Canonical Correlation Analysis

Canonical Correlation analysis (CCA) aims to quantify the associations between two sets of variables, each set consists of two or more variables. The association is measured by correlation, thus is on linear relationships of the two groups of variables.

Examples

Do last year's labor data relate or explain this year's labor situation?

Are the multi-subject math science performances of the students related to their reading performances?

Are there associations between government policies and economic variables?

For decathlon athletes, is their track event performance predictive of their performance in field events?

CCA data and objectives

The data for canonical correlation analysis are of n observations, each observation is a vector with $p+q$ components. The vector components form two groups of interest.

$$\left[\begin{array}{cc} \underbrace{\begin{matrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jp} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{matrix}}_{p \text{ variables}} & \underbrace{\begin{matrix} y_{11} & y_{12} & \cdots & y_{1q} \\ y_{21} & y_{22} & \cdots & y_{2q} \\ \vdots & \vdots & \vdots & \vdots \\ y_{j1} & y_{j2} & \cdots & y_{jq} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nq} \end{matrix}}_{q \text{ variables}} \end{array}\right] \begin{matrix} \leftarrow \text{1st observation} \\ \leftarrow \text{2nd observation} \\ \vdots \\ \leftarrow j\text{th observation} \\ \vdots \\ \leftarrow n\text{th observation} \end{matrix}$$

In the data matrix, each row can be viewed as the measurement of a sample point in $\mathbb{R}^{p+q}$ draw from a random vector, it's transpose can be expressed as

$$[X_1 \ \cdots \ X_p \ \ Y_1 \ \cdots \ Y_q]$$

We can think of the data as two data matrices of dimensions $n \times p$ and $n \times q$, observed together.
In the following, we assume the n measurements are independent observations from $[X_1 \ \cdots \ X_p \ \ Y_1 \ \cdots \ Y_q]$.

The objective of canonical correlation analysis is to quantify the association between two variable groups $\{X_1, \cdots, X_p\}$ and $\{Y_1, \cdots, Y_q\}$, in terms of linear correlation.

1 Population canonical variates and canonical correlations

Notations

In theoretical development of the method, we will consider the corresponding population random vectors $\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix}$, where

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = [X_1 \ X_2 \ \cdots \ X_p]' \in \mathbb{R}^p, \qquad \boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{bmatrix} = [Y_1 \ Y_2 \ \cdots \ Y_q]' \in \mathbb{R}^q.$$

Remarks on notations

Another common notation for $\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix}$ is $\begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}$, as in Johnson and Wichern, using double subscripts to index the component variables,

$\boldsymbol{X}_1 = (X_{11}, \cdots, X_{1p})'$, p-variate random vector,
$\boldsymbol{X}_2 = (X_{21}, \cdots, \cdots, X_{2q})'$, q-variate random vector.

We use $\boldsymbol{X}$ and $\boldsymbol{Y}$ notation to simplify the subscripts. Note that the $\boldsymbol{X}$ (component) variables and the $\boldsymbol{Y}$ variables are treated equally in this analysis, meaning not necessarily one set is the input and the other set is the output.

The formulation of the canonical correlation analysis (CCA)

We investgate the canonical correlation structure on the two population vectors $\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$ and $\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_q \end{bmatrix}$.

Goal of CCA:

The first and foremost objective for CCA is to find a pair of variables, $\boldsymbol{a}'\boldsymbol{X}$ and $\boldsymbol{b}'\boldsymbol{Y}$, each a linear combination of the original variables in their respective group,

$$\begin{cases} \boldsymbol{a}'\boldsymbol{X} = & a_1 X_1 + \cdots + a_p X_p \\ \boldsymbol{b}'\boldsymbol{Y} = & b_1 Y_1 + \cdots + b_q Y_q \end{cases}$$

so that the correlation of the new pair of variables represents the maximum correlation between the two groups of original variables. (Notes: Here the notations a_i, b_j simply denote the scalar components of generic vectors $\boldsymbol{a}$ and $\boldsymbol{b}$.)

In other words, the goal is to find

$$\text{argmax}_{\boldsymbol{a},\boldsymbol{b}} Corr(\boldsymbol{a}'\boldsymbol{X}, \boldsymbol{b}'\boldsymbol{Y}) \qquad \text{among all } \boldsymbol{a} \in \mathbb{R}^p, \ \boldsymbol{b} \in \mathbb{R}^q.$$

After finding such a pair, if needed, the derivation process can continue to find the next maximum-correlated pair of linear combinations among all pairs uncorrelated with the first selected pair, and so on. There are at most $\min\{p, q\}$ such pairs.

Conventions

These successively chosen pairs of linear combinations are called **canonical variables** or **canonical variates**, the correlations between canonical variate pairs are termed **canonical correlations**.

Without loss of generality, $p \le q$ is often used as a convenient assumption on the dimensions of the two sets of variables.

Treating the two sets of random variables jointly as

$$\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} = [\boldsymbol{X}' \ \boldsymbol{Y}']' = [X_1 \ \cdots \ X_p \ Y_1 \ \cdots \ Y_q]'$$

Their expectation vector is

$$\mathbb{E} \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}_{(p+q)\times 1}$$

and their covariance matrix is

$$Cov \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}_{(p+q)\times(p+q)}$$

Note that matrix Σ_{12} is of dimensions $p \times q$, matrix Σ_{21} is $q \times p$, they are transposes of each other, $\Sigma_{21}' = \Sigma_{12}$.

