

## EX 1 Dual CCA

 $N \leq d_1$ 

- a) Show, that it is always possible to find an optimal solution in the span of the data, that is,  $w_x = X\alpha_x$ ,  $w_y = Y\alpha_y$ .  
with some coefficient vectors  $\alpha_x \in \mathbb{R}^N$  and  $\alpha_y \in \mathbb{R}^N$ .

Let  $U = w_x^T X$  and  $V = w_y^T Y$  denote the projection of  $X \in \mathbb{R}^{d_1 \times N}$  by  $w_x \in \mathbb{R}^{d_1}$  and that of  $Y \in \mathbb{R}^{d_2 \times N}$  by  $w_y \in \mathbb{R}^{d_2}$ , respectively.  
The objective of linear CCA is to find <sup>linear</sup> projections that maximize Pearson's correlation between  $U$  and  $V$ .

To express the optimization problem in its dual form, we rewrite the solutions  $w_x, w_y$  as linear combinations of their corresponding training datapoints. The vectors  $w_x, w_y$  are expressed in their dual form by using new coordinate vectors  $\alpha_x \in \mathbb{R}^N$  and  $\alpha_y \in \mathbb{R}^N$ , so that  $w_x = X\alpha_x$  and  $w_y = Y\alpha_y$ .

We can write the above expressions, since ~~U and V are the projections of points X onto a direction  $w_x$  (and points Y onto a direction  $w_y$ , respectively)~~.

~~the dimensionality of the space is larger than the number of data points in the training set.~~  
From the properties of projections, directions can be expressed as linear combination of points ~~in the dual space~~ ~~expressing the~~

~~$w_x = \sum_{i=1}^n \alpha_i x_i = X\alpha_x$~~

~~because  $w_x$  and  $w_y$  will lie in the column space of  $X$  and  $Y$ .~~

Besides the properties of projections, we can use also the key property of the kernel methods and ~~RKHS~~ RKHS (reproducing kernel Hilbert spaces) that the ~~space~~ ~~is~~ ~~closed under linear combinations~~ ~~in the feature space~~ kernel scale is linear transformation in the feature space by expressing the weight vectors as linear combinations of the training data

$$w_x = \Phi_x \alpha = \sum_{i=1}^n \alpha_i \phi_x(x_i) = \sum_{i=1}^n \alpha_i \Phi_x(x_i) = X\alpha_x$$

$$\text{for } K_X = \langle \Phi_x, \Phi_x \rangle = X^T X$$

(see Representer Theorem)

the optimal direction vector is contained in the subspace that is spanned by the data points

Ex 1 b) Show that the dual optimization problem is equivalent to finding the solution of the generalized eigenvalue problem.

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

in  $dx, dy$   
where  $A = X^T X$  and  
 $B = Y^T Y$ .

The primal optimization problem is:

$$\begin{aligned} \text{Find } w_x \in \mathbb{R}^{d_1}, w_y \in \mathbb{R}^{d_2} \text{ maximizing } w_x^T C_{xy} w_y \\ \text{s.t. } w_x^T C_{xx} w_x = 1 \quad w_y^T C_{yy} w_y = 1 \\ C_{xx} = X^T X, \quad C_{xy} = X^T Y, \quad C_{yy} = Y^T Y \end{aligned}$$

The corresponding Lagrangian is:

$$L = w_x^T C_{xy} w_y - \frac{1}{2} \lambda_1 (w_x^T C_{xx} w_x - 1) - \frac{1}{2} \lambda_2 (w_y^T C_{yy} w_y - 1)$$

Substituting  $w_x = X \alpha_x$  and  $w_y = Y \alpha_y$ :

$$\begin{aligned} L &= \alpha_x^T X^T X Y^T Y \alpha_y - \frac{1}{2} \lambda_1 (\alpha_x^T X^T X X^T X \alpha_x - 1) - \frac{1}{2} \lambda_2 (\alpha_y^T Y^T Y Y^T Y \alpha_y - 1) \\ &= \alpha_x^T A \cdot B \alpha_y - \frac{1}{2} \lambda_1 (\alpha_x^T A \cdot A \alpha_x - 1) - \frac{1}{2} \lambda_2 (\alpha_y^T B \cdot B \alpha_y - 1) \end{aligned}$$

Taking derivatives wrt.  $\alpha_x$  and  $\alpha_y$  we obtain:

~~$$\frac{\partial L}{\partial \alpha_x} = A \cdot B \alpha_y - \lambda_1 A^2 \alpha_x = 0 \quad (1)$$~~

~~$$\frac{\partial L}{\partial \alpha_y} = B \cdot A \alpha_x - \lambda_2 B^2 \alpha_y = 0 \quad (2)$$~~

Subtracting  $\alpha_y^T$  times the second equation from  $\alpha_x^T$  times the first we have:

~~$$\begin{aligned} (3) \quad \alpha_x^T A \cdot B \alpha_y - \alpha_x^T \lambda_1 A^2 \alpha_x - \alpha_y^T B \cdot A \alpha_x + \alpha_y^T \lambda_2 B^2 \alpha_y = \\ = \lambda_2 \underbrace{\alpha_y^T B^2 \alpha_y}_{=1} - \lambda_1 \underbrace{\alpha_x^T A^2 \alpha_x}_{=1} = 0 \end{aligned}$$~~

