Abteilung Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 2

Recall: For a sample of d_1- and d_2-dimensional data of size N, given as two data matrices $X \in \mathbb{R}^{d_1 \times N}$, $Y \in \mathbb{R}^{d_2 \times N}$ (assumed to be centered), canonical correlation analysis (CCA) finds a one-dimensional projection maximizing the cross-correlation for constant auto-correlation. The primal optimization problem is:

$$\begin{aligned} \text{Find } w_x \in \mathbb{R}^{d_1}, w_y \in \mathbb{R}^{d_2} \text{ maximizing } \quad & w_x^\top C_{xy} w_y \\ \text{subject to } \quad & w_x^\top C_{xx} w_x &= 1 \\ & w_y^\top C_{yy} w_y &= 1, \end{aligned} \tag{1}$$

where $C_{xx} = \frac{1}{N} X X^\top \in \mathbb{R}^{d_1 \times d_1}$ and $C_{yy} = \frac{1}{N} Y Y^\top \in \mathbb{R}^{d_2 \times d_2}$ are the auto-covariance matrices of X resp. Y, and $C_{xy} = \frac{1}{N} X Y^\top \in \mathbb{R}^{d_1 \times d_2}$ is the cross-covariance matrix of X and Y.

Exercise 1: Primal CCA $(10 + 5$ P)

We have seen in the lecture that a solution of the canonical correlation analysis can be found in some eigenvector of the generalized eigenvalue problem:

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

(a) *Show* that among all eigenvectors (w_x, w_y) the solution is the one associated to the highest eigenvalue.

(b) *Show* that if (w_x, w_y) is a solution, then $(-w_x, -w_y)$ is also a solution of the CCA problem.

Exercise 2: Dual CCA $(10 + 15 + 5 + 5$ P)

In this exercise, we would like to derive the dual optimization problem.

(a) *Show*, that it is always possible to find an optimal solution in the span of the data, that is,

$$w_x = X \alpha_x \ , \quad w_y = Y \alpha_y$$

with some coefficient vectors $\alpha_x \in \mathbb{R}^N$ and $\alpha_y \in \mathbb{R}^N$.

(b) *Show* that the solution of the dual optimization problem is found in an eigenvector of the generalized eigenvalue problem

$$\begin{bmatrix} 0 & A \cdot B \\ B \cdot A & 0 \end{bmatrix} \begin{bmatrix} \alpha_x \\ \alpha_y \end{bmatrix} = \rho \cdot \begin{bmatrix} A^2 & 0 \\ 0 & B^2 \end{bmatrix} \begin{bmatrix} \alpha_x \\ \alpha_y \end{bmatrix}$$

where $A = X^\top X$ and $B = Y^\top Y$.

(c) *Show* that the solution of the dual is given by the eigenvector associated to the highest eigenvalue.

(d) *Show* how a solution to the original problem can be obtained from the solution of the generalized eigenvalue problem of the dual.

Exercise 3: CCA and Least Square Regression (20 P)

Consider some supervised dataset with the inputs stored in a matrix $X \in \mathbb{R}^{D \times N}$ and the targets stored in a vector $y \in \mathbb{R}^N$. We assume that both our inputs and targets are centered. The least squares regression optimization problem is:

$$\min_{v \in \mathbb{R}^D} \quad \|X^\top v - y\|^2$$

We would like to relate least square regression and CCA, specifically, their respective solutions v and (w_x, w_y).

(a) *Show* that if X and y are the two modalities of CCA (i.e. $X \in \mathbb{R}^{D \times N}$ and $y \in \mathbb{R}^{1 \times N}$), the first part of the solution of CCA (i.e. the vector w_x) is equivalent to the solution v of least square regression up to a scaling factor.

Exercise 4: Programming (30 P)

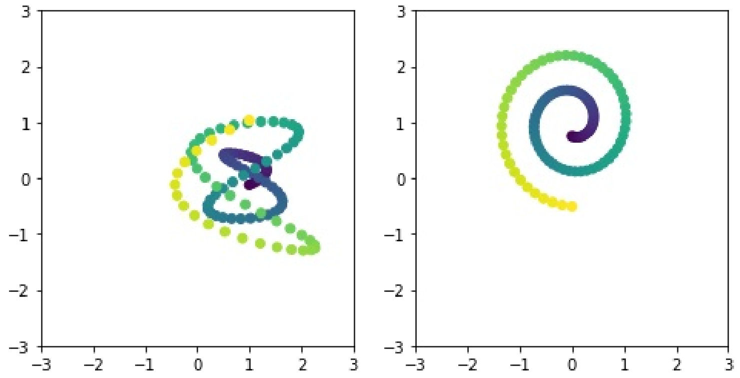
Download the programming files on ISIS and follow the instructions.

