Hilbert space, the inner product of two kernel functions can be obtained by kernel value, i.e., $\langle \kappa(s,\cdot), \kappa(t,\cdot)\rangle = \kappa(s,t)$, which makes the computation for KCCA much less demanding.

6 Appendix: nonlinear canonical correlation analysis

Consider a probability space $(\mathcal{X}, \mathcal{B}, P)$ such that the Hilbert space $L_2(P)$ is separable.[2] Let $X = [X^{(1)}; X^{(2)}]$ be a random vector on $(\mathcal{X}, \mathcal{B}, P)$ with marginals denoted by P_1 and P_2, respectively. We also assume that $\mathcal{X}$ is compact. In this appendix we will give a short review for the theory of nonlinear canonical correlation analysis (NLCCA) studied by Dauxois and co-authors. The NLCCA of $X^{(1)}$ and $X^{(2)}$ can be interpreted in terms of canonical angles between two Hilbert subspaces $L_2(P_1)$ and $L_2(P_2)$ in $L_2(P)$, where canonical angles can be obtained by spectral analysis of certain compact operators. We will also review the approximation of NLCCA by a sequence of suitable LCCAs, as well as the estimation problem of NLCCA and its asymptotic distribution. See mainly Dauxois and Nkiet, (1998), also consult Dauxois and Nkiet (2002), Dauxois, Romain and Viguier (1993), Dauxois and Nkiet (1997), Dauxois, Nkiet and Romain (2004). Their works have laid the theoretical foundation for KCCA, as the KCCA can be regarded as a special case of the NLCCA using kernel bases for approximation.

6.1 NLCCA and measures of association

Let $L_2'(P_1) = \{f \in L_2(P_1) : Ef(X^{(1)}) = 0\}$ and $L_2'(P_2) = \{g \in L_2(P_2) : Eg(X^{(2)}) = 0\}$ denote the centered L_2 subspaces of marginals. The NLCCA is the search of two variates $f_1(X^{(1)})$ ($f_1 \in L_2'(P_1)$) and $g_1(X^{(2)})$ ($g_1 \in L_2'(P_2)$) such that the pair (f_1, g_1) maximizes $\langle f, g\rangle_{L_2(P)}$, where $f \in L_2'(P_1)$ and $g \in L_2'(P_2)$, under the constraints $\|f\|_{L_2(P_1)} = 1$ and $\|g\|_{L_2(P_2)} = 1$.[3] For $\nu \geq 2$, one searches for two variates $f_\nu \in L_2'(P_1)$ and $g_\nu \in L_2'(P_2)$ that maximizes the above-mentioned constrained optimization problem with the additional iterations of orthonormality constraints $\langle f_k, f_\nu\rangle_{L_2(P_1)} = 0$ and $\langle g_k, g_\nu\rangle_{L_2(P_2)} = 0$ for all $k = 1, \ldots, \nu - 1$. Denote $\rho_\nu = \langle f_\nu(X^{(1)}), g_\nu(X^{(2)})\rangle_{L_2(P)}$. Trivial canonical terms $\rho_0 = 1$ and $f_0(X^{(1)}) = g_0(X^{(2)}) = 1$ can be added to the canonical analysis for completeness. However, they do not give any information about independence of X and Y. Define

[2] That is, it has a countable dense subset. In a separable Hilbert space countable orthonormal systems are used to expand any element as an infinite sum.

[3] Note that, for $f \in L_2'(P_1)$ and $g \in L_2'(P_2)$, we have $\langle f, g\rangle_{L_2(P)} = \langle f, g\rangle_{L_2'(P)}$, $\|f\|_{L_2(P_1)} = \|f\|_{L_2'(P_1)}$ and $\|g\|_{L_2(P_2)} = \|g\|_{L_2'(P_2)}$.

