(Kernel) Canonical Correlation Analysis

CS-E4830 - Kernel Methods in Machine Learning: Unsupervised Learning Algorithms 2

Viivi Uurtio

Department of Computer Science, Aalto University
Helsinki Institute for Information Technology HIIT

March 20, 2019

Aalto University
School of Science

HELSINKI
INSTITUTE FOR
INFORMATION
TECHNOLOGY

Canonical correlation methods find multivariate relations from two-view datasets

relation: a set of ordered pairs (2-tuples)

The set $\{(x, y) : x^2 + y^2 = 1\}$ is the set of all ordered pairs (x, y) for which $x^2 + y^2 = 1$.

multivariate relation: a set of ordered tuples of more than two elements

The set $\{(x, y, z) : x^2 + y^2 + z^2 = 1\}$ is the set of all ordered tuples (x, y, z) for which $x^2 + y^2 + z^2 = 1$.

function: a relation for which every element in the domain maps on a single element in the codomain

Example of a function: $(x, y) : y = x^2$.

Canonical correlation methods find multivariate relations from two-view datasets

view: a matrix in $\mathbb{R}^{n \times p}$ where n and p denote the observations and variables respectively

two-view dataset: every observation is described by $p + q$ variables, that is the dataset consists of two matrices $\mathbf{X}_a \in \mathbb{R}^{n \times p}$ and $\mathbf{X}_b \in \mathbb{R}^{n \times q}$

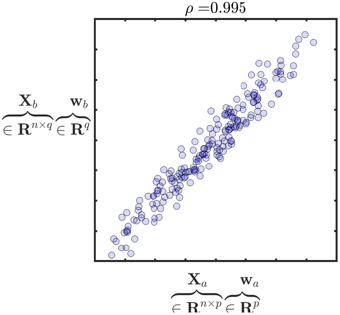
Given a set of $p + q$ variables, or vectors in $\mathbb{R}^n$, canonical correlation methods find subsets of variables that are ordered tuples

Canonical correlation methods find multivariate relations from two-view datasets

Example. Let $\mathbf{X}_a \in \mathbb{R}^{200\times10}$ and $\mathbf{X}_b \in \mathbb{R}^{200\times10}$ denote the two views of the data. Let the first two variables in view $\mathbf{X}_a$ be linearly related with the first two variables of view $\mathbf{X}_b$, that is $\mathbf{x}_b^1 + \mathbf{x}_b^2 = \mathbf{x}_a^1 + \mathbf{x}_a^2 + \boldsymbol{\xi}$ where $\boldsymbol{\xi}$ denotes normal noise.

The related variables are determined from the values of the entries of $\mathbf{w}_a$ and $\mathbf{w}_b$.

Canonical correlation methods find the $\mathbf{w}_a$ and $\mathbf{w}_b$ that maximize the canonical correlation ρ between $\mathbf{X}_a\mathbf{w}_a$ and $\mathbf{X}_b\mathbf{w}_b$.



This lecture covers standard canonical correlation analysis (CCA) and kernel CCA

CCA

$$\boxed{\max_{\mathbf{w}_a,\mathbf{w}_b} \quad \frac{\langle \mathbf{X}_a\mathbf{w}_a, \mathbf{X}_b\mathbf{w}_b\rangle}{||\mathbf{X}_a\mathbf{w}_a||_2||\mathbf{X}_b\mathbf{w}_b||_2}}$$

Kernel CCA

$$\boxed{\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad \frac{\langle \mathbf{K}^x\boldsymbol{\alpha}, \mathbf{K}^y\boldsymbol{\beta}\rangle}{\|\mathbf{K}^x\boldsymbol{\alpha}\|_2\|\mathbf{K}^y\boldsymbol{\beta}\|_2}}$$

$\rightarrow$ How these optimisation problems are solved

$\rightarrow$ How we interpret the solution to the CCA and KCCA problems

$\rightarrow$ When kernel CCA is more useful than CCA

A CCA model is assessed on test data

Training:

- Learn the coefficients
-Tune the hyperparameters

Testing:
Are the learnt relations predictive?