Canonical Correlation Analysis

In this exercise, we consider canonical correlation analysis (CCA) on two simple problems, one in low dimensions and one in high dimensions. The goal is to implement the primal and dual versions of CCA to handle these two different cases. The first dataset consists of two trajectories in two dimensions. The dataset is extracted and plotted below. The first data points are shown in dark blue, and the last ones are shown in yellow.

In [1]:

```
import numpy
import matplotlib
%matplotlib inline
from matplotlib import pyplot as plt
import utils

X,Y   = utils.getdata()
p1,p2 = utils.plotdata(X,Y)
```



For these two trajetories, that can be understood as two different modalities of the same data, we would like determine under which projections they appear maximally correlated.

Implementing Primal CCA

As stated in the lecture, the CCA problem in its primal form consists of maximizing the cross-correlation objective:
$$J(w_x,w_y) = w_x^\top C_{xy} w_y$$

subject to autcorrelation constraints $w_x^\top C_{xx} w_x = 1$ and $w_y^\top C_{yy} w_y = 1$. Using the method of Lagrange multipliers, this optimization problem can be reduced to finding the first eigenvector of the generalized eigenvalue problem:
$$ \begin{bmatrix}0 & C_{xy}\\C_{yx} & 0\end{bmatrix} \begin{bmatrix}w_x\\w_y\end{bmatrix} = \lambda \begin{bmatrix}C_{xx} & 0\\0 & C_{yy}\end{bmatrix} \begin{bmatrix}w_x\\w_y\end{bmatrix} $$

Your first task is to write a function that solves the CCA problem in the primal (i.e. that solves the generalized eigenvalue problem above). The function you need to implement receives two matrices `X` and `Y` of size `N` $\times$ `d1` and `N` $\times$ `d2` respectively. It returns two vectors of size `d1` and `d2` corresponding to the projections associated to the modalities `X` and `Y`. *(Hint: Note that the data matrices `X` and `Y` have not been centered yet.)*

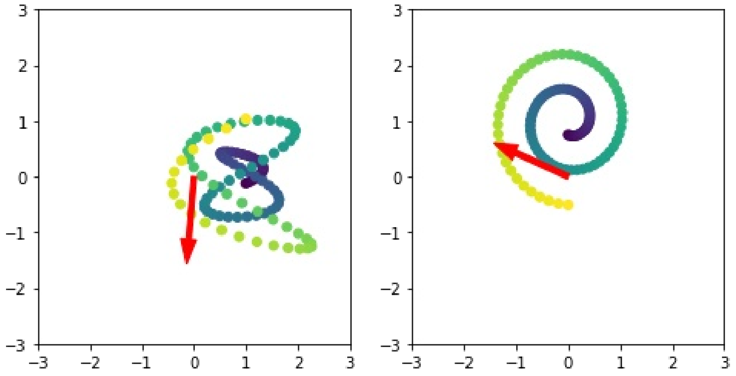
In [2]:

```
import numpy

def CCAprimal(X,Y):

    ## -------------------------------
    ## TODO: replace by your solution
    ## -------------------------------
    import solution
    wx,wy = solution.CCAprimal(X,Y)
    ## -------------------------------

    return wx,wy
```

The function can now be called with our dataset. The learned projection vectors w_x and w_y are plotted as red arrows.

In [3]:

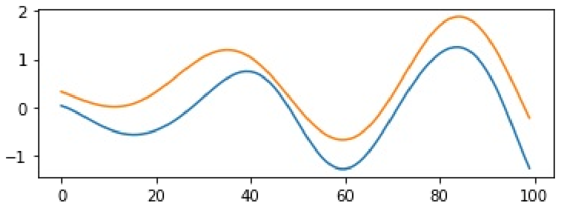
```
wx,wy = CCAprimal(X,Y)

p1,p2 = utils.plotdata(X,Y)
p1.arrow(0,0,1*wx[0],1*wx[1],color='red',width=0.1)
p2.arrow(0,0,1*wy[0],1*wy[1],color='red',width=0.1)
plt.show()
```



In each modality, the arrow points in a specific direction (note that the optimal CCA directions are defined up to a sign flip of both w_x and w_y). Furthermore, we can verify CCA has learned a meaningful solution by projecting the data on it.

