MACHINE LEARNING

Kernel Canonical Correlation Analysis

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Structure of today's and next week's class

1) Briefly go through one extension of principal component analysis, namely Canonical Correlation Analysis (CCA).

2) Derive the non-linear version of CCA, kernel CCA (kCCA).

3) Make an exercise to understand the modulation of the space generated by CCA and kCCA.

Canonical Correlation Analysis (CCA)

$x \in \mathbb{R}^{N_x}$ $y \in \mathbb{R}^{N_y}$

$\{x^1, y^1\}$

$$\max_{w^x, w^y} corr\left(w_x^T x, w_y^T y\right)$$

$\{x^2, y^2\}$

Video description **Audio description**

Determine features in two (or more) separate descriptions of the dataset that best explain each datapoint.

Extract hidden structure that maximize correlation across two different projections.

Canonical Correlation Analysis (CCA)

Pair of multidimensional zero mean variables

$$X = \left\{ x^i \in \mathbb{R}^{N_x} \right\}_{i=1}^{M}, Y = \in \left\{ y^i \in \mathbb{R}^{N_y} \right\}_{i=1}^{M}$$

We have *M* instances of the pairs.

$$\left\{ x^1, y^1 \right\}$$

$$\max_{w^x, w^y} corr\left(w_x^T x, w_y^T y \right)$$

Search two projections w_x and w_y :

$$z_x = w_x^T X \quad \text{and} \quad z_y = w_y^T Y$$

$$\left\{ x^2, y^2 \right\}$$

solutions of:

$$\max \rho = \max_{w_x, w_y} \operatorname{corr}\left(z_x, z_y \right)$$

Canonical Correlation Analysis (CCA)

Search two projections w_x and w_y :

$z_x = w_x^T X$ and $z_y = w_y^T Y$

$$\max_{w^x, w^y} corr\left(w_x^T x, w_y^T y\right)$$

solutions of:

$$\max \rho = \max_{w_x, w_y} \operatorname{corr}\left(\mathrm{z}_x, \mathrm{z}_y\right)$$

$$= \max_{w_x, w_y} \frac{\mathrm{w}_\mathrm{x}^\mathrm{T} E\left\{XY^T\right\} \mathrm{w}_\mathrm{y}}{\left\|\mathrm{w}_\mathrm{x}^\mathrm{T} X\right\| \left\|\mathrm{w}_\mathrm{y}^T Y\right\|} = \max_{w_x, w_y} \frac{\mathrm{w}_\mathrm{x}^\mathrm{T} C_{xy} \mathrm{w}_\mathrm{y}}{\sqrt{\mathrm{w}_\mathrm{x}^\mathrm{T} C_{xx} w_x \mathrm{w}_\mathrm{y}^\mathrm{T} C_{yy} w_y}}$$

With X and Y zero mean, i.e.
$E\{X\} = E\{Y\} = 0$

Canonical Correlation Analysis (CCA)

Crosscovariance matrix

C_{xy} is $N_x \times N_y$

Measure crosscorrelation between X and Y.

Covariance matrices

$C_{xx} = \mathrm{E}\left\{XX^T\right\}: N_x \times N_x$

$C_{yy} = \mathrm{E}\left\{YY^T\right\}: N_y \times N_y$

solutions of:

$$\max \rho = \max_{w_x, w_y} \operatorname{corr}\left(z_x, z_y\right)$$

$$= \max_{w_x, w_y} \frac{w_x^T E\left\{XY^T\right\} w_y}{\left\|w_x^T X\right\| \left\|w_y^T Y\right\|} = \max_{w_x, w_y} \frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} w_x w_y^T C_{yy} w_y}}$$

With X and Y zero mean, i.e.

$E\{X\} = E\{Y\} = 0$

Canonical Correlation Analysis (CCA)

