Max–Planck–Institut für biologische Kybernetik
Max Planck Institute for Biological Cybernetics

MAX-PLANCK-GESELLSCHAFT

Technical Report No. 108

The Geometry Of Kernel Canonical Correlation Analysis

Malte Kuss[1] and Thore Graepel[2]

May 2003

[1] Max Planck Institute for Biological Cybernetics, Dept. Schölkopf,
Spemannstrasse 38, 72076 Tübingen, Germany, email: malte.kuss@tuebingen.mpg.de
[2] Microsoft Research Ltd, Roger Needham Building,
7 J J Thomson Avenue, Cambridge CB3 0FB, U.K, email: thoreg@microsoft.com

The Geometry Of Kernel Canonical Correlation Analysis

Malte Kuss, Thore Graepel

Abstract. Canonical correlation analysis (CCA) is a classical multivariate method concerned with describing linear dependencies between sets of variables. After a short exposition of the linear sample CCA problem and its analytical solution, the article proceeds with a detailed characterization of its geometry. Projection operators are used to illustrate the relations between canonical vectors and variates. The article then addresses the problem of CCA between spaces spanned by objects mapped into kernel feature spaces. An exact solution for this kernel canonical correlation (KCCA) problem is derived from a geometric point of view. It shows that the expansion coefficients of the canonical vectors in their respective feature space can be found by linear CCA in the basis induced by kernel principal component analysis. The effect of mappings into higher dimensional feature spaces is considered critically since it simplifies the CCA problem in general. Then two regularized variants of KCCA are discussed. Relations to other methods are illustrated, e.g., multicategory kernel Fisher discriminant analysis, kernel principal component regression and possible applications thereof in blind source separation.

1 Introduction

Kernel methods attract a great deal of attention in the machine learning field of research initially due to the success of support vector machines. A common principle of these methods is to construct nonlinear variants of linear algorithms by substituting the linear inner product by kernel functions. Under certain conditions these kernel functions can be interpreted as representing the inner product of data objects implicitly mapped into a nonlinear related feature space (see for example Schölkopf and Smola (2002)).

Let $\mathbf{x}_i \in \mathcal{X}\ i = 1, \dots, m$ denote input space objects and consider a feature space mapping $\boldsymbol{\phi} : \mathcal{X} \rightarrow \mathcal{F}$ where the feature space $\mathcal{F}$ is an inner product space. The "kernel trick" is to calculate the inner product in $\mathcal{F}$,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle_{\mathcal{F}} \ , \tag{1}$$

using a kernel function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ of input space objects while avoiding explicit mappings $\boldsymbol{\phi}$. If an algorithm can be restated such that the data objects only appear in terms of inner products, one substitutes the linear dot product by such a kernel function[1]. Though mappings $\boldsymbol{\phi}$ will be used as an auxiliary concept during the construction of geometric algorithms, they never have to be constructed explicitly. The resulting kernel algorithm can be interpreted as running the original algorithm on the feature space mapped objects $\boldsymbol{\phi}(\mathbf{x}_i)$.

This construction has been used to derive kernel variants of various methods originated in multivariate statistics. Prominent examples are kernel principal component analysis (Schölkopf et al. 1998), kernel discriminant analysis (Mika et al. 1999) and variants of chemometric regression methods like kernel principal component regression, kernel ridge regression and kernel partial least squares regression (Rosipal and Trejo 2001). Furthermore, several authors have studied the construction of a kernel variant of CCA and proposed quite different algorithms (Lai and Fyfe 2000; Melzer et al. 2001; van Gestel et al. 2001; Bach and Jordan 2002).

Although CCA is a well known concept in mathematical statistics, it is seldom used in statistical practice. For this reason the following section starts with an introduction to sample linear CCA and describes the solution from a geometric point of view. We then go further into the question of how the canonical correlation between configurations of points mapped into kernel feature spaces can be determined while preserving the geometry of the original method. Afterwards we consider regularized variants of this problem and discuss their advantages. Finally, we illustrate relations to other methods, e.g. kernel principal component regression, blind source separation and multicategory kernel discriminant analysis.

[1] In the examples below we use polynomial kernels of the form $k(\mathbf{x}_i, \mathbf{x}_j) = \left(\langle \mathbf{x}_i, \mathbf{x}_j \rangle_{\mathcal{X}} + \theta\right)^d$ and Gaussian radial basis function (rbf) kernels $k(\mathbf{x}_i, \mathbf{x}_j) = exp\left(-\frac{1}{2\sigma^2}||\mathbf{x}_i - \mathbf{x}_j||^2\right)$.

2 Linear Canonical Correlation Analysis

Canonical correlation analysis (CCA) as introduced by Hotelling (1935,1936) is concerned with describing linear relations between sets of variables. Let $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i)$ for $i = 1, \ldots, m$ denote samples of measurements on m objects where $\mathbf{x}_i$ and $\mathbf{y}_i$ are meant to describe different aspects of these objects. A classical example—also illustrating the origin of CCA—would be to think of a psychological experiment collecting n_x measurements of reading ability $\mathbf{x}_i$ and n_y quantities describing the analytical ability $\mathbf{y}_i$ of m individuals. From a machine learning perspective, it may be more familiar to think of $\mathbf{x}_i$ as describing the ith observation while the corresponding $\mathbf{y}_i$ describes aspects of the class affiliation of this object. Even if the latter example suggests a directional relation between the sets, in general CCA handles the sets symmetrically. The data is compactly written using a partitioned matrix $\mathbf{Z} := \left[\begin{array}{cc} \mathbf{X} & \mathbf{Y} \end{array}\right]$ such that $\mathbf{z}_i$ corresponds to the ith row of $\mathbf{Z}$. We initially presume $m \gg n_x + n_y$ and a full column rank of $\mathbf{X}$ and $\mathbf{Y}$. Throughout the paper, we also implicitly assume the data $\mathbf{Z}$ to be column centered.

To gain insight into the geometry of the method it is advantageous to contemplate the CCA solution with respect to the spaces spanned by the rows and columns of the matrices $\mathbf{X}$ and $\mathbf{Y}$. Just to illustrate the notation used let $\mathbf{A}$ be an arbitrary $[m \times n]$ matrix then $\mathcal{L}\{\mathbf{A}\} := \{\mathbf{A}\boldsymbol{\alpha} \,|\, \boldsymbol{\alpha} \in \mathbb{R}^n\}$ will be referred to as the column-space and $\mathcal{L}\{\mathbf{A}'\} := \{\mathbf{A}'\boldsymbol{\alpha} \,|\, \boldsymbol{\alpha} \in \mathbb{R}^m\}$ the row-space of $\mathbf{A}$ (Harville 1997, 4.1).