In [4]:

```
plt.figure(figsize=(6,2))
plt.plot(numpy.dot(X,wx))
plt.plot(numpy.dot(Y,wy))
plt.show()
```



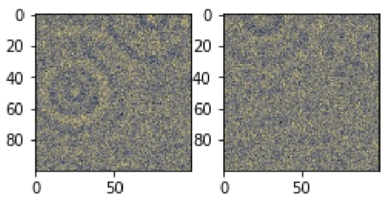
Clearly, the data is correlated in the projected space.

Implementing Dual CCA

In the second part of the exercise, we consider the case where the data is high dimensional (with `d` $\gg$ `N`). Such high-dimensionality occurs for example, when input data are images. We consider the scenario where sources emit spatially, and two (noisy) receivers measure the spatial field at different locations. We would like to identify signal that is common to the two measured locations, e.g. a given source emitting at a given frequency. We first load the data and show one example.

In [5]:

```
X,Y = utils.getHDdata()

utils.plotHDdata(X[0],Y[0])
plt.show()
```



Several sources can be perceived, however, there is a significant level of noise. Here again, we will use CCA to find subspaces where the two modalities are maximally correlated. In this example, because there are many more dimensions than there are data points, it is more advantageous to solve CCA in the dual. Your task is to implement a CCA dual solver that receives two data matrices of size `N` $\times$ `d1` and `N` $\times$ `d2` respectively as input, and returns the associate CCA directions (two vectors of respective sizes `d1` and `d2`).

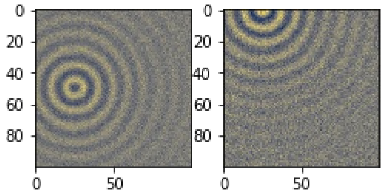
In [6]:

```
def CCAdual(X,Y):

    ## -------------------------------
    ## TODO: replace by your solution
    ## -------------------------------
    import solution
    wx,wy = solution.CCAdual(X,Y)
    ## -------------------------------

    return wx,wy
```

We now call the function we have implemented with a training sequence of 100 pairs of images. Because the returned solution is of same dimensions as the inputs, it can be rendered in a similar fashion.

In [7]:

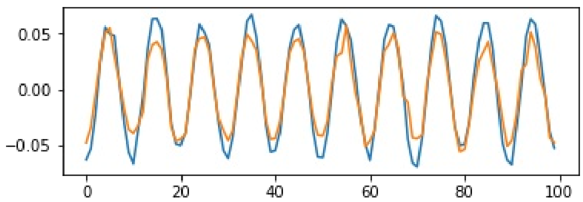
```
wx,wy = CCAdual(X[:100],Y[:100])

utils.plotHDdata(wx,wy)
plt.show()
```



Here, we can clearly see a common factor that has been extracted between the two fields, specifically a point source emitting at a particular frequency. A test sequence of 100 pairs of images can now be projected on these two filters:

In [8]:

```
plt.figure(figsize=(6,2))
plt.plot(numpy.dot(X[100:],wx))
plt.plot(numpy.dot(Y[100:],wy))
plt.show()
```



Clearly the two projected signals are correlated and the input noise has been strongly reduced.

Exercise Sheet 2

1 Primal CCA

a) $$(w_x^T \quad w_y^T)\begin{pmatrix} 0 & C_{xy} \\ C_{yx} & 0 \end{pmatrix}\begin{pmatrix} w_x \\ w_y \end{pmatrix} = \lambda (w_x \quad w_y)\begin{pmatrix} C_{xx} & 0 \\ 0 & C_{yy} \end{pmatrix}\begin{pmatrix} w_x \\ w_y \end{pmatrix}$$

$$\Leftrightarrow (w_y^T C_{yx} \quad w_x^T C_{xy})\begin{pmatrix} w_x \\ w_y \end{pmatrix} = \lambda (w_x^T C_{xx} \quad w_y C_{yy})\begin{pmatrix} w_x \\ w_y \end{pmatrix}$$

$$\Leftrightarrow w_y^T C_{yx} w_x + w_x^T C_{xy} w_y = \lambda (w_x^T C_{xx} w_x + w_y^T C_{yy} w_y)$$

$$\Leftrightarrow w_y^T C_{yx} w_x + w_x^T C_{xy} w_y = 2\lambda$$

$$\Leftrightarrow (C_{yx} w_x)^T w_y + w_x^T C_{xy} w_y = 2\lambda$$

$$\Leftrightarrow w_x^T C_{yx}^T w_y + w_x C_{xy} w_y = 2\lambda$$

$$\Leftrightarrow 2 w_x^T C_{xy} w_y = 2\lambda$$

$$\Leftrightarrow w_x^T C_{xy} w_y = \lambda \quad \rightarrow \max w_x^T C_{xy} w_y = \max \lambda$$