$\theta_\nu = \arccos(\rho_\nu)$, $0 \le \theta_\nu \le \pi/2$, $\nu = 0, 1, 2, \ldots$, named canonical angles (arranged in increasing order) between the two Hilbert subspaces $L_2(P_1)$ and $L_2(P_2)$. The canonical angles can be obtained by spectral analysis (solving for eigenvalues and associated eigenvectors) of one of the following self-adjoint operators:

$$\Pi_1\Pi_2\Pi_1,\ \Pi_2\Pi_1\Pi_2,$$

where Π_i is the orthogonal projection from $L_2(P)$ onto $L_2(P_i)$. Take $T = \Pi_1\Pi_2\Pi_1$ for instance. Assume T is compact. The spectrum of T can be obtained as

$$T : L_2(P) \mapsto L_2(P_1) \text{ is of the form } T = \sum_{\nu=0}^{\infty} \rho_\nu^2 f_\nu \otimes f_\nu,\ f_\nu \in L_2(P_1) \text{ and } \{f_\nu\} \text{ are orthonormal,}$$

where $(f_\nu \otimes f_\nu)(h) = \langle f_\nu, h\rangle_{L_2(P)} f_\nu$, $\forall h \in L_2(P)$. Let $g_\nu = \rho_\nu^{-1}\Pi_2 f_\nu \in L_2(P_2)$. Then

$$\Pi_2\Pi_1\Pi_2 g_\nu = \rho_\nu^{-1}\Pi_2\Pi_1\Pi_2 f_\nu = \rho_\nu^{-1}\Pi_2\Pi_1\Pi_2\Pi_1 f_\nu = \rho_\nu^{-1}\Pi_2(\rho_\nu^2 f_\nu) = \rho_\nu^2 g_\nu,$$

and

$$\langle g_\nu, g_\mu\rangle_{L_2(P)} = (\rho_\nu\rho_\mu)^{-1}\langle\ \Pi_2\Pi_1 f_\nu, \Pi_2\Pi_1 f_\mu\ \rangle_{L_2(P)} = (\rho_\nu\rho_\mu)^{-1}\langle\ f_\nu, \Pi_1\Pi_2^2\Pi_1 f_\mu\ \rangle_{L_2(P)} = \delta_{\nu\mu}.$$

Therefore, we have

$$\Pi_2\Pi_1\Pi_2 = \sum_{\nu=0}^{\infty} \rho_\nu^2 g_\nu \otimes g_\nu,\quad g_\nu \in L_2(P_2) \text{ and } \{g_\nu\} \text{ are orthonormal.}$$

It is also easy to see that canonical variates have the property:

$$\langle f_\nu, g_\mu\rangle_{L_2(P)} = \langle \Pi_1 f_\nu, \rho_\mu^{-1}\Pi_2 f_\mu\rangle_{L_2(P)} = \langle f_\nu, \rho_\mu^{-1}\Pi_1\Pi_2\Pi_1 f_\mu\rangle_{L_2(P)} = \rho_\nu\delta_{\nu\mu},\quad \nu, \mu = 0, 1, 2, \ldots. \quad (14)$$

The NLCCA consists of a triple:

$$\{(\rho_\nu, f_\nu(X^{(1)}), g_\nu(X^{(2)})), \nu = 0, 1, 2, \ldots\}.$$

The numbers $0 \le \rho_\nu \le 1$ are arranged in decreasing order and are termed nonlinear canonical correlation coefficients. The trivial canonical terms can be easily avoided by restricting the NLCCA to centered variables.

Based on the sequence of canonical correlation coefficients $\{\rho_\nu\}_{\nu=1}^{\infty}$, various measures of association are constructed as functions of $\{\rho_\nu\}_{\nu=1}^{\infty}$, denoted by $Assoc(\rho_1, \rho_2, \ldots)$. Note that these ρ_ν's implicitly depend on $X^{(1)}$ and $X^{(2)}$. The association measure (10) in Section 4 takes the following form

$$r(X^{(1)}, X^{(2)}) = Assoc(\rho_1, \rho_2, \ldots) = \max\{\rho_1, \rho_2, \ldots\}, \quad (15)$$

while the association measure (11) takes the form:

$$r(X^{(1)}, X^{(2)}) = Assoc(\rho_1, \rho_2, \ldots) = -\sum_{\nu=1}^{\infty} \log(1 - \rho_\nu^2). \quad (16)$$

Proposition 1 (Proposition 3.1-(2) of Dauxois and Nkiet, 1998.) *Let $r(X^{(1)}, X^{(2)})$ be defined as either (15) or (16). We have $r(X^{(1)}, X^{(2)}) = 0$ if and only if $X^{(1)}$ and $X^{(2)}$ are stochastically independent.*

