Notes on Statistical and Machine Learning

Canonical Correlation Analysis

Chapter: *24* **Prepared by:** *Chenxi Zhou*

This note is prepared based on

- *Chapter 7, Linear Dimensionality Reduction* in Izenman (2009),
- *Chapter 15, Latent Variable Models for Blind Source Separation* in Izenman (2009), and
- *Chapter 8, Sparse Multivariate Methods* in Hastie, Tibshirani, and Wainwright (2015).

I. Canonical Variates and Canonical Correlations

1. **Introduction:** *Canonical correlation analysis* (CCA) is a method for studying linear relationships between two vector variates, $X = (X_1, \cdots, X_p)^\top$ and $Y = (Y_1, \cdots, Y_s)^\top$.

2. **Basic Setup:** Let
$$\begin{pmatrix} X \\ Y \end{pmatrix}$$
be a collection of $p + s$ variables partitioned to two disjoint sub-collections. Furthermore, assume X and Y are jointly distributed with mean
$$\mathbb{E}\left[\begin{pmatrix} X \\ Y \end{pmatrix}\right] = \begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{pmatrix}$$
and the covariance matrix
$$\mathrm{Cov}\left[\begin{pmatrix} X \\ Y \end{pmatrix}\right] = \mathbb{E}\left[\begin{pmatrix} X - \boldsymbol{\mu}_X \\ Y - \boldsymbol{\mu}_Y \end{pmatrix}\begin{pmatrix} X - \mu_X \\ Y - \mu_Y \end{pmatrix}^\top\right] = \begin{pmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY} \end{pmatrix}.$$
We assume $\boldsymbol{\Sigma}_{XX}$ and $\boldsymbol{\Sigma}_{YY}$ are nonsingular.

3. **Main Idea of CCA:** CCA seeks to replace the two sets of correlated variables, X and Y, by t pairs of new variables
$$(\xi_j, \omega_j), \qquad \text{for } j = 1, \cdots, t, \text{ with } t \le \min\{p, s\}.$$
Here, for all $j = 1, \cdots, t$,
$$\begin{aligned} \xi_j &:= \mathbf{g}_j^\top X = g_{j,1}X_1 + g_{j,2}X_2 + \cdots + g_{j,p}X_p, \\ \omega_j &:= \mathbf{h}_j^\top Y = h_{j,1}Y_1 + h_{j,2}Y_2 + \cdots + h_{j,s}Y_s, \end{aligned}$$
are linear projections of X and Y, respectively. The j-th pair of coefficient vectors, $\mathbf{g}_j := (g_{j,1}, \cdots, g_{j,p})^\top \in \mathbb{R}^p$ and $\mathbf{h}_j := (h_{j,1}, \cdots, h_{j,s})^\top \in \mathbb{R}^s$ are chosen so that

(a) the pairs $\{(\xi_j, \omega_j)\}_{j=1}^t$ are ranked in importance through their correlations

$$\rho_j := \mathrm{Cor}(\xi_j, \omega_j) = \frac{\mathrm{Cov}(\xi_j, \omega_j)}{\sqrt{\mathrm{Var}[\xi_j]} \cdot \sqrt{\mathrm{Var}[\omega_j]}} = \frac{\mathbf{g}_j^\top \mathbf{\Sigma}_{XY} \mathbf{h}_j}{\sqrt{\mathbf{g}_j^\top \mathbf{\Sigma}_{XX} \mathbf{g}_j} \cdot \sqrt{\mathbf{h}_j^\top \mathbf{\Sigma}_{YY} \mathbf{h}_j}}, \tag{1}$$

which are listed in the descending order of magnitude, i.e., $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_t$;

(b) ξ_j is uncorrelated with all previously derived ξ_i's, that is,

$$\mathrm{Cov}(\xi_j, \xi_i) = g_j^\top \mathbf{\Sigma}_{XX} g_i = 0, \qquad \text{for all } i < j; \tag{2}$$

(c) ω_j is uncorrelated with all previously derived ω_i's, that is,

$$\mathrm{Cov}(\omega_j, \omega_i) = h_j^\top \mathbf{\Sigma}_{YY} h_i = 0, \qquad \text{for all } i < j. \tag{3}$$

The pairs $\{(\xi_j, \omega_j)\}_{j=1}^t$ are called the *first t pairs of canonical variates of X and Y* and their correlations (1) are called the *t largest canonical correlations*.

Remark. If the correlation is regarded as the primary determinant of information in the system of variables, then CCA is a major tool for reducing the dimensionality of the original two sets of variables.

II. Least-Squares Optimality of CCA

1. **Setup and Goal:** Let $\mathbf{G} \in \mathbb{R}^{t \times p}$ and $\mathbf{H} \in \mathbb{R}^{t \times s}$, where $1 \leq t \leq \min\{p, s\}$, be the matrices of weights such that X and Y are linear projected into new vector variates, that is,

$$\boldsymbol{\xi} = \mathbf{G}X, \qquad \text{and} \qquad \boldsymbol{\omega} = \mathbf{H}Y, \tag{4}$$

respectively, where $\boldsymbol{\xi} := (\xi_1, \xi_2, \cdots, \xi_t)^\top$ and $\boldsymbol{\omega} := (\omega_1, \omega_2, \cdots, \omega_t)^\top$. Consider the problem of finding $\boldsymbol{\nu} \in \mathbb{R}^t$, $\mathbf{G}$ and $\mathbf{H}$ to minimize

$$\mathrm{trace}\left\{ \mathbb{E}\Big[\big(\mathbf{H}Y - \boldsymbol{\nu} - \mathbf{G}X\big)\big(\mathbf{H}Y - \boldsymbol{\nu} - \mathbf{G}X\big)^\top \Big] \right\}, \tag{5}$$

where we assume that the covariance matrix of $\mathbf{H}Y$ is $\mathbf{\Sigma}_{\boldsymbol{\omega}\boldsymbol{\omega}} := \mathbf{H}\mathbf{\Sigma}_{YY}\mathbf{H}^\top = \mathbf{I}_t$. In other words, we are trying to find $\boldsymbol{\nu} \in \mathbb{R}^t$, $\mathbf{G}$ and $\mathbf{H}$ such that

