Sheet 3

EX 1 Dual CCA $\boxed{N \ll d_i}$

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 linear 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).

~~When the dimensionality of the space is larger than the number of data cases in the~~ training set.

From the properties of projections, directions ~~in the dimensionality of the space is larger than the~~ can be expressed as linear combination of points

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

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 (reproducing kernel Hilbert spaces) that the kernel seeks linear transformation in the feature space by expressing the weight vectors as linear combinations of the training data

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

for $K_x = \langle \Phi_x, \Phi_x \rangle = X^T X$

(see Representer Theorem)

or in other words: 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.

$$\begin{bmatrix} 0 & A\cdot B \\ B\cdot A & 0 \end{bmatrix}\begin{bmatrix} \alpha_x \\ \alpha_y \end{bmatrix} = \rho \begin{bmatrix} A^2 & 0 \\ 0 & B^2 \end{bmatrix}\begin{bmatrix} \alpha_x \\ \alpha_y \end{bmatrix}$$ in α_x, α_y where $A = X^TX$ and $B = Y^TY$.

The primal optimization problem is:

Find $w_x \in \mathbb{R}^{d_1}, w_y \in \mathbb{R}^{d_2}$ maximizing $w_x^T C_{xy} w_y$

s.t. $w_x^T C_{xx} w_x = 1 \quad w_y^T C_{yy} w_y = 1$

$C_{xx} = XX^T, \; C_{xy} = XY^T, \; C_{yy} = YY^T$

The corresponding Lagrangian is:

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

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

$$\mathcal{L} = \alpha_x^T X^T X Y^T Y \alpha_y - \frac{1}{2}\lambda_1 (\alpha_x^T X^T X X^T X \alpha_x - 1) - \frac{1}{2}\lambda_2 (\alpha_y^T Y^T Y Y^T Y \alpha_y - 1)$$

$$= \alpha_x^T A\cdot B\, \alpha_y - \frac{1}{2}\lambda_1 (\alpha_x^T A\cdot A\, \alpha_x - 1) - \frac{1}{2}\lambda_2 (\alpha_y^T B B\, \alpha_y - 1)$$

Taking derivatives wrt. α_x and α_y we obtain:

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

$$\frac{\partial \mathcal{L}}{\partial \alpha_y} = BA\alpha_x - \lambda_2 B^2 \alpha_y = 0 \quad (2)$$

Substracting α_y^T times the second equation from α_x^T times the first we have:

$$(3)\quad \alpha_x^T A\cdot B\alpha_y - \alpha_x^T \lambda_1 A^2 \alpha_x - \alpha_y^T BA\alpha_x + \alpha_y^T \lambda_2 B^2 \alpha_y =$$

$$= \lambda_2 \underbrace{\alpha_y^T B^2 \alpha_y}_{=1} - \lambda_1 \underbrace{\alpha_x^T A^2 \alpha_x}_{=1} = 0$$

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

$w_x^T C_{xx} w_x = 1 \quad w_x^T XX^T w_x = 1 \quad \alpha_x^T X^T XX^T X\alpha_x = 1 \quad \alpha_x^T A^2 \alpha_x = 1$

$w_y^T C_{yy} w_y = 1 \quad w_y^T YY^T w_y = 1 \quad \alpha_y^T Y^T YY^T Y\alpha_y = 1 \quad \alpha_y^T B^2 \alpha_y = 1$

What together with (3) implies:

$\lambda_1 - \lambda_2 = 0$. 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)):

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

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

From $\alpha_y^T B^2 \alpha_y = 1$

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

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

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

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

$\rightarrow \lambda = (B^2 \alpha_y)^{-1} BA\alpha_x$

(4) $\lambda = (B\alpha_y)^{-1} A\alpha_x = (B\alpha_y)^T A\alpha_x = \alpha_y^T B^T A \alpha_x =$

$= \alpha_x^T AB \alpha_y = w_x^T C_{xy} w_y = \rho$ (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.

Ex1

c)

A solution to the original problem can be obtained from a solution of the generalized eigenvalue problem by transforming α_x and α_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 kernalized?

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.

$$w_x = \sum_i \alpha_{xi} \Phi(x_i) = \Phi \alpha_x \qquad w_y = \sum_i \alpha_{yi} \Psi(y_i) = \Psi \alpha_y$$

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 $\phi : \mathbb{R}^{d_1} \to \mathcal{H}_x$ and $\psi : \mathbb{R}^{d_2} \to \mathcal{H}_y$ transform the original data X, Y to feature vectors in RKHS $\mathcal{H}_x$ and $\mathcal{H}_y$.
Inner-product kernels for $\mathcal{H}_x$ and $\mathcal{H}_y$ are defined as:

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

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

$$[K_x]_{nn'} = k_x(x_n, x_{n'}) \qquad [K_y]_{nn'} = k_y(y_n, y_{n'})$$

Similar to CCA, the first pair of dual vectors ($\alpha_x, \alpha_y \in \mathbb{R}^N$) can be determined by solving

$$\max_{\alpha, \beta} \alpha_x^T K_x K_y \alpha_y \qquad \text{s.t. } \alpha_x^T K_x^2 \alpha_x = 1, \; \alpha_y^T K_y^2 \alpha_y = 1$$

which can be obtained from CCA primal problem by substituting: $w_x = \Phi \alpha_x$, $w_y = \Psi \alpha_y$ and replacing 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 kernalize the dual problem (and its GEV formulation).

Ex 2 b) Explain how the solution of the resulting kernelized CCA are to be interpreted, and under which condition the solution can/cannot be expressed as directions in the input spaces $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$.