6.2 Approximation to NLCCA by kernel bases

To put the NLCCA in practice and for implementation by computer, one has to work on discretization of NLCCA. Such a discretization is made by approximating the NLCCA by an essentially increasing sequence of finite linear independent systems for $L_2(P_1)$ and $L_2(P_2)$, respectively. For $k \in \mathbb{N}$, let $\{\phi_\nu^k\}_{1\le\nu\le p_k}$ be a linearly independent system in $L_2(P_1)$ and $\{\psi_\mu^k\}_{1\le\mu\le q_k}$ be a linearly independent system in $L_2(P_2)$. Let $\mathcal{V}_1^k$ and $\mathcal{V}_2^k$ be, respectively, the subspaces spanned by the systems $\{\phi_\nu^k\}_{1\le\nu\le p_k}$ and $\{\psi_\mu^k\}_{1\le\mu\le q_k}$. Assume that these sequences of subspaces are essentially increasing and dense in $L_2(P_1)$ and $L_2(P_2)$, respectively, in the following sense:

$$\limsup_k \mathcal{V}_i^k := \lim_{K\to\infty} \cup_{k\ge K} \mathcal{V}_i^k \text{ is dense in } L_2(P_i),\ i = 1, 2. \tag{17}$$

Proposition 2 *Suppose the linear independent systems satisfy the essential density condition (17). Without loss of generality, take $T = \Pi_1\Pi_2\Pi_1$ (or resp. $T = \Pi_2\Pi_1\Pi_2$). Then the approximation operator $T_k = \Pi_{\mathcal{V}_1^k}\Pi_{\mathcal{V}_2^k}\Pi_{\mathcal{V}_1^k}$ (or resp. $T_k = \Pi_{\mathcal{V}_2^k}\Pi_{\mathcal{V}_1^k}\Pi_{\mathcal{V}_2^k}$) converges uniformly to T as $k \to \infty$. (Here the uniform convergence is referring to sup-norm.)*

Proof: Since $\mathcal{V}_1^k$ and $\mathcal{V}_2^k$ satisfy condition (17), we have $\|\Pi_1 - \Pi_{\mathcal{V}_1^k}\|_\infty$ and $\|\Pi_2 - \Pi_{\mathcal{V}_2^k}\|_\infty$ converge to zero as $k \to \infty$. It is then straightforward to get that $\|T_k - T\|_\infty \to 0$. □

The NLCCA can be approximated (in terms of uniform convergence of a certain underlying sequence of linear operators proved in Proposition 2) by the LCCA of random vectors

$$\Phi^k(X^{(1)}) := [\phi_1^k(X^{(1)}); \ldots; \phi_{p_k}^k(X^{(1)})] \tag{18}$$

and

$$\Psi^k(X^{(2)}) := [\psi_1^k(X^{(2)}); \ldots; \psi_{q_k}^k(X^{(2)})]. \tag{19}$$

That is to say, the NLCCA of $X^{(1)}$ and $X^{(2)}$ can be approximated by a sequence of suitable LCCAs of Φ^k and Ψ^k. The underlying sequence of LCCA approximations depends on the choice of linear independent systems $\{\phi_\nu^k\}_{1\le\nu\le p_k}$ and $\{\psi_\mu^k\}_{1\le\mu\le q_k}$. Choices of the systems $\{\phi_\nu^k\}_{1\le\nu\le p_k}$ and $\{\psi_\mu^k\}_{1\le\mu\le q_k}$ can be step functions, or B-splines (Dauxois and Nkiet, 1998). In this article we use kernel functions as our choice of linear independent systems. The KCCA takes some special ϕ- and ψ-functions:

$$\phi_\nu^k(x^{(1)}) = \kappa_{1,\sigma_1}(x^{(1)}, z_\nu^{(1)}),\ z_\nu^{(1)} \in \mathcal{X}_1,\ \nu = 1, \ldots, p_k, \tag{20}$$

$$\psi_\mu^k(x^{(2)}) = \kappa_{2,\sigma_2}(x^{(2)}, z_\mu^{(2)}),\ z_\mu^{(2)} \in \mathcal{X}_2,\ \mu = 1, \ldots, q_k, \tag{21}$$

where σ_1 and σ_2 are kernel window widths and $\{z_\nu^{(1)}\}_{\nu=1}^{p_k}$ and $\{z_\mu^{(2)}\}_{\mu=1}^{q_k}$ are often (but not necessarily) taken to be subsets of full data, described as the reduced set approach in the Implementation Section.