The aim of sample canonical correlation analysis is to determine vectors $\mathbf{v}_j \in \mathcal{L}\{\mathbf{X}'\}$ and $\mathbf{w}_j \in \mathcal{L}\{\mathbf{Y}'\}$ such that the variates $\mathbf{a}_j := \mathbf{X}\mathbf{v}_j$ and $\mathbf{b}_j := \mathbf{Y}\mathbf{w}_j$ are maximally correlated.

$$\operatorname{cor}(\mathbf{a}_j, \mathbf{b}_j) := \frac{\langle \mathbf{a}_j, \mathbf{b}_j \rangle}{\|\mathbf{a}_j\| \, \|\mathbf{b}_j\|} \tag{2}$$

Usually, this is formulated as a constraint optimization problem

$$\begin{gathered} \underset{\mathbf{v}_j \in \mathcal{L}\{\mathbf{X}'\}, \mathbf{w}_j \in \mathcal{L}\{\mathbf{Y}'\}}{\operatorname{argmax}} \quad \mathbf{v}_j'\mathbf{X}'\mathbf{Y}\mathbf{w}_j \\ \text{subject to } \mathbf{v}_j'\mathbf{X}'\mathbf{X}\mathbf{v}_j = \mathbf{w}_j'\mathbf{Y}'\mathbf{Y}\mathbf{w}_j = 1 \end{gathered} \tag{3}$$

whereby the constraint is arbitrary in some respect as the lengths of $\mathbf{a}_j \in \mathcal{L}\{\mathbf{X}\}$ and $\mathbf{b}_j \in \mathcal{L}\{\mathbf{Y}\}$ do not affect the correlation (2) while $\|\mathbf{a}_j\|, \|\mathbf{b}_j\| > 0$ holds. The solution of (3) gives the first pair of canonical vectors $(\mathbf{v}_1, \mathbf{w}_1)$, and $\mathbf{a}_1 = \mathbf{X}\mathbf{v}_1$ and $\mathbf{b}_1 = \mathbf{Y}\mathbf{w}_1$ are the corresponding canonical variates. Up to $r = \min(\dim \mathcal{L}\{\mathbf{X}\}, \dim \mathcal{L}\{\mathbf{Y}\})$ pairs of canonical vectors $(\mathbf{v}_j, \mathbf{w}_j)$ can be recursively defined maximizing (3) subject to corresponding variates being orthogonal to previously found pairs. Referring to the examples above, CCA can be interpreted as constructing pairs of factors (or call them features) from $\mathbf{X}$ and $\mathbf{Y}$ respectively by linear combination of the respective variables, such that linear dependencies between the sets of variables are summarized.

Analytically, the maximization of (3) leads to the eigenproblems

$$\begin{aligned} \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}\left(\mathbf{Y}'\mathbf{Y}\right)^{-1}\mathbf{Y}'\mathbf{X}\mathbf{v}_j &= \lambda_j^2\mathbf{v}_j \qquad (4) \\ \left(\mathbf{Y}'\mathbf{Y}\right)^{-1}\mathbf{Y}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}\mathbf{w}_j &= \lambda_j^2\mathbf{w}_j \qquad (5) \end{aligned}$$

describing the canonical vectors $(\mathbf{v}_j, \mathbf{w}_j)$ as eigenvectors corresponding to the major r non-zero eigenvalues $1 \geq \lambda_1^2 \geq \ldots \geq \lambda_r^2 > 0$. Note that the eigenvalues equal the squared canonical correlation coefficients such that $\lambda_j = \operatorname{cor}(\mathbf{a}_j, \mathbf{b}_j)$. Usually but not necessarily $\mathbf{v}_j$ and $\mathbf{w}_j$ are scaled such that $\|\mathbf{a}_j\| = \|\mathbf{b}_j\| = 1$ as in (3), which will be assumed in the following.

We now turn to the geometry of the canonical variates and vectors which is more illustrative than the algebraic solution. When constructing CCA between kernel feature spaces in the following section, understanding the geometry will help us to verify the correctness of the solution.

At first a column-space point of view of the geometry will be described (Afriat 1957; Kockelkorn 2000). By examining (2) we find that the canonical correlation coefficient $\lambda_j = \operatorname{cor}(\mathbf{a}_j, \mathbf{b}_j)$ equals the cosine of the angle between the variates $\mathbf{a}_j$ and $\mathbf{b}_j$. Maximizing this cosine can be interpreted as minimizing the angle between $\mathbf{a}_j$ and $\mathbf{b}_j$, which in turn is equivalent to minimizing the distance for variates of equal length,

$$\begin{gathered} \underset{\mathbf{a}_j \in \mathcal{L}\{\mathbf{X}\}, \mathbf{b}_j \in \mathcal{L}\{\mathbf{Y}\}}{\operatorname{argmin}} \quad \|\mathbf{a}_j - \mathbf{b}_j\| \\ \text{subject to } \|\mathbf{a}_j\| = \|\mathbf{b}_j\| = 1\,, \end{gathered} \tag{6}$$

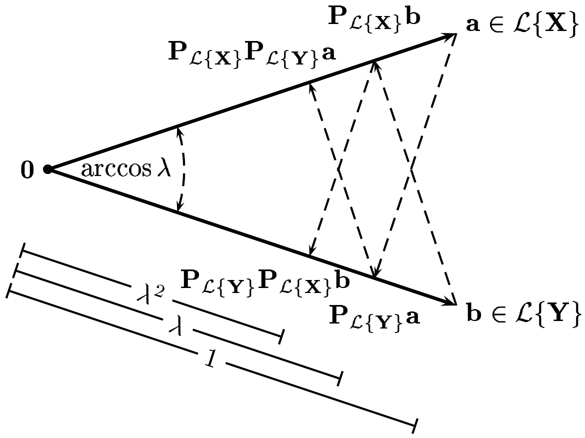
$\mathbf{P}_{\mathcal{L}\{\mathbf{X}\}}\mathbf{b}$
$\mathbf{a} \in \mathcal{L}\{\mathbf{X}\}$
$\mathbf{P}_{\mathcal{L}\{\mathbf{X}\}}\mathbf{P}_{\mathcal{L}\{\mathbf{Y}\}}\mathbf{a}$
$\mathbf{0}$
$\arccos\lambda$
λ^2
$\mathbf{P}_{\mathcal{L}\{\mathbf{Y}\}}\mathbf{P}_{\mathcal{L}\{\mathbf{X}\}}\mathbf{b}$
λ
$\mathbf{P}_{\mathcal{L}\{\mathbf{Y}\}}\mathbf{a}$
$\mathbf{b} \in \mathcal{L}\{\mathbf{Y}\}$
1

Figure 1: Illustration of the column-space geometry of the CCA solution. The canonical variates are the vectors $\mathbf{a} \in \mathcal{L}\{\mathbf{X}\}$ and $\mathbf{b} \in \mathcal{L}\{\mathbf{Y}\}$ that minimize their enclosed angle. The image of the orthogonal projection of $\mathbf{a}$ onto $\mathcal{L}\{\mathbf{Y}\}$ is $\lambda\mathbf{b}$ and likewise $\mathbf{P}_{\mathcal{L}\{\mathbf{X}\}}\mathbf{b} = \lambda\mathbf{a}$. Projecting these back onto the respective other space leads to relations (7) and (8).

again enforcing orthogonality with respect to previously found pairs. Let $\mathbf{P}_{\mathcal{L}\{\mathbf{X}\}} := \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'$ and $\mathbf{P}_{\mathcal{L}\{\mathbf{Y}\}} := \mathbf{Y}\left(\mathbf{Y}'\mathbf{Y}\right)^{-}\mathbf{Y}'$ denote the orthogonal projections onto the respective column-spaces $\mathcal{L}\{\mathbf{X}\}$ and $\mathcal{L}\{\mathbf{Y}\}$ (Harville 1997, 12.3). In view of these projections, the eigenproblems (4) and (5) give an obvious geometric characterization of the solution

$$\mathbf{P}_{\mathcal{L}\{\mathbf{X}\}}\mathbf{P}_{\mathcal{L}\{\mathbf{Y}\}}\mathbf{a}_j = \lambda_j^2\mathbf{a}_j \tag{7}$$

$$\mathbf{P}_{\mathcal{L}\{\mathbf{Y}\}}\mathbf{P}_{\mathcal{L}\{\mathbf{X}\}}\mathbf{b}_j = \lambda_j^2\mathbf{b}_j\,. \tag{8}$$

The column-space geometry of the first pair of canonical variates is illustrated in Figure 1.