Population canonical variates

Consider a pair of linear combinations of the variables:

$$U = \boldsymbol{a}'\boldsymbol{X}, \quad V = \boldsymbol{b}'\boldsymbol{Y}, \qquad \boldsymbol{a} \in \mathbb{R}^p, \quad \boldsymbol{b} \in \mathbb{R}^q.$$

Then U and V are univariate random variables, with

$$\begin{aligned} cov(U, V) &= \boldsymbol{a}'Cov(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{b} = \boldsymbol{a}'\Sigma_{12}\boldsymbol{b} \\ var(U) &= \boldsymbol{a}'Cov(\boldsymbol{X})\boldsymbol{a} = \boldsymbol{a}'\Sigma_{11}\boldsymbol{a} \\ var(V) &= \boldsymbol{b}'Cov(\boldsymbol{Y})\boldsymbol{b} = \boldsymbol{b}'\Sigma_{22}\boldsymbol{b} \end{aligned}$$

Canonical correlation analysis seeks $\boldsymbol{a}$ and $\boldsymbol{b}$ to maximize the correlation

$$corr(U, V) = \frac{cov(U, V)}{\sqrt{var(U)}\sqrt{var(V)}} = \frac{\boldsymbol{a}'\Sigma_{12}\boldsymbol{b}}{\sqrt{\boldsymbol{a}'\Sigma_{11}\boldsymbol{a}}\ \sqrt{\boldsymbol{b}'\Sigma_{22}\boldsymbol{b}}} \tag{1}$$

Note that replacing $\boldsymbol{a}$ and $\boldsymbol{b}$ by their constant multiples will not change the correlation (1). Hence for convenience of this derivation, we normalize $\boldsymbol{a}$ and $\boldsymbol{b}$ by the constraints

$$var(U) = \boldsymbol{a}'\Sigma_{11}\boldsymbol{a} = 1, \qquad var(V) = \boldsymbol{b}'\Sigma_{22}\boldsymbol{b} = 1.$$

Maximizing correlation in (1) can be described as

$$\text{Maximize} \quad \rho = \rho_{\boldsymbol{a},\boldsymbol{b}} = \boldsymbol{a}'\Sigma_{12}\boldsymbol{b} \qquad \text{under constraints} \quad \boldsymbol{a}'\Sigma_{11}\boldsymbol{a} = 1, \quad \boldsymbol{b}'\Sigma_{22}\boldsymbol{b} = 1.$$

Just like the way we derived principal components, Lagrange Multiplier method can be used to find the solutions.

In the following we will go through the procedure to find canonical variates $\boldsymbol{a}$ and $\boldsymbol{b}$ which achieve the maximization of canonical correlation in (1). The steps utilize the Lagrange Multiplier method, properties of symmetric and positive semidefinite matrices, and the singular value decomposition of matrices.

Proof. **Derivation of canonical variates** (The leading pair)

The Lagrange multiplier setup

The optimization problem of CCA can be formulated by the method of Lagrange Multipliers by setting the Lagrangian function

$$L = \boldsymbol{a}'\Sigma_{12}\boldsymbol{b} - \frac{1}{2}\lambda(\boldsymbol{a}'\Sigma_{11}\boldsymbol{a} - 1) - \frac{1}{2}\gamma(\boldsymbol{b}'\Sigma_{22}\boldsymbol{b} - 1).$$

Optimal $\boldsymbol{a}$ and $\boldsymbol{b}$ that maximize $\rho = \boldsymbol{a}'\Sigma_{12}\boldsymbol{b}$ must occur at a critical point at which $\dfrac{\partial L}{\partial \boldsymbol{a}} = \boldsymbol{0}_p$, $\dfrac{\partial L}{\partial \boldsymbol{b}} = \boldsymbol{0}_q$.

Differentiating L with respect to $\boldsymbol{a}, \boldsymbol{b}$ and equating the derivatives to zero vectors,

$$\begin{cases} \dfrac{\partial L}{\partial \boldsymbol{a}} = \Sigma_{12}\boldsymbol{b} - \lambda\Sigma_{11}\boldsymbol{a} = \boldsymbol{0}_p \\ \dfrac{\partial L}{\partial \boldsymbol{b}} = \Sigma_{21}\boldsymbol{a} - \gamma\Sigma_{22}\boldsymbol{b} = \boldsymbol{0}_q \end{cases} \tag{2}$$

In the second equation, we applied

$$\Sigma_{21} = \Sigma_{12}' \qquad and \qquad \rho = \boldsymbol{a}'\Sigma_{12}\boldsymbol{b} = (\boldsymbol{a}'\Sigma_{12}\boldsymbol{b})' = \boldsymbol{b}'\Sigma_{21}\boldsymbol{a}$$

The values of Lagrange multipliers

To find out the values of λ and γ satisfying (2), left multiply $\boldsymbol{a}'$ and $\boldsymbol{b}'$ to the two equations in (2) respectively. Use $\rho = \boldsymbol{a}'\Sigma_{12}\boldsymbol{b} = \boldsymbol{b}'\Sigma_{21}\boldsymbol{a}$ again, we obtain

$$\begin{aligned} \boldsymbol{a}'\Sigma_{12}\boldsymbol{b} - \lambda\boldsymbol{a}'\Sigma_{11}\boldsymbol{a} = \rho - \lambda = 0 \quad &\Rightarrow \quad \lambda = \rho \qquad (since\ \ \boldsymbol{a}'\Sigma_{11}\boldsymbol{a} = 1) \\ \boldsymbol{b}'\Sigma_{21}\boldsymbol{a} - \gamma\boldsymbol{b}'\Sigma_{22}\boldsymbol{b} = \rho - \gamma = 0 \quad &\Rightarrow \quad \gamma = \rho \qquad (since\ \ \boldsymbol{b}'\Sigma_{22}\boldsymbol{b} = 1) \end{aligned}$$

So (2) is simplified to

$$\begin{cases} \Sigma_{12}\boldsymbol{b} - \rho\Sigma_{11}\boldsymbol{a} = \boldsymbol{0}_p \\ \Sigma_{21}\boldsymbol{a} - \rho\Sigma_{22}\boldsymbol{b} = \boldsymbol{0}_q \end{cases} \tag{3}$$

Next we need to solve for $\boldsymbol{a}$ and $\boldsymbol{b}$. Note that the two equations above are in different dimensions.