$$\mathbf{H}Y \approx \boldsymbol{\nu} + \mathbf{G}X.$$

2. **Derivation of the Minimizer:** Note that

$$\begin{aligned} f(\boldsymbol{\nu}, \mathbf{G}, \mathbf{H}) := \mathbb{E}\Big[\big(\mathbf{H}Y - \boldsymbol{\nu} - \mathbf{G}X\big)\big(\mathbf{H}Y - \boldsymbol{\nu} - \mathbf{G}X\big)^\top \Big] \\ = \mathbf{H}(\mathbf{\Sigma}_{YY} + \boldsymbol{\mu}_Y \boldsymbol{\mu}_Y^\top)\mathbf{H}^\top - \mathbf{H}\boldsymbol{\mu}_Y \boldsymbol{\nu}^\top - \mathbf{H}(\mathbf{\Sigma}_{YX} + \boldsymbol{\mu}_Y \boldsymbol{\mu}_X^\top)\mathbf{G}^\top \\ - \boldsymbol{\nu}\boldsymbol{\mu}_Y \mathbf{H}^\top + \boldsymbol{\nu}\boldsymbol{\nu}^\top + \boldsymbol{\nu}\boldsymbol{\mu}_X \mathbf{G}^\top \\ - \mathbf{G}(\mathbf{\Sigma}_{XY} + \boldsymbol{\mu}_X \boldsymbol{\mu}_Y^\top)\mathbf{H}^\top + \mathbf{G}\boldsymbol{\mu}_X \boldsymbol{\nu}^\top + \mathbf{G}(\mathbf{\Sigma}_{XX} + \boldsymbol{\mu}_X \boldsymbol{\mu}_X^\top)\mathbf{G}^\top. \end{aligned} \tag{6}$$

We first fix $\mathbf{G}$ and $\mathbf{H}$ and minimize over $\boldsymbol{\nu}$:

$$\begin{aligned}
\operatorname{trace}\big\{f(\boldsymbol{\nu}, \mathbf{G}, \mathbf{H})\big\} = \operatorname{trace}\Big\{&\big(\boldsymbol{\nu} - (\mathbf{H}\boldsymbol{\mu}_Y - \mathbf{G}\boldsymbol{\mu}_X)\big)\big(\boldsymbol{\nu} - (\mathbf{H}\boldsymbol{\mu}_Y - \mathbf{G}\boldsymbol{\mu}_X)\big)^\top \\
&- (\mathbf{H}\boldsymbol{\mu}_Y - \mathbf{G}\boldsymbol{\mu}_X)(\mathbf{H}\boldsymbol{\mu}_Y - \mathbf{G}\boldsymbol{\mu}_X)^\top \\
&+ \mathbf{H}(\boldsymbol{\Sigma}_{YY} + \boldsymbol{\mu}_Y\boldsymbol{\mu}_Y^\top)\mathbf{H}^\top - \mathbf{H}(\boldsymbol{\Sigma}_{YX} + \boldsymbol{\mu}_Y\boldsymbol{\mu}_X^\top)\mathbf{G}^\top \\
&- \mathbf{G}(\boldsymbol{\Sigma}_{XY} + \boldsymbol{\mu}_X\boldsymbol{\mu}_Y^\top)\mathbf{H}^\top + \mathbf{G}(\boldsymbol{\Sigma}_{XX} + \boldsymbol{\mu}_X\boldsymbol{\mu}_X^\top)\mathbf{G}^\top\Big\} \\
\geq \operatorname{trace}\Big\{&- (\mathbf{H}\boldsymbol{\mu}_Y^\top - \mathbf{G}\boldsymbol{\mu}_X^\top)(\mathbf{H}\boldsymbol{\mu}_Y^\top - \mathbf{G}\boldsymbol{\mu}_X^\top)^\top \\
&+ \mathbf{H}(\boldsymbol{\Sigma}_{YY} + \boldsymbol{\mu}_Y\boldsymbol{\mu}_Y^\top)\mathbf{H}^\top - \mathbf{H}(\boldsymbol{\Sigma}_{YX} + \boldsymbol{\mu}_Y\boldsymbol{\mu}_X^\top)\mathbf{G}^\top \\
&- \mathbf{G}(\boldsymbol{\Sigma}_{XY} + \boldsymbol{\mu}_X\boldsymbol{\mu}_Y^\top)\mathbf{H}^\top + \mathbf{G}(\boldsymbol{\Sigma}_{XX} + \boldsymbol{\mu}_X\boldsymbol{\mu}_X^\top)\mathbf{G}^\top\Big\} \\
= \operatorname{trace}\Big\{&\mathbf{H}\boldsymbol{\Sigma}_{YY}\mathbf{H}^\top - \mathbf{H}\boldsymbol{\Sigma}_{YX}\mathbf{G}^\top - \mathbf{G}\boldsymbol{\Sigma}_{XY}\mathbf{H}^\top + \mathbf{G}\boldsymbol{\Sigma}_{XX}\mathbf{G}^\top\Big\},
\end{aligned}$$

where the inequality becomes an equality if and only if

$$\boldsymbol{\nu} = \mathbf{H}\boldsymbol{\mu}_Y - \mathbf{G}\boldsymbol{\mu}_X. \tag{7}$$

We next minimizer over the matrix $\mathbf{G}$ by noticing that

$$\begin{aligned}
&\operatorname{trace}\Big\{\mathbf{H}\boldsymbol{\Sigma}_{YY}\mathbf{H}^\top - \mathbf{H}\boldsymbol{\Sigma}_{YX}\mathbf{G}^\top - \mathbf{G}\boldsymbol{\Sigma}_{XY}\mathbf{H}^\top + \mathbf{G}\boldsymbol{\Sigma}_{XX}\mathbf{G}^\top\Big\} \\
&= \operatorname{trace}\Big\{\big(\mathbf{G}\boldsymbol{\Sigma}_{XX}^{1/2} - \mathbf{H}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-\frac{1}{2}}\big)\big(\mathbf{G}\boldsymbol{\Sigma}_{XX}^{1/2} - \mathbf{H}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-\frac{1}{2}}\big) \\
&\qquad - \mathbf{H}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\mathbf{H}^\top + \mathbf{H}\boldsymbol{\Sigma}_{YY}\mathbf{H}^\top\Big\} \\
&\geq \operatorname{trace}\big\{\mathbf{H}\boldsymbol{\Sigma}_{YY}\mathbf{H}^\top - \mathbf{H}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\mathbf{H}^\top\big\} \\
&= t - \sum_{j=1}^{t}\lambda_j\big(\mathbf{H}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\mathbf{H}^\top\big),
\end{aligned}$$

where the inequality becomes an equality if and only if

$$\mathbf{G} = \mathbf{H}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}. \tag{8}$$