Basically, the canonical variates $\mathbf{a}_j$ and $\mathbf{b}_j$ for $j = 1, \ldots, r$ are the elements of their respective column-spaces minimizing the angle between them with respect to the implied orthogonality $\mathbf{a}_j \perp \mathbf{a}_l$ and $\mathbf{b}_j \perp \mathbf{b}_l$ towards previously found pairs $l < j$.

So the column-space perspective provides an elegant and illuminating description of the CCA solution. However, for the construction of geometric algorithms the row-space geometry is the more common point of view and will therefore be considered here as well. Again, let $\mathbf{v}_j$ and $\mathbf{w}_j$ be a pair of canonical vectors and $\mathbf{a}_j$ and $\mathbf{b}_j$ the corresponding canonical variates. If we project $\mathbf{x}_i$ and $\mathbf{y}_i$ onto the respective canonical vectors we obtain

$$\mathbf{P}_{\mathcal{L}\{\mathbf{v}_j\}}\mathbf{x}_i = a_{ji}\frac{\mathbf{v}_j}{\left\|\mathbf{v}_j\right\|^2} \tag{9}$$

$$\mathbf{P}_{\mathcal{L}\{\mathbf{w}_j\}}\mathbf{y}_i = b_{ji}\frac{\mathbf{w}_j}{\left\|\mathbf{w}_j\right\|^2} \tag{10}$$

where a_{ij} and b_{ij} denote the scores of ith observation on the jth canonical variates. Figure 2 illustrates the row-space geometry.

Another appealing description of CCA can be motivated by a least square regression problem which also has been introduced by Hotelling (1935). Given $\mathbf{X}$ and $\mathbf{Y}$, the problem is to find the linear combination of the columns of the respective other matrix which can be most accurately predicted by a least square regression. These “most predictable criteria” turn out to be the canonical variates. Further details on CCA and its applications can be found in Gittins (1985) and Mardia et al. (1979). Björck and Golub (1973) provide a detailed study of the computational aspects of CCA.

3 Kernel Canonical Correlation Analysis

We now describe how to determine canonical variates for spaces spanned by kernel feature space mapped objects. Therefore let $\phi_{\mathcal{X}} : \mathcal{X} \to \mathcal{F}_{\mathcal{X}}$ and $\phi_{\mathcal{Y}} : \mathcal{Y} \to \mathcal{F}_{\mathcal{Y}}$ denote feature space mappings corresponding to possibly different kernel functions $k_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_j) := \langle\phi_{\mathcal{X}}(\mathbf{x}_i), \phi_{\mathcal{X}}(\mathbf{x}_j)\rangle$ and $k_{\mathcal{Y}}(\mathbf{y}_i, \mathbf{y}_j) := \langle\phi_{\mathcal{Y}}(\mathbf{y}_i), \phi_{\mathcal{Y}}(\mathbf{y}_j)\rangle$. We use a

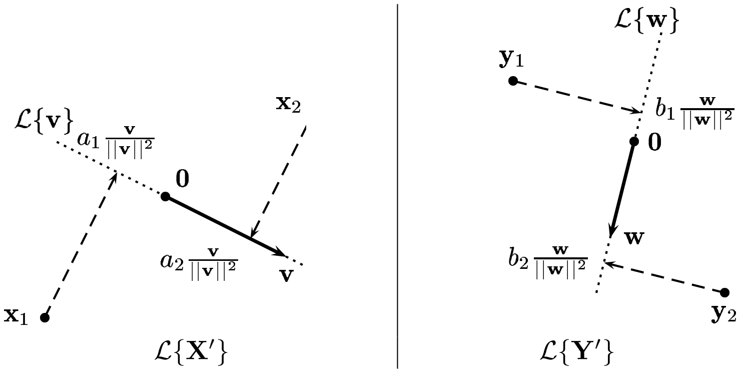
$\mathcal{L}\{\mathbf{v}\}$
$a_1 \frac{\mathbf{v}}{||\mathbf{v}||^2}$
$\mathbf{x}_2$
$\mathbf{0}$
$a_2 \frac{\mathbf{v}}{||\mathbf{v}||^2}$
$\mathbf{v}$
$\mathbf{x}_1$
$\mathcal{L}\{\mathbf{X}'\}$
$\mathcal{L}\{\mathbf{w}\}$
$\mathbf{y}_1$
$b_1 \frac{\mathbf{w}}{||\mathbf{w}||^2}$
$\mathbf{0}$
$\mathbf{w}$
$b_2 \frac{\mathbf{w}}{||\mathbf{w}||^2}$
$\mathbf{y}_2$
$\mathcal{L}\{\mathbf{Y}'\}$

Figure 2: Illustration of the row-space geometry of the canonical vectors. The left and right part have to be seen separately and respectively show the canonical vectors $\mathbf{v} \in \mathcal{L}\{\mathbf{X}'\}$ and $\mathbf{w} \in \mathcal{L}\{\mathbf{Y}'\}$ and two exemplary observations $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i)$ $i = 1, 2$. The correlation of the variates is indicated by $a_1, b_1 < 0$ and $a_2, b_2 > 0$.

compact representation of the objects in feature spaces $\mathbf{\Phi}_\mathcal{X} := [\boldsymbol{\phi}_\mathcal{X}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}_\mathcal{X}(\mathbf{x}_m)]'$ and likewise $\mathbf{\Phi}_\mathcal{Y} := [\boldsymbol{\phi}_\mathcal{Y}(\mathbf{y}_1), \ldots, \boldsymbol{\phi}_\mathcal{Y}(\mathbf{y}_m)]'$. These configurations span the spaces $\mathcal{L}\{\mathbf{\Phi}_\mathcal{X}\}$ and $\mathcal{L}\{\mathbf{\Phi}_\mathcal{Y}\}$ which will be referred to as effective feature spaces. As usual $\mathbf{K}_\mathcal{X} := \mathbf{\Phi}_\mathcal{X}\mathbf{\Phi}'_\mathcal{X}$ and $\mathbf{K}_\mathcal{Y} := \mathbf{\Phi}_\mathcal{Y}\mathbf{\Phi}'_\mathcal{Y}$ denote the $[m \times m]$ kernel inner product matrices, also known as kernel Gram matrices, which can be constructed element-wise as $(\mathbf{K}_\mathcal{X})_{ij} := k_\mathcal{X}(\mathbf{x}_i, \mathbf{x}_j)$ and $(\mathbf{K}_\mathcal{Y})_{ij} := k_\mathcal{Y}(\mathbf{y}_i, \mathbf{y}_j)$ for $i, j = 1, \ldots, m$. A notable advantage of the kernel approach—and thus of the method considered below—is the ability to handle various data types, e.g. strings and images, by using an appropriate kernel function.