Correlation not affected by rescaling the norm of the vectors,

$\Rightarrow$ we can ask that $\mathrm{w}_{\mathrm{x}}^{\mathrm{T}} C_{xx} w_x = \mathrm{w}_{\mathrm{y}}^{\mathrm{T}} C_{yy} w_y = 1$

$$\boxed{\begin{aligned}&\max \rho = \max_{w_x, w_y} \ \mathrm{w}_{\mathrm{x}}^{\mathrm{T}} C_{xy} \mathrm{w}_{\mathrm{y}} \\ &\text{u. c. } \mathrm{w}_{\mathrm{x}}^{\mathrm{T}} C_{xx} w_x = \mathrm{w}_{\mathrm{y}}^{\mathrm{T}} C_{yy} w_y = 1\end{aligned}}$$

solutions of:

$$\max \rho = \max_{w_x, w_y} \ \mathrm{corr}\left(\mathrm{z}_x, \mathrm{z}_y\right)$$

$$= \max_{w_x, w_y} \frac{\mathrm{w}_{\mathrm{x}}^{\mathrm{T}} E\left\{XY^T\right\} \mathrm{w}_{\mathrm{y}}}{\left\|\mathrm{w}_{\mathrm{x}}^{\mathrm{T}} X\right\| \left\|\mathrm{w}_{\mathrm{y}}^{T} Y\right\|} = \max_{w_x, w_y} \frac{\mathrm{w}_{\mathrm{x}}^{\mathrm{T}} C_{xy} \mathrm{w}_{\mathrm{y}}}{\sqrt{\mathrm{w}_{\mathrm{x}}^{\mathrm{T}} C_{xx} w_x \mathrm{w}_{\mathrm{y}}^{\mathrm{T}} C_{yy} w_y}}$$

Canonical Correlation Analysis (CCA)

Correlation not affected by rescaling the norm of the vectors,

$\Rightarrow$ we can ask that $\mathrm{w}_{\mathrm{x}}^{\mathrm{T}} C_{xx} w_x = \mathrm{w}_{\mathrm{y}}^{\mathrm{T}} C_{yy} w_y = 1$

$$\boxed{\begin{aligned} &\max \rho = \max_{w_x, w_y} \ \mathrm{w}_{\mathrm{x}}^{\mathrm{T}} C_{xy} \mathrm{w}_{\mathrm{y}} \\ &\text{u. c. } \mathrm{w}_{\mathrm{x}}^{\mathrm{T}} C_{xx} w_x = \mathrm{w}_{\mathrm{y}}^{\mathrm{T}} C_{yy} w_y = 1 \end{aligned}}$$

To determine the optimum (maximum) of ρ, solve by Lagrange:

$$L\left(w_x, w_y, \lambda_x, \lambda_y\right) = w_x^T C_{xy} w_y - \lambda_x \left(w_x^T C_{xx} w_x - 1\right) - \lambda_y \left(w_y^T C_{yy} w_y - 1\right)$$

Taking the partial derivatives over w_x, w_y

$\Rightarrow \lambda_x = \lambda_y := \lambda / 2$

Canonical Correlation Analysis (CCA)

Replacing λ and write the set of equations gives:

$$\begin{pmatrix} 0 & C_{xy} \\ C_{yx} & 0 \end{pmatrix}\begin{pmatrix} w_x \\ w_y \end{pmatrix} = \lambda \begin{pmatrix} C_{xx} & 0 \\ 0 & C_{yy} \end{pmatrix}\begin{pmatrix} w_x \\ w_y \end{pmatrix}$$

Generalized Eigenvalue Problem;
It can be reduced to a classical eigenvalue problem if C_{xx} is invertible

$\Rightarrow$ Which can be rewritten as

$$\boxed{C_{xy} C_{yy}^{-1} C_{yx} w_x = \lambda^2 C_{xx} w_x}$$

Solving for w_y gives:

$$C_{yx} C_{xx}^{-1} C_{xy} w_y = \lambda^2 C_{yy} w_y$$

If C_{yy} is invertible, it becomes an eigenvalue problem as for w_y.