Vectors a and b as eigenvectors of matrices

First, we solve for $\boldsymbol{a}$ in (3). To cancel out the $\boldsymbol{b}$ terms (then simplify the $\boldsymbol{a}$ term in the first equation), left multiplying the first equation in (3) by $\rho\Sigma_{11}^{-1}$, then left multiply the second equation by $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}$,

$$\begin{cases} \rho\Sigma_{11}^{-1}\Sigma_{12}\boldsymbol{b} - \rho^2\boldsymbol{a} = \boldsymbol{0}_p \\ \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\boldsymbol{a} - \rho\Sigma_{11}^{-1}\Sigma_{12}\boldsymbol{b} = \boldsymbol{0}_p \end{cases}$$

Now the two equations are of the same dimensions, we can add the two equations to cancel out the $\boldsymbol{b}$ terms.

$$\Rightarrow \qquad \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\boldsymbol{a} - \rho^2\boldsymbol{a} = \boldsymbol{0}_p.$$

Write the equation as

$$(\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \rho^2\boldsymbol{I}_p)\boldsymbol{a} = \boldsymbol{0}_p$$

If there are ρ^2 and $\boldsymbol{a} \neq \boldsymbol{0}_p$ satisfying the equation, then ρ^2 is an eigenvalue of $p \times p$ matrix $\boldsymbol{A} = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, and $\boldsymbol{a}$ is an eigenvector of $\boldsymbol{A}$ with eigenvalue ρ^2.

Analogous derivation yields

$$(\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \rho^2\boldsymbol{I}_q)\boldsymbol{b} = \boldsymbol{0}_q$$

Then $\boldsymbol{b}$ is an eigenvector of the $q \times q$ matrix $\boldsymbol{B} = \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ with eigenvalue ρ^2, if there are ρ^2 and $\boldsymbol{b} \neq \boldsymbol{0}_p$ satisfying the equation.

To obtain exact expressions for our objective vectors $\boldsymbol{a}, \boldsymbol{b}$, we need to obtain properties of eigenvalues and eigenvectors of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

Eigenvalues of matrices A and B

We have obtained

$$\begin{cases} (\boldsymbol{A} - \rho^2\boldsymbol{I}_p)\boldsymbol{a} = \boldsymbol{0}_p \\ (\boldsymbol{B} - \rho^2\boldsymbol{I}_q)\boldsymbol{b} = \boldsymbol{0}_q \end{cases}$$

What are possible eigenvalues of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, and what are their relationships?

Not that matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are not necessarily symmetric. There existence of real eigenvalues and their ranges are not readily established.

In the above, the eigenvalue is of the form ρ^2, hinting the non-negativeness of the eigenvalues, which is a property of positive semidefinite matrices. To relate $\boldsymbol{A}$ and $\boldsymbol{B}$ with symmetric, positive semidefinite matrices, first recall that any symmetric positive (semi)definite matrix can be written as the square of a symmetric positive (semi)definite matrix, which can be denoted symbolically as the "square root" matrix of the original one. Assume that the components of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are not linear combinations of each other, then Σ_{11}, Σ_{22} are positive definite, consequently so are their inverse matrices. We may write

$$\Sigma_{11} = \Sigma_{11}^{1/2}\Sigma_{11}^{1/2}, \qquad \Sigma_{22} = \Sigma_{22}^{1/2}\Sigma_{22}^{1/2}, \qquad \Sigma_{11}^{-1} = \Sigma_{11}^{-1/2}\Sigma_{11}^{-1/2}, \qquad \Sigma_{22}^{-1} = \Sigma_{22}^{-1/2}\Sigma_{22}^{-1/2}$$

where $\Sigma_{ii}^{1/2}$ is the symbol for a symmetric positive definite that its square is Σ_{ii}, etc. Noe that such matrices decompositions are not necessarily unique. Now rewrite

$$\boldsymbol{A} = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Sigma_{11}^{-1/2}(\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2})(\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2})\Sigma_{11}^{1/2} = \Sigma_{11}^{-1/2}(\boldsymbol{C}\boldsymbol{C}')\Sigma_{11}^{1/2} \tag{4}$$

with

$$\boldsymbol{C} = \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}$$

a p-by-q matrix. Similarly

$$\boldsymbol{B} = \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = \Sigma_{22}^{-1/2}(\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2})(\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2})\Sigma_{22}^{1/2} = \Sigma_{22}^{-1/2}(\boldsymbol{C}'\boldsymbol{C})\Sigma_{22}^{1/2} \tag{5}$$

Recall from matrix algebra that if a matrix $M_1 = Q^{-1}M_2Q$ for some invertible matrix Q, then matrices M_1 and M_2 are "similar", often denoted as $M_1 \sim M_2$. Similar matrices share the same eigenvalues.

Expression (4) shows that $\boldsymbol{A} \sim \boldsymbol{C}\boldsymbol{C}'$, thus $\boldsymbol{A}$ has the same eigenvalues as $\boldsymbol{C}\boldsymbol{C}'$.

Similarly, (5) implies that $\boldsymbol{B} \sim \boldsymbol{C}'\boldsymbol{C}$, thus $\boldsymbol{B}$ has the same eigenvalues as $\boldsymbol{C}'\boldsymbol{C}$.

To achieve our goal, it is efficient and convenient to focus on the properties of $\boldsymbol{C}\boldsymbol{C}'$ and $\boldsymbol{C}'\boldsymbol{C}$ and their eigenvalues and eigenvectors first.

The eigenvectors of $\boldsymbol{C}\boldsymbol{C}'$ and $\boldsymbol{C}'\boldsymbol{C}$ will lead to eigenvectors of $\boldsymbol{A}$ and $\boldsymbol{B}$, which are the objectives of this derivation of the CCA.

Common non-zero eigenvalues of $\boldsymbol{C}\boldsymbol{C}'$ and $\boldsymbol{C}'\boldsymbol{C}$: ρ_i^{*2}

Matrices $\boldsymbol{C}\boldsymbol{C}'$ and $\boldsymbol{C}'\boldsymbol{C}$ have several nice algebraic properties:

- They are symmetric matrices (while $\boldsymbol{A}, \boldsymbol{B}$ are not necessarily symmetric). Therefore all of their eigenvalues are real, and the eigenvectors can be chosen to be mutually orthogonal.
- They are positive semi-definite matrices. Therefore the eigenvalues are non-negative.
- In addition, $\boldsymbol{C}'\boldsymbol{C}$ and $\boldsymbol{C}\boldsymbol{C}'$ share the same nonzero eigenvalues.