Since we know the canonical vectors $\mathbf{v}_j \in \mathcal{L}\{\mathbf{\Phi}'_\mathcal{X}\}$ and $\mathbf{w}_j \in \mathcal{L}\{\mathbf{\Phi}'_\mathcal{Y}\}$ to lie in the spaces spanned by the feature space mapped objects we can represent them as linear combinations $\mathbf{v}_j = \mathbf{\Phi}'_\mathcal{X}\boldsymbol{\alpha}_j$ and $\mathbf{w}_j = \mathbf{\Phi}'_\mathcal{Y}\boldsymbol{\beta}_j$ using $\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j \in \mathbb{R}^m$ as expansion coefficients. Accordingly, the canonical variates are $\mathbf{a}_j = \mathbf{\Phi}_\mathcal{X}\mathbf{v}_j = \mathbf{K}_\mathcal{X}\boldsymbol{\alpha}_j$ and likewise $\mathbf{b}_j = \mathbf{\Phi}_\mathcal{Y}\mathbf{w}_j = \mathbf{K}_\mathcal{Y}\boldsymbol{\beta}_j$. As in the linear method the feature space configurations $\mathbf{\Phi}_\mathcal{X}$ and $\mathbf{\Phi}_\mathcal{Y}$ are assumed to be centered which can be realized by a subsequent column and row centering of the kernel Gram matrices (Schölkopf et al. 1998).

As in the linear case, the aim of kernel canonical correlation analysis (KCCA) is to find canonical vectors in terms of expansion coefficients $\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j \in \mathbb{R}^m$. Formulated as a constraint optimization problem this leads to

$$\underset{\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j \in \mathbb{R}^m}{\operatorname{argmax}} \; \boldsymbol{\alpha}'_j \mathbf{K}_\mathcal{X}\mathbf{K}_\mathcal{Y}\boldsymbol{\beta}_j \tag{11}$$
$$\text{subject to } \boldsymbol{\alpha}'_j\mathbf{K}_\mathcal{X}\mathbf{K}_\mathcal{X}\boldsymbol{\alpha}_j = \boldsymbol{\beta}'_j\mathbf{K}_\mathcal{Y}\mathbf{K}_\mathcal{Y}\boldsymbol{\beta}_j = 1$$

again for $j = 1, \ldots, \min(\dim \mathcal{L}\{\mathbf{\Phi}_\mathcal{X}\}, \dim \mathcal{L}\{\mathbf{\Phi}_\mathcal{Y}\})$ and with respect to orthogonality towards previously found pairs. Note that in case the Gramians are singular the expansion coefficients corresponding to the canonical vectors are not unique and one cannot proceed straightforward as in the linear case.

From a geometric point of view the effective feature spaces are identical to the spaces spanned by the kernel Gram matrices.

$$\mathcal{L}\{\mathbf{\Phi}_\mathcal{X}\} = \mathcal{L}\{\mathbf{\Phi}_\mathcal{X}\mathbf{\Phi}'_\mathcal{X}\} = \mathcal{L}\{\mathbf{K}_\mathcal{X}\} \tag{12}$$
$$\mathcal{L}\{\mathbf{\Phi}_\mathcal{Y}\} = \mathcal{L}\{\mathbf{\Phi}_\mathcal{Y}\mathbf{\Phi}'_\mathcal{Y}\} = \mathcal{L}\{\mathbf{K}_\mathcal{Y}\} \tag{13}$$

So the canonical variates $\mathbf{a}_j \in \mathcal{L}\{\mathbf{K}_\mathcal{X}\}$ and $\mathbf{b}_j \in \mathcal{L}\{\mathbf{K}_\mathcal{Y}\}$ can be considered elements of the column-spaces of the Gramians and therefore can be described using bases of these spaces.

For this purpose we use kernel principal components which constitute particular orthogonal bases of the effective feature spaces (Schölkopf et al. 1998). Here we restrict ourselves to the description of how to find the principal components for $\mathbf{\Phi}_\mathcal{X}$. Afterwards it should be obvious how the principal components for $\mathbf{\Phi}_\mathcal{Y}$ can be analogously determined. The first $i = 1, \ldots, d$ principal components $\mathbf{u}_i \in \mathcal{L}\{\mathbf{\Phi}'_\mathcal{X}\}$ combined in a matrix $\mathbf{U}_\mathcal{X} = [\mathbf{u}_1, \ldots, \mathbf{u}_d]$ form an orthonormal basis of a d-dimensional subspace $\mathcal{L}\{\mathbf{U}_\mathcal{X}\} \subseteq \mathcal{L}\{\mathbf{\Phi}'_\mathcal{X}\}$ and can therefore also be described as linear combinations $\mathbf{U}_\mathcal{X} = \mathbf{\Phi}'_\mathcal{X}\mathbf{A}_\mathcal{X}$ where the $[m \times d]$ matrix $\mathbf{A}_\mathcal{X}$ holds the expansion coefficients. From a geometric point of view $\mathbf{A}_\mathcal{X}$ is chosen to minimize the sum of squared distances between $\mathbf{\Phi}'_\mathcal{X}$ and the projection of $\mathbf{\Phi}'_{\mathcal{X}}$ onto $\mathcal{L}\{\mathbf{U}_{\mathcal{X}}\}$ given by $\mathbf{P}_{\mathcal{L}\{\mathbf{U}\}}\mathbf{\Phi}'_{\mathcal{X}} = \mathbf{U}_{\mathcal{X}}\mathbf{U}'_{\mathcal{X}}\mathbf{\Phi}'_{\mathcal{X}}$.

$$\begin{gathered}\operatorname*{argmin}_{\mathbf{A}\in\mathbb{R}^{m\times d}} \left\|\mathbf{\Phi}'_{\mathcal{X}} - \mathbf{U}_{\mathcal{X}}\mathbf{U}'_{\mathcal{X}}\mathbf{\Phi}'_{\mathcal{X}}\right\|^2 \\ \text{subject to } \mathbf{U}'_{\mathcal{X}}\mathbf{U}_{\mathcal{X}} = \mathbf{I}_d\end{gathered} \tag{14}$$

Analytically, the optimal $\mathbf{A}_{\mathcal{X}}$ is found using the eigendecomposition $\mathbf{K}_{\mathcal{X}} = \mathbf{V\Lambda V}'$ of the p.s.d. kernel Gram matrix such that $\mathbf{A}_{\mathcal{X}}$ consists of the first d columns of $\mathbf{V\Lambda}^{-\frac{1}{2}}$. So the principal components are $\mathbf{U}_{\mathcal{X}} = \mathbf{\Phi}'_{\mathcal{X}}\mathbf{A}_{\mathcal{X}}$ and the coordinates of the $\mathbf{\Phi}_{\mathcal{X}}$ with respect to the principal components as a basis are $\mathbf{C}_{\mathcal{X}} = \mathbf{\Phi}_{\mathcal{X}}\mathbf{U}_{\mathcal{X}} = \mathbf{K}_{\mathcal{X}}\mathbf{A}_{\mathcal{X}}$.

