STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 301—310

METHODS OF REPRESENTATION FOR KERNEL CANONICAL CORRELATION ANALYSIS

Mirosław Krzyśko[1], Łukasz Waszak[2]

ABSTRACT

Classical canonical correlation analysis seeks the associations between two data sets, i.e. it searches for linear combinations of the original variables having maximal correlation. Our task is to maximize this correlation. This problem is equivalent to solving the generalized eigenvalue problem. The maximal correlation coefficient (being a solution of this problem) is the first canonical correlation coefficient. In this paper we construct nonlinear canonical correlation analysis in reproducing kernel Hilbert spaces. The new kernel generalized eigenvalue problem always has the solution equal to one, and this is a typical case of over-fitting. We present methods to solve this problem and compare the results obtained by classical and kernel canonical correlation analysis.

Key words: Canonical correlation analysis, generalized eigenvalue problem, reproducing kernel Hilbert space.

1. Introduction

The classical tool for studying the association between a dependent variable Y and a set of p explanatory variables $\boldsymbol{X} = (X_1, \dots, X_p)'$ is multiple regression. Often we are interested in a more complicated interaction, i.e. an interaction between a set of q dependent variables $\boldsymbol{Y} = (Y_1, \dots, Y_q)$ and a set of p explanatory variables $\boldsymbol{X} = (X_1, \dots, X_p)'$. This method was proposed by Hotelling (1936), and is referred to in the literature as canonical correlation analysis.

Our task is to find the strength of association between two vectors $\boldsymbol{Y}$ and $\boldsymbol{X}$. For this purpose we construct new variables V and U, being a linear combination of the original vectors $\boldsymbol{Y}$ and $\boldsymbol{X}$, i.e. $U = \boldsymbol{a}'\boldsymbol{X}$ and $V = \boldsymbol{b}'\boldsymbol{Y}$, where $\boldsymbol{a} \in \mathrm{R}^p, \boldsymbol{b} \in \mathrm{R}^q$. The variables U and V obtained in this way are real one-

1 Faculty of Mathematics and Computer Science at Adam Mickiewicz University in Poznań, mkrzysko@amu.edu.pl.
2 Faculty of Mathematics and Computer Science at Adam Mickiewicz University in Poznań, lwaszak@amu.edu.pl.

dimensional variables ($U,V \in \mathbf{R}$). The dependence between $\boldsymbol{Y}$ and $\boldsymbol{X}$ can be expressed as the classical correlation coefficient $\rho(U,V)$ between V and U.

2. Classical Canonical Correlation Analysis (CCA)

Suppose $\tilde{\boldsymbol{X}} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_N) \in \mathrm{R}^{p\times N}$ and $\tilde{\boldsymbol{Y}} = (\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_N) \in \mathrm{R}^{p\times N}$ are two centred (standardized) N-element data sets, i.e. $E\tilde{\boldsymbol{X}} = \boldsymbol{0}_N$, $Cov\tilde{\boldsymbol{X}} = \boldsymbol{1}_N$ and $\boldsymbol{X}_i = (X_{1i}, \dots, X_{pi})' \in \mathrm{R}^p$, $\boldsymbol{Y}_i = (Y_{1i}, \dots, Y_{qi})' \in \mathrm{R}^q$, for $i = 1, \dots, N$.

The classical task of canonical correlation analysis (CCA) is to find the linear combinations of original vectors $\boldsymbol{Y} \in \tilde{\boldsymbol{Y}}$ and $\boldsymbol{X} \in \tilde{\boldsymbol{X}}$ maximizing $\rho(U,V)$, i.e.:

$$\max \rho(U,V),$$

where $U = \boldsymbol{a}'\boldsymbol{X}$ and $V = \boldsymbol{b}'\boldsymbol{Y}$, where $\boldsymbol{a} \in \mathrm{R}^p, \boldsymbol{b} \in \mathrm{R}^q$. Obviously:

$$\rho(U,V) = \frac{cov(U,V)}{\sqrt{Var(U)Var(V)}},$$

$$cov(U,V) = E(UV) - E(U)E(V),$$
$$Var(U) = E(U^2) - E^2(U),$$
$$Var(V) = E(V^2) - E^2(V).$$