Finally, we minimize over the matrix $\mathbf{H}$. Let $\mathbf{U}^\top := \mathbf{H}\boldsymbol{\Sigma}_{YY}^{\frac{1}{2}}$ so that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_t$. By the Poincaré Separation Theorem[1], we have

$$t - \sum_{j=1}^{t}\lambda_j(\mathbf{U}^\top\mathbf{R}\mathbf{U}) \geq t - \sum_{j=1}^{t}\lambda_j(\mathbf{R}), \tag{9}$$

[1] The *Poincaré Separation Theorem* says the following: if $\mathbf{A}$ is an $(n\times n)$-matrix and $\mathbf{U}$ is an $(n\times k)$-matrix, where $k \leq n$, such that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k$. Then,

$$\lambda_j(\mathbf{U}^\top\mathbf{A}\mathbf{U}) \leq \lambda_j(\mathbf{A}),$$

with equality if the columns of $\mathbf{U}$ are the first k eigenvectors of $\mathbf{A}$.

where

$$\mathbf{R} := \mathbf{\Sigma}_{YY}^{-\frac{1}{2}} \mathbf{\Sigma}_{YX} \mathbf{\Sigma}_{XX}^{-1} \mathbf{\Sigma}_{XY} \mathbf{\Sigma}_{YY}^{-\frac{1}{2}}, \tag{10}$$

with equality being held if and only if the columns of $\mathbf{U}$ are the eigenvectors associated with the first t eigenvalues of $\mathbf{R}$.

Therefore, the $\boldsymbol{\nu}$, $\mathbf{G}$ and $\mathbf{H}$ that minimize (5) are given by

$$\begin{aligned}
\boldsymbol{\nu}^{(t)} &:= \mathbf{H}^{(t)} \boldsymbol{\mu}_Y - \mathbf{G}^{(t)} \boldsymbol{\mu}_X, \\
\mathbf{G}^{(t)} &:= \mathbf{V}^{(t)} \mathbf{\Sigma}_{YY}^{-\frac{1}{2}} \mathbf{\Sigma}_{YX} \mathbf{\Sigma}_{XX}^{-1}, \\
\mathbf{H}^{(t)} &:= \mathbf{V}^{(t)} \mathbf{\Sigma}_{YY}^{-\frac{1}{2}},
\end{aligned}$$

where

$$\mathbf{V}^{(t)} := \begin{pmatrix} \mathbf{v}_1^\top \\ \vdots \\ \mathbf{v}_t^\top \end{pmatrix}$$

is a $(t \times s)$-matrix with the j-th row $\mathbf{v}_j$ being the eigenvector associated with the j-th largest eigenvalue of $\mathbf{R}$, for all $j = 1, \cdots, t$.

3. **Canonical Variate Score:** Let $\mathbf{g}_j := (g_{j,1}, \cdots, g_{j,p})^\top \in \mathbb{R}^p$ and $\mathbf{h}_j := (h_{j,1}, \cdots, h_{j,s})^\top \in \mathbb{R}^s$ be the j-th rows of $\mathbf{G}^{(t)}$ and $\mathbf{H}^{(t)}$, respectively, for all $j = 1, 2, \cdots, t$. The *j-th pair of canonical variate score*, (ξ_j, ω_j), is given by

$$\xi_j = \mathbf{g}_j^\top X, \qquad \omega_j = \mathbf{h}_j^\top Y, \tag{11}$$

where

$$\mathbf{g}_j = \mathbf{\Sigma}_{XX}^{-1} \mathbf{\Sigma}_{XY} \mathbf{\Sigma}_{YY}^{-\frac{1}{2}} \mathbf{v}_j, \qquad \text{and} \qquad \mathbf{h}_j = \mathbf{\Sigma}_{YY}^{-\frac{1}{2}} \mathbf{v}_j, \tag{12}$$

for all $j = 1, \cdots, t$.

4. **Covariance Matrix of Canonical Variate Scores:** Consider the vector of the canonical variate scores

$$\boldsymbol{\xi}^{(t)} = \mathbf{G}^{(t)} X, \qquad \text{and} \qquad \boldsymbol{\omega}^{(t)} = \mathbf{H}^{(t)} Y. \tag{13}$$

Then, the covariance matrix between $\boldsymbol{\xi}^{(t)}$ and $\boldsymbol{\omega}^{(t)}$ is

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{\xi}^{(t)}, \boldsymbol{\omega}^{(t)}) &= \mathrm{Cov}(\mathbf{G}^{(t)} X, \mathbf{H}^{(t)} Y) = \mathbf{G}^{(t)} \mathrm{Cov}(X, Y) {\mathbf{H}^{(t)}}^\top \\
&= \mathbf{V}^{(t)} \mathbf{\Sigma}_{YY}^{-\frac{1}{2}} \mathbf{\Sigma}_{YX} \mathbf{\Sigma}_{XX}^{-1} \mathbf{\Sigma}_{XY} \mathbf{\Sigma}_{YY}^{-\frac{1}{2}} {\mathbf{V}^{(t)}}^\top \\
&= \mathbf{V}^{(t)} \mathbf{R} {\mathbf{V}^{(t)}}^\top \\
&= \mathbf{\Lambda}^{(t)} \\
&= \mathrm{diag}(\lambda_1(\mathbf{R}), \cdots, \lambda_t(\mathbf{R})).
\end{aligned} \tag{14}$$

The covariance matrix of $\boldsymbol{\xi}^{(t)}$ is

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{\xi}^{(t)}, \boldsymbol{\xi}^{(t)}) &= \mathbf{G}^{(t)} \,\mathrm{Cov}(X, X) \mathbf{G}^{(t)\top} \\ &= \mathbf{V}^{(t)} \boldsymbol{\Sigma}_{YY}^{-\frac{1}{2}} \boldsymbol{\Sigma}_{YX} \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{XX} \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{XY} \boldsymbol{\Sigma}_{YY}^{-\frac{1}{2}} \mathbf{V}^{(t)\top} \\ &= \mathbf{V}^{(t)} \boldsymbol{\Sigma}_{YY}^{-\frac{1}{2}} \boldsymbol{\Sigma}_{YX} \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{XY} \boldsymbol{\Sigma}_{YY}^{-\frac{1}{2}} \mathbf{V}^{(t)\top} \\ &= \boldsymbol{\Lambda}^{(t)}. \end{aligned} \tag{15}$$