The constraints of the primal problem can be rewritten as:  
(substitute:  $w_x = X \alpha_x$   $w_y = Y \alpha_y$ )

$$\begin{aligned} w_x^T C_{xx} w_x = 1 &\quad w_x^T X X^T w_x = 1 &\quad \alpha_x^T X^T X X^T X \alpha_x = 1 &\quad \alpha_x^T A^2 \alpha_x = 1 \\ w_y^T C_{yy} w_y = 1 &\quad w_y^T Y Y^T w_y = 1 &\quad \alpha_y^T Y^T Y Y^T Y \alpha_y = 1 &\quad \alpha_y^T B^2 \alpha_y = 1 \end{aligned}$$

What together with (3) implies:

$$\lambda_1 - \lambda_2 = 0 \quad . \quad \text{Let } \lambda := \lambda_1 = \lambda_2$$

Considering the case where the matrices  $A$  and  $B$  are invertible.  
( $X, Y$  of full rank), we have (from (2)):

~~$$\alpha_y = B^{-1} B^{-1} A^{-1} \alpha_x \quad \alpha_x = A^{-1} A^{-1} \alpha_y$$~~

$$BA\alpha_x = \lambda B^2\alpha_y$$

$$\text{From } \alpha_y^T B^2 \alpha_y = 1$$

$$\alpha_y^T B B^T \alpha_y = 1$$

$$(\alpha_y^T B)(\alpha_y^T B)^T = 1$$

$$(\alpha_y^T B)^T = (\alpha_y^T B)^{-1}$$

~~W = A\alpha\_x~~ ~~B\alpha\_y~~ ~~W = X\alpha\_x~~

~~J = X - \mu X~~ ~~J = Y - \mu Y~~

~~J = X - \mu X~~ ~~J = Y - \mu Y~~

and  $B = B^T$  since  $B$  is a square matrix  $N \times N$

$$\lambda = (B\alpha_y)^{-1} BA\alpha_x$$

$$(4) \quad \lambda = (B\alpha_y)^{-1} A\alpha_x = (B\alpha_y)^T A\alpha_x = \alpha_y^T B^T A\alpha_x = \\ = \alpha_x^T A B \alpha_y = w_x^T C_{xy} w_y = \rho \quad (\text{from the definition of Pearson's correlation coefficient})$$

Therefore by combining equations (1) and (2) and (4)

$$A \cdot B \alpha_y - \lambda A^2 \alpha_x = 0$$

$$B \cdot A \alpha_x - \lambda B^2 \alpha_y = 0$$

we can formulate the generalized eigenvalue problem (GEP)

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

where  $A, B$  are symmetric positive semi-definite.

Ex 1  
c)

A solution to the original problem can be obtained from a solution of the generalized eigenvalue problem by transforming  $\alpha_x$  and  $\alpha_y$  to  $w_x$  and  $w_y$ .

The solution in the input space :  $w_x = X\alpha_x$

$$w_y = Y\alpha_y$$

EX2 a) Describe how the CCA problem and its GEV formulation can be kernelized?

The starting point is to map the data-cases to feature vectors  $\Phi(x_i)$  and  $\Psi(y_i)$ . When the dimensionality of the space is larger than the number of data-cases in the training-set, then the solution must lie in the span of data-cases i.e.

$$w_x = \sum_i \alpha_i \Phi(x_i) = \Phi \alpha_x \quad w_y = \sum_i \alpha_i \Psi(y_i) = \Psi \alpha_y$$

In Kernel CCA, we suppose that the original data are mapped into a feature space via nonlinear functions. Then linear CCA is applied in the feature space.

Nonlinear functions  $\phi: \mathbb{R}^{d_1} \rightarrow H_x$  and  $\psi: \mathbb{R}^{d_2} \rightarrow H_y$  transform the original data  $X, Y$  to feature vectors in RKHS  $H_x$  and  $H_y$ .

Inner-product kernels for  $H_x$  and  $H_y$  are defined as:

$$k_x(x, x') = \langle \Phi(x), \Phi(x') \rangle$$

$$k_y(y, y') = \langle \Psi(y), \Psi(y') \rangle \quad (\text{Gram matrices})$$

$K_x$  and  $K_y$  are  $N \times N$  kernel matrices defined as

$$[K_x]_{nn'} = k_x(x_n, x_{n'}) \quad [K_y]_{nn'} = k_y(y_n, y_{n'}) \quad (\alpha_x, \alpha_y \in \mathbb{R}^N)$$

Similar to CCA, the first pair of dual vectors can be determined by solving

$$\max_{\alpha, \beta} \alpha^T K_x K_y \beta \quad \text{s.t. } \alpha^T K_x^2 \alpha = 1 \quad \alpha^T K_y \beta = 1$$

which can be obtained from CCA primal problem by substituting:

$w_x = \Phi \alpha_x$ ,  $w_y = \Psi \alpha_y$  and replacing ~~the original vectors  $x_i$~~  by their feature vectors  $\Phi(x_i)$  and  $y_i$  by  $\Psi(y_i)$ , respectively.

Moreover, it is needed to apply "kernel trick":  $\Phi(x_i)^T \Phi(x_j) = k_{ij}$  (i.e. implicitly mapping to some higher-dimensional feature space)

Exactly the same strategy applies to kernelize the dual problem (and its GEV formulation).