Proposition 3 *Assume that κ_1 and κ_2 are continuous and translation type kernels. Also assume (1) the sequences $\{z_\nu^{(1)}\}_{\nu=1}^{p_k}$ and $\{z_\mu^{(2)}\}_{\mu=1}^{q_k}$ are dense in $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively; and (2) $\sigma_1, \sigma_2 \to 0$, as $k \to \infty$. Then $\mathcal{V}_1^k$ and $\mathcal{V}_2^k$ are essentially dense in $L_2(P_1)$ and $L_2(P_2)$, respectively.*

Proposition 3 can be easily obtained by Lemma 1 below.

Lemma 1 *Let ω be a positive continuous function in a compact set $\Omega \subset \Re^q$ such that $\int_\Omega \omega(t)dt = 1$. Let $\omega_h(t) = h^{-q}\omega(t/h)$ and let ω_h* denote the convolution operator in $L_2(\Omega)$. Then $\omega_h * f$ converges to f in $L_2(\Omega)$. (See, e.g., §6.6, Aubin, 1979).*

6.3 Estimates and asymptotic distribution

Let $\{[x_j^{(1)}; x_j^{(2)}]\}_{j=1}^n$ be iid sample having the same distribution as $[X^{(1)}; X^{(2)}]$. Assume, without loss of generality, that $p_k \leq q_k$. Let $\rho_\nu^{k,n}$ and $(f_\nu^{k,n}, g_\nu^{k,n})$, $\nu = 1, \ldots, p_k$, be the sample correlation coefficients and pairs of canonical variates, respectively, based on data $\{\Phi^k(x_1^{(1)}), \ldots, \Phi^k(x_n^{(1)})\}$ and $\{\Psi^k(x_1^{(2)}), \ldots, \Psi^k(x_n^{(2)})\}$. Let $r_{k,n}$ be the sample estimate of (11) based on approximation by the linear independent systems $\{\phi_\nu^k\}_{1 \leq \nu \leq p_k}$ and $\{\psi_\mu^k\}_{1 \leq \mu \leq q_k}$. Also let

$$T_{k,n} = \sum_{\nu=1}^{p_k} (\rho_\nu^{k,n})^2 f_\nu^{k,n} \otimes f_\nu^{k,n}. \tag{22}$$

Proposition 4 (Dauxois and Nkiet, 1998) *Let $T_{k,n}$ be the sample estimate of T_k based on iid sample $\{[x_j^{(1)}; x_j^{(2)}]\}_{j=1}^n$ as given in (22). Then $T_{k,n}$ converges a.s. to T_k, as $n \to \infty$.*

Proposition 4 can be obtained by the law of large numbers.

Proposition 5 (Dauxois and Nkiet, 1998) *Suppose the systems are chosen to be essentially increasing sequences and are eventually dense in $L_2(P_1)$ and $L_2(P_2)$, respectively. Then, under the null hypothesis that $X^{(1)}$ and $X^{(2)}$ are independent, $nr_{k,n}$ converges in distribution, as $n \to \infty$, to $\chi^2_{(p_k-1)(q_k-1)}$.*

Dauxois and Nkiet's Proposition is stated under the conditions that the sequence of pairs of independent systems $\{\phi_\nu^k\}_{1 \leq \nu \leq p_k}$ and $\{\psi_\mu^k\}_{1 \leq \mu \leq q_k}$ are increasing and their unions are dense in $L_2(P_1)$ and $L_2(P_2)$, respectively. Proposition 5 is instead under the essential density condition (17). However, Dauxois and Nkiet's arguments still follow.

Acknowledgment

The authors thank Hwang, Chii-Ruey for helpful discussion.

References

Akaho, S. (2001). A kernel method for canonical correlation analysis. *International Meeting of Psychometric Society (IMPS2001).*

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68, 337–404.

Aubin, J.P. (1979). *Applied Functional Analysis.* John Wiley & Sons, New York.

Bach, F.R. and Jordan, M.I. (2002). Kernel independent component analysis. *J. Mach. Learning Res.*, 3, 1–48.

Bartlett, M.S. (1947a). Multivariate analysis. *Supp. J. Roy. Statist. Soc.*, 9, 176–197.

Bartlett, M.S. (1947b). The general canonical correlation distribution. *Ann. Math. Statist.*, 18, 1–17.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Kluwer Academic Publisher, Boston.

Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, 20, 273–279.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, UK.

Dauxois, J. and Nkiet, G.M. (1997). Canonical analysis of two Euclidean subspaces and its applications. *Linear Algebra Appl.*, 264, 355–388.

Dauxois, J. and Nkiet, G.M. (1998). Nonlinear canonical analysis and independence tests. *Ann. Statist.*, 26, 1254–1278.

Dauxois, J. and Nkiet, G.M. (2002). Measure of association for Hilbert subspaces and some applications. *J. Multivariate. Anal.*, 82, 263–298.

Dauxois, J., Nkiet, G.M. and Romain, Y. (2004). Canonical analysis relative to a closed subspace. *Linear Algebra Appl.*, 388, 119–145.

Dauxois, J., Romain, Y. and Viguier, S. (1993). Comparison of two factor subspaces. *J. Multivariate Anal.*, 44, 160–178.

Eubank, R. and Hsing, T. (2005). Canonical correlation for stochastic processes. *preprint*.

Fukumizu, K., Bach, F.R. and Jordan, M.I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Mach. Learning Res.*, 5, 73–99.

Gretton, A., Herbrich, R. and Smola, A. (2003). The kernel mutual information. Technical Report, MPI for Biological Cybernetics, Tuebingen,Germany.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O. and Schölkopf, B. (2005). Kernel methods for measuring independence. *J. Machine Learning Research*, 6, 2075–2129.

Hardoon, D.R., Szedmak, S. and Shawe-Taylor, J. (2004). Canonical correlation analysis: an overview with application to learning methods. *Neural Computation*, 16, 2639–2664.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York.

Herbrich, R. (2002). *Learning Kernel Classifiers*. MIT Press, Cambridge, MA.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28, 321–377.

Hsing, T., Liu, L.-Y., Brun, M. and Dougherty, E. R. (2005). The coefficient of intrinsic dependence. *Pattern Recognition*, 38, 623-636.

Huang, S.Y. and Hwang, C.R. (2005). Kernel Fisher discriminant analysis in Gaussian reproducing kernel Hilbert space. Technical report, Institute of Statistical Science, Academia Sinica, Taiwan. `http://www.stat.sinica.edu.tw/syhuang`.

Huang, C.M, Lee, Y.J., Lin, D.K.J. and Huang, S.Y. (2006). Model selection for support vector machine via uniform design. `http://dmlab1.csie.ntust.edu.tw/downloads`.

Jensen, D.R. and Mayer, L.S. (1977). Some variational results and their applications in multiple inference. *Ann. Statist.*, 5, 922–931.

Kuss, M. and Graepel, T. (2003). The geometry of kernel canonical correlation analysis. Technical report, Max Planck Institute for Biological Cybernetics, Germany.

Lee, Y.J. and Huang, S.Y. (2006). Reduced support vector machines: a statistical theory. *IEEE Trans. Neural Networks*, accepted. `http://dmlab1.csie.ntust.edu.tw/downloads`.

Lee, Y.J., Lo, H.Y. and Huang, S.Y. (2003). Incremental reduced support vector machine. In *International Conference on Informatics Cybernetics and Systems (ICICS 2003)*, Kaohsiung, Taiwan. `http://dmlab1.csie.ntust.edu.tw/downloads`.

Lee, Y.J. and Mangasarian, O.L. (2001). RSVM: reduced support vector machines. *Proceeding 1st International Conference on Data Mining*, SIAM.

Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis*. Springer.

Schölkopf, B. and Smola, A.J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA.

Smola, A. and Schölkopf, B. (2000). Sparse greedy matrix approximation for machine learning. In *Proc. 17th International Conf. on Machine Learning*, 911–918. Morgan Kaufmann, San Francisco, CA.

Snelson, E. and Ghahramani, Z. (2005). Sparse Gaussian processes using pseudo-inputs. In Y. Weiss, B. Schölkopf and J. Platt, editors, *Advances in Neural Information Processing Systems*, 18, MIT Press, Cambridge, MA.

Vapnik, V.N. (1998). *Statistical Learning Theory*. Wiley, New York.

Wang, J., Neskovic, P. and Cooper, L.N. (2005). Training data selection for support vector machines. In Lipo Wang, Ke Chen and Yew-Soon Ong, editors, *Advances in Natural Computation: Proceedings, Part I, First International Conference*, Lecture Notes in Computer Science 3610, 554–564, Springer-Verlag, Berlin.