Because $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ are centred, $E(U) = 0 = E(V)$. Moreover, it can be assumed without loss of generality that $E(U^2) = 1 = E(V^2)$. Consequently we obtain $\rho(U,V) = E(UV)$. Now we want to maximize $E(UV)$, i.e.:

$$\max \rho(U,V) = \max E(UV)$$

To maximize $E(UV)$ we construct a Lagrangian on $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$:

$$F(\boldsymbol{A},\boldsymbol{B}) = \boldsymbol{A}'\boldsymbol{S}_{XY}\boldsymbol{B} - \frac{\lambda}{2}(\boldsymbol{A}'\boldsymbol{S}_{XX}\boldsymbol{A} - N) - \frac{\mu}{2}(\boldsymbol{B}'\boldsymbol{S}_{YY}\boldsymbol{B} - N) \tag{1}$$

where $\boldsymbol{A} = (\boldsymbol{a}_i)$, $\boldsymbol{a}_i \in \mathbb{R}^p, \boldsymbol{B} = (\boldsymbol{b}_i)$, $\boldsymbol{b}_i \in \mathbb{R}^q, i = 1, \dots, N$ and $\boldsymbol{S}_{XX} = \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}'$, $\boldsymbol{S}_{YY} = \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}'$, $\boldsymbol{S}_{XY} = \tilde{\boldsymbol{X}}\tilde{\boldsymbol{Y}}' = \boldsymbol{S}_{YX}'$ and $\frac{\lambda}{2}, \frac{\mu}{2}$ are Lagrange multipliers.

Taking the derivates of the components of vectors $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ and equating to zero we obtain:

$$\frac{\partial F}{\partial \boldsymbol{A}} = \boldsymbol{S}_{XY}\boldsymbol{B} - \lambda\boldsymbol{S}_{XX}\boldsymbol{A} = \boldsymbol{0} \tag{2}$$
$$\frac{\partial F}{\partial \boldsymbol{B}} = \boldsymbol{S}_{XY}'\boldsymbol{A} - \mu\boldsymbol{S}_{YY}\boldsymbol{B} = \boldsymbol{0} \tag{3}$$

where $\frac{\partial F}{\partial \boldsymbol{A}} = \left(\frac{\partial F}{\partial \boldsymbol{a}_i}\right), \frac{\partial F}{\partial \boldsymbol{B}} = \left(\frac{\partial F}{\partial \boldsymbol{b}_i}\right)$.

Because $Var(U) = Var(V) = 1$, we arrive at $\lambda = \mu = \boldsymbol{A}'\boldsymbol{S}_{XY}\boldsymbol{B} = \rho$. Therefore (2) and (3) are equivalent respectively to (4) and (5):

$$\boldsymbol{S}_{XY}\boldsymbol{B} = \rho \boldsymbol{S}_{XX}\boldsymbol{A} \tag{4}$$

$$\boldsymbol{S}'_{XY}\boldsymbol{A} = \rho \boldsymbol{S}_{YY}\boldsymbol{B} \tag{5}$$

Using a matrix representation, equations (4) and (5) are equivalent to:

$$\boldsymbol{S}_1\boldsymbol{C} = \rho \boldsymbol{S}_2\boldsymbol{C} \tag{6}$$

where

$$\boldsymbol{S}_1 = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{S}_{XY} \\ \boldsymbol{S}'_{XY} & \boldsymbol{0} \end{bmatrix},\ \boldsymbol{S}_2 = \begin{bmatrix} \boldsymbol{S}_{XX} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}_{YY} \end{bmatrix},\ \boldsymbol{C} = (\boldsymbol{c}_i),\ \boldsymbol{c}_i = \begin{bmatrix} \boldsymbol{a}_i \\ \boldsymbol{b}_i \end{bmatrix}.$$

Equation (6) is a classical case of the generalized eigenvalue problem. The matrix $\boldsymbol{S}_2$ is non-singular and can be presented as the classical eigenvalue problem:

$$(\boldsymbol{S} - \rho \boldsymbol{I})\boldsymbol{C} = \boldsymbol{0} \tag{7}$$

where $\boldsymbol{S} = \boldsymbol{S}_2^{-1}\boldsymbol{S}_1$.

The solutions $\rho_i(\boldsymbol{S})$ of equation (7) $\det(\boldsymbol{S} - \rho \boldsymbol{I}) = 0$ determine the *i*-th canonical correlation coefficient, and the corresponding vector $\boldsymbol{c}_i(\boldsymbol{S})$ the *i*-th pair of canonical variables $(U_i, V_i) = (\boldsymbol{a}_i{}'\boldsymbol{X}, \boldsymbol{b}_i{}'\boldsymbol{Y})$.

3. Introduction to reproducing kernel Hilbert spaces (RKHS)

We introduce some facts about reproducing kernel Hilbert spaces (RKHS) which will be used in our analysis (Preda (2006)). Let $H \subseteq \mathrm{R}^d$ be a set and $\mathcal{H}$ be a Hilbert space of functions on H. Denote by $\langle\cdot,\cdot\rangle$ the inner product of $\mathcal{H}$. A bivariate real-valued function $\boldsymbol{K}$ on H is said to be a reproducing kernel for $\mathcal{H}$ if:
(RK1) $\forall x \in H$: $\boldsymbol{K}(\cdot, x) \in H$
(RK2) (reproducing property) $\forall x \in H\ \forall F \in \mathcal{H}$: $F(x) = \langle F, \boldsymbol{K}(\cdot, x)\rangle$.