And, finally, the covariance matrix of $\boldsymbol{\omega}^{(t)}$ is

$$\begin{aligned}\mathrm{Cov}(\omega^{(t)}, \omega^{(t)}) &= \mathbf{H}^{(t)} \,\mathrm{Cov}(Y, Y) \mathbf{H}^{(t)\top} \\ &= \mathbf{H}^{(t)} \boldsymbol{\Sigma}_{YY} \mathbf{H}^{(t)\top} \\ &= \mathbf{I}_t. \end{aligned}$$

It follows that the correlation matrix between $\boldsymbol{\xi}^{(t)}$ and $\boldsymbol{\omega}^{(t)}$ is

$$\mathrm{Cor}(\boldsymbol{\xi}^{(t)}, \boldsymbol{\omega}^{(t)}) = \boldsymbol{\Lambda}^{\frac{1}{2}}. \tag{16}$$

Consequently, we have

$$\mathrm{Cor}(\xi_j^{(t)}, \xi_k^{(t)}) = \mathrm{Cor}(\omega_j^{(t)}, \omega_k^{(t)}) = \delta_{j,k}, \qquad \mathrm{Cor}(\xi_j^{(t)}, \omega_k^{(t)}) = \rho_j \cdot \delta_{j,k}, \tag{17}$$

where $\rho_j := \sqrt{\lambda_j(\mathbf{R})}$ and $\delta_{j,k}$ is the Kronecker delta, for all $j, k = 1, \cdots, t$.

5. **Interpretation of the Coefficients $\{g_{i,j}\}$ and $\{h_{i,j}\}$:** We choose the coefficients $\{g_{i,j}\}_{i\in\{1,\cdots,t\}, j\in\{1,\cdots,p\}}$ and $\{h_{i,j}\}_{i\in\{1,\cdots,t\}, j\in\{1,\cdots,s\}}$ so that

   (a) the first pair (ξ_1, ω_1) has the largest possible correlation among all linear combinations of X and Y;

   (b) For $j = 2, \cdots, t$, the j-th pair (ξ_j, ω_j) has the largest possible correlation ρ_j among all linear combinations of X and Y in which ξ_j is uncorrelated with $\xi_1, \xi_2, \cdots, \xi_{j-1}$ and ω_j is uncorrelated with $\omega_1, \omega_2, \cdots, \omega_{j-1}$.

   It follows that

   $$1 > \rho_1 > \rho_2 > \cdots > \rho_t > 0. \tag{18}$$

   Here, the correlation coefficient, ρ_j, between ξ_j and ω_j, is called the *canonical correlation coefficient* associated with the j-th pair of canonical variates for all $j = 1, \cdots, t$.

6. **Special Case — when $s = 1$:** When $s = 1$, the matrix $\mathbf{R}$ reduces to the squared multiple correlation coefficient of Y with the best linear predictor of Y using $X_1, \cdots, X_p$:

   $$R = \frac{\boldsymbol{\sigma}_{YX}^\top \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\sigma}_{XY}}{\sigma_Y^2},$$

   where σ_Y^2 is the variance of Y and $\boldsymbol{\sigma}_{XY} \in \mathbb{R}^p$ is the covariance between X and Y.

   The j-th canonical correlation coefficient, ρ_j, can be interpreted as the multiple correlation coefficient of either $\xi_j = \mathbf{g}_j^\top X$ with Y or $\omega_j = \mathbf{h}_j^\top Y$ with X.

7. **Special Case — when $s = p = 1$:** When $s = p = 1$, $\mathbf{R}$ is the *squared correlation coefficient* between Y and X,

$$\mathbf{R} = \rho^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2},$$

where σ_X^2 and σ_Y^2 are the variances of X and Y, respectively, and σ_{XY} is the covariance between X and Y.

8. **Invariance:** Canonical correlations are *invariant* under simultaneous nonsingular linear transformation of the random vectors X and Y. Let $\mathbf{D} \in \mathbb{R}^{p\times p}$ and $\mathbf{F} \in \mathbb{R}^{s\times s}$ be two nonsingular matrices and let

$$X' = \mathbf{D}X, \qquad Y' = \mathbf{F}Y.$$

Then, the canonical correlations of $\mathbf{D}X$ and $\mathbf{F}Y$ are identical to those of X and Y.

Remark. One consequence of this invariance property is that the canonical correlations obtained from the *covariance* matrix and those from the *correlation* matrix are identical.

III. CCA as a Correlation-Maximization Technique

1. **Goal:** We derive CCA by maximizing the correlation between linear combinations of X and those of Y.

2. **Assumption:** Assume $\mathbb{E}[X] = \mathbf{0}_p$ and $\mathbb{E}[Y] = \mathbf{0}_s$. Consequently, all ξ_j's and ω_j's, for $j = 1, 2, \cdots, t$, have zero mean.

3. **Derivation of (ξ_1, ω_1):** Consider an arbitrary linear projections $\xi = \mathbf{g}^\top X$ and $\omega = \mathbf{h}^\top Y$, and assume that they both have unit variances

$$\mathrm{Var}(\xi) = \mathbf{g}^\top \mathbf{\Sigma}_{XX}\mathbf{g} = 1, \qquad \text{and} \qquad \mathrm{Var}(\omega) = \mathbf{h}^\top \mathbf{\Sigma}_{YY}\mathbf{h} = 1. \tag{19}$$

Then, we find vectors $\mathbf{g}$ and $\mathbf{h}$ such that the random variables ξ and ω have maximal correlations

$$\mathrm{Cor}(\xi, \omega) = \mathbf{g}^\top \mathbf{\Sigma}_{XY}\mathbf{h} \tag{20}$$

among all linear functions of X and Y.

In other words, we solve the following optimization problem

$$\begin{aligned} &\underset{\mathbf{g},\mathbf{h}}{\text{maximize}} \ \mathbf{g}^\top \mathbf{\Sigma}_{XY}\mathbf{h} \\ &\text{subject to} \ \mathbf{g}^\top \mathbf{\Sigma}_{XX}\mathbf{g} = \mathbf{h}^\top \mathbf{\Sigma}_{YY}\mathbf{h} = 1. \end{aligned} \tag{21}$$