Williams, C.K.I. and Seeger, M. (2001). Using the Nyström method to speed up kernel machines. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems*, 13, 682-688, Cambridge, MA, MIT Press.

Table 1: Correlation for the first and the second pairs of canonical variates and kernel canonical variates for Example 2.

average correlation coefficient	LCCA (s.e.)	KCCA (s.e.)
between 1st pair cv's	0.0573 (0.0046)	0.9926 (0.0001)
between 2nd pair cv's	0.0132 (0.0020)	0.9646 (0.0005)

Table 2: Power estimates (or type-I error estimate in III-1 case) for testing independence of (X, Y) for several distributions based on 100 replicate runs. The significance level is $\alpha = 0.05$ The D&N columns are power estimates taken from Dauxois and Nkiet (1998) using order 2 and 3 splines for approximation. (For case I, the D&N has used sample sizes 200 and 400, instead of 500.)

case	n	LCCA (s.e.)	KCCA (s.e.)	D&N spline2	D&N spline3
I	500*	0.42 (0.049)	1.00 (0.000)	1.00	1.00
II	500	0.02 (0.014)	1.00 (0.000)	0.99	0.99
III-1, $\rho = 0$	500	0.06 (0.024)	0.04 (0.020)	NA	NA
III-2, $\rho = 0.2$	500	0.99 (0.010)	0.96 (0.020)	NA	NA
III-3, $\rho = 0.5$	500	1.00 (0.000)	1.00 (0.000)	NA	NA
III-4, $\rho = 0.8$	500	1.00 (0.000)	1.00 (0.000)	0.99	0.73
IV	500	0.62 (0.049)	1.00 (0.000)	1.00	1.00
V	500	0.37 (0.048)	1.00 (0.000)	1.00	1.00
VI	500	0.09 (0.029)	1.00 (0.000)	NA	NA