If we choose $d_{\mathcal{X}} = \dim\mathcal{L}\{\mathbf{\Phi}_{\mathcal{X}}\} = \operatorname{rk}\mathbf{K}_{\mathcal{X}}$ then the $[m \times d_{\mathcal{X}}]$ matrix $\mathbf{C}_{\mathcal{X}}$ of principal component transformed data constitutes a basis such that $\mathcal{L}\{\mathbf{\Phi}_{\mathcal{X}}\} = \mathcal{L}\{\mathbf{C}_{\mathcal{X}}\}$. Analogously, consider the $[m\times d_{\mathcal{Y}}]$ matrix $\mathbf{C}_{\mathcal{Y}}$ of coordinates describing $\mathbf{\Phi}_{\mathcal{Y}}$ in the kernel principal component basis $\mathbf{U}_{\mathcal{Y}}$ such that $\mathcal{L}\{\mathbf{\Phi}_{\mathcal{Y}}\} = \mathcal{L}\{\mathbf{C}_{\mathcal{Y}}\}$.

The problem of finding canonical correlations between kernel feature spaces thus reduces to linear CCA between kernel principal component scores.

$$\left(\mathbf{C}'_{\mathcal{X}}\mathbf{C}_{\mathcal{X}}\right)^{-1}\mathbf{C}'_{\mathcal{X}}\mathbf{C}_{\mathcal{Y}}\left(\mathbf{C}'_{\mathcal{Y}}\mathbf{C}_{\mathcal{Y}}\right)^{-1}\mathbf{C}'_{\mathcal{Y}}\mathbf{C}_{\mathcal{X}}\boldsymbol{\psi}_j = \lambda_j^2\boldsymbol{\psi}_j \tag{15}$$

$$\left(\mathbf{C}'_{\mathcal{Y}}\mathbf{C}_{\mathcal{Y}}\right)^{-1}\mathbf{C}'_{\mathcal{Y}}\mathbf{C}_{\mathcal{X}}\left(\mathbf{C}'_{\mathcal{X}}\mathbf{C}_{\mathcal{X}}\right)^{-1}\mathbf{C}'_{\mathcal{X}}\mathbf{C}_{\mathcal{Y}}\boldsymbol{\xi}_j = \lambda_j^2\boldsymbol{\xi}_j \tag{16}$$

Then the canonical vectors are given by $\mathbf{v}_j = \mathbf{\Phi}_{\mathcal{X}}\mathbf{A}_{\mathcal{X}}\boldsymbol{\psi}_j$ and $\mathbf{w}_j = \mathbf{\Phi}_{\mathcal{Y}}\mathbf{A}_{\mathcal{Y}}\boldsymbol{\xi}_j$ or referring to above notation $\boldsymbol{\alpha}_j = \mathbf{A}_{\mathcal{X}}\boldsymbol{\psi}_j$ and $\boldsymbol{\beta}_j = \mathbf{A}_{\mathcal{Y}}\boldsymbol{\xi}_j$. So the corresponding kernel canonical variates are $\mathbf{a}_j = \mathbf{K}_{\mathcal{X}}\mathbf{A}_{\mathcal{X}}\boldsymbol{\psi}_j$ and $\mathbf{b}_j = \mathbf{K}_{\mathcal{Y}}\mathbf{A}_{\mathcal{Y}}\boldsymbol{\xi}_j$. An example is given in Figure 3. Scores on the kernel canonical vectors for previously unseen objects $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ can easily be calculated by computing the score on the particular kernel principal vectors and weighting them with $\boldsymbol{\psi}_j$ or $\boldsymbol{\xi}_j$ respectively.

Applying a principal component transformation to the data, also seems to be a common procedure when singular covariance matrices occure in linear CCA (see for example Khatri (1976)). Note that the values of the non-null canonical correlation coefficients λ_j^2 are not affected by this, since the resulting eigenproblem is similar. The procedure can also be understood as constructing Moore-Penrose inverses in the projections occuring in (7) and (8).

Using a subset of kernel principal components as basis vectors, e.g., by omitting those corresponding to smaller eigenvalues, can still lead to highly correlated features and often has a smoothing effect. But since the directions of the major canonical vectors are not necessarily related to those of the major principal components, this has to be handled with caution. Theoretical optimality of the canonical vectors can only be assured by using complete bases. Computationally this leads to the problem of estimating the dimensions of the effective feature spaces by looking at the eigenspectra of the kernel Gramians during the calculation of KCCA. Fortunately, for some widely used kernel functions, e.g. polynomial and RBF kernels, general propositions about the dimensionality of the corresponding feature spaces are available.

As shown, the canonical correlation between $\mathcal{L}\{\mathbf{\Phi}_{\mathcal{X}}\}$ and $\mathcal{L}\{\mathbf{\Phi}_{\mathcal{Y}}\}$ can be exactly determined—at least theoretically. But the effect of mapping the data into higher dimensional spaces has to be critically reconsidered. The sample canonical correlation crucially depends on the relation between the sample size and the dimensionalities of the spaces involved. Feature space mappings usually considered in kernel methods share the property of mapping into higher dimensional spaces such that the dimension of the effective feature space is larger than that of the input space. If the spaces $\mathcal{L}\{\mathbf{\Phi}_{\mathcal{X}}\}$ and $\mathcal{L}\{\mathbf{\Phi}_{\mathcal{Y}}\}$ share a common subspace of dimension $h = \dim\left(\mathcal{L}\{\mathbf{\Phi}_{\mathcal{X}}\} \cap \mathcal{L}\{\mathbf{\Phi}_{\mathcal{Y}}\}\right)$, then $\mathbf{a}_j = \mathbf{b}_j$ and therefore $\operatorname{cor}(\mathbf{a}_j, \mathbf{b}_j) = 1$ for $j = 1, \ldots, h$ (see Figure 1). If $\dim\mathcal{L}\{\mathbf{K}_{\mathcal{X}}\} + \dim\mathcal{L}\{\mathbf{K}_{\mathcal{Y}}\} > m$ the effective feature spaces will share a common subspace. Especially in case of the frequently used Gaussian radial basis function kernel the Gramians $\mathbf{K}_{\mathcal{X}}$ and $\mathbf{K}_{\mathcal{Y}}$ are nonsingular so that the effective feature spaces are identical and the CCA problem becomes trivial. In general mappings into higher dimensional spaces are most likely to increase the canonical correlation coefficient relative to linear CCA between the input spaces. Therefore the kernel canonical correlation coefficient has to be interpreted with caution and KCCA should rather be considered as a geometric algorithm to construct highly correlated features.

The proposed method includes linear CCA as special case when using linear kernel functions for which the mappings $\boldsymbol{\phi}_{\mathcal{X}}$ and $\boldsymbol{\phi}_{\mathcal{Y}}$ are the identity mappings.