The Lagrangian function of this problem is

$$L_1(\mathbf{g}, \mathbf{h}, \lambda, \mu) := \mathbf{g}^\top \mathbf{\Sigma}_{XY}\mathbf{h} - \frac{\lambda}{2}(\mathbf{g}^\top \mathbf{\Sigma}_{XX}\mathbf{g} - 1) - \frac{\mu}{2}(\mathbf{h}^\top \mathbf{\Sigma}_{YY}\mathbf{h} - 1), \tag{22}$$

where $\lambda > 0$ and $\mu > 0$ are the Lagrangian multipliers. We differentiate L_1 with respect to $\mathbf{g}$ and $\mathbf{h}$ and set derivatives to zero, and obtain

$$\frac{\partial L_1}{\partial \mathbf{g}} = \mathbf{\Sigma}_{XY}\mathbf{h} - \lambda \mathbf{\Sigma}_{XX}\mathbf{g} = \mathbf{0}_p, \tag{23}$$

$$\frac{\partial L_1}{\partial \mathbf{h}} = \mathbf{\Sigma}_{YX}\mathbf{g} - \mu \mathbf{\Sigma}_{YY}\mathbf{h} = \mathbf{0}_s. \tag{24}$$

Two observations:

- If we multiply (23) on the left by $\mathbf{g}^\top$ and (24) on the left by $\mathbf{h}^\top$, we have

$$\mathbf{g}^\top \mathbf{\Sigma}_{XY}\mathbf{h} - \lambda\, \mathbf{g}^\top \mathbf{\Sigma}_{XX}\mathbf{g} = 0,$$
$$\mathbf{h}^\top \mathbf{\Sigma}_{YX}\mathbf{g} - \mu\, \mathbf{h}^\top \mathbf{\Sigma}_{YY}\mathbf{h} = 0;$$

  that is,

$$\mathbf{g}^\top \mathbf{\Sigma}_{XY}\mathbf{h} = \lambda = \mu. \tag{25}$$

- Since (23) and (24) can be written equivalently as

$$-\lambda \mathbf{\Sigma}_{XX}\mathbf{g} + \mathbf{\Sigma}_{XY}\mathbf{h} = \mathbf{0}_p, \qquad \text{and} \qquad \mathbf{\Sigma}_{YX}\mathbf{g} - \lambda \mathbf{\Sigma}_{YY}\mathbf{h} = \mathbf{0}_s.$$

  Pre-multiplying the first equation by $\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}$ and substitute into the second one, we have

$$(\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY} - \lambda^2 \mathbf{\Sigma}_{YY})\mathbf{h} = \mathbf{0}_s, \tag{26}$$

  or equivalently, we have

$$\left(\mathbf{\Sigma}_{YY}^{-\frac{1}{2}}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-\frac{1}{2}} - \lambda^2 \mathbf{I}_s\right)\mathbf{\Sigma}_{YY}^{\frac{1}{2}}\mathbf{h} = \mathbf{0}_s. \tag{27}$$

  It follows that, at the optimality, we must have

$$\mathbf{g}_1 := \frac{1}{\lambda}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-\frac{1}{2}}\mathbf{v}_1, \qquad \text{and} \qquad \mathbf{h}_1 := \mathbf{\Sigma}_{YY}^{-\frac{1}{2}}\mathbf{v}_1, \tag{28}$$

  where $\mathbf{v}_1$ is the eigenvector of the matrix

$$\mathbf{R} = \mathbf{\Sigma}_{YY}^{-\frac{1}{2}}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-\frac{1}{2}}$$

  associated with the largest eigenvalue.

Hence, the first pair of canonical variates is $(\xi_1, \omega_1) = (\mathbf{g}_1^\top X, \mathbf{h}_1^\top Y)$ and the maximal correlation is the square root of the largest eigenvalue of $\mathbf{R}$, i.e.,

$$\mathrm{Cor}(\xi_1, \omega_1) = \mathbf{g}_1^\top \mathbf{\Sigma}_{XY}\mathbf{h}_1 = \sqrt{\lambda_1(\mathbf{R})}. \tag{29}$$

Here, note the largest eigenvalue of $\mathbf{R}$, $\lambda_1(\mathbf{R})$, and the optimal Lagrangian multipliers, λ^* and μ^*, are related by $\lambda_1(\mathbf{R}) = \lambda^{*2} = \mu^{*2}$.

4. **Derivation of (ξ_2, ω_2):** Given (ξ_1, ω_1), let $\xi = \mathbf{g}^\top X$ and $\omega = \mathbf{h}^\top Y$ be a second pair of arbitrary linear projections. We require

   - ξ and ω have the largest correlation among all such linear combinations of X and Y;
   - ξ is uncorrelated with ξ_1 and ω is uncorrelated with ω_1:
   $$\mathbf{g}^\top \mathbf{\Sigma}_{XX} \mathbf{g}_1 = 0, \qquad \mathbf{h}^\top \mathbf{\Sigma}_{YY} \mathbf{h}_1 = 0; \tag{30}$$
   - ξ and ω have unit variance:
   $$\mathbf{g}^\top \mathbf{\Sigma}_{XX} \mathbf{g} = 1, \qquad \mathbf{h}^\top \mathbf{\Sigma}_{YY} \mathbf{h} = 1; \tag{31}$$
   - ξ is uncorrelated with ω_1 and ω is uncorrelated with ξ_1:
   $$\mathrm{Cor}(\xi, \omega_1) = \mathbf{g}^\top \mathbf{\Sigma}_{XY} \mathbf{h}_1 \overset{(*)}{=} \lambda_1 \mathbf{g}^\top \mathbf{\Sigma}_{XX} \mathbf{g}_1 = 0, \tag{32}$$
   $$\mathrm{Cor}(\omega, \xi_1) = \mathbf{h}^\top \mathbf{\Sigma}_{YX} \mathbf{g}_1 \overset{(**)}{=} \lambda_1 \mathbf{h}^\top \mathbf{\Sigma}_{YY} \mathbf{h}_1 = 0, \tag{33}$$
   where $(*)$ and $(**)$ follow from (23) and (24), respectively.

   Then, we solve the following optimization problem
   $$\begin{aligned} \underset{\mathbf{g},\mathbf{h}}{\text{maximize}} \ & \mathrm{Cor}(\mathbf{g}^\top X, \mathbf{h}^\top Y) = \mathbf{g}^\top \mathbf{\Sigma}_{XY} \mathbf{h} \\ \text{subject to} \ & \mathbf{g}^\top \mathbf{\Sigma}_{XX} \mathbf{g} = 1, \\ & \mathbf{h}^\top \mathbf{\Sigma}_{YY} \mathbf{h} = 1, \\ & \mathbf{g}^\top \mathbf{\Sigma}_{XX} \mathbf{g}_1 = 0, \\ & \mathbf{h}^\top \mathbf{\Sigma}_{YY} \mathbf{h}_1 = 0. \end{aligned} \tag{34}$$

   The Lagrangian function of this preceding optimization problem is
   $$\begin{aligned} L_2(\mathbf{g}, \mathbf{h}, \lambda, \mu, \eta, \nu) := & \mathbf{g}^\top \mathbf{\Sigma}_{XY} \mathbf{h} - \frac{\lambda}{2}(\mathbf{g}^\top \mathbf{\Sigma}_{XX} \mathbf{g} - 1) - \frac{\mu}{2}(\mathbf{h}^\top \mathbf{\Sigma}_{YY} \mathbf{h} - 1) \\ & - \frac{\eta}{2} \mathbf{g}^\top \mathbf{\Sigma}_{XX} \mathbf{g}_1 - \frac{\nu}{2} \mathbf{h}^\top \mathbf{\Sigma}_{YY} \mathbf{h}_1, \end{aligned} \tag{35}$$
   where $\lambda > 0$, $\mu > 0$, $\eta > 0$ and $\nu > 0$ are the Lagrangian multipliers.