Label and sort the p eigenvalues of the $p \times p$ matrix $\boldsymbol{C}\boldsymbol{C}'$ in descending order as

$$\rho_1^{*2} \geq \rho_2^{*2} \geq \cdots \geq \rho_r^{*2} \geq 0, \qquad r = \min\{p, q\} \qquad (\rho_i^* \geq 0)$$

The top r eigenvalues $\rho_i^{*2}, i = 1, \cdots, r$, equal to the r largest eigenvalues of the $q \times q$ matrix $\boldsymbol{C}'\boldsymbol{C}$. Other eigenvalues of $\boldsymbol{C}'\boldsymbol{C}$ or $\boldsymbol{C}\boldsymbol{C}'$ must be zero.

By the similarity relationship in (4) and (5), eigenvalues $\rho_1^{*2}, \cdots, \rho_r^{*2}$, are ordered r largest eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{B}$ as well. All other eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{B}$ must be zero.

Orthonormal eigenvectors of $\boldsymbol{C}\boldsymbol{C}'$ and $\boldsymbol{C}'\boldsymbol{C}$: $\boldsymbol{e}_i$ and $\boldsymbol{f}_i$

Below we obtain related eigenvectors of $\boldsymbol{C}\boldsymbol{C}'$ and $\boldsymbol{C}'\boldsymbol{C}$, which will lead to eigenvectors of $\boldsymbol{A}$ and $\boldsymbol{B}$.

Let $\boldsymbol{e}_i$ be the p-variate unit eigenvector of $\boldsymbol{C}\boldsymbol{C}'$ corresponding to eigenvalue ρ_i^{*2}, $i = 1, \cdots, r$.

$$\boldsymbol{C}\boldsymbol{C}'\boldsymbol{e}_i = \rho_i^{*2}\boldsymbol{e}_i.$$

Then

$$\boldsymbol{C}'\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{e}_i) = \boldsymbol{C}'(\boldsymbol{C}\boldsymbol{C}'\boldsymbol{e}_i) = \boldsymbol{C}'(\rho_i^{*2}\boldsymbol{e}_i) = \rho_i^{*2}(\boldsymbol{C}'\boldsymbol{e}_i)$$

which means the q-variate vector $\boldsymbol{C}'\boldsymbol{e}_i$ and its multiples are eigenvectors of $\boldsymbol{C}'\boldsymbol{C}$ with eigenvalue ρ_i^{*2}. This also shows that $\boldsymbol{C}\boldsymbol{C}'$ and $\boldsymbol{C}'\boldsymbol{C}$ share the same non-zero eigenvalues.

In particular, we choose $\boldsymbol{f}_i$ to be the q-variate unit eigenvector of $\boldsymbol{C}\boldsymbol{C}'$ corresponding to eigenvalue ρ_i^{*2}. Since $\boldsymbol{f}_i$ and $\boldsymbol{C}'\boldsymbol{e}_i$ are eigenvectors of the same eigenvalue, we may let $\boldsymbol{f}_i = c_i\boldsymbol{C}'\boldsymbol{e}_i$ for some constant c_i, and $\boldsymbol{f}_i'\boldsymbol{f}_i = 1$.

By the symmetric, positive semi-definiteness of $\boldsymbol{C}\boldsymbol{C}'$, $\{\boldsymbol{e}_i\}_{i=1}^p$ can be obtained and can be chosen to be orthonormal:

$$\boldsymbol{e}_i'\boldsymbol{e}_j = \delta_{ij} = \begin{cases} 1 & \textit{when } i = j, \\ 0 & \textit{otherwise.} \end{cases} \qquad i, j = 1, \cdots, p.$$

To determine the constatn c_i in $\boldsymbol{f}_i$, use the desired orthogonal condition

$$\boldsymbol{f}_i'\boldsymbol{f}_j = \delta_{ij} = \begin{cases} 1 & \textit{when } i = j, \\ 0 & \textit{otherwise.} \end{cases}$$

That is,

$$\boldsymbol{f}_i'\boldsymbol{f}_j = (c_i\boldsymbol{C}'\boldsymbol{e}_i)'(c_j\boldsymbol{C}'\boldsymbol{e}_j) = c_ic_j\boldsymbol{e}_i'\boldsymbol{C}\boldsymbol{C}'\boldsymbol{e}_j = c_ic_j\boldsymbol{e}_i'\rho_j^{*2}\boldsymbol{e}_j = c_ic_j\rho_j^{*2}\delta_{ij} = \delta_{ij} \quad \Longleftrightarrow \quad c_i = 1/\rho_i^*$$

which leads to

$$\boldsymbol{f}_i = (1/\rho_i^*)\boldsymbol{C}'\boldsymbol{e}_i$$

Orthnormal eigenvectors can be established for zero eigenvalues.

Now both $\{\boldsymbol{e}_i\}_{i=1}^p$ $\{\boldsymbol{f}_i\}_{i=1}^q$ are chosen to be orthonormal:

$$\boldsymbol{e}_i'\boldsymbol{e}_j = \begin{cases} 1 & \textit{when } i = j, \\ 0 & \textit{when } i \neq j. \end{cases} \qquad i, j = 1, \cdots, p; \qquad \boldsymbol{f}_i'\boldsymbol{f}_j = \begin{cases} 1 & \textit{when } i = j, \\ 0 & \textit{when } i \neq j. \end{cases} \qquad i, j = 1, \cdots, q.$$

Eigenvectors of $\boldsymbol{A}$ and $\boldsymbol{B}$: $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$

Let

$$\boldsymbol{a}_i = \Sigma_{11}^{-1/2}\boldsymbol{e}_i, \qquad \boldsymbol{b}_i = \Sigma_{22}^{-1/2}\boldsymbol{f}_i.$$