Note that we can also find directions of maximum covariance between kernel feature spaces in a similar way. Referring to the above notation, the problem is to maximize

$$\operatorname{cov}(\mathbf{a}_j, \mathbf{b}_j) := \frac{\langle \mathbf{a}_j, \mathbf{b}_j \rangle}{\|\mathbf{v}_j\|\,\|\mathbf{w}_j\|} \tag{17}$$

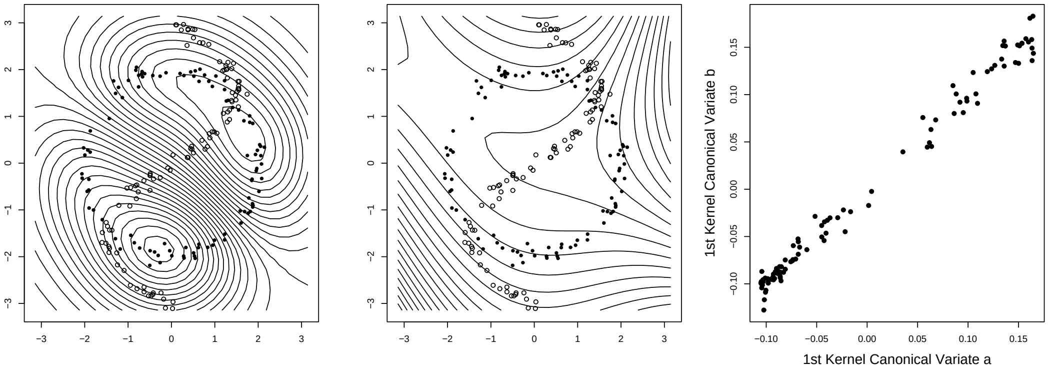

Figure 3: Kernel canonical correlation example. The data consists of two sets of 100 points each. For $\mathbf{X}$ the points are lying on a circle (solid points) while $\mathbf{Y}$ (circles) describe a sine curve (points correspond by arclength). For $\mathbf{X}$ we used a RBF kernel $(\sigma = 1)$ and for $\mathbf{Y}$ a homogeneous polynomial kernel of degree $(d = 2)$. The lines plotted describe regions of equal score on the first canonical vectors, which can be thought of as orthogonal (see Schölkopf et al. (1998)). This is shown for $\mathbf{v}_1 \in \mathcal{L}\{\mathbf{\Phi}'_{\mathcal{X}}\}$ (upper) and for $\mathbf{w}_1 \in \mathcal{L}\{\mathbf{\Phi}'_{\mathcal{Y}}\}$ (middle). The bottom plot shows the first pair of kernel canonical variates $(\mathbf{a}_1, \mathbf{b}_1)$ showing that $\langle \phi(\mathbf{x}_i), \mathbf{v}_1 \rangle_{\mathcal{F}}$ and $\langle \phi(\mathbf{y}_i), \mathbf{w}_1 \rangle_{\mathcal{F}}$ are highly correlated for $i = 1, \ldots, m$.

subject to orthogonality with previously found pairs as in the CCA derivation. In short, the solution is characterized by the eigenproblems

$$\begin{aligned} \mathbf{C}'_{\mathcal{X}}\mathbf{C}_{\mathcal{Y}}\mathbf{C}'_{\mathcal{Y}}\mathbf{C}_{\mathcal{X}}\boldsymbol{\psi}_j &= \lambda_j \boldsymbol{\psi}_j && (18)\\ \mathbf{C}'_{\mathcal{Y}}\mathbf{C}_{\mathcal{X}}\mathbf{C}'_{\mathcal{X}}\mathbf{C}_{\mathcal{Y}}\boldsymbol{\xi}_j &= \lambda_j \boldsymbol{\xi}_j && (19) \end{aligned}$$

again using the kernel principal components as bases of the effective feature spaces.

4 Regularized Variants

In previous approaches the kernel CCA problem (11) had been handled analogously to the linear CCA problem (3) by optimizing (11) in $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$ directly (e.g. Melzer et al. (2001)). An obvious drawback of this procedure is that kernel Gram matrices $\mathbf{K}_{\mathcal{X}}$ and $\mathbf{K}_{\mathcal{Y}}$ have to be inverted at some point during the derivation and they are not necessarily nonsingular. This is caused by not using a minimal basis for the description of canonical vectors. To overcome this problem, it has been suggested to add small multiples of the identity matrix $\gamma_{\mathcal{X}}\mathbf{I}$ and $\gamma_{\mathcal{Y}}\mathbf{I}$ to the kernel Gram matrices. This approach, which will be referred to as regularized kernel correlation, leads to a unique solution described by the eigenproblems

$$\begin{aligned} \left(\mathbf{K}^2_{\mathcal{X}} + \gamma_{\mathcal{X}}\mathbf{I}\right)^{-1}\mathbf{K}_{\mathcal{X}}\mathbf{K}_{\mathcal{Y}}\left(\mathbf{K}^2_{\mathcal{Y}} + \gamma_{\mathcal{Y}}\mathbf{I}\right)^{-1}\mathbf{K}_{\mathcal{Y}}\mathbf{K}_{\mathcal{X}}\boldsymbol{\alpha}_j &= \lambda_j^2 \boldsymbol{\alpha}_j\\ \left(\mathbf{K}^2_{\mathcal{Y}} + \gamma_{\mathcal{Y}}\mathbf{I}\right)^{-1}\mathbf{K}_{\mathcal{Y}}\mathbf{K}_{\mathcal{X}}\left(\mathbf{K}^2_{\mathcal{X}} + \gamma_{\mathcal{X}}\mathbf{I}\right)^{-1}\mathbf{K}_{\mathcal{X}}\mathbf{K}_{\mathcal{Y}}\boldsymbol{\beta}_j &= \lambda_j^2 \boldsymbol{\beta}_j\,. \end{aligned}$$

The so found pairs of vectors $(\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j)$ maximize the regularized criterion

$$\frac{\langle \mathbf{a}_j, \mathbf{b}_j \rangle}{\sqrt{\|\mathbf{a}_j\|^2 + \gamma_{\mathcal{X}}\|\boldsymbol{\alpha}_j\|^2}\sqrt{\|\mathbf{b}_j\|^2 + \gamma_{\mathcal{Y}}\|\boldsymbol{\beta}_j\|^2}} \qquad (20)$$

instead of maximizing the correlation coefficient $\mathrm{cor}(\mathbf{a}_j, \mathbf{b}_j)$ (2). The solution neither shows the geometry of the kernel canonical vectors nor gives an optimal correlation of the variates. On the other hand, the additional ridge parameters $\gamma_{\mathcal{X}}$ and $\gamma_{\mathcal{Y}}$ induce a beneficial control of over-fitting and enhance the numerical stability of the solution. In many experiments the solution of this regularized problem shows a better generalization ability than the kernel canonical vectors, in the sense of giving higher correlated scores for new objects. It also avoids the problem of estimating the dimensionality of the effective feature spaces.

These merits motivate a regularization of the kernel CCA method proposed in the previous section. Then the criterion to maximize is

$$\frac{\langle \mathbf{a}_j, \mathbf{b}_j \rangle}{\sqrt{\|\mathbf{a}_j\|^2 + \gamma_{\mathcal{X}} \|\boldsymbol{\psi}_j\|^2} \sqrt{\|\mathbf{b}_j\|^2 + \gamma_{\mathcal{Y}} \|\boldsymbol{\xi}_j\|^2}} \tag{21}$$

which in the context of linear CCA has been introduced by Vinod (1976) under the name "canonical ridge". Maximizing (21) in $\boldsymbol{\psi}_j$ and $\boldsymbol{\xi}_j$ leads to the eigenproblems

$$\begin{aligned} \left(\mathbf{C}'_{\mathcal{X}}\mathbf{C}_{\mathcal{X}} + \gamma_{\mathcal{X}}\mathbf{I}\right)^{-1} \mathbf{C}'_{\mathcal{X}}\mathbf{C}_{\mathcal{Y}} \left(\mathbf{C}'_{\mathcal{Y}}\mathbf{C}_{\mathcal{Y}} + \gamma_{\mathcal{Y}}\mathbf{I}\right)^{-1} \mathbf{C}'_{\mathcal{Y}}\mathbf{C}_{\mathcal{X}}\boldsymbol{\psi}_j &= \lambda_j^2 \boldsymbol{\psi}_j \\ \left(\mathbf{C}'_{\mathcal{Y}}\mathbf{C}_{\mathcal{Y}} + \gamma_{\mathcal{Y}}\mathbf{I}\right)^{-1} \mathbf{C}'_{\mathcal{Y}}\mathbf{C}_{\mathcal{X}} \left(\mathbf{C}'_{\mathcal{X}}\mathbf{C}_{\mathcal{X}} + \gamma_{\mathcal{X}}\mathbf{I}\right)^{-1} \mathbf{C}'_{\mathcal{X}}\mathbf{C}_{\mathcal{Y}}\boldsymbol{\xi}_j &= \lambda_j^2 \boldsymbol{\xi}_j \,. \end{aligned}$$

In experiments the so–obtained feature space vectors were often found to give higher correlated features compared to the regularized kernel correlation solution.