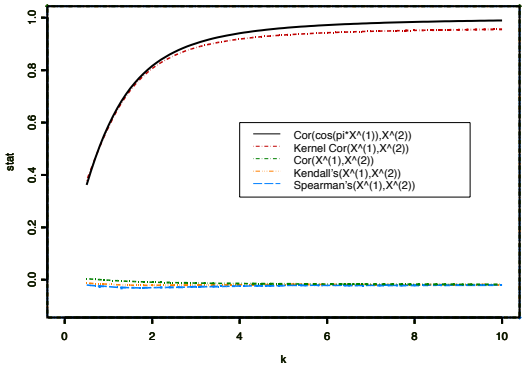

Figure 1: (a) Example 1, model I

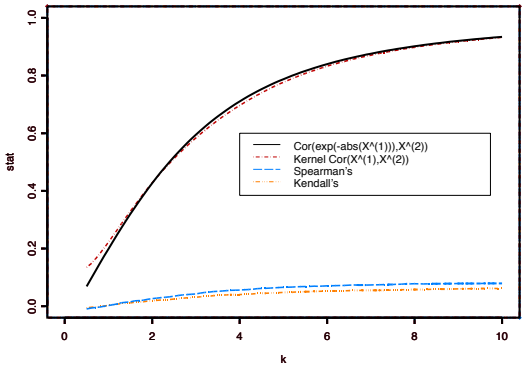

Figure 1: (b) Example 1, model II

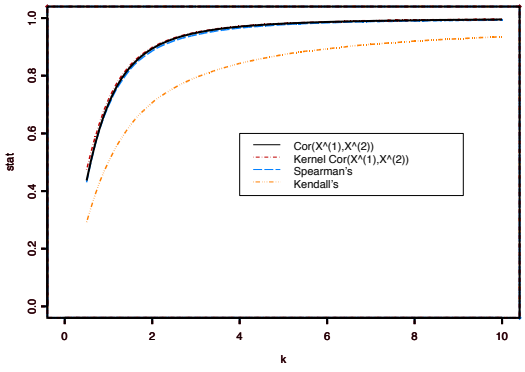

Figure 1: (c) Example 1, model III

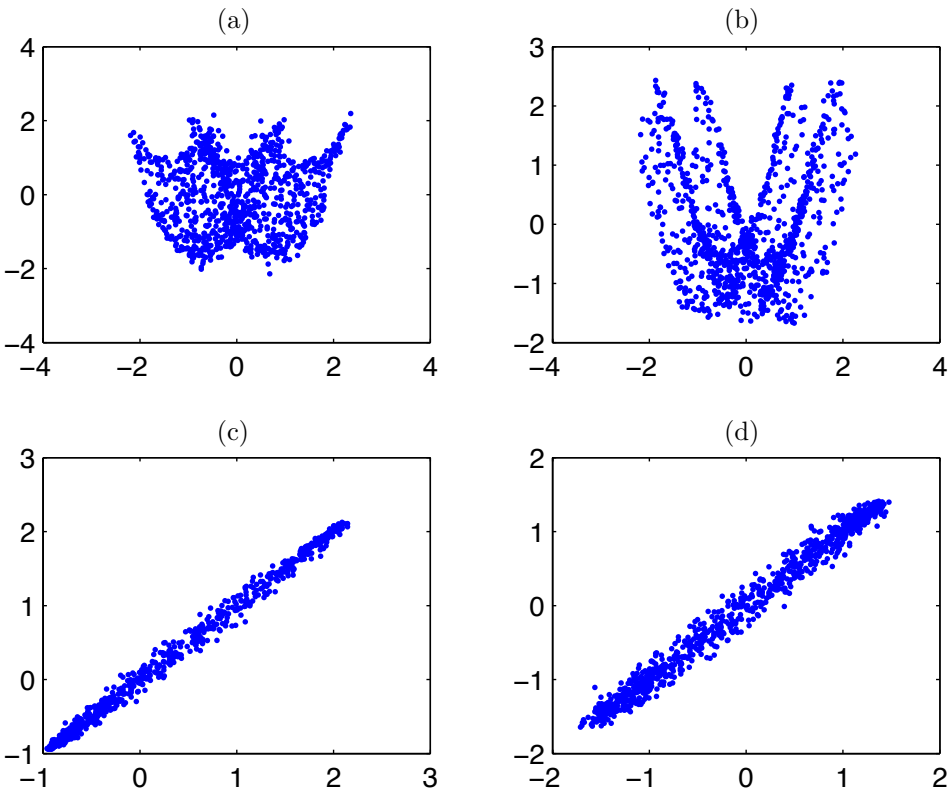

Figure 2: Scatter plots of the first and second pairs of linear canonical variates ((a) and (b)) and kernel canonical variates ((c) and (d)) for Example 2.