Then

$$\boldsymbol{A}\boldsymbol{a}_i = \boldsymbol{A}(\Sigma_{11}^{-1/2}\boldsymbol{e}_i) = \Sigma_{11}^{-1/2}\boldsymbol{C}\boldsymbol{C}'\Sigma_{11}^{1/2}(\Sigma_{11}^{-1/2}\boldsymbol{e}_i) = \Sigma_{11}^{-1/2}\boldsymbol{C}\boldsymbol{C}'\boldsymbol{e}_i = \Sigma_{11}^{-1/2}(\rho_i^{*2}\boldsymbol{e}_i) = \rho_i^{*2}(\Sigma_{11}^{-1/2}\boldsymbol{e}_i) = \rho_i^{*2}\boldsymbol{a}_i$$

Therefore $\boldsymbol{a}_i = \Sigma_{11}^{-1/2}\boldsymbol{e}_i$ is an eigenvector of $\boldsymbol{A}$ with eigenvalue ρ_i^{*2} for $i = 1, \cdots, p$.

Analogously, we can show that $\boldsymbol{b}_i = \Sigma_{22}^{-1/2}\boldsymbol{f}_i$ is an eigenvector of $\boldsymbol{B}$ with eigenvalue ρ_i^{*2} for $i = 1, \cdots, q$.

Optimal a and b maximizing canonical correlation

The canonical variate pairs $\boldsymbol{a}_i, \boldsymbol{b}_i$ are eigenvectors of matrices $\boldsymbol{A}, \boldsymbol{B}$ with the same eigenvalue ρ_i^{*2}. The pair of linear combinations $\boldsymbol{a}_i'\boldsymbol{X}$ and $\boldsymbol{b}_i'\boldsymbol{Y}$ has correlation

$$\begin{aligned}\rho = corr(\boldsymbol{a}_i'\boldsymbol{X}, \boldsymbol{b}_i'\boldsymbol{Y}) = cov(\boldsymbol{a}_i'\boldsymbol{X}, \boldsymbol{b}_i'\boldsymbol{Y}) &= \boldsymbol{a}_i'\Sigma_{12}\boldsymbol{b}_i = (\Sigma_{11}^{-1/2}\boldsymbol{e}_i)'\Sigma_{12}(\Sigma_{22}^{-1/2}\boldsymbol{f}_i)\\ &= \boldsymbol{e}_i'(\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2})(\boldsymbol{f}_i)\\ &= \boldsymbol{e}_i'\boldsymbol{C}\boldsymbol{f}_i\\ &= \boldsymbol{e}_i'\boldsymbol{C}(1/\rho_i^*)\boldsymbol{C}'\boldsymbol{e}_i\\ &= (1/\rho_i^*)\boldsymbol{e}_i'\boldsymbol{C}\boldsymbol{C}'\boldsymbol{e}_i\\ &= (1/\rho_i^*)\boldsymbol{e}_i'\rho_i^{*2}e_i = \rho_i^*\boldsymbol{e}_i'e_i\\ &= \rho_i^*\end{aligned}$$

There are at most $r = \min\{p, q\}$ such pairs, up to a factor of (-1) of $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$. Conventionally $\rho \geq 0$.

Choose $\boldsymbol{a}_1, \boldsymbol{b}_1$, the first pair of canonical variables. Now we can achieve $\rho = \rho_1^*$, the common largest eigenvalue of $\boldsymbol{A}$ and $\boldsymbol{B}$, which maximizes (1), by a pair of linear combinations $U = \boldsymbol{a}_1'\boldsymbol{X}$ and $V = \boldsymbol{b}_1'\boldsymbol{Y}$.

This concludes the derivation of the leading pair of canonical variates.

□

The above procedure produces

$$U_1 = \boldsymbol{a}_1'\boldsymbol{X}, \qquad V_1 = \boldsymbol{b}_1'\boldsymbol{Y}$$

capturing maximum correlation

$$corr(U_1, V_1) = \rho_1^*$$

among all possible linear combinations $\boldsymbol{a}'\boldsymbol{X}$ and $\boldsymbol{b}'\boldsymbol{Y}$ (under constraints $\boldsymbol{a}'\Sigma_{11}\boldsymbol{a} = 1,\ \boldsymbol{b}'\Sigma_{22}\boldsymbol{b} = 1$).

If one wants to explore the rest correlations between $\boldsymbol{X}$ and $\boldsymbol{Y}$, the procedure can go on to find the second pair $U_2 = \boldsymbol{a}_2'\boldsymbol{X}$ and $V_2 = \boldsymbol{b}_2'\boldsymbol{Y}$ with maximum $corr(U_2, V_2)$ under constraints. We may define the Lagrangian function

$$L_2 = \boldsymbol{a}_2{}'\Sigma_{12}\boldsymbol{b}_2 - \frac{1}{2}\lambda(\boldsymbol{a}_2{}'\Sigma_{11}\boldsymbol{a}_2 - 1) - \frac{1}{2}\gamma(\boldsymbol{b}_2{}'\Sigma_{22}\boldsymbol{b}_2 - 1) - \delta\, \boldsymbol{a}_2{}'\Sigma_{11}\boldsymbol{a}_1 - \nu\, \boldsymbol{b}_2{}'\Sigma_{22}\boldsymbol{b}_1 - \beta\, \boldsymbol{a}_1{}'\Sigma_{12}\boldsymbol{b}_2 - \alpha\, \boldsymbol{b}_1{}'\Sigma_{21}\boldsymbol{a}_2$$

Setting

$$\begin{cases}\dfrac{\partial L_2}{\partial \boldsymbol{a}_2} = \boldsymbol{0}_p\\[2ex] \dfrac{\partial L_2}{\partial \boldsymbol{b}_2} = \boldsymbol{0}_q\end{cases}$$

With analogous derivations (exercise) we can obtain the second pair of canonical covariates (U_2, V_2) under the normalization $var(U_2) = var(V_2) = 1$ and the orthogonality conditions, achieving the strongest correlation $Cov(U_2, V_2) = \rho_2^*$, and are uncorrelated to the first canonical pair (U_1, V_1) by satisfying the constraints

$$cov(U_1, U_2) = 0, \quad cov(V_1, V_2) = 0, \quad cov(U_1, V_2) = 0, \quad cov(U_2, V_1) = 0$$

The process may continue, if desired, to obtain all pairs of canonical variates.