Learning methods:
Standard Eigenvalue Problem (CCA, Kernel CCA)
Generalised Eigenvalue Problem (CCA, Kernel CCA)
Singular Value Decomposition (CCA)

CCA is based on linear transformations

We only know the transformations (the data matrices)

$$\underbrace{\mathbf{X}_a}_{\mathbb{R}^{n\times p}}\underbrace{\mathbf{w}_a}_{\mathbb{R}^{p}} = \underbrace{\mathbf{z}_a}_{\mathbb{R}^{n}} \quad \underbrace{\mathbf{X}_b}_{\mathbb{R}^{n\times q}}\underbrace{\mathbf{w}_b}_{\mathbb{R}^{q}} = \underbrace{\mathbf{z}_b}_{\mathbb{R}^{n}}$$

We want to find the positions $\mathbf{w}_a$ and $\mathbf{w}_b$ and their images $\mathbf{z}_a$ and $\mathbf{z}_b$ such that the cosine of the angle between the images is maximized:

$$\cos\theta_r = \max\langle \mathbf{z}_a^r, \mathbf{z}_b^r\rangle$$

$$||\mathbf{z}_a^r||_2 = 1 \quad ||\mathbf{z}_b^r||_2 = 1$$

$$\langle \mathbf{z}_a^r, \mathbf{z}_a^j\rangle = 0 \quad \langle \mathbf{z}_b^r, \mathbf{z}_b^j\rangle = 0$$

$$\forall j \neq r : j, r = 1, 2, \ldots, \min(p, q)$$

Consecutive pairs of images with greater enclosing angles are found in the orthogonal complements.

The CCA problem can be formulated in terms of the data matrices and $\mathbf{w}_a$ and $\mathbf{w}_b$

Let $\mathbf{C}_{ab} = \frac{1}{n-1}\mathbf{X}_a^\top \mathbf{X}_b$, $\mathbf{C}_{ba} = \frac{1}{n-1}\mathbf{X}_b^\top \mathbf{X}_a$, $\mathbf{C}_{aa} = \frac{1}{n-1}\mathbf{X}_a^\top \mathbf{X}_a$, and $\mathbf{C}_{bb} = \frac{1}{n-1}\mathbf{X}_b^\top \mathbf{X}_b$ denote the empirical between-set and within-set covariance matrices.

We can formulate the CCA problem:

$$\cos\theta = \max_{\mathbf{z}_a, \mathbf{z}_b} \langle \mathbf{z}_a, \mathbf{z}_b \rangle = \max_{\mathbf{w}_a, \mathbf{w}_b} \mathbf{w}_a^\top \mathbf{C}_{ab} \mathbf{w}_b$$

$$||\mathbf{z}_a||_2 = \sqrt{\mathbf{w}_a \mathbf{C}_{aa} \mathbf{w}_a} = 1 \quad ||\mathbf{z}_b||_2 = \sqrt{\mathbf{w}_b \mathbf{C}_{bb} \mathbf{w}_b} = 1$$

The norm constraints are generally expressed in squared form, $\mathbf{w}_a \mathbf{C}_{aa} \mathbf{w}_a = 1$ and $\mathbf{w}_b \mathbf{C}_{bb} \mathbf{w}_b = 1$.

The standard eigenvalue problem is obtained through the Lagrange multiplier technique

Let $L = \mathbf{w}_a^\top \mathbf{C}_{ab}\mathbf{w}_b - \frac{\rho_1}{2}(\mathbf{w}_a^\top \mathbf{C}_{aa}\mathbf{w}_a - 1) - \frac{\rho_2}{2}(\mathbf{w}_b^\top \mathbf{C}_{bb}\mathbf{w}_b - 1)$, where ρ_1 and ρ_2 denote the Lagrange multipliers.

$$\frac{\delta L}{\delta \mathbf{w}_a} = \mathbf{C}_{ab}\mathbf{w}_b - \rho_1 \mathbf{C}_{aa}\mathbf{w}_a = \mathbf{0} \text{ and } \frac{\delta L}{\delta \mathbf{w}_b} = \mathbf{C}_{ba}\mathbf{w}_a - \rho_2 \mathbf{C}_{bb}\mathbf{w}_b = \mathbf{0}$$

$$\mathbf{w}_a^\top \mathbf{C}_{ab}\mathbf{w}_b - \rho_1 \mathbf{w}_a^\top \mathbf{C}_{aa}\mathbf{w}_a = 0 \text{ and } \mathbf{w}_b^\top \mathbf{C}_{ba}\mathbf{w}_a - \rho_1 \mathbf{w}_b^\top \mathbf{C}_{bb}\mathbf{w}_b = 0$$

Since $\mathbf{w}_a^\top \mathbf{C}_{aa}\mathbf{w}_a = \mathbf{w}_b^\top \mathbf{C}_{bb}\mathbf{w}_b = 1$ we have $\rho_1 = \rho_2 = \rho$.