   Differentiating L_2 with respect to $\mathbf{g}$ and $\mathbf{h}$ and setting the derivatives to 0 yields
   $$\frac{\partial L_2}{\partial \mathbf{g}} = \mathbf{\Sigma}_{XY} \mathbf{h} - \lambda \mathbf{\Sigma}_{XX} \mathbf{g} - \eta \mathbf{\Sigma}_{XX} \mathbf{g}_1 = \mathbf{0}_p, \tag{36}$$
   $$\frac{\partial L_2}{\partial \mathbf{h}} = \mathbf{\Sigma}_{YX} \mathbf{g} - \mu \mathbf{\Sigma}_{YY} \mathbf{h} - \mu \mathbf{\Sigma}_{YY} \mathbf{h}_1 = \mathbf{0}_s. \tag{37}$$

   By solving this linear system, we have that the second pair of the canonical variate is $(\xi_2, \omega_2) = (\mathbf{g}_2^\top X, \mathbf{h}_2^\top Y)$ is
   $$\mathbf{g}_2 = \mathbf{\Sigma}_{XX}^{-1} \mathbf{\Sigma}_{XY} \mathbf{\Sigma}_{YY}^{-\frac{1}{2}} \mathbf{v}_2, \qquad \mathbf{h}_2 = \mathbf{\Sigma}_{YY}^{-\frac{1}{2}} \mathbf{v}_2, \tag{38}$$

where $\mathbf{v}_2$ is the eigenvector of $\mathbf{R}$ associated with the second largest eigenvalue of $\mathbf{R}$, and their correlation is

$$\mathrm{Cor}(\xi_2, \omega_2) = \mathbf{g}_2^\top \mathbf{\Sigma}_{XY} \mathbf{h}_2 = \sqrt{\lambda_2(\mathbf{R})}. \tag{39}$$

5. **Derivation of (ξ_j, ω_j) for $j \geq 3$:** The remaining canonical variates (ξ_j, ω_j), for $j \geq 3$, can be obtained by choosing coefficients $\mathbf{g}_j$ and $\mathbf{h}_j$ such that (ξ_j, ω_j) has the largest correlation among all pairs of linear combinations of X and Y that are also uncorrelated with all previously derived pairs, $\{(\xi_i, \omega_i)\}_{i=1}^{j-1}$, until no further solution can be found.

IV. Sample Estimates

1. **Setup:** Let $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ be i.i.d observations from (X, Y). Let

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{pmatrix} \in \mathbb{R}^{p\times n} \qquad \text{and} \qquad \mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_n \end{pmatrix} \in \mathbb{R}^{s\times n}.$$

2. **Estimation of Coefficient Matrix:** We estimate $\mathbf{G}$ and $\mathbf{H}$ by

$$\widehat{\mathbf{G}}^{(t)} := \widehat{\mathbf{V}}^{(t)} \widehat{\mathbf{\Sigma}}_{YY}^{-\frac{1}{2}} \widehat{\mathbf{\Sigma}}_{YX} \widehat{\mathbf{\Sigma}}_{XX}^{-1}, \tag{40}$$

$$\widehat{\mathbf{H}}^{(t)} := \widehat{\mathbf{V}}^{(t)} \widehat{\mathbf{\Sigma}}_{YY}^{-\frac{1}{2}}, \tag{41}$$

respectively, and

$$\widehat{\mathbf{V}}^{(t)} := \begin{pmatrix} \hat{\mathbf{v}}_1^\top \\ \vdots \\ \hat{\mathbf{v}}_t^\top \end{pmatrix} \tag{42}$$

where $\hat{\mathbf{v}}_j$ is the eigenvector associated with the j-th largest eigenvalue of

$$\widehat{\mathbf{R}} = \widehat{\mathbf{\Sigma}}_{YY}^{-\frac{1}{2}} \widehat{\mathbf{\Sigma}}_{YX} \widehat{\mathbf{\Sigma}}_{XX}^{-1} \widehat{\mathbf{\Sigma}}_{XY} \widehat{\mathbf{\Sigma}}_{YY}^{-\frac{1}{2}}, \tag{43}$$

for $j = 1, \cdots, t$.

3. **Estimation of Canonical Variates:** The j-th row of $\hat{\boldsymbol{\xi}} := \widehat{\mathbf{G}}^{(t)}\mathbf{X}$ and $\hat{\boldsymbol{\omega}} := \widehat{\mathbf{H}}^{(t)}\mathbf{Y}$ together form the j-th pair of *sample canonical variates* $(\hat{\boldsymbol{\xi}}_j, \hat{\boldsymbol{\omega}}_j)$, where

$$\hat{\boldsymbol{\xi}}_j = \hat{\mathbf{g}}_j^\top \mathbf{X}, \qquad \text{and} \qquad \hat{\boldsymbol{\omega}}_j = \hat{\mathbf{h}}_j^\top \mathbf{Y} \tag{44}$$

with *canonical variate scores* of

$$\hat{\xi}_{i,j} = \hat{\mathbf{g}}_j^\top \mathbf{x}_i, \qquad \text{and} \qquad \hat{\omega}_{i,j} = \hat{\mathbf{h}}_j^\top \mathbf{y}_i, \tag{45}$$

for $i = 1, \cdots, n$, where

$$\hat{\mathbf{g}}_j^\top = \hat{\mathbf{v}}_j^\top \widehat{\mathbf{\Sigma}}_{YY}^{-\frac{1}{2}} \widehat{\mathbf{\Sigma}}_{YX} \widehat{\mathbf{\Sigma}}_{XX}^{-1} \tag{46}$$

is the j-th row of $\widehat{\mathbf{G}}^{(t)}$, and

$$\hat{\mathbf{h}}_j^\top = \hat{\mathbf{v}}_j^\top \widehat{\mathbf{\Sigma}}_{YY}^{-\frac{1}{2}} \tag{47}$$

is the j-row of $\widehat{\mathbf{H}}^{(t)}$.