If $\mathcal{H}$ admits an reproducing kernel $\boldsymbol{K}$, then $\boldsymbol{K}$ has the following properties (Aronszajn, 1950):
(K1) $\boldsymbol{K}$ is the unique reproducing kernel for $\mathcal{H}$.
(K2) $\boldsymbol{K}$ is symmetric and non-negative definite.
(K3) Elements of the form
$\sum_{i=1}^{n} a_i \boldsymbol{K}(t_i, \cdot),\ n \in \mathrm{N},\ \{a_i \in \mathrm{R},\ i = 1, \dots, n\}, \{t_i \in H,\ i = 1, \dots, n\}$ are dense in $\mathcal{H}$.

In view of *(K3)*, if $\boldsymbol{K}$ is a symmetric and non-negative definite function, one can construct a Hilbert space $\mathcal{H}_{\boldsymbol{K}}$ which is the completion of all functions on H of the form $\sum_{i=1}^{n} a_i \boldsymbol{K}(t_i,\cdot)$ under the inner product:

$$\langle \sum_{i=1}^{n} a_i \boldsymbol{K}(t_i,\cdot), \sum_{i=1}^{n} b_j \boldsymbol{K}(s_j,\cdot) \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \boldsymbol{K}(t_i, s_j)$$

Thus, $\mathcal{H}_{\boldsymbol{K}}$ is an RKHS with reproducing kernel $\boldsymbol{K}$ and we have the well known result (Aronszajn, 1950):

Moore–Aronszajn Theorem. To every non-negative definite function $\boldsymbol{K}$ *on* $H \times H$ *there is a corresponding unique RKHS* $\mathcal{H}_{\boldsymbol{K}}$ *of real-valued functions on* H *and vice versa.*

Examples of reproducing kernels:

Gaussian kernel: $\boldsymbol{k}_G(\boldsymbol{x},\boldsymbol{y}) = \exp(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2}) = \exp(-c \parallel \boldsymbol{x}-\boldsymbol{y} \parallel^2), c = \frac{1}{2\sigma^2} > 0.$

Polynomial kernel: $\boldsymbol{k}_P(\boldsymbol{x},\boldsymbol{y}) = (b\langle \boldsymbol{x},\boldsymbol{y}\rangle + c\,)^d = (b\boldsymbol{x}'\boldsymbol{y} + c)^d,\ b, c \in \mathrm{R}, d \in \mathrm{N}.$

Theorem. Let $F \in \mathcal{H}$ *and* $\widetilde{\boldsymbol{K}}$ *be the kernel matrix for non-centred data. Then the kernel matrix* $\boldsymbol{K}$ *for centred data is expressed by the formula:*

$$\boldsymbol{K} = \boldsymbol{P}\widetilde{\boldsymbol{K}}\boldsymbol{P},$$

where $\boldsymbol{P} = \boldsymbol{I}_N - \frac{1}{n}\boldsymbol{1}_N\boldsymbol{1}_N'$ *and* $\boldsymbol{1}_N$ *is N-dimensional vector consisting of one.*

4. Kernel Canonical Correlation Analysis (KCCA)

Let space $\mathbb{R}^p$ be mapped to a reproducing kernel Hilbert space: $\boldsymbol{\varphi}: \mathbb{R}^p \longrightarrow H(k_x), \boldsymbol{X}_i \mapsto \boldsymbol{\varphi}(\boldsymbol{X}_i), i = 1, \dots, N,$ and space $\mathbb{R}^q$ to a reproducing kernel Hilbert space $H(k_y)$: $\boldsymbol{\psi}: \mathbb{R}^q \longrightarrow H(k_y), \boldsymbol{Y}_i \mapsto \boldsymbol{\psi}(\boldsymbol{Y}_i), i = 1, \dots, N.$ We formulate a new CCA task on the sets $\boldsymbol{\varphi}(\widetilde{\mathbb{X}})$ and $\boldsymbol{\psi}(\widetilde{\mathbb{Y}})$, i.e. on elements $\boldsymbol{\varphi}(\boldsymbol{X}_i) \in \boldsymbol{\varphi}(\widetilde{\mathbb{X}})$ and $\boldsymbol{\psi}(\boldsymbol{Y}_i) \in \boldsymbol{\psi}(\widetilde{\mathbb{Y}})$, for $i = 1, \dots, N$, assuming $\boldsymbol{\varphi}(\widetilde{\mathbb{X}})$ and $\boldsymbol{\psi}(\widetilde{\mathbb{Y}})$ are centred. Then each of the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is in the subspaces stretched respectively on $\boldsymbol{\varphi}(\boldsymbol{X}_i)$ and on $\boldsymbol{\psi}(\boldsymbol{Y}_i)$:

$$\boldsymbol{a} = \sum_{i=1}^{N} \alpha_i \boldsymbol{\varphi}(\boldsymbol{X}_i)$$

$$\boldsymbol{b} = \sum_{i=1}^{N} \beta_i \boldsymbol{\psi}(\boldsymbol{Y}_i)$$

The kernelized problem of CCA has the same form (KCCA):

$$\max \rho(U,V) = \max E(UV)$$

where

$$U = \boldsymbol{a}'\boldsymbol{\varphi}(X) = \sum_{i=1}^{N} \alpha_i \boldsymbol{\varphi}'(X_i)\boldsymbol{\varphi}(X)$$

$$V = \boldsymbol{b}'\boldsymbol{\psi}(Y) = \sum_{i,j=1}^{N} \beta_i \boldsymbol{\psi}'(Y_i)\boldsymbol{\psi}(Y)$$

To maximize $E(UV)$ we construct analogously to CCA a new Lagrangian on $\boldsymbol{\varphi}(\widetilde{X})$ and $\boldsymbol{\psi}(\widetilde{Y})$:

$$F_H(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{N}\left(\sum_{j=1}^{N} \alpha_j \boldsymbol{\varphi}'(X_j)\right)\boldsymbol{\varphi}(X_i)\boldsymbol{\psi}'(Y_i)\left(\sum_{k=1}^{N} \beta_k\, \boldsymbol{\psi}(Y_k)\right)$$

$$-\frac{\lambda}{2}\left[\sum_{i=1}^{N}\left(\sum_{j=1}^{N} \alpha_j \boldsymbol{\varphi}'(X_j)\right)\boldsymbol{\varphi}(X_i)\boldsymbol{\varphi}'(X_i)\left(\sum_{j=1}^{N} \alpha_j \boldsymbol{\varphi}(X_i)\right) - N\right]$$

$$-\frac{\mu}{2}\left[\sum_{i=1}^{N}\left(\sum_{j=1}^{N} \beta_j \boldsymbol{\psi}'(Y_j)\right)\boldsymbol{\psi}(Y_i)\boldsymbol{\psi}'(Y_i)\left(\sum_{k=1}^{N} \beta_k\, \boldsymbol{\psi}(Y_k)\right) - N\right] \quad (8)$$

Using the Moore-Aronszajn theorem and kernel trick, i.e. $\boldsymbol{k}(\boldsymbol{X}_i, \boldsymbol{X}_j) = \langle \boldsymbol{\varphi}(\boldsymbol{X}_i), \boldsymbol{\varphi}(\boldsymbol{X}_j)\rangle$ and $\boldsymbol{k}(\boldsymbol{Y}_i, \boldsymbol{Y}_j) = \langle \boldsymbol{\psi}(\boldsymbol{Y}_i), \boldsymbol{\psi}(\boldsymbol{Y}_j)\rangle$, (8) can be presented as:

$$F_H(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{\alpha}' \boldsymbol{K}_X \boldsymbol{K}_Y\, \boldsymbol{\beta} - \frac{\lambda}{2}\left(\boldsymbol{\alpha}' \boldsymbol{K}_X^2 \boldsymbol{\alpha} - N\right) - \frac{\mu}{2}\left(\boldsymbol{\beta}' \boldsymbol{K}_Y^2 \boldsymbol{\beta} - N\right) \quad (9)$$

where $\boldsymbol{K}_X = [\boldsymbol{k}(\boldsymbol{X}_i, \boldsymbol{X}_j)]$ and $\boldsymbol{K}_Y = [\boldsymbol{k}(\boldsymbol{Y}_i, \boldsymbol{Y}_j)]$ are kernel matrices and where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_N)'$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_N)'$.

Taking the derivates of the components of vectors $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ and equating to zero we obtain:

$$\frac{\partial F_H}{\partial \boldsymbol{\alpha}} = \boldsymbol{K}_X \boldsymbol{K}_Y\, \boldsymbol{\beta} - \lambda \boldsymbol{K}_X^2 \boldsymbol{\alpha} = \boldsymbol{0} \quad (10)$$

$$\frac{\partial F_H}{\partial \boldsymbol{\beta}} = \boldsymbol{K}_Y \boldsymbol{K}_X \boldsymbol{\alpha} - \mu \boldsymbol{K}_Y^2 \boldsymbol{\beta} = \boldsymbol{0} \quad (11)$$

where $\frac{\partial F}{\partial \boldsymbol{\alpha}} = \left(\frac{\partial F}{\partial \alpha_i}\right), \frac{\partial F}{\partial \boldsymbol{\beta}} = \left(\frac{\partial F}{\partial \beta_i}\right)$.