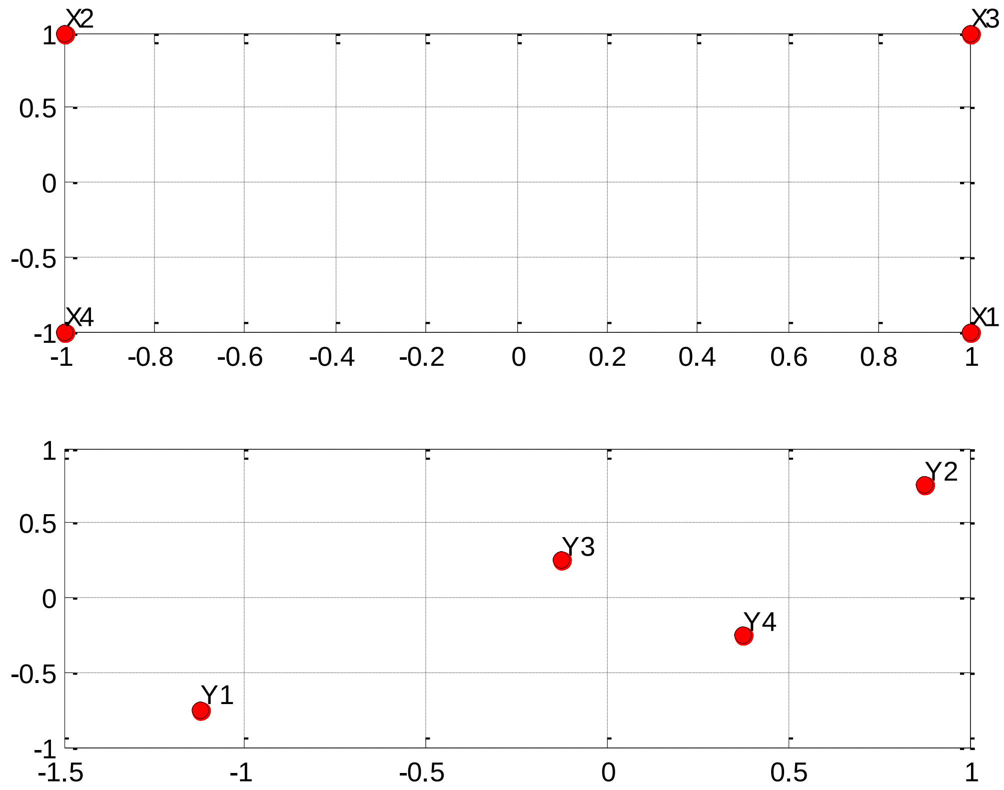
These two eigenvalue problems yield a pair of q vectors $\left\{ w_x^i, w_y^i \right\}_{i=1..q}$, where $q = \min(N_x, N_y)$

$w_x^i \in \mathbb{R}^{N_x}, w_y^i \in \mathbb{R}^{N_y}$

CCA: Exercise I

Consider the example below of a dataset of 4 points with 2-dimensional coordinates in both X and Y.

- Determine by hand the directions found by CCA in each space.
- Contrast to the directions found by PCA.



Kernel Canonical Correlation Analysis

CCA finds basis vectors, s.t. the correlation between the projections (of all datapoints in *X* and *Y*) is mutually maximized.

CCA is a generalized version of PCA for two or more multi-dimensional datasets, but unlike PCA it does have the constraint to find orthogonal vectors.

Assumes a linear correlation. If correlation non-linear
→ Kernel CCA.

Kernel Canonical Correlation Analysis (kCCA)

$x \in \mathbb{R}^{N_x}$ $\qquad$ $y \in \mathbb{R}^{N_y}$

$\{x^1, y^1\}$

$$\max_{w^x, w^y} corr\left(w_x^T \phi_x(x), w_y^T \phi_y(y)\right)$$

$\{x^2, y^2\}$

Assume two transformations

ϕ_x $\quad$ ϕ_y

Video description $\qquad$ **Audio description**

And then perform correlation analysis in feature space across the two feature spaces.

From CCA to Kernel CCA

$$X = \left\{ x^i \in \mathbb{R}^{N_x} \right\}_{i=1}^{M}, Y = \in \left\{ y^i \in \mathbb{R}^{N_y} \right\}_{i=1}^{M}$$

Send into two separate feature spaces for data in *X* and in *Y.*

$$\left\{ \phi_x\left(x^i\right)\right\}_{i=1}^{M} \text{ and } \left\{ \phi_y\left(y^i\right)\right\}_{i=1}^{M}, \quad \text{with } \sum_{i=1}^{M} \phi_x\left(x^i\right) = 0 \text{ and } \sum_{i=1}^{M} \phi_y\left(y^i\right) = 0$$

Construct associated kernel matrices:

$$K_x = F_x^T F_x,\ K_y = F_y^T F_y, \quad \text{columns of } F_x, F_y \text{ are } \phi_x\left(x^i\right),\ \phi_y\left(y^i\right)$$

From CCA to Kernel CCA

In Linear CCA, we were solving for:

$$\max_{w_x, w_y} \mathrm{w}_x^T C_{xy} \mathrm{w}_y$$

$$\text{u.c.} \quad \mathrm{w}_x^T C_{xx} \underline{w_x} = \mathrm{w}_y^T C_{yy} \underline{w_y} = 1$$