Nevertheless, the regularized variants constructed in this section do not exhibit the exact geometry of the canonical correlation. From a geometric point of view, the effect of the ridge terms can be interpreted as distortions of the projections resulting in a suboptimal solution regarding the correlation of obtained variates. For a given sample and ridge parameters $\gamma_{\mathcal{X}}, \gamma_{\mathcal{Y}} > 0$ the maximum value of (21) is smaller than the kernel CCA coefficient obtained by (15, 16) but always larger or equal to the value of (20), which also holds for the correlation of the corresponding variates. For $\gamma_{\mathcal{X}}, \gamma_{\mathcal{Y}} \to 0$ all three approaches become equivalent which can be interpreted analogously to the limit description of the Moore-Penrose inverse (Harville 1997, 20.7). Figure 4 illustrates a toy-example comparing the presented methods on the open-closed-book dataset provided by Mardia et al. (1979).

5 Relations Towards Other Methods

Canonical correlation analysis embodies various other multivariate methods which arise as special cases for certain restrictions on the kind and number of utilized variables (Gittins 1985; Mardia et al. 1979). Although CCA is a symmetric method from a conceptual point of view, in these cases it is mostly used in a directed sense by considering $\mathbf{X}$ as input and $\mathbf{Y}$ as target variables. It is then that CCA shows its least square regression character.

From the "most predictable criterion" property it can easily be derived that if $\mathbf{y}$ is a centered $[m \times 1]$ vector and a linear kernel for $\mathbf{y}$ is used then the KCCA solution gives the estimator of the least square regression estimator of $\mathbf{C}_{\mathcal{X}}$ onto $\mathbf{y}$ which is equivalent to the kernel principal component regression estimator (Rosipal and Trejo 2001). As in the linear case, the squared kernel canonical correlation coefficient λ^2 describes the proportion of the sums of squares explained by the regression.

Linear CCA also includes Fisher's linear discriminant analysis as a special case. Since the geometry of linear CCA is preserved in the kernel variant this relation also holds for the kernel methods (Mika et al. 1999). Thereby the KCCA formulation provides an elegant solution to the general multicategory case. Let $\mathbf{X} = \left[\mathbf{X}'_1, \ldots, \mathbf{X}'_g\right]'$ be an $[m \times n]$ matrix of input space samples partitioned into g classes. We then construct an $[m \times g]$ indicator matrix $\mathbf{Y}$

$$\mathbf{Y}_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ belongs to class } j \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

of binary dummy variables. By computing the canonical correlation between $\mathcal{L}\{\boldsymbol{\Phi}_{\mathcal{X}}\}$ and $\mathcal{L}\{\mathbf{Y}\}$, the canonical vectors $\mathbf{v}_j \in \mathcal{L}\{\boldsymbol{\Phi}'_{\mathcal{X}}\}$ for $j = 1, \ldots, g$ are equivalent to the kernel Fisher discriminant (KFD) vectors. Figure 5 provides two examples for the well known IRIS data set using linear and polynomial kernels. Note that this formulation of KFD can go without a regularization parameter. The regularized forms of KCCA can be shown to include kernel ridge regression and regularized kernel Fisher discriminant analysis as special cases analogously to the relations described above.

The idea of relating two kernel feature spaces $\mathcal{L}\{\boldsymbol{\Phi}_{\mathcal{X}}\}$ and $\mathcal{L}\{\boldsymbol{\Phi}_{\mathcal{Y}}\}$ has recently been considered more generally in the kernel dependency estimation framework by Weston et al. (2002). The objective of their approach is to learn mappings from objects of $\mathcal{L}\{\boldsymbol{\Phi}'_{\mathcal{X}}\}$ to targets in $\mathcal{L}\{\boldsymbol{\Phi}'_{\mathcal{Y}}\}$. KCCA and in particular its special case KFD can be embedded in this framework.

Several authors studied applications of canonical correlation analysis in the context of blind source separation problems. A linear approach by Borga and Knutsson (2001) uses CCA to find an approximate diagonalization of the autocorrelation matrix of a set of signals. Given a linear mixture $\mathbf{X} = \mathbf{S}\mathbf{A}$ of highly autocorrelated but otherwise uncorrelated sources $\mathbf{S}$ the authors compute CCA between the signals $\mathbf{X}$ and time delayed signals $\mathbf{X}[\tau]$

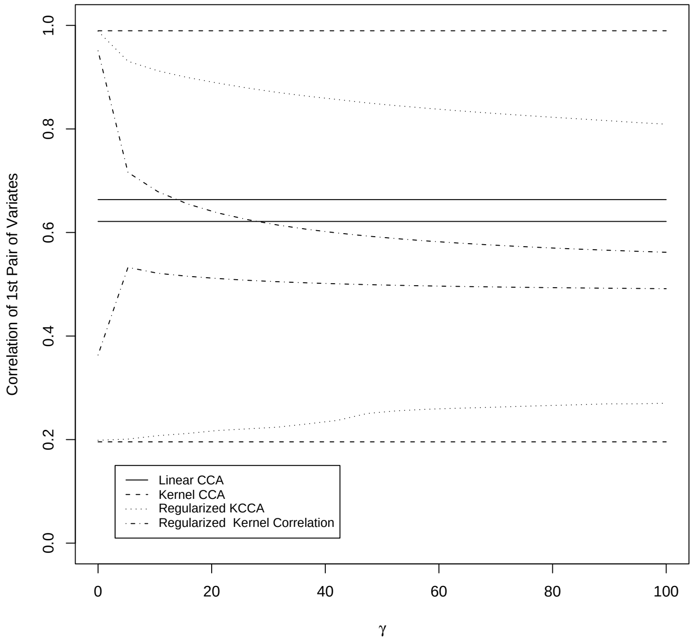

Figure 4: Example comparison of CCA variants. The dataset consists of 88 observations on 5 variables of which the first two constitute $\mathbf{X}$ and the remaining three $\mathbf{Y}$. For $\mathbf{X}$ a RBF kernel $(\sigma = 1)$ and for $\mathbf{Y}$ a polynomial kernel $(d = 4)$ was used. The plot shows correlation coefficients of the obtained variates with respect to a ridge parameter $\gamma = \gamma_{\mathcal{X}} = \gamma_{\mathcal{Y}}$. A cross validation procedure was used and the correlation coefficients were averaged. The respective upper line shows the averaged correlation of the first pair of variates constructed from the training sets while the lower lines give the correlation of features constructed from the test sets.