Analogously, we arrive at $\lambda = \mu = \boldsymbol{\alpha}' \boldsymbol{K}_X \boldsymbol{K}_Y\, \boldsymbol{\beta} = \rho$. Therefore, (10) and (11) are equivalent to (12) and (13) respectively:

$$\boldsymbol{K}_X \boldsymbol{K}_Y\, \boldsymbol{\beta} = \rho \boldsymbol{K}_X^2 \boldsymbol{\alpha} \quad (12)$$

$$\boldsymbol{K}_Y \boldsymbol{K}_X \boldsymbol{\alpha} = \rho \boldsymbol{K}_Y^2 \boldsymbol{\beta} \quad (13)$$

Using a matrix representation, equations (12) and (13) are equivalent to:

$$\boldsymbol{K}_1\boldsymbol{C}(\boldsymbol{k}) = \rho(\boldsymbol{k})\boldsymbol{K}_2\boldsymbol{C}(\boldsymbol{k}) \tag{14}$$

where

$$\boldsymbol{K}_1 = \begin{bmatrix} \mathbf{0} & \boldsymbol{K}_X\boldsymbol{K}_Y \\ \boldsymbol{K}_Y\boldsymbol{K}_X & \mathbf{0} \end{bmatrix},\ \boldsymbol{K}_2 = \begin{bmatrix} \boldsymbol{K}_X^2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{K}_Y^2 \end{bmatrix},\ \boldsymbol{C}(\boldsymbol{k}) = (\boldsymbol{c}_i(\boldsymbol{k})), \boldsymbol{c}_i(\boldsymbol{k}) = \begin{bmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \end{bmatrix}.$$

Equation (14) is the generalized eigenvalue problem and could again be presented as the classical eigenvalue problem:

$$(\boldsymbol{K} - \rho(\boldsymbol{k})\boldsymbol{I})\boldsymbol{C}(\boldsymbol{k}) = \mathbf{0} \tag{15}$$

where $\boldsymbol{K} = \boldsymbol{K}_2^{-1}\boldsymbol{K}_1$. The matrix $\boldsymbol{K}_2$ is non-negative determined and can include singularity values. The generalized Moore–Penrose pseudoinverse of the matrix is ambiguous.

In order to solve this problem we first apply the idea used in **ridge regression**, namely regularization of the matrix $\boldsymbol{K}_2$, i.e.

$$\boldsymbol{K}_2 \mapsto \boldsymbol{K}_2 + \varepsilon\boldsymbol{I} \tag{16}$$

where $\varepsilon > 0$; it is enough to take $\varepsilon = 10^{-5}$.

Consequently, the matrix $\boldsymbol{K}$ has the form:

$$\boldsymbol{K} = (\boldsymbol{K}_2 + \varepsilon\boldsymbol{I})^{-1}\boldsymbol{K}_1 \tag{17}$$

A second concept for solving this problem is the idea of using **SVD decomposition** of the matrix $\boldsymbol{K}_2$, i.e.

$$\boldsymbol{K}_2 = \boldsymbol{QGQ}' \tag{18}$$

where $\boldsymbol{Q}$ is the orthogonal ($\boldsymbol{QQ}' = \boldsymbol{I}$) matrix and $\boldsymbol{G}$ is the diagonal matrix of degree t, where t is the number of eigen-values of matrix $\boldsymbol{K}_2$ greater then 10^{-6}.

As a result, we obtain the matrix $\boldsymbol{K}$:

$$\boldsymbol{K} = \boldsymbol{K}_2^{-1}\boldsymbol{K}_1 = (\boldsymbol{QGQ}')^{-1}\boldsymbol{K}_1 = \boldsymbol{QG}^{-1}\boldsymbol{Q}'\boldsymbol{K}_1 \tag{19}$$

The solutions $\rho_i(\boldsymbol{K})$ of equation (15) $\det(\boldsymbol{K} - \rho\boldsymbol{I}) = 0$ determine the *i*-th kernel canonical correlation coefficient, and the corresponding vector $\boldsymbol{c}_i(\boldsymbol{K})$ the *i*-th pair of kernel canonical variables $(U_i, V_i) = (\boldsymbol{a}_i\boldsymbol{K}_X, \boldsymbol{b}_i\boldsymbol{K}_Y)$.