In kernel CCA, we solve for:

$$\max_{w_x, w_y} \alpha_x^T \underbrace{F_x^T F_x}_{K_x} \underbrace{F_y^T F_y}_{K_y} \alpha_y$$

$$\text{u.c.} \quad \alpha_x^T \underbrace{F_x^T F_x F_x^T F_x}_{K_x} \alpha_x = \alpha_y^T \underbrace{F_y^T F_y F_y^T F_y}_{K_y} \alpha_y = 1$$

Express the projection vectors as a linear combination of images of datapoints in feature space (as in kPCA):

$w_x = F_x \alpha_x$ and $w_y = F_y \alpha_y$

$\Rightarrow w_x = \sum_{i=1}^{M} \alpha_{x,i} \phi_x \left(x^i\right)$ and $w_x = \sum_{i=1}^{M} \alpha_{y,i} \phi_y \left(y^i\right)$

Replace the covariance and crosscovariance matrices by the product of the projection vectors in feature space (as in kPCA):

$C_{xx} = F_x F_x^T$

$C_{yy} = F_y F_y^T$

$C_{xy} = F_x F_y^T$

Kernel CCA

In summary, in kernel CCA, we search the projection vectors w_x, w_y (that live in feature space) so as to maximize:

$$\max_{w_x, w_y} \rho = \max_{w_x, w_y} corr\left(w_x \phi_x(x), w_y \phi_y(y)\right)$$

$$\max_{w_x, w_y} \rho = \max_{\alpha_x, \alpha_y} \alpha_x^T K_x K_y \alpha_y$$

$$u.c. \left(\alpha_x^T K_x^2 \alpha_x\right) = \left(\alpha_y^T K_y^2 \alpha_y\right) = 1$$

This is again a generalized eigenvalue problem with α_x, α_y the dual eigenvectors (as dual vectors in kPCA), see documentation in annexes for derivation.

Generalized eigenvalue problem:

$$\begin{pmatrix} 0 & K_x K_y \\ K_y K_x & 0 \end{pmatrix} \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix} = \lambda \begin{pmatrix} K_x^2 & 0 \\ 0 & K_y^2 \end{pmatrix} \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}$$

Kernel CCA

If the intersection between the spaces spanned by $K_x\alpha_x,\ K_y\alpha_y$ is non-zero, then the problem has a trivial solution, as $\rho \sim \cos\left(K_x\alpha_x, K_y\alpha_y\right) = 1$ (see solution to the exercises).

$$\max_{w_x, w_y} \rho = \max_{\alpha_x, \alpha_y} \alpha_x^T K_x K_y \alpha_y$$

$$u.c.\ \left(\alpha_x^T K_x^2 \alpha_x\right) = \left(\alpha_y^T K_y^2 \alpha_y\right) = 1$$

Generalized eigenvalue problem:

$$\begin{pmatrix} 0 & K_x K_y \\ K_y K_x & 0 \end{pmatrix} \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix} = \lambda \begin{pmatrix} K_x^2 & 0 \\ 0 & K_y^2 \end{pmatrix} \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}$$

Kernel CCA

Add a regularization term to increase the rank of the matrix and make it invertible (to avoid the trivial solution)

$$K_x^2 \rightarrow \left(K_x + \frac{M\kappa}{2} \mathrm{I} \right)^2, \quad \kappa > 0$$

Generalized eigenvalue problem:

$$\begin{pmatrix} 0 & K_x K_y \\ K_y K_x & 0 \end{pmatrix} \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix} = \lambda \begin{pmatrix} K_x^2 & 0 \\ 0 & K_y^2 \end{pmatrix} \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}$$

Several methods have been proposed to choose carefully the regularization term so as to get projections that are as close as possible to the “true” projections.

Kernel CCA

$$\underbrace{\begin{pmatrix} 0 & K_x K_y \\ K_y K_x & 0 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}}_{\alpha} = \lambda \underbrace{\begin{pmatrix} \left(K_x + \frac{M\kappa}{2} \mathrm{I}\right)^2 & 0 \\ 0 & \left(K_y + \frac{M\kappa}{2} \mathrm{I}\right)^2 \end{pmatrix}}_{B} \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}$$

Set: $B = C^T C$ and $\beta = C\alpha$

Becomes a classical eigenvalue problem → $\boxed{\left(C^T\right)^{-1} A C^{-1} \beta = \lambda \beta}$