- The first pair is the most desirable one. In all normalized (variance = 1) linear combination variables

  $$U_1 = \boldsymbol{a}_1'\boldsymbol{X}, \quad V_1 = \boldsymbol{b}_1'\boldsymbol{Y}, \quad var(U_1) = 1, \quad var(V_1) = 1.$$

  the (U_1, V_1) pair that maximizes $Corr(U_1, V_1)$ in (1) among all $\boldsymbol{a}_1, \boldsymbol{b}_1$ are attained at

  $$\boldsymbol{a}_1 = \Sigma_{11}^{-1/2}\boldsymbol{e}_1, \quad \boldsymbol{b}_1 = \Sigma_{22}^{-1/2}\boldsymbol{f}_1,$$

  so

  $$U_1 = \boldsymbol{e}_1'\Sigma_{11}^{-1/2}\boldsymbol{X}, \quad V_1 = \boldsymbol{f}_1'\Sigma_{22}^{-1/2}\boldsymbol{Y}, \qquad Cov(U_1, V_1) = \rho_1^*.$$

  It verifies that

  $$var(U_1) = \boldsymbol{e}_1'\Sigma_{11}^{-1/2}var(\mathbf{X})\Sigma_{11}^{-1/2}\boldsymbol{e}_1 = \boldsymbol{e}_1'\boldsymbol{e}_1 = 1, \qquad var(V_1) = \cdots = \boldsymbol{f}_1'\boldsymbol{f}_1 = 1.$$

- The second pair of normalized canonical variables are

  $$U_2 = \boldsymbol{a}_2'\boldsymbol{X}, \quad V_2 = \boldsymbol{b}_2'\boldsymbol{Y}, \quad var(U_2) = var(V_2) = 1, \qquad Cov(U_2, V_2) = \rho_2^*.$$

  Additional constraints are

  $$\boldsymbol{a}_2'\Sigma_{11}\boldsymbol{a}_1 = 0, \quad \boldsymbol{b}_2'\Sigma_{22}\boldsymbol{b}_1 = 0, \quad \boldsymbol{a}_1'\Sigma_{12}\boldsymbol{b}_2 = 0, \quad \boldsymbol{a}_2'\Sigma_{12}\boldsymbol{b}_1 = 0,$$

  which imply that

  $$cov(U_1, U_2) = 0, \qquad cov(V_1, V_2) = 0, \qquad cov(U_1, V_2) = 0, \qquad cov(U_2, V_1) = 0.$$

- The kth pair of canonical variables are

  $$U_k = \boldsymbol{e}_k'\Sigma_{11}^{-1/2}\boldsymbol{X}, \quad V_k = \boldsymbol{f}_k'\Sigma_{22}^{-1/2}\boldsymbol{Y} \qquad with \quad var(U_k) = var(V_k) = 1, \quad Cov(U_k, V_k) = \rho_k^*.$$

  All satisfies the condition that, for $\ell < k$,

  $$\boldsymbol{a}_k'\Sigma_{11}\boldsymbol{a}_\ell = 0, \quad \boldsymbol{b}_k'\Sigma_{22}\boldsymbol{b}_\ell = 0, \quad \boldsymbol{a}_\ell\Sigma_{12}\boldsymbol{b}_k = 0, \quad \boldsymbol{a}_k'\Sigma_{12}\boldsymbol{b}_\ell = 0.$$

- The process stops at $k = \min\{p, q\}$.

Summary of properties of canonical variables $U_i = \boldsymbol{a}_i'\boldsymbol{X}$, $V = \boldsymbol{b}_i'\boldsymbol{Y}$

- $var(U_k) = var(V_k) = 1$
- $cov(U_k, U_\ell) = 0, cov(V_k, V_\ell) = 0$ for $k \neq \ell$.
- $cov(U_k, V_\ell) = 0$ for $k \neq \ell$
- $cov(U_i, V_i) = \rho_i^*$,
  where ρ_i^{*2} is the common ith largest eigenvalue of the following matrices:
  - $p \times p$ matrix $\boldsymbol{C}\boldsymbol{C}' = \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$
  - $q \times q$ matrix $\boldsymbol{C}'\boldsymbol{C} = \Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$
  - $p \times p$ matrix $\boldsymbol{A} = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Sigma_{11}^{-1/2}\boldsymbol{C}\boldsymbol{C}'\Sigma_{11}^{1/2}$
  - $q \times q$ matrix $\boldsymbol{B} = \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = \Sigma_{22}^{-1/2}\boldsymbol{C}'\boldsymbol{C}\Sigma_{22}^{1/2}$

and

$$\boldsymbol{A}\boldsymbol{a}_i = \rho_i^{*2}\boldsymbol{a}, \qquad \boldsymbol{B}\boldsymbol{b}_i = \rho_i^{*2}\boldsymbol{b}$$

The proofs are a repeated use of extended Cauchy-Schwarz inequality and its corollary the maximization lemma.

Covariance matrix of canonical variables

The covariance matrix of the canonical variables has neat form.

$$Cov(U_1, \cdots, U_r, V_1, \cdots, V_r) = \begin{bmatrix} I_r & R_r \\ R_r & I_r \end{bmatrix}, \qquad R_r = diag[\rho_1^*, \cdots, \rho_r^*]$$

Sort in canonical variate pairs,

$$Cov(U_1, V_1, \cdots, U_r, V_r) = diag[D_1, \cdots, D_r], \qquad D_k = \begin{bmatrix} 1 & \rho_k^* \\ \rho_k^* & 1 \end{bmatrix}$$

To scale or not to scale

If there are large variations in the magnitudes of component variable in multivariate data, we know that principal component analysis on the original variables and that on scaled data with $var(X_i) = 1$ may yield very different results.

Does scaling matter in canonical correlation analysis?