b) maximize $w_x^T C_{xy} w_y$ with $w_x^T C_{xx} w_x = 1 = w_y^T C_{yy} w_y$

$w_x \rightarrow -w_x, \quad w_y \Rightarrow -w_y$:

$$(-w_x)^T C_{xy} (-w_y) \rightarrow w_x^T C_{xy} w_y$$
$$(-w_x)^T C_{xx} (-w_x) = 1 = w_x^T C_{xx} w_x$$
$$(-w_y)^T C_{yy} (-w_y) = 1 = w_y^T C_{yy} w_y$$

The maximization problem and the constraints stay the same

$\rightarrow w_x, w_y$ yields the same solution as $-w_x, -w_y$

2 Dual CCA

a) $w_x = S_x + n_x$, $w_y = S_y + n_y$

max $(S_x + n_x)^T C_{xy} (S_y + n_y)$ st $w_x^T C_{xy} w_y = (S_x + n_x)^T X Y^T (S_y + n_y)$

$n_x + n_y$ have no effect. We can write S_x as $X\alpha_x$ and S_y as $Y\alpha_y$ $\rightarrow$ weighted combination of our data will give us dual information

b)
$$\begin{pmatrix} 0 & XY^T \\ YX^T & 0 \end{pmatrix}\begin{pmatrix} X\alpha_x \\ Y\alpha_y \end{pmatrix} = \lambda \begin{pmatrix} XX^T & 0 \\ 0 & YY^T \end{pmatrix}\begin{pmatrix} X\alpha_x \\ Y\alpha_y \end{pmatrix}$$

$$= \begin{pmatrix} X^T & 0 \\ 0 & Y^T \end{pmatrix}\begin{pmatrix} 0 & XY^T \\ YX^T & 0 \end{pmatrix}\begin{pmatrix} X\alpha_x \\ Y\alpha_y \end{pmatrix} = \lambda \begin{pmatrix} X^T & 0 \\ 0 & Y^T \end{pmatrix}\begin{pmatrix} XX^T & 0 \\ 0 & YY^T \end{pmatrix}\begin{pmatrix} X\alpha_x \\ Y\alpha_y \end{pmatrix}$$

$$= \begin{pmatrix} 0 & X^TX\,Y^TY \\ Y^TY\,X^TX & 0 \end{pmatrix}\begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix} = \lambda \begin{pmatrix} X^TXX^TX & 0 \\ 0 & Y^TYY^TY \end{pmatrix}\begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}$$

$$= \begin{pmatrix} 0 & AB \\ BA & 0 \end{pmatrix}\begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix} = \lambda \begin{pmatrix} AA & 0 \\ 0 & BB \end{pmatrix}\begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}$$

Better:

$$\begin{pmatrix} 0 & C_{xy}Y \\ C_{yx}X & 0 \end{pmatrix}\begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix} = \lambda \begin{pmatrix} C_{xx}X & 0 \\ 0 & C_{yy}Y \end{pmatrix}\begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}$$

$$\Leftrightarrow \begin{pmatrix} 0 & XY^TY \\ YX^TX & 0 \end{pmatrix}\begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix} = \lambda \begin{pmatrix} XX^TX & 0 \\ 0 & YY^TY \end{pmatrix}\begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}$$

$$\Leftrightarrow \begin{pmatrix} X^T & 0 \\ 0 & Y^T \end{pmatrix}\begin{pmatrix} 0 & XY^TY \\ YX^TX & 0 \end{pmatrix}\begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix} = \lambda \begin{pmatrix} X^T & 0 \\ 0 & Y^T \end{pmatrix}\begin{pmatrix} XX^TX & 0 \\ 0 & YY^TY \end{pmatrix}\begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}$$

$$\Leftrightarrow \begin{pmatrix} 0 & X^TXY^TY \\ Y^TYX^TX & 0 \end{pmatrix}\begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix} = \lambda \begin{pmatrix} X^TXX^TX & 0 \\ 0 & Y^TYY^TY \end{pmatrix}\begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}$$

$$\Leftrightarrow \begin{pmatrix} 0 & AB \\ BA & 0 \end{pmatrix}\begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix} = \lambda \begin{pmatrix} A^2 & 0 \\ 0 & B^2 \end{pmatrix}\begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}$$

c) $\begin{pmatrix} 0 & AB \\ BA & 0 \end{pmatrix} \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix} = \rho \begin{pmatrix} A^2 & 0 \\ 0 & B^2 \end{pmatrix} \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}$