Kernel CCA

$$X = \left\{ x^i \in \mathbb{R}^{N_x} \right\}_{i=1}^{M}, Y = \in \left\{ y^i \in \mathbb{R}^{N_y} \right\}_{i=1}^{M}$$

Two datasets case

Can be extended to multiple datasets:

L datasets: $X_1, ..., X_L$ with M observations each

Dimensions $N_1, ...N_L$:, i.e. $X_i : N_i \times M$

Applying non-linear transformation ϕ_1, to $X_1, ...X_L$

$\rightarrow$ construct L Gram matrices: $K_1, ..., K_L$

Kernel CCA

$$X = \left\{ x^i \in \mathbb{R}^{N_x} \right\}_{i=1}^{M}, Y = \in \left\{ y^i \in \mathbb{R}^{N_y} \right\}_{i=1}^{M}$$

Two datasets case

Can be extended to multiple datasets:

L datasets: $X_1, \ldots., X_L$ with M observations each

Dimensions $N_1, \ldots. N_L$:, i.e. $X_i : N_i \times M$

$$\begin{pmatrix} 0 & K_1K_2 & \ldots\ldots. & K_1K_L \\ K_2K_1 & 0 & \ldots\ldots.. & K_2K_L \\ . & & & \\ . & & & \\ K_LK_1 & K_LK_2 & \ldots\ldots. & 0 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ . \\ . \\ \alpha_L \end{pmatrix} = \lambda \begin{pmatrix} \left(K_1 + \frac{M\kappa}{2} \mathrm{I} \right)^2 & & 0 \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ 0 & & \left(K_L + \frac{M\kappa}{2} \mathrm{I} \right)^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ . \\ . \\ \alpha_L \end{pmatrix}$$

Kernel CCA

$$X = \left\{ x^i \in \mathbb{R}^{N_x} \right\}_{i=1}^{M}, Y = \in \left\{ y^i \in \mathbb{R}^{N_y} \right\}_{i=1}^{M}$$

Two datasets case

had as solution the following generalized eigenvalue problem:

$$\underbrace{\begin{pmatrix} 0 & K_x K_y \\ K_y K_x & 0 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}}_{\alpha} = \lambda \underbrace{\begin{pmatrix} \left(K_x + \frac{M\kappa}{2} \mathrm{I}\right)^2 & 0 \\ 0 & \left(K_y + \frac{M\kappa}{2} \mathrm{I}\right)^2 \end{pmatrix}}_{B} \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}$$

→ Can be extended to multiple datasets (MKCCA)

$$\begin{pmatrix} 0 & K_1 K_2 & \ldots\ldots\ldots & K_1 K_L \\ K_2 K_1 & 0 & \ldots\ldots\ldots & K_2 K_L \\ . & & & \\ . & & & \\ K_L K_1 & K_L K_2 & \ldots\ldots\ldots & 0 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ . \\ . \\ \alpha_L \end{pmatrix} = \lambda \begin{pmatrix} \left(K_1 + \frac{M\kappa}{2} \mathrm{I}\right)^2 & & 0 \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ 0 & & \left(K_L + \frac{M\kappa}{2} \mathrm{I}\right)^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ . \\ . \\ \alpha_L \end{pmatrix}$$

CCA: Exercise II

Consider the following kernel matrices :

$$K_x = \begin{bmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad K_y = \begin{bmatrix} 1 & 0.8 & 0 \\ 0.8 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

a) How many datapoints do you have? What are the dimensions of the two feature spaces?

b) Assume a RBF kernel with same kernel width for K_x and K_y, draw the distribution of points and give the shape of the dual vectors α_x and α_y, solutions of $\max_{\alpha_x,\alpha_y} \rho = \max_{\alpha_x,\alpha_y} \alpha_x^T K_x K_y \alpha_y$ with $\alpha_x^T K_x^2 \alpha_x = \alpha_y^T K_y^2 \alpha_y = 1$.