Recall the relationship between covariance matrix Σ and correlation matrix $\boldsymbol{R}$,

$$\Sigma = D\boldsymbol{R}D, \qquad \boldsymbol{R} = D^{-1}\Sigma D^{-1}$$

where D is the diagonal matrix with ith diagonal element $\sqrt{\sigma_{ii}}$. We may write

$$D = [diag(\Sigma)]^{1/2}$$

Let $(\boldsymbol{X}, \boldsymbol{Y})$ be the pair of the original random vectors, $(\boldsymbol{X}^*, \boldsymbol{Y}^*)$ be the scaled pair with each component variable with variance $= 1$. The relationship can be expressed as

$$(\boldsymbol{X}^*, \boldsymbol{Y}^*) = (D_1^{-1}\boldsymbol{X}, D_2^{-1}\boldsymbol{Y})$$

with

$$D_1 = [diag(\Sigma_{11})]^{1/2}, \qquad D_2 = [diag(\Sigma_{22})]^{1/2}.$$

If $(U_k, V_k) = (\boldsymbol{a}_k'\boldsymbol{X}, \boldsymbol{b}_k'\boldsymbol{Y})$ is the kth pair of canonical correlation variates of the original variables, with

$$cov(U_k, V_k) = cov(\boldsymbol{a}_k'\boldsymbol{X}, \boldsymbol{b}_k'\boldsymbol{Y}) = \rho_k^*$$

and if $(U_k^*, V_k^*) = (\boldsymbol{a}_k^{*\prime}\boldsymbol{X}^*, \boldsymbol{b}_k^{*\prime}\boldsymbol{Y}^*)$ is the kth pair of canonical correlation variates of the scaled variance-one variables, then from the derivation process, we have

$$\boldsymbol{a}_k^* = D_1\boldsymbol{a}_k, \qquad \boldsymbol{b}_k^* = D_2\boldsymbol{b}_k$$

and

$$\begin{aligned} cov(U_k^*, V_k^*) &= cov(\boldsymbol{a}_k^{*\prime}\boldsymbol{X}^*, \boldsymbol{b}_k^{*\prime}\boldsymbol{Y}^*) \\ &= cov\left((D_1\boldsymbol{a}_k)'D_1^{-1}\boldsymbol{X}, (D_2\boldsymbol{b}_k)'D_2^{-1}\boldsymbol{Y}\right) \\ &= cov(\boldsymbol{a}_k{}'D_1'D_1^{-1}\boldsymbol{X}, \boldsymbol{b}_k{}'D_2'D_2^{-1}\boldsymbol{Y}) \\ &= cov(\boldsymbol{a}_k'\boldsymbol{X}, \boldsymbol{b}_k'\boldsymbol{Y}) \\ &= cov(U_k, V_k) = \rho_k^* \end{aligned}$$

Therefore, the canonical coefficients for the scaled variance-one variables are simply related to the canonical coefficients attached to the original variables, by component-wise scaling.

The canonical correlations, the ρ_k^*'s, are unchanged by the scaling. If the strength of the correlation between the two groups of variables is the main focus (as it often is), the scaling does not matter.

Of course there are the usual non-uniqueness of $\boldsymbol{a}_k, \boldsymbol{b}_k$ when $\rho_k^* = \rho_{k+1}^*$ happens.

Remarks on canonical variables and CCA

- Matrix $\boldsymbol{C} = \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}$, sometimes called the canonical correlation matrix, can be viewed as a generalization of the correlation coefficient of two univariate random variables to two multivariate random vectors.
- The pair of matrices

  $$\boldsymbol{C}\boldsymbol{C}' = \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2} \quad \text{and} \quad \boldsymbol{C}'\boldsymbol{C} = \Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$$

  called the matrices of multivariate coefficients of determination, are natural generalizations of population coefficient of determination (square of the correlation between Y and its best linear predictor, may not be covered in this course) in linear regression.
- The derivation of canonical variates shows that the $\rho*_i$'s are correlations, thus the matrices $\boldsymbol{A}$, $\boldsymbol{B}$, positive semi-definite matrices $\boldsymbol{C}\boldsymbol{C}'$ and $\boldsymbol{C}'\boldsymbol{C}$ all have eigenvalues $\in [0, 1]$.
- In the computation of CCA, a condensed Singular value decomposition (SVD) is conducted on $\boldsymbol{C}$.

  As the above population CCA derivation has demonstrated, in the case $r = p \leq q$,

  $$\boldsymbol{C} = UDV = [\boldsymbol{e}_1 \ \cdots \ \boldsymbol{e}_p]\, diag(\rho_1^*, \cdots, \rho_p^*) \begin{bmatrix} \boldsymbol{f}_1' \\ \vdots \\ \boldsymbol{f}_p' \end{bmatrix} \quad \in \mathbb{R}^{p \times q}$$

  Based on this SVD of matrix $\boldsymbol{C}$, all of the desired vectors $\boldsymbol{a}_i = \Sigma_{11}^{-1/2}\boldsymbol{e}_i$, $\boldsymbol{b}_i = \Sigma_{22}^{-1/2}\boldsymbol{f}_i$, and their canonical variables $U_i = \boldsymbol{a}_i'\boldsymbol{X}$, $V_i = \boldsymbol{b}_i'\boldsymbol{Y}$ can be obtained accordingly.

Comparisons of CCA and PCA

Canonical correlation analysis (CCA) can be viewed as a multivariate statistical technique similar in spirit to principal component analysis (PCA).

In Principal Component Analysis, from $\boldsymbol{X} = [X_1 \ X_2 \ \cdots \ X_p]'$, we derived principal components $\boldsymbol{a}_i\boldsymbol{X}$, $i = 1, \cdots, p$. The uncorrelated PC variables capture the variation in the data and reduce the dimension of the data effectively. However certain structures of the data can be obscured or lost in the process.

CCA evaluates a particular structure in $[\boldsymbol{X}' \ \boldsymbol{Y}']' = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix}$. This linear structure (canonical correlation) can be described as a type of linear correlations between two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$.