4. **Estimation of Sample Canonical Correlation Coefficient:** The *sample canonical correlation coefficient* for the j-th pair of sample canonical variates $(\hat{\boldsymbol{\xi}}_j, \hat{\boldsymbol{\omega}}_j)$ is

$$\hat{\rho}_j = \frac{\hat{\mathbf{g}}_j^\top \widehat{\boldsymbol{\Sigma}}_{XY} \hat{\mathbf{h}}_j}{\sqrt{\hat{\mathbf{g}}_j^\top \widehat{\boldsymbol{\Sigma}}_{XX} \hat{\mathbf{g}}_j} \cdot \sqrt{\hat{\mathbf{h}}_j^\top \widehat{\boldsymbol{\Sigma}}_{YY} \hat{\mathbf{h}}_j}}, \tag{48}$$

for all $j = 1, \cdots, t$.

V. Kernel CCA

1. **Overview:** The kernel CCA uses

   (a) a *nonlinear* transformation, $\Phi_1 : \mathbb{R}^p \to \mathcal{H}_1$, of one set of input data, $\mathbf{x}_i \in \mathbb{R}^p$, for all $i = 1, 2, \cdots, n$, and

   (b) another nonlinear transformation, $\Phi_2 : \mathbb{R}^s \to \mathcal{H}_2$, of a second set of input data, $\mathbf{y}_i \in \mathbb{R}^s$, for all $j = 1, 2, \cdots, n$.

   Here, for each $j = 1, 2$, $\mathcal{H}_j$ is a reproducing kernel Hilbert space (RKHS).

   Then, we carry out CCA between two transformed sets of input data $\{\Phi_1(\mathbf{x}_i)\}_{i=1}^n$ and $\{\Phi_2(\mathbf{y}_i)\}_{i=1}^n$, where we assume that both sets of transformed data have been centered.

2. **Goal:** We wish to find $f_1 \in \mathcal{H}_1$ and $f_2 \in \mathcal{H}_2$ such that the features $f_1(X) = \langle \Phi_1(X), f_1 \rangle_{\mathcal{H}}$ and $f_2(Y) = \langle \Phi_2(Y), f_2 \rangle_{\mathcal{H}}$ have the maximal correlation.

3. **Naive Kernel CCA:** We consider to maximize the correlation of transformed X and Y, i.e., we maximize

$$\hat{\rho}_{\text{kernel}}(f_1, f_2) := \frac{\widehat{\mathrm{Cov}}(f_1(X), f_2(Y))}{\sqrt{\widehat{\mathrm{Var}}[f_1(X)]} \cdot \sqrt{\widehat{\mathrm{Var}}[f_2(Y)]}}, \tag{49}$$

   subject to

$$\begin{aligned} f_1 &\in \mathrm{Span}\big(\Phi_1(\mathbf{x}_1), \Phi_1(\mathbf{x}_2), \cdots, \Phi_1(\mathbf{x}_n)\big) \\ f_2 &\in \mathrm{Span}\big(\Phi_2(\mathbf{y}_1), \Phi_2(\mathbf{y}_2), \cdots, \Phi_2(\mathbf{y}_n)\big), \end{aligned}$$

   where

$$\begin{aligned} \widehat{\mathrm{Cov}}(f_1(X), f_2(Y)) &= \frac{1}{n} \sum_{i=1}^n f_1(\mathbf{x}_i) f_2(\mathbf{y}_i) \\ &= \frac{1}{n} \sum_{i=1}^n \langle f_1, \Phi_1(\mathbf{x}_i) \rangle_{\mathcal{H}_1} \langle f_2, \Phi_2(\mathbf{y}_i) \rangle_{\mathcal{H}_2} \\ &= \frac{1}{n} \boldsymbol{\alpha}_1^\top \mathbf{K}_1 \mathbf{K}_2 \boldsymbol{\alpha}_2, &(50) \\ \widehat{\mathrm{Var}}[f_1(X)] &= \frac{1}{n} \boldsymbol{\alpha}_1^\top \mathbf{K}_1^2 \boldsymbol{\alpha}_1, &(51) \\ \widehat{\mathrm{Var}}[f_2(Y)] &= \frac{1}{n} \boldsymbol{\alpha}_2^\top \mathbf{K}_2^2 \boldsymbol{\alpha}_2. &(52) \end{aligned}$$

In the equations above, $\boldsymbol{\alpha}_1 \in \mathbb{R}^n$, $\boldsymbol{\alpha}_2 \in \mathbb{R}^n$, the matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ are the $n \times n$ Gram matrices associated with $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ and $\{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n\}$, respectively. In other words, the (i, j)-th entry of $\mathbf{K}_1$ is

$$\langle \Phi_1(\mathbf{x}_i), \Phi_1(\mathbf{x}_j) \rangle_{\mathcal{H}_1},$$

and the (i, j)-th entry of $\mathbf{K}_2$ is

$$\langle \Phi_2(\mathbf{y}_i), \Phi_2(\mathbf{y}_j) \rangle_{\mathcal{H}_2}.$$

It follows that (49) becomes

$$\hat{\rho}_{\text{kernel}}(f_1, f_2) = \frac{\boldsymbol{\alpha}_1^\top \mathbf{K}_1 \mathbf{K}_2 \boldsymbol{\alpha}_2}{\sqrt{(\boldsymbol{\alpha}_1^\top \mathbf{K}_1^2 \boldsymbol{\alpha}_1) \cdot (\boldsymbol{\alpha}_2^\top \mathbf{K}_2^2 \boldsymbol{\alpha}_2)}}, \tag{53}$$

and we maximize over $\boldsymbol{\alpha}_1 \in \mathbb{R}^n$ and $\boldsymbol{\alpha}_2 \in \mathbb{R}^n$.

4. **Solution to (53):** Differentiating (53) with respect to $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ and setting the results to zero yield the generalized eigen-equations

$$\mathbf{K}\boldsymbol{\alpha} = \lambda \mathbf{D}\boldsymbol{\alpha},$$

where

$$\mathbf{K} = \begin{pmatrix} \mathbf{0} & \mathbf{K}_1\mathbf{K}_2 \\ \mathbf{K}_2\mathbf{K}_1 & \mathbf{0} \end{pmatrix}, \qquad \mathbf{D} = \begin{pmatrix} \mathbf{K}_1^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_2^2 \end{pmatrix}, \qquad \boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix}.$$

It turns out that all pairs of "kernel canonical variates" in feature space are *perfectly correlated*. The reason is essentially due to overfitting.