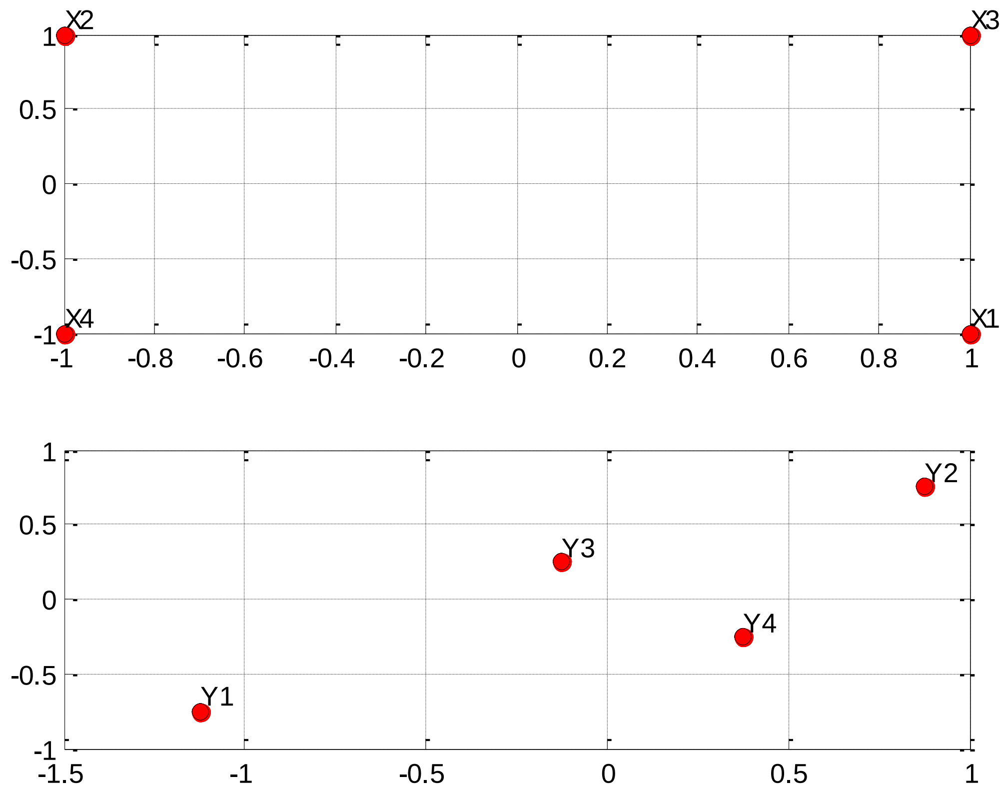
c) What is the effect of changing the kernel width on K_x and K_y and on α_x and α_y ?

d) Do (b) when considering a polynomial kernel. Assume then same distribution of points as for RBF kernel. What are the Gram matrices?

CCA: Exercise III

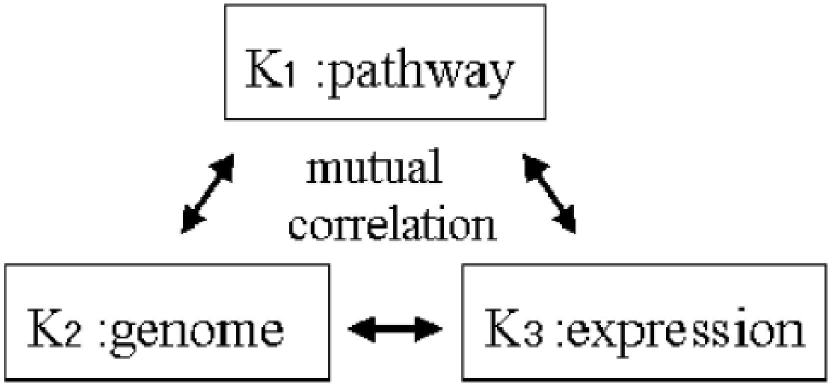
Consider the example below of a dataset of 4 points with 2-dimensional coordinates in both X and Y.