Then $\mathbf{w}_a = \frac{\mathbf{C}_{aa}^{-1}\mathbf{C}_{ab}\mathbf{w}_b}{\rho}$ and $\frac{1}{\rho}\mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}\mathbf{C}_{ab}\mathbf{w}_b - \rho \mathbf{C}_{bb}\mathbf{w}_b = 0.$

If $\mathbf{C}_{bb}^{-1}$ is invertible we have $\mathbf{C}_{bb}^{-1}\mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}\mathbf{C}_{ab}\mathbf{w}_b = \rho^2 \mathbf{w}_b.$

The canonical correlations are the square roots of the eigenvalues of the matrix $\mathbf{C}_{bb}^{-1}\mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}\mathbf{C}_{ab}$, the eigenvectors correspond to $\mathbf{w}_b$ and the $\mathbf{w}_a = \frac{\mathbf{C}_{aa}^{-1}\mathbf{C}_{ab}\mathbf{w}_b}{\rho}$.

The generalised eigenvalue problem is obtained from simultaneous equations

$$\frac{\delta L}{\delta \mathbf{w}_a} = \mathbf{C}_{ab}\mathbf{w}_b - \rho_1 \mathbf{C}_{aa}\mathbf{w}_a = \mathbf{0} \text{ and } \frac{\delta L}{\delta \mathbf{w}_b} = \mathbf{C}_{ba}\mathbf{w}_a - \rho_2 \mathbf{C}_{bb}\mathbf{w}_b = \mathbf{0}$$

$$\mathbf{C}_{ab}\mathbf{w}_b = \rho \mathbf{C}_{aa}\mathbf{w}_a \text{ and } \mathbf{C}_{ba}\mathbf{w}_a = \rho \mathbf{C}_{bb}\mathbf{w}_b$$

$$\begin{pmatrix} \mathbf{0} & \mathbf{C}_{ab} \\ \mathbf{C}_{ba} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{pmatrix} = \rho \begin{pmatrix} \mathbf{C}_{aa} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{bb} \end{pmatrix} \begin{pmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{pmatrix}$$

The generalised eigenvalues come in pairs

$$\{\rho_1, -\rho_1, \rho_2, -\rho_2, \ldots, \rho_p, -\rho_p, 0\}$$

where $p < q$.

The positive generalised eigenvalues correspond to the canonical correlations.

The SVD can be applied on a rectangular matrix of covariances

$$\mathbf{C}_{aa} = \mathbf{C}_{aa}^{1/2}\mathbf{C}_{aa}^{1/2} \quad \text{and} \quad \mathbf{C}_{bb} = \mathbf{C}_{bb}^{1/2}\mathbf{C}_{bb}^{1/2}$$

$$\begin{pmatrix} \mathbf{C}_{aa}^{-1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{bb}^{-1/2} \end{pmatrix} \begin{pmatrix} \mathbf{C}_{aa} & \mathbf{C}_{ab} \\ \mathbf{C}_{ba} & \mathbf{C}_{bb} \end{pmatrix} \begin{pmatrix} \mathbf{C}_{aa}^{-1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{bb}^{-1/2} \end{pmatrix} =$$

$$\begin{pmatrix} \mathbf{I}_q & \mathbf{C}_{aa}^{-1/2}\mathbf{C}_{ab}\mathbf{C}_{bb}^{-1/2} \\ \mathbf{C}_{bb}^{-1/2}\mathbf{C}_{ba}\mathbf{C}_{aa}^{-1/2} & \mathbf{I}_p \end{pmatrix}$$

$$\mathbf{C}_{aa}^{-1/2}\mathbf{C}_{ab}\mathbf{C}_{bb}^{-1/2} = \mathbf{U}^\top\mathbf{S}\mathbf{V}$$

$\rightarrow$ The columns of the matrices $\mathbf{U}$ and $\mathbf{V}$ correspond to the sets of orthonormal left and right singular vectors respectively.

$\rightarrow$ The positions $\mathbf{w}_a$ and $\mathbf{w}_b$ are obtained from

$\mathbf{w}_a = \mathbf{C}_{aa}^{-1/2}\mathbf{U} \quad \mathbf{w}_b = \mathbf{C}_{bb}^{-1/2}\mathbf{V}$

$\rightarrow$ The singular values of matrix $\mathbf{S}$ correspond to the canonical correlations.