5. **Regularized Kernel CCA:** We apply the regularization to solve the kernel CCA problem. More precisely, we penalize the $\mathcal{H}_1$-norm of f_1 and the $\mathcal{H}_2$-norm of f_2 each by the same small constant value $\kappa > 0$, and replace $\mathbf{K}_1^2$ by $(\mathbf{K}_1 + \kappa \mathbf{I}_n)^2$ and $\mathbf{K}_2^2$ by $(\mathbf{K}_2 + \kappa \mathbf{I}_n)^2$ in $\mathbf{D}$.

   (a) *Justification:* suppose $\theta > 0$ is the regularization parameter, then

$$\begin{aligned} \widehat{\mathrm{Var}}[f_1(X)] + \theta \|f_1\|_{\mathcal{H}_1}^2 &= \frac{1}{n} \boldsymbol{\alpha}_1^\top \mathbf{K}_1^2 \boldsymbol{\alpha}_1 + \theta \boldsymbol{\alpha}_1^\top \mathbf{K}_1 \boldsymbol{\alpha}_1 \approx \frac{1}{n} \boldsymbol{\alpha}_1^\top (\mathbf{K}_1 + \kappa \mathbf{I}_n)^2 \boldsymbol{\alpha}_1, \\ \widehat{\mathrm{Var}}[f_2(Y)] + \theta \|f_2\|_{\mathcal{H}_2}^2 &= \frac{1}{n} \boldsymbol{\alpha}_2^\top \mathbf{K}_2^2 \boldsymbol{\alpha}_2 + \theta \boldsymbol{\alpha}_2^\top \mathbf{K}_2 \boldsymbol{\alpha}_2 \approx \frac{1}{n} \boldsymbol{\alpha}_2^\top (\mathbf{K}_2 + \kappa \mathbf{I}_n)^2 \boldsymbol{\alpha}_2, \end{aligned}$$

   where we can see $\kappa = \frac{1}{2} n \theta$.

   (b) *Optimization Problem:* The regularized optimization problem (53) becomes

$$\tilde{\rho}_{\text{kernel}}(f_1, f_2; \kappa) = \frac{\boldsymbol{\alpha}_1^\top \mathbf{K}_1 \mathbf{K}_2 \boldsymbol{\alpha}_2}{\sqrt{(\boldsymbol{\alpha}_1^\top (\mathbf{K}_1 + \kappa \mathbf{I}_n)^2 \boldsymbol{\alpha}_1) \cdot (\boldsymbol{\alpha}_2^\top (\mathbf{K}_2 + \kappa \mathbf{I}_n)^2 \boldsymbol{\alpha}_2)}}, \tag{54}$$

(c) *Effects of κ:* The value of κ determines the weight to be placed upon the penalty terms compared with the variance terms. In particular,

   i. as κ gets close to zero, the variance term dominates, whereas

   ii. as κ gets larger, the variance term becomes more affected by the amount of roughness allowed by the penalty term.

(d) *Solution:* Differentiating (54) with respect to $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ and setting the results to zero yield

$$\mathbf{K}\boldsymbol{\alpha} = \lambda \mathbf{D}^{(\kappa)}\boldsymbol{\alpha}, \tag{55}$$

where

$$\mathbf{D}^{(\kappa)} := \begin{pmatrix} (\mathbf{K}_1 + \kappa \mathbf{I}_n)^2 & \mathbf{0} \\ \mathbf{0} & (\mathbf{K}_2 + \kappa \mathbf{I}_n)^2 \end{pmatrix}.$$

Again, this is a generalized eigen-equation, which has $2n$ pairs of eigenvalues

$$\lambda_1, -\lambda_1, \cdots, \lambda_n, -\lambda_n.$$

(e) *Equivalent Generalized Eigen-equation:* The generalized eigen-equation (55) can be written as

$$\mathbf{K}^{(\kappa)}\boldsymbol{\alpha} = (1+\lambda)\mathbf{D}^{(\kappa)}\boldsymbol{\alpha}, \tag{56}$$

where

$$\mathbf{K}^{(\kappa)} = \begin{pmatrix} (\mathbf{K}_1 + \kappa \mathbf{I}_n)^2 & \mathbf{K}_1\mathbf{K}_2 \\ \mathbf{K}_2\mathbf{K}_1 & (\mathbf{K}_2 + \kappa \mathbf{I}_n)^2 \end{pmatrix}.$$

Then, (56) has the following pairs of eigenvalues

$$1+\lambda_1, 1-\lambda_1, \cdots, 1+\lambda_n, 1-\lambda_n.$$

Note that (56) can be equivalently written as

$$\widetilde{\mathbf{K}}^{(\kappa)}\tilde{\boldsymbol{\alpha}} = \tilde{\lambda}\tilde{\boldsymbol{\alpha}},$$

where

$$\begin{aligned}
\widetilde{\mathbf{K}}^{(\kappa)} &= [\mathbf{D}^{(\kappa)}]^{-\frac{1}{2}}\mathbf{K}^{(\kappa)}[\mathbf{D}^{(\kappa)}]^{-\frac{1}{2}} = \begin{pmatrix} \mathbf{I}_n & \widetilde{\mathbf{K}}_1^{(\kappa)}\widetilde{\mathbf{K}}_2^{(\kappa)} \\ \widetilde{\mathbf{K}}_2^{(\kappa)}\widetilde{\mathbf{K}}_1^{(\kappa)} & \mathbf{I}_n \end{pmatrix}, \\
\widetilde{\mathbf{K}}_1^{(\kappa)} &:= (\mathbf{K}_1 + \kappa \mathbf{I}_n)^{-1}\mathbf{K}_1, \\
\widetilde{\mathbf{K}}_2^{(\kappa)} &:= (\mathbf{K}_2 + \kappa \mathbf{I}_n)^{-1}\mathbf{K}_2, \\
\tilde{\boldsymbol{\alpha}} &:= [\mathbf{D}^{(\kappa)}]^{-\frac{1}{2}}\boldsymbol{\alpha}, \\
\tilde{\lambda} &:= 1 + \lambda.
\end{aligned}$$

References

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC. ISBN: 1498712169.

Izenman, Alan J (Mar. 2009). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. en. Springer Science & Business Media.