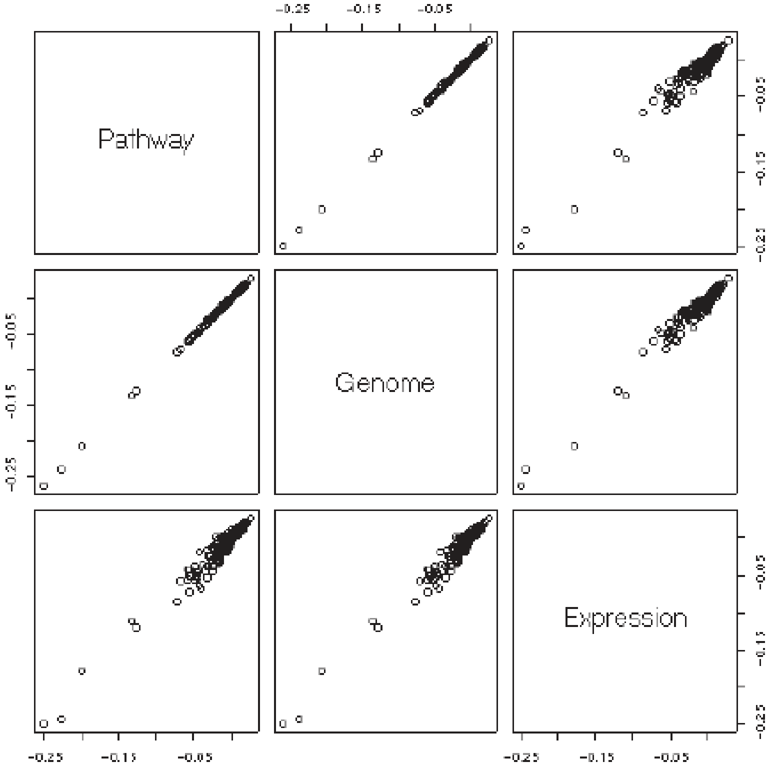
- What is the shape of the kernel matrices and dual eigenvectors and draw the isolines when considering a RBF kernel.
- Do the same with a polynomial kernel.



Applications of Kernel CCA

Goal: To measure correlation between heterogeneous datasets and to extract sets of genes which share similarities with respect to multiple biological attributes





Kernel matrices K1, K2 and K3 correspond to gene-gene similarities in pathways, genome position, and microarray expression data resp.
Use RBF kernel with fixed kernel width.

Correlation scores in MKCCA: pathway vs. genome vs. expression.

Y Yamanishi, JP Vert, A Nakaya, M Kanehisa - Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis, Bioinformatics, 2003

Applications of Kernel CCA

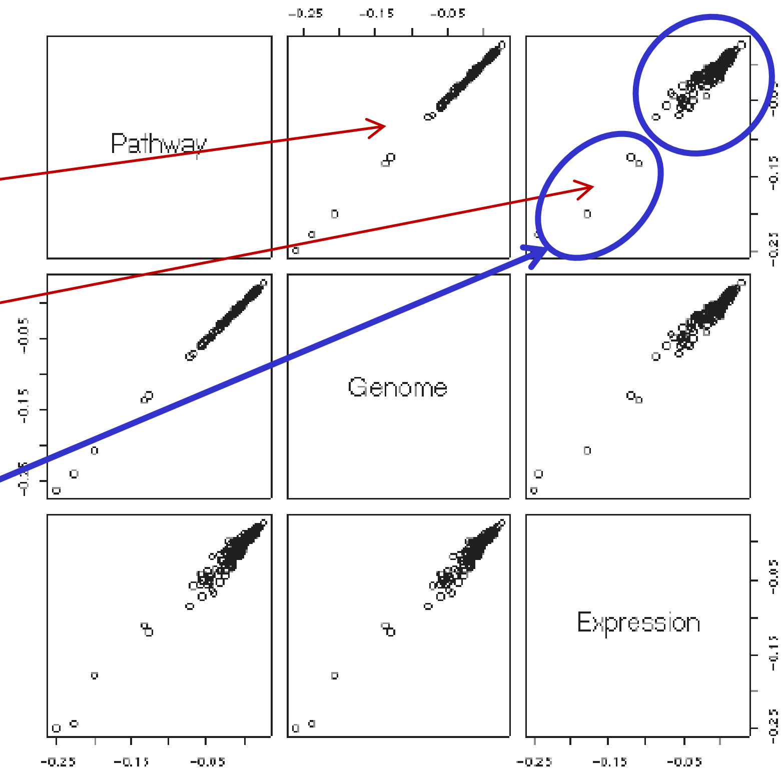
Goal: To measure correlation between heterogeneous datasets and to extract sets of genes which share similarities with respect to multiple biological attributes

Gives pairwise correlation between K1, K2

Gives pairwise correlation between K1, K3

Two clusters correspond to genes close to each other with respect to their positions in the pathways, in the genome, and to their expression

A readout of the entries with equal projection onto the first canonical vectors α give the genes which belong to each cluster