$\Leftrightarrow \begin{pmatrix} 0 & X^T X Y^T Y \\ Y^T Y X^T \cancel{X} & 0 \end{pmatrix} \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix} = \rho \begin{pmatrix} X^T X X^T X & 0 \\ 0 & Y^T Y Y^T Y \end{pmatrix} \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}$

$\Leftrightarrow \begin{pmatrix} X^T X Y^T Y \alpha_y \\ Y^T Y X^T X \alpha_x \end{pmatrix} = \rho \begin{pmatrix} X^T X X^T X \alpha_x \\ Y^T Y Y^T Y \alpha_y \end{pmatrix}$

$\Leftrightarrow \begin{pmatrix} X^T X Y^T w_y \\ Y^T Y X^T w_x \end{pmatrix} = \rho \begin{pmatrix} X^T X X^T w_y \\ Y^T Y Y^T w_y \end{pmatrix}$

$\Leftrightarrow (\alpha_x^T \quad \alpha_y^T) \begin{pmatrix} X^T C_{xy} w_y \\ Y^T C_{yx} w_x \end{pmatrix} = \rho (\alpha_x^T \quad \alpha_y^T) \begin{pmatrix} X^T C_{xx} w_x \\ Y^T C_{yy} w_y \end{pmatrix}$

$\Leftrightarrow w_x^T C_{xy} w_y + w_y^T C_{yx} w_x = \rho (w_x^T C_{xx} w_x + w_y^T C_{yy} w_y)$

$\Leftrightarrow 2 w_x^T C_{xy} w_y = 2\rho \Leftrightarrow w_x^T C_{xy} w_y = \rho$ $\rightarrow$ pick EV with highest EVa

d) $\begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix} \rightarrow \begin{pmatrix} w_x \\ w_y \end{pmatrix} = \begin{pmatrix} X \alpha_x \\ Y \alpha_y \end{pmatrix}$

1) $\begin{pmatrix} 0 & AB \\ BA & 0 \end{pmatrix} \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix} = \lambda \begin{pmatrix} A^2 & 0 \\ 0 & B^2 \end{pmatrix} \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}$

2) $\begin{pmatrix} 0 & AB \\ BA & 0 \end{pmatrix} \begin{pmatrix} w_x \\ w_y \end{pmatrix} = \lambda \begin{pmatrix} A^2 & 0 \\ 0 & B^2 \end{pmatrix} \begin{pmatrix} w_x \\ w_y \end{pmatrix}$

$\begin{pmatrix} 0 & AB \\ BA & 0 \end{pmatrix} \begin{pmatrix} X \alpha_x \\ Y \alpha_y \end{pmatrix} = \lambda \begin{pmatrix} A^2 & 0 \\ 0 & B^2 \end{pmatrix} \begin{pmatrix} X \alpha_x \\ Y \alpha_y \end{pmatrix}$

3 CCA and Least Square Regression

a) $\min_{v} \| X^T v - y \|^2 = \min_{v} (X^T v - y)^T (X^T v - y)$

$= \min_{v} (v^T X - y)(X^T v - y) = \min_{v} v^T X X^T v - v^T X y - y X^T v + y^2$

$= \min_{v} v^T X X^T v - v^T X y - (X y)^T V = \min v^T X X^T v - 2 v^T X y + y^2$

Find minimum: $\frac{\partial}{\partial v} v^T X X^T v - 2 v^T X y = 0$

$\Leftrightarrow \frac{\partial}{\partial b} a^T X X^T b + \frac{\partial}{\partial a} a^T X X^T b + \frac{\partial}{\partial v} - 2 v^T X y = 0$

$\Leftrightarrow v^T X X^T + v^T X X^T - 2 y^T X^T = 0$

$\Leftrightarrow 2 (X X^T)^T v - 2 X y = 0 \Leftrightarrow 2 X X^T v - 2 X y = 0$

$\Leftrightarrow X X^T v = X y \Leftrightarrow v = (X X^T)^{-1} X y$

From 1), $C_{xy} w_y = \lambda C_{xx} w_x \Leftrightarrow X_y w_y = \lambda X X^T w_x$

$\Leftrightarrow w_x = \frac{w_y}{\lambda} (X X^T)^{-1} X_y \Leftrightarrow w_x = \frac{1}{\lambda} w_y V$