5. Quasi Kernel Canonical Correlation Analysis (Q-KCCA)

In this case let only the space R^p be mapped to a reproducing kernel Hilbert space: $\boldsymbol{\varphi}: \mathrm{R}^p \longrightarrow H(k_x)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_q) \in \mathrm{R}^q$ (Zheng et al. (2006)). We present a concatenation of the methods proposed in section 3 and 4.

$$U = \alpha' \varphi(X) = \sum_{i=1}^{N} \alpha_i \varphi'(X_i)\varphi(X)$$
$$V_k = b'Y = \sum_{i=1}^{N} b_i Y_i$$

Using the Moore-Aronszajn theorem and kernel trick again, i.e. $k(X_i, X_j) = \langle \varphi(X_i), \varphi(X_j) \rangle$, we obtain:

$$F_H(\alpha, B) = \alpha' K_X \tilde{Y} B - \frac{\lambda}{2}\left(\alpha' K_X^2 \alpha - N\right) - \frac{\mu}{2}\left(B' \tilde{Y}\tilde{Y}' B - N\right) \quad (20)$$

Taking the derivates of the components of vectors α_i and b_i and equating to zero we obtain:

$$\frac{\partial F_H}{\partial \alpha} = K_X \tilde{Y} B - \lambda K_X^2 \alpha = 0 \quad (21)$$

$$\frac{\partial F_H}{\partial B} = \tilde{Y}' K_X \alpha - \mu \tilde{Y}\tilde{Y}' B = 0 \quad (22)$$

where $\frac{\partial F}{\partial \alpha} = \left(\frac{\partial F}{\partial \alpha_i}\right), \frac{\partial F}{\partial B} = \left(\frac{\partial F}{\partial b_i}\right)$.

Therefore, (21) and (22) are equivalent to (23) and (24) respectively:

$$K_X \tilde{Y}\, b = \lambda K_X^2 \alpha \quad (23)$$

$$\tilde{Y}' K_X \alpha = \mu \tilde{Y}\tilde{Y}' b \quad (24)$$

Using a matrix representation, equations (23) and (24) are equivalent to:

$$A_1 A(k) = \rho(k) A_2 A(k) \quad (25)$$

where

$$A_1 = \begin{bmatrix} 0 & K_X \tilde{Y} \\ \tilde{Y}' K_X & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} K_X^2 & 0 \\ 0 & \tilde{Y}\tilde{Y}' \end{bmatrix}, \quad A(k) = (a_i(k)),\ a_i(k) = \begin{bmatrix} \alpha_i \\ b_i \end{bmatrix}.$$

Again, we can transform problem (25) again into the classical eigenvalue problem:

$$(A - \rho(k) I) A(k) = 0 \quad (26)$$

Using regularization of the matrix K_X^2 we obtain:

$$A = \begin{bmatrix} K_X^2 + \varepsilon I & 0 \\ 0 & \tilde{Y}' \end{bmatrix}^{-1} \begin{bmatrix} 0 & K_X \tilde{Y}' \\ \tilde{Y} K_X & 0 \end{bmatrix} \quad (27)$$

Analogously, using SVD decomposition of the matrix K_X^2 we obtain

$$A = \begin{bmatrix} Q G^{-1} Q' & 0 \\ 0 & \left(\tilde{Y}\tilde{Y}'\right)^{-1} \end{bmatrix} \begin{bmatrix} 0 & K_X \tilde{Y} \\ \tilde{Y}' K_X & 0 \end{bmatrix} \quad (28)$$

6. Example - Car Marks Data Set

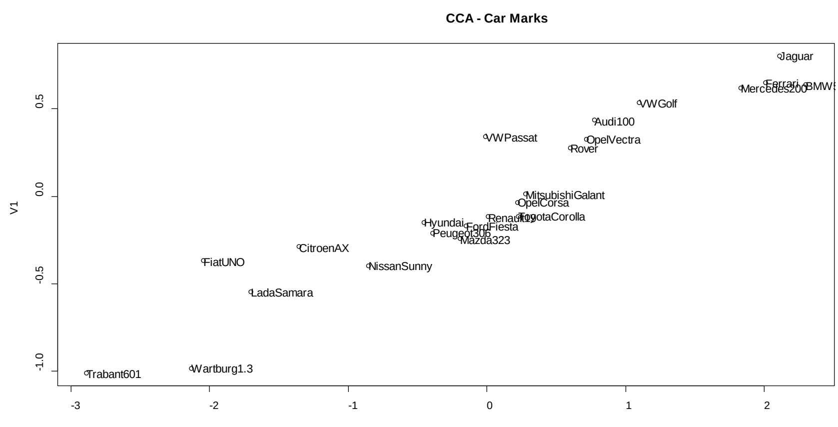
These data come from the book *Applied Multivariate Statistical Analysis* by Wolfgang Härdle and Léopold Simar (2007). They are averaged marks for 24 car types from a sample of 40 persons. The marks range from 1 (very good) to 6 (very bad), on the pattern of German school marks. The variables are (X_1) Economy, (X_2) Service, (X_3) Non-depreciation of value, (X_4) Price (mark 1 for very cheap cars), (X_5) Design, (X_6) Sporty car, (X_7) Safety, (X_8) Easy handling. The exact description of the data set is found in the book in section B.7 Car Marks (p. 434-435). In particular, we would like to investigate the relation between the two variables representing non-depreciation of value and price of the car, and all other variables, i.e. $\boldsymbol{Y} = (X_3, X_4)'$ and $\boldsymbol{X} = (X_1, X_2, X_5, X_6, X_7, X_8)'$. All variables are standardized.