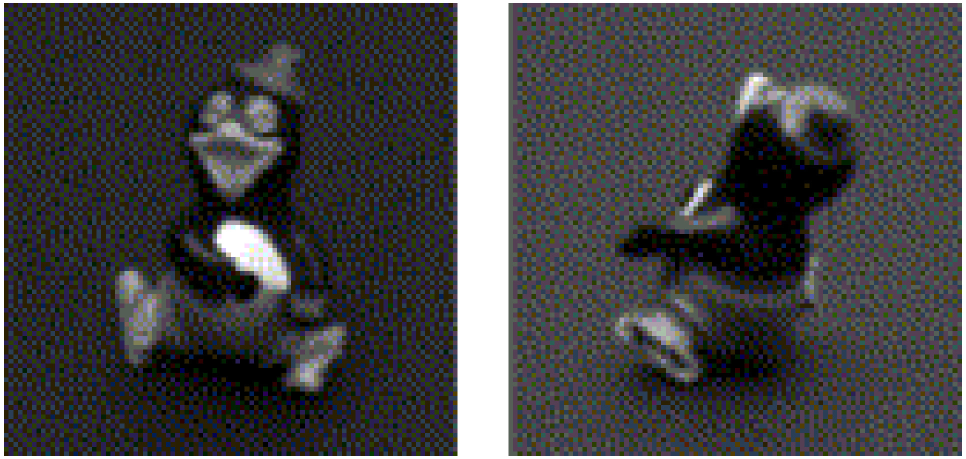


Correlation scores in MKCCA: pathway vs. genome vs. expression.

Y Yamanishi, JP Vert, A Nakaya, M Kanehisa - Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis, Bioinformatics, 2003

Applications of Kernel CCA

Goal: To construct appearance models for estimating an object's pose from raw brightness images.



Example of two image datapoints with different poses

X: Set of images

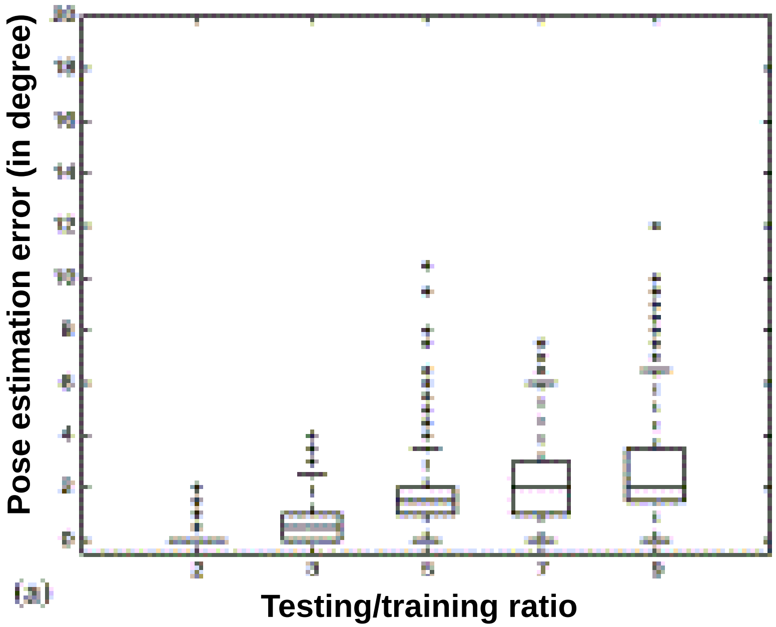
Y: Pose parameters (pan and tilt angle of the object w.r.t. the camera in degrees)

Method: used linear kernel on X and RBF kernel on Y and compared performance to applying PCA on the (X, Y) dataset directly.

T. Melzer, M. Reiter and H. Bischof, Appearance models based on kernel canonical correlation analysis, Pattern Recognition 36 (2003) p. 1961–1971

Applications of Kernel CCA

Goal: To construct appearance models for estimating an object's pose from raw brightness images



kernel-CCA performs better than kPCA, especially for small testing/training ratio (i.e., for larger training sets).

The kernel-CCA estimators tend to produce less outliers, i.e., gross errors, and consequently yield a smaller standard deviation of the pose estimation error than their PCA-based counterparts.

For very small training sets, the performance of both approaches becomes similar

T. Melzer, M. Reiter and H. Bischof, Appearance models based on kernel canonical correlation analysis, Pattern Recognition 36 (2003) p. 1961–1971

Summary

- CCA is an excellent means to discover appropriate projections when your data is multi-modal.
- In each modality (separately), CCA finds projections that highlight features common to the datapoints as a whole.
- It generates projections that are different from performing PCA on each modality separately.
- The non-linear version of CCA, kernel CCA, generates sets of projections different from linear CCA and from kPCA.
- These projections highlight sets of modalities that are common to groups of datapoints. It is a good pre-processing method before clustering.