Jupyter Exercise 1: canonical correlations through the standard eigenvalue problem

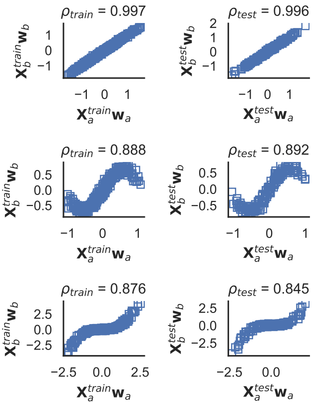
Let $\mathbf{X}_a^{\text{train}} \in \mathbb{R}^{500\times 10}$ and $\mathbf{X}_b^{\text{train}} \in \mathbb{R}^{500\times 10}$ where the columns of both matrices are generated from a random normal distribution with zero mean and unit variance. The following relations are simulated:

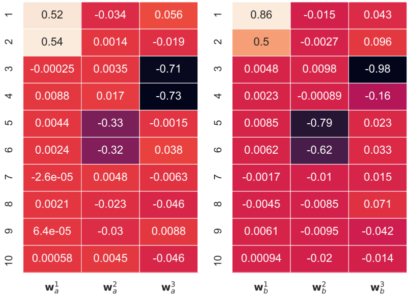
$$\begin{array}{rcl} \mathbf{X}_b^1 + \mathbf{X}_b^2 & = & \mathbf{X}_a^1 + \mathbf{X}_a^2 + \boldsymbol{\xi} \\ \mathbf{X}_b^2 + \mathbf{X}_b^4 & = & (\mathbf{X}_a^2 + \mathbf{X}_4^2)^3 + \boldsymbol{\xi} \\ \mathbf{X}_b^5 + \mathbf{X}_b^6 & = & \sin(\mathbf{X}_a^5 + \mathbf{X}_a^6) + \boldsymbol{\xi} \\ \text{where } \boldsymbol{\xi} & \sim & \mathcal{N}(0, 0.1) \end{array}$$

We compute the square roots of the eigenvalues of the matrix $\mathbf{C}_{bb}^{-1}\mathbf{C}_{ba}\mathbf{C}_{aa}^{-1}\mathbf{C}_{ab}$ and obtain

$$\{\mathbf{0.997}, \mathbf{0.888}, \mathbf{0.876}, 0.172, 0.168, 0.127, 0.1020.094, 0.022, 0.002\}$$

Jupyter Exercise 1: canonical correlations through the standard eigenvalue problem





The score plots show the forms of the relations

The related variables are determined from $\mathbf{w}_a$ and $\mathbf{w}_b$

If there are more variables than examples the within-set covariance matrix becomes singular

$\rightarrow$ The inverses of $\mathbf{C}_{aa}$ and/or $\mathbf{C}_{aa}$ cannot be computed if $p > n$ or $q > n$

$\rightarrow$ This problem is addressed with regularization: we add small positive constants to the diagonal of the within-set covariance matrix.

Regularised standard eigenvalue problem:

$$\left(\mathbf{C}_{bb} + c_b\mathbf{I}\right)^{-1}\mathbf{C}_{ba}\left(\mathbf{C}_{aa} + c_a\mathbf{I}\right)^{-1}\mathbf{C}_{ab}\mathbf{w}_b = \rho^2\mathbf{w}_b.$$

Regularised generalised eigenvalue problem:

$$\begin{pmatrix} \mathbf{0} & \mathbf{C}_{ab} \\ \mathbf{C}_{ba} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{pmatrix} = \rho \begin{pmatrix} \mathbf{C}_{aa} + c_a\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{bb} + c_b\mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{pmatrix}$$

Kernel CCA (KCCA) is CCA on Hilbert Space Objects

The observations are transformed to Hilbert spaces $\mathcal{H}_a$ and $\mathcal{H}_b$ using symmetric positive semi-definite kernels

$$\mathbf{K}_a(\mathbf{x}_a^i, \mathbf{x}_a^j) = \langle \boldsymbol{\phi}_a(\mathbf{x}_a^i), \boldsymbol{\phi}_a(\mathbf{x}_a^j) \rangle_{\mathcal{H}_a} \text{ and } \mathbf{K}_b(\mathbf{x}_b^i, \mathbf{x}_b^j) = \langle \boldsymbol{\phi}_b(\mathbf{x}_b^i), \boldsymbol{\phi}_b(\mathbf{x}_b^j) \rangle_{\mathcal{H}_b}$$

where $i, j = 1, 2, \ldots, n$. In KCCA, the data matrices, $\mathbf{X}_a \in \mathbb{R}^{n \times p}$ and $\mathbf{X}_b \in \mathbb{R}^{n \times q}$, are substituted by the Gram matrices $\mathbf{K}_a \in \mathbb{R}^{n \times n}$ and $\mathbf{K}_b \in \mathbb{R}^{n \times n}$.