In the classical case of CCA we obtain two non-zero correlation coefficients $\rho_1 = 0.9792$ and $\rho_2 = 0.8851$. For the largest correlation coefficient ρ_1 we obtain vectors $\boldsymbol{a}_1' = [-0.3632, 0.1504] \in \mathbf{R}^2$ and $\boldsymbol{b}_1' = [0.0039, 0.4574, 0.2429, 0.2064, 0.6216, -0.3850] \in \mathbf{R}^6$ which correspond to the canonical variables (U_1, V_1), where:

$$(U_1^i)_{i=1,\dots,=N} = (\boldsymbol{a}_1' \boldsymbol{X}_i)_{i=1,\dots,=N}$$

$$(V_1^i)_{i=1,\dots,=N} = (\boldsymbol{b}_1' \boldsymbol{Y}_i)_{i=1,\dots,=N}$$

A projection (CCA) into the coordinate system of the canonical variables corresponding to the canonical coefficients ρ_1 is shown below:



In the kernel case (KCCA) we can use the polynomial kernel $k_P(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}'\boldsymbol{y} + 1)^2$. We obtain 26 non-zero correlation coefficients. We present only the six largest:

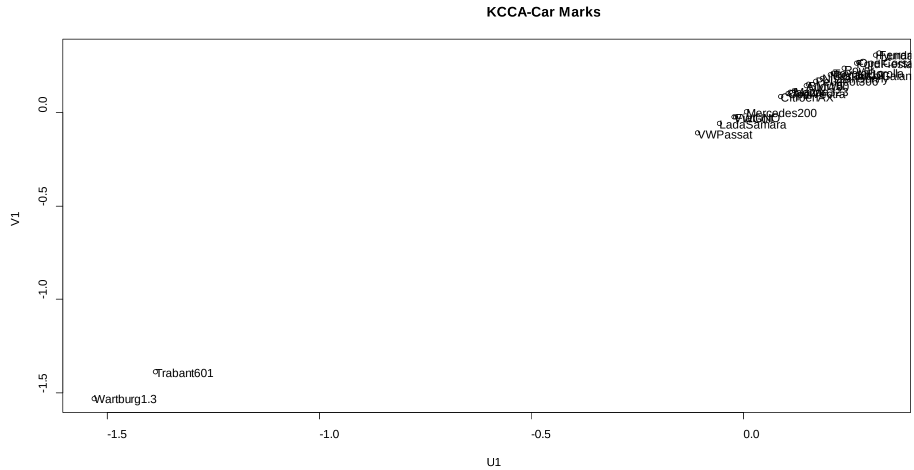
$\rho_1 = 1.000000, \rho_2 = 0.999998, \rho_3 = 0.999980, \rho_4 = 0.999385, \rho_5 = 0.998763, \rho_6 = 0.000002$.

For the largest correlation coefficient ρ_1 we obtain vectors $\boldsymbol{a}_1' = [0.0664, \dots, 0.0536] \in \mathbf{R}^{23}$ and $\boldsymbol{b}_1' = [-0.0234, \dots, 0.2440] \in \mathbf{R}^{23}$ which correspond to the canonical variables (U_1, V_1), where:

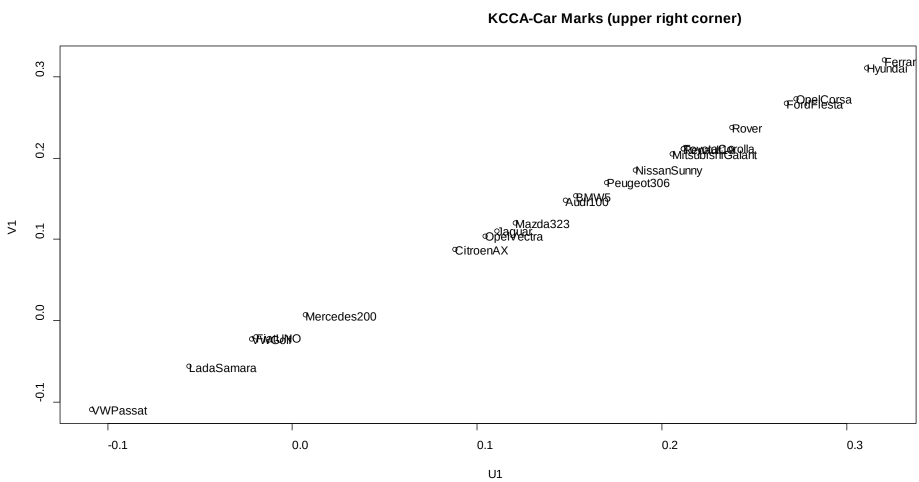
$$(U_1^i)_{i=1,\dots,=N} = \boldsymbol{a}_1' \boldsymbol{K}_X$$