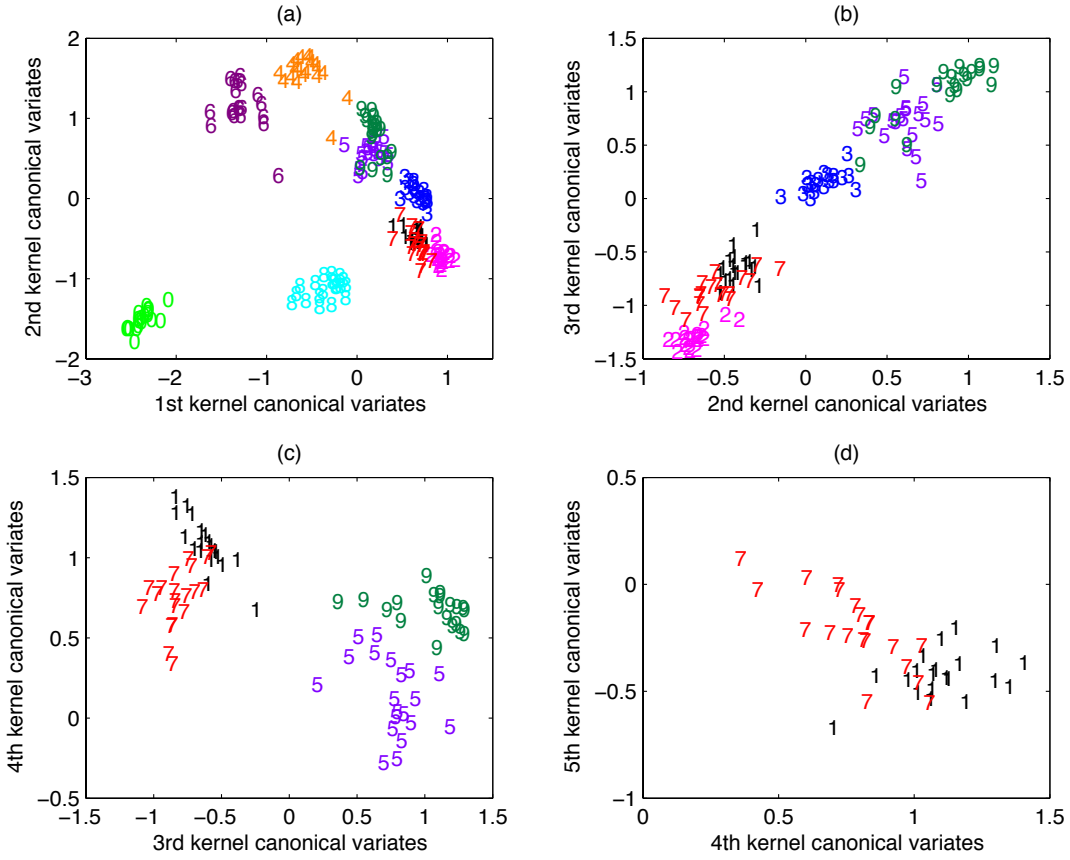

Figure 3: Scatter plots of pendigits over the 1st-and-2nd, 2nd-and-3rd, 3rd-and-4th, and 4th-and-5th kernel canonical variates.

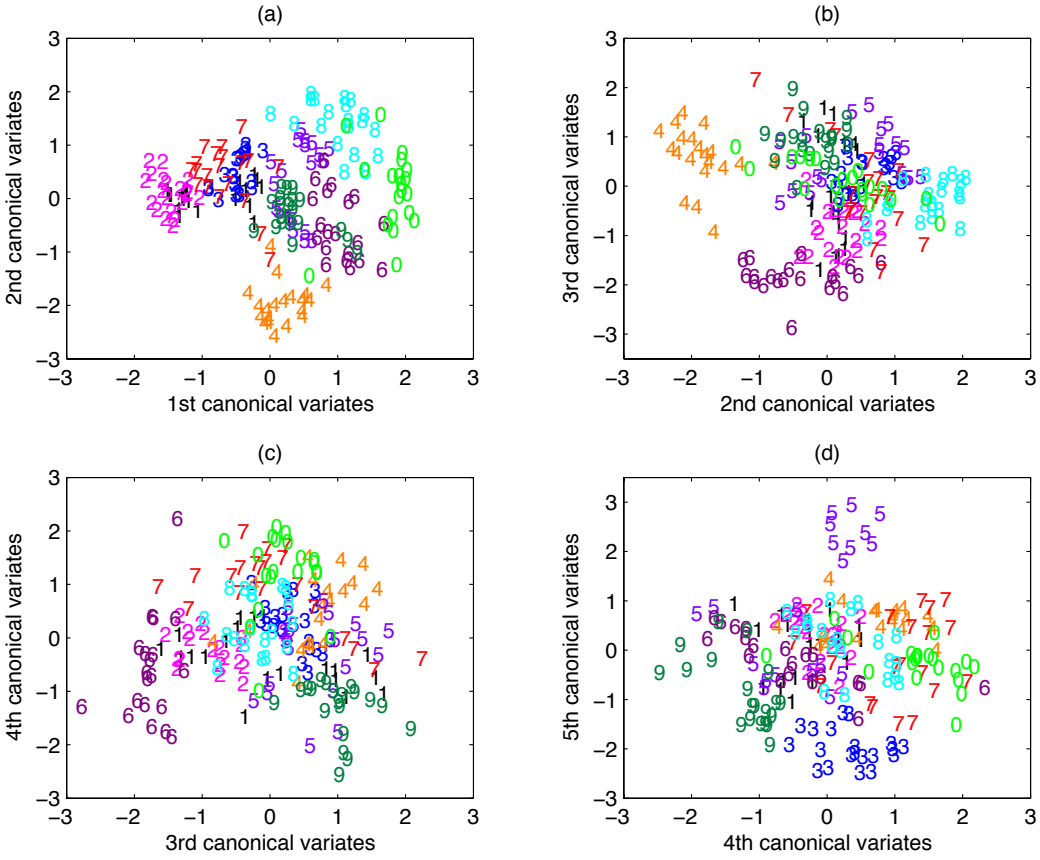

Figure 4: Scatter plots of pendigits over the 1st-and-2nd, 2nd-and-3rd, 3rd-and-4th, and 4th-and-5th canonical variates.