Let $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ denote the positions in the kernel space $\mathbb{R}^n$ that have the images $\mathbf{z}_a = \mathbf{K}_a \boldsymbol{\alpha}$ and $\mathbf{z}_b = \mathbf{K}_b \boldsymbol{\beta}$. The KCCA optimisation problem becomes:

$$\max_{\mathbf{z}_a, \mathbf{z}_b \in \mathbb{R}^n} \langle \mathbf{z}_a, \mathbf{z}_b \rangle = \boldsymbol{\alpha}^\top \mathbf{K}_a^\top \mathbf{K}_b \boldsymbol{\beta}$$

$$||\mathbf{z}_a||_2 = \sqrt{\boldsymbol{\alpha}^\top \mathbf{K}_a^2 \boldsymbol{\alpha}} = 1 \quad ||\mathbf{z}_b||_2 = \sqrt{\boldsymbol{\beta}^\top \mathbf{K}_b^2 \boldsymbol{\beta}} = 1$$

Pen-and-Paper Exercise 2: Derive the KCCA problem.

To find non-spurious correlations, kernel CCA needs to be regularised

KCCA is regularised in similar manner as CCA.

Regularised kernelised standard eigenvalue problem:

$$\left(\mathbf{K}_b + c_a \mathbf{I}\right)^{-2} \mathbf{K}_b \mathbf{K}_a \left(\mathbf{K}_a + c_b \mathbf{I}\right)^{-2} \mathbf{K}_a \mathbf{K}_b \boldsymbol{\alpha} = \rho^2 \boldsymbol{\alpha}$$

Regularised kernelised generalised eigenvalue problem:

$$\underbrace{\begin{pmatrix} \mathbf{0} & \mathbf{K}_a \mathbf{K}_b \\ \mathbf{K}_b \mathbf{K}_a & \mathbf{0} \end{pmatrix}}_{\mathbf{A}} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = \rho \underbrace{\begin{pmatrix} \left(\mathbf{K}_a + c_a \mathbf{I}\right)^2 & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{K}_b + c_b \mathbf{I}\right)^2 \end{pmatrix}}_{\mathbf{B}} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix}$$

The hyperparameters $c_a > 0$ and $c_b > 0$ can be determined through cross-validation. In general, a small positive value, such as 0.02, can be used [1].

Pen-and-Paper Exercise 1: Show that KCCA needs to be regularised.

Linear KCCA is the same as CCA

Canonical correlation: $\rho_{\text{cca}} \quad = \quad \frac{\langle \mathbf{X}_a \mathbf{w}_a, \mathbf{X}_b \mathbf{w}_b \rangle}{||\mathbf{X}_a \mathbf{w}_a||_2 ||\mathbf{X}_b \mathbf{w}_b||_2}.$

Kernel canonical correlation: $\rho_{\text{kcca}} = \frac{\langle \mathbf{K}_a \boldsymbol{\alpha}, \mathbf{K}_b \boldsymbol{\beta} \rangle}{\|\mathbf{K}_a \boldsymbol{\alpha}\|_2 \|\mathbf{K}_b \boldsymbol{\beta}\|_2}$

Let $\mathbf{K}_a = \mathbf{X}_a \mathbf{X}_a^\top$ and $\mathbf{K}_b = \mathbf{X}_b \mathbf{X}_b^\top$.

$$\rho_{\text{kcca}} = \frac{\langle \mathbf{X}_a \mathbf{X}_a^\top \boldsymbol{\alpha}, \mathbf{X}_b \mathbf{X}_b^\top \boldsymbol{\beta} \rangle}{\|\mathbf{X}_a \mathbf{X}_a^\top \boldsymbol{\alpha}\|_2 \|\mathbf{X}_b \mathbf{X}_b^\top \boldsymbol{\beta}\|_2}$$

Denote $\mathbf{w}_a = \mathbf{X}_a^\top \boldsymbol{\alpha}$ and $\mathbf{w}_b = \mathbf{X}_b^\top \boldsymbol{\beta}$.