$$(V_1^i)_{i=1,\dots,=N} = \boldsymbol{b}_1' \boldsymbol{K}_Y$$

A projection (KCCA) into the coordinate system of the canonical variables corresponding to the canonical coefficients ρ_1 is shown below:



Because in this image the elements in the upper right corner are not clearly visible, we present below a magnification of the upper part:

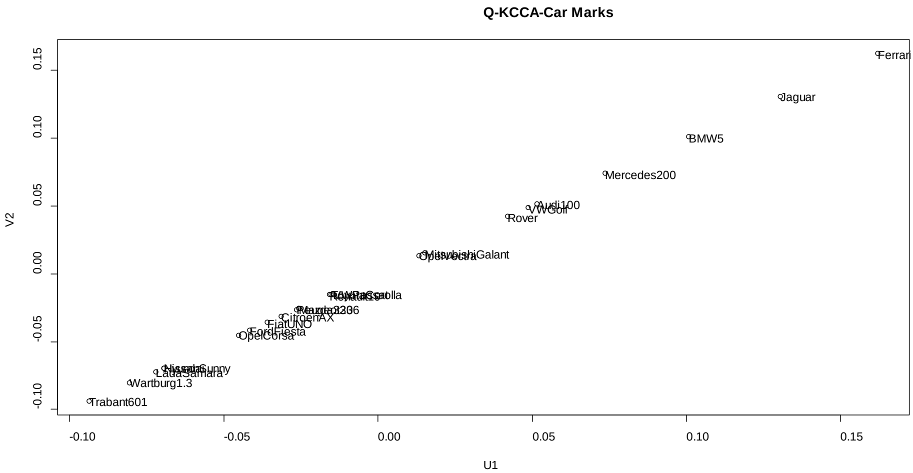


Finally, in the quasi kernel case (Q-KCCA) we obtain two non-zero correlation coefficients $\rho_1 = 0.99995$ and $\rho_2 = 0.99935$. For the largest correlation coefficient ρ_1 we obtain vectors $\boldsymbol{a}_1' = [-0.0916, \dots, 0.0153] \in \mathbb{R}^{23}$ and $\boldsymbol{b}_1' = [-0.0101, -0.0774] \in \mathbb{R}^2$ which correspond to the canonical variables (U_1, V_1), where:

$$(U_1^i)_{i=1,\dots,=N} = \boldsymbol{a}_1' \boldsymbol{K_X}$$

$$(V_1^i)_{i=1,\dots,=N} = (\boldsymbol{b}_1' \boldsymbol{Y}_i)_{i=1,\dots,=N}$$

A projection (Q-KCCA) into the coordinate system of the canonical variables corresponding to the canonical coefficients ρ_1 is shown below:



7. Conclusions

Kernel methods compared to the classical methods give a correlation coefficient close to 1, i.e. two data sets can be presented as 100% correlated data sets. In the discussed example it can be seen that the largest correlation coefficients for Q-KCCA and KCCA are similar (respectively 0.99995 and 1.00000). However, intuitively the projection into the coordinate system of the canonical variables (corresponding to the largest canonical coefficients) in the case of Q-KCCA is more similar to CCA then KCCA. Therefore, Q-KCCA can be recommended as one of the best methods in nonlinear canonical correlation analysis.

REFERENCES

ARONSZAJN, N. (1950): Theory of reproducing kernels, Transactions of the American Mathematical Society 68, 337–404.

HARDLE, W., SIMAR, L. (2007): Applied Multivariate Statistical Analysis, Springer, 321-330 and 434-435.

HOTELLING, H. (1936): Relation between two sets of variates, Biometrika 28, 321-377.

PREDA, C. (2006): Regression models for functional data by reproducing kernel Hilbert spaces methods, Journal of Statistical Planning and Inference 137, 831.

ZHENG, W., ZHOU, X., ZOU, C., ZHAO L. (2006): Facial Expression Recognition Using Kernel Canonical Correlation Analysis, IEEE Transaction on Neural Networks 17(1), 233.