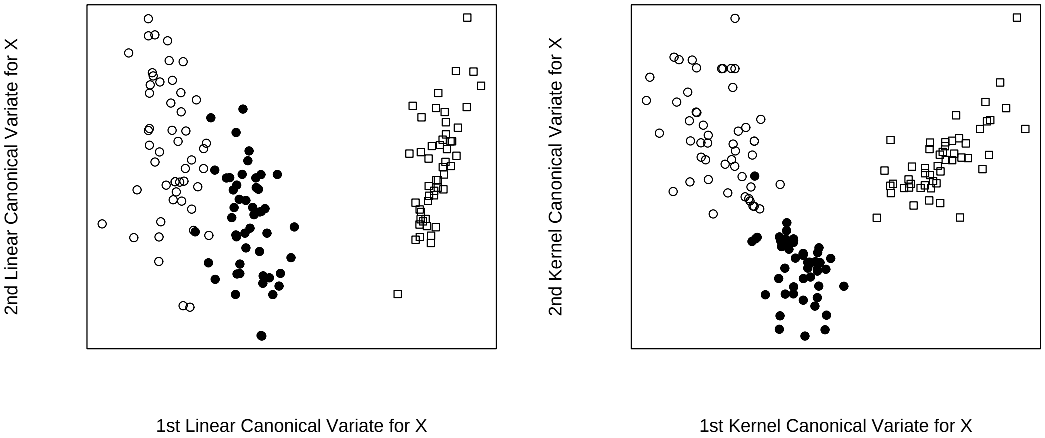

Figure 5: Kernel Fisher discriminant analysis as special case of KCCA: For illustration purposes we used Fisher's famous IRIS data set consisting of 4 measurements on 150 flowers taken from three different iris populations ("Iris setosa" (squares), "Iris versicolor" (dots), "Iris virginica" (circles)). The plots show the first two canonical variates $\mathbf{a}_1$ and $\mathbf{a}_2$ found by kernel canonical correlation between $\mathbf{\Phi}_{\mathcal{X}}$ and an indicator matrix $\mathbf{Y}$ (22). First we used a linear kernel and obtained the well known linear discriminant solution. For the second plot we used a homogeneous polynomial kernel $(d = 4)$.

for several lags τ. Afterwards the matrix of canonical vectors for $\mathbf{X}$ is used as an estimator for $\mathbf{S}^{-1}$ showing notable performance. Using KCCA, a nonlinear transformation of the data can be incorporated into this method. However, in numerous experiments for nonlinear mixtures, it proved to be difficult to find a kernel which only approximately unmixed the signals.

Regularized kernel correlation has recently been used as criterion of independence in kernel approaches to independent component analysis methods (Bach and Jordan 2002). The basic idea is that independence is equivalent to uncorrelatedness under all continuous transformations of the random variables. Instead of considering all continuous transformations the criterion is approximated by regularized kernel canonical correlation on transformations of the random variables restricted to the function space induced by the kernel. An early reference in this context is Hannan (1961).

6 Discussion

As shown, canonical correlations between kernel feature spaces can be exactly analyzed. Geometric concepts can be used to interpret the canonical solution. In general, relations like $\mathcal{L}\{\mathbf{\Phi}_{\mathcal{X}}\} = \mathcal{L}\{\mathbf{K}_{\mathcal{X}}\}$ illustrate that solutions of kernel variants of linear algorithms can be geometrically identical to solutions of the corresponding original linear algorithm by simply using kernel principal component transformed data. Previous approaches did not consider the geometry of CCA, e.g. Lai and Fyfe (2000), and the proposed methods were similar to regularized kernel correlation (van Gestel et al. 2001; Melzer et al. 2001; Bach and Jordan 2002).

The tendency of KCCA to overfit the data and numerical difficulties suggest the use of a regularized approximative variant. We described regularized kernel correlation and a regularized form of KCCA, which gave higher correlated features on training and often on test data.

Kernel principal component regression and an elegant formulation of multicategory kernel discriminant analysis can be shown to be special cases of the proposed methods. Note that while this article only considered CCA between two sets of variables, a generalization towards more than two sets can be constructed as described by Kettenring (1971) using kernel principal component scores instead of the raw input space data.

References

Afriat, S. N. (1957). Orthogonal and oblique projectors and the characteristics of pairs of vector spaces. *Proceedings of the Cambridge Philosophical Society 53*(4), 800–816.

Bach, F. R. and M. I. Jordan (2002). Kernel independent component analysis. *Journal of Machine Learning Research 3*, 1–48.

Björck, A. and G. H. Golub (1973). Numerical methods for computing angles between linear subspaces. *Mathematics of Computation 27*(123), 579–594.

Borga, M. and H. Knutsson (2001). A canonical correlation approach to blind source separation. Technical Report LiU-IMT-EX-0062, Department of Biomedical Engineering, Linköping University, Sweden.

Gittins, R. (1985). *Canonical Analysis - A review with applications in ecology*. Berlin: Springer.

Hannan, E. J. (1961). The general theory of canonical correlation and its relation to functional analysis. *The Journal of the Australian Mathematical Society 2*, 229–242.

Harville, D. A. (1997). *Matrix Algebra From a Statistican's Perspective*. New York: Springer.

Hotelling, H. (1935). The most predictable criterion. *The Journal of Educational Psychology 26*(2), 139–143.

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

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika 58*(3), 433–451.

Khatri, C. G. (1976). A note on multiple and canonical correlation for a singular covariance matrix. *Psychometrika 41*(4), 465–470.

Kockelkorn, U. (2000). *Lineare Statistische Methoden*. München: Oldenbourg.

Lai, P. L. and C. Fyfe (2000). Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems 10*(5), 365–377.

Mardia, K. V., J. Kent, and J. M. Bibby (1979). *Multivariate Analysis*. London: Academic Press.

Melzer, T., M. Reiter, and H. Bischof (2001). Nonlinear feature extraction using generalized canonical correlation analysis. In G. Dorffner, H. Bischof, and K. Hornik (Eds.), *Proceedings of the International Conference on Artificial Neural Networks*, Berlin, pp. 353–360. Springer.

Mika, S., G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller (1999). Fisher discriminant analysis with kernels. In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas (Eds.), *Neural Networks for Signal Processing IX*, pp. 41–48. IEEE.

Rosipal, R. and L. J. Trejo (2001). Kernel partial least squares regression in reproducing kernel Hilbert space. *Journal of Machine Learning Research 2*, 97–123.

Schölkopf, B., A. Smola, and K.-R. Müller (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation 10*, 1299–1319.

Schölkopf, B. and A. J. Smola (2002). *Learning with Kernels*. Cambridge, Massachusetts: The MIT press.

van Gestel, T., J. A. K. Suykens, J. D. Brabanter, B. D. Moor, and J. Vandewalle (2001). Kernel canonical correlation analysis and least squares support vector machines. In G. Dorffner, H. Bischof, and K. Hornik (Eds.), *Proceedings of the International Conference on Artificial Neural Networks*, Berlin, pp. 381–386. Springer.

Vinod, H. D. (1976). Canonical ridge and econometrics of joint production. *Journal of Econometrics 4*(2), 147–166.

Weston, J., O. Chapelle, A. Elisseeff, B. Schölkopf, and V. Vapnik (2002). Kernel dependency estimation. Technical Report 098, Max Planck Institute for Biological Cybernetics, Tübingen, Germany.