We obtain $\rho_{kcca} = \rho_{cca}$.

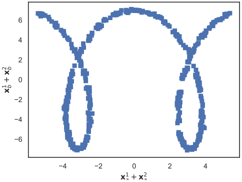
Jupyter Exercise 2: Quadratic KCCA finds non-monotonic trigonometric relations

The homogeneous quadratic polynomial kernel $\mathbf{K} = \langle \mathbf{X}, \mathbf{X} \rangle^2$ finds periodic trigonometric relations.

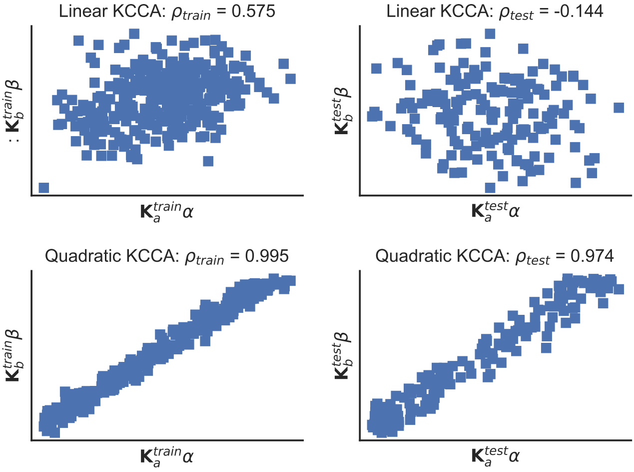
Example simulated data: Let every entry of $\boldsymbol{\theta} \in \mathbb{R}^n \sim U[-2\pi, 2\pi]$.

$$\mathbf{X}_a = \begin{pmatrix} 3 \cdot \sin \frac{\mathbf{x}_a^2}{1} + \boldsymbol{\xi} & \cdots & \boldsymbol{\theta}_p \end{pmatrix}$$

$$\mathbf{X}_b = \begin{pmatrix} 3 \cdot \cos \frac{\mathbf{x}_a^2}{1} + \boldsymbol{\xi} & 6 \cdot \cos \frac{\mathbf{x}_a^2}{0.4} + \boldsymbol{\xi} & \cdots & \boldsymbol{\theta}_q \end{pmatrix}$$



Jupyter Exercise 2: Quadratic KCCA finds non-monotonic trigonometric relations



Exercise: Complete the generalized eigenvalue problem for KCCA and apply it on this dataset.

Wrap-up: CCA is an eigenvalue-based method that finds multivariate relations from two-view datasets

CCA:

$\rightarrow$ we solve the CCA problem through a standard or generalized eigenvalue problem or by applying the SVD.

$\rightarrow$ The related variables are determined from the entries of $\mathbf{w}_a$ and $\mathbf{w}_b$.

$\rightarrow$ The value of the training and test correlation is obtained from $\rho = \frac{\langle \mathbf{X}_a^t \mathbf{w}_a, \mathbf{X}_b^t \mathbf{w}_b \rangle}{||\mathbf{X}_a^t \mathbf{w}_a|| ||\mathbf{X}_b^t \mathbf{w}_b||}$ where t is either train or test data.

$\rightarrow$ Training correlation shows whether learning occurs, test correlation shows if the relation is predictive

$\rightarrow$ The form of the underlying relation is seen from the score plot.

Wrap-up: KCCA is an eigenvalue-based method that finds multivariate relations from two-view datasets

KCCA:

$\rightarrow$ we map the observations to a Hilbert space and solve the CCA problem there, through the standard or generalised eigenvalue problem.

$\rightarrow$ with non-linear kernels, we cannot extract the related variables from the dual coefficient vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

$\rightarrow$ a high test (non-linear) kernel canonical correlation tells that the data contains non-linear relations

References

- Francis R Bach and Michael I Jordan. "Kernel independent component analysis". In: *Journal of machine learning research* 3.Jul (2002), pp. 1–48.
- Viivi Uurtio et al. "A Tutorial on Canonical Correlation Methods". In: *ACM Comput. Surv.* 50.6 (Nov. 2017), 95:1–95:33. ISSN: 0360-0300. DOI: 10.1145/3136624. URL: http://doi.acm.org/10.1145/3136624.