Classical CCA looks for linear mappings $w_x^T X$ and $w_y^T Y$ that achieve maximum correlation. Kernel CCA extends this approach by looking for functions $\Phi \in H_X$ and $\Psi \in H_Y$ such that the random variables $\Phi(X)$ and $\Psi(Y)$ have maximal correlation. The nonlinear mappings allow to clarify the dependency between X and Y, even though it cannot be captured by classical CCA (if they have no linear correlation). 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.

Therefore, analogously to classical CCA, where solutions w_x, w_y are the weights of the linear combinations which maximize the correlation between the variables X and Y, (i.e. in input space), then in Kernel CCA the solutions α_x and α_y are interpreted as the weights that maximize the correlation between the (nonlinearly) transformed variables (i.e. in the feature space).
The solutions α_x and α_y can be expressed as directions in the input spaces $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$ if they are transformed to

$$w_X = \Phi(X)\alpha_x \quad \text{and} \quad w_Y = \Psi(Y)\alpha_y.$$

The sample KCCA crucially depends on the relation between the sample size and the dimensionalities of the space involved. Mappings into higher dimensional spaces are most likely to increase the Kernel Canonical Correlation coefficient relative to linear CCA between the input spaces. Therefore the KCCA has to be interpreted with caution (and should rather be considered as a geometrical algorithm to construct highly correlated features).

3. (b)

unconstrained objective

$$\max_{\phi_x, \phi_y, w_x, w_y} \; w_x^T C_{xy} w_y + \alpha \cdot [\min(0, 1 - w_x^T C_{xx} w_x + \min(0, 1 - w_y^T C_{yy} w_y)]$$

for $w_x \in R^{h_1}$ and $w_y \in R^{h_2}$

original CCA objective

$$\max_{w_x, w_y} \; w_x^T C_{xy} w_y$$

s.t. $w_x^T C_{xx} w_x = 1$

$w_y^T C_{yy} w_y = 1$

⇒ As, here the function $\phi_x(x; \phi_x)$ and $\phi_y(x; \phi_y)$ is nothing but passing the input views in a stacked non-linear representation. So, on the equation given above it is simply calculating CCA over the outputs of ϕ functions.

This arrangement represents exactly the CCA form it performed linearly, and can be represented.

$$(\phi_x^*, \phi_y^*, w_x, w_y) = \max_{\phi_x, \phi_y, w_x, w_y} \text{corr}\left(w_x \phi(x, \phi_x), w_y \phi(y, \phi_y)\right)$$

To work with the constraints :-

orig. CCA = $\max_{w_x, w_y} w_x^T C_{xy} w_y$

$$w_x^T C_{xx} w_x = 1$$
$$w_y^T C_{yy} w_y = 1$$

replace co-variance matrices from linear to non-linear.

$$C_{xx} = \mathbb{E}[\phi_x \phi_x^T]$$
$$C_{xy} = \mathbb{E}[\phi_y \phi_y^T]$$
$$C_{xy} = \mathbb{E}[\phi_x \phi_y^T]$$

unconstrained environment

$$w_x^T C_{xx} w_x - 1 = 0$$
$$w_y^T C_{yy} w_y - 1 = 0$$

$\Rightarrow \min[0, 1 - w_x^T C_{xx} w_x] = 0 = \lambda_x \min$
$\Rightarrow \min[0, 1 - w_y^T C_{yy} w_y] = 0 = \lambda_y \min$

} apply

so, $\max_{\phi_x, \phi_y, w_x, w_y} w_x^T C_{xy} w_y + \alpha^* \lambda_x \min + \alpha^* \lambda_y \min$

$= \max_{\phi_x \phi_y, w_x w_y} \dfrac{w_x^T C_{xy} w_y + \alpha[\min(0, 1 - w_x^T C_{xx} w_x) + \min(0, 1 - w_y^T C_{yy} w_y)]}{}$

exactly same as DCCA

or CCA in non-linear manner.

3. b

Gradient as a function of Jacobian matrix.

$$f = \omega_x^T C_{xy} \omega_y + \alpha \left[\min(0, 1 - \omega_x^T C_{xx} \omega_x) + \min(0, 1 - \omega_y^T C_{yy} \omega_y) \right]$$

gradient of f with respect to ϕ_x

$$= \omega_x^T \frac{C_{xy}}{d\phi_x} \omega_y + \alpha \left(- \omega_x^T \frac{C_{xx}}{d\phi_x} \omega_x \right) = 0$$

$$\omega_x^T \frac{E[\phi_x \phi_y^T]}{d\phi_x} \omega_y + - \alpha \left(\omega_x^T \frac{E(\phi_x \phi_x^T)}{d\phi_x} \omega_x \right) = 0$$

$$\omega_x^T E\left(\frac{d\phi_x}{d\phi_x}, \frac{d\phi_y}{d\phi_x} \right) \omega_y - \alpha \left(\omega_x^T E\left(\frac{d\phi_x}{d\phi_x} \frac{d\phi_x^T}{d\phi_x} \right) \omega_x \right) = 0$$

ϕ_x and ϕ_y are independent

$$\omega_x^T E\left(\frac{d\phi_x}{d\phi_x} \right) E\left(\frac{d\phi_y}{d\phi_x} \right) \omega_y = - \alpha \left(\omega_x^T E\left(\frac{2 \, d\phi_x}{d\phi_x} \right) \omega_x \right)$$

$$= 0 - 2\alpha \, \omega_x^T E\left(\frac{d\phi_x}{d\phi_x} \right) \omega_x$$

$$\Delta_{\phi_x} = - 2\alpha \, \omega_x^T E\left(\frac{d\phi_x}{d\phi_x} \right) \omega_x \Rightarrow \text{Solution.}$$