At each step, while PCA works with a single random vector and maximizes the variance of projections of the data, CCA works with a pair of random vectors (or generalized to a set of m random vectors) and maximizes correlation between sets of projections.

2 Sample CCA

Similar to the practical usage of principal component analysis, in applications, the observed data are often treated as the population to carry out canonical correlation analysis directly, without probability distribution assumptions on the data generating mechanism.

- In the place of population random vectors, we work with sample data

$$[\ \boldsymbol{X}_{n\times p} \ \boldsymbol{Y}_{n\times q}\ ]$$

- In the place to Σ_{ij} are sample covariance matrices. The joint sample covariance matrix of all $p+q$ variables can be in terms of sample covariance matrices of x and y variables, as well as their covariance.

$$\boldsymbol{S} = \begin{bmatrix} S_x & S_{xy} \\ S_{yx} & S_y \end{bmatrix}$$

- The first pair of sample canonical-variate vectors are given be

$$(\hat{\boldsymbol{a}}_1, \hat{\boldsymbol{b}}_1) = \text{argmax}_{\boldsymbol{a}\in\mathbb{R}^p, \boldsymbol{b}\in\mathbb{R}^q}\{\boldsymbol{a}' S_{xy} \boldsymbol{b} : \ \boldsymbol{a}' S_x \boldsymbol{a} = 1, \ \boldsymbol{b}' S_y \boldsymbol{b} = 1\}$$

- The second pair of sample canonical-variate vectors are given be

$$(\hat{\boldsymbol{a}}_2, \hat{\boldsymbol{b}}_2) = \text{argmax}_{\boldsymbol{a},\boldsymbol{b}}\{\boldsymbol{a}' S_{xy} \boldsymbol{b} : \ \boldsymbol{a}' S_x \boldsymbol{a} = \boldsymbol{b}' S_y \boldsymbol{b} = 1, \boldsymbol{a}' S_x \boldsymbol{a}_1 = \boldsymbol{a}' S_{xy} \boldsymbol{b}_1 = \boldsymbol{b}' S_{yx} \boldsymbol{a}_1 = \boldsymbol{b}' S_y \boldsymbol{b}_1 = 0\}$$

- For $k \le \min\{p,q\}$,the kth pair of sample canonical-variate vectors are given be

$$(\hat{\boldsymbol{a}}_k, \hat{\boldsymbol{b}}_k) = \text{argmax}_{\boldsymbol{a},\boldsymbol{b}}\{\boldsymbol{a}' S_{xy} \boldsymbol{b} : \ \boldsymbol{a}' S_x \boldsymbol{a} = \boldsymbol{b}' S_y \boldsymbol{b} = 1, \boldsymbol{a}' S_x \boldsymbol{a}_i = \boldsymbol{a}' S_{xy} \boldsymbol{b}_i = \boldsymbol{b}' S_{yx} \boldsymbol{a}_i = \boldsymbol{b}' S_y \boldsymbol{b}_i = 0, i = 1, \cdots, k-1.\}$$

- The achieved kth sample canonical correlation is

$$\hat{\rho}_k^* = \hat{\boldsymbol{a}}_k' S_{xy} \hat{\boldsymbol{b}}_k, \qquad k = 1, \cdots, \min\{p,q\}$$

  where

$$\hat{\rho_1^*}^2 \ge \hat{\rho_2^*}^2 \ge \cdots \ge \hat{\rho_r^*}^2 \ge 0, \qquad r = \min\{p,q\} \qquad (\hat{\rho}_i^* \ge 0)$$

  are top r common non-zero eigenvalues of four matrices, $\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{C}}\hat{\boldsymbol{C}}', \hat{\boldsymbol{C}}'\hat{\boldsymbol{C}}$., with

$$\hat{\boldsymbol{A}} = S_x^{-1} S_{xy} S_y^{-1} S_{yx}, \quad \hat{\boldsymbol{B}} = S_y^{-1} S_{21y} S_x^{-1} S_{xy}, \qquad \hat{\boldsymbol{C}} = S_x^{-1/2} S_{xy} S_y^{-1/2}$$

  As the conventional notation assumes, a square root matrix is a matrix that it's square matrix is the original matrix.

- For computational efficiency, square rood matrices and inverse matrices are avoided. In practice, a condensed SVD is obtained for $\hat{\boldsymbol{C}}$. The rest steps follow, analogous to the steps described for the computation of the population CCA.

3 Large sample inferences

Are the two sets of variables, or two populations, actually uncorrelated? If so, CCA is unnecessary.

We consider statistical test to check the existence of the correlation. As before, we impose normality assumptions to obtain probabilistic inference. Suppose

$$\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} \sim N_{p+q}(\boldsymbol{\mu}, \Sigma), \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}_{(p+q)\times(p+q)}$$

The interest is to test if the cross-covariance matrix is a zero-matrix,

$$\begin{cases} H_0 : & \Sigma_{12} = \boldsymbol{O}_{p\times q} \\ H_a : & \Sigma_{12} \ne \boldsymbol{O}_{p\times q} \end{cases}$$

Likelihood Ratio test

For multivariate normal with n observations, the maximum likelihood has a simple form (as we have derived),

$$\max_{\boldsymbol{\mu},\Sigma} L(\boldsymbol{\mu}, \Sigma) = L(\hat{\boldsymbol{\mu}}, \hat{\Sigma}) = \frac{1}{(2\pi)^{np/2}|\hat{\Sigma}|^{n/2}} e^{-np/2}$$

where $|A|$ denotes the determinant of matrix A,

$$\hat{\Sigma} = \boldsymbol{S}_n = \frac{n-1}{n}\boldsymbol{S},$$

and $\boldsymbol{S}$ is the sample variance-covariance matrix. Write

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{bmatrix}$$

The maximum likelihood under $H_0 : \Sigma_{12} = \boldsymbol{O}_{p\times q}$ has the expression

$$\frac{1}{(2\pi)^{np/2}\left(\frac{n-1}{n}|\boldsymbol{S}_{11}| \times |\boldsymbol{S}_{22}|\right)^{n/2}} e^{-np/2}$$

The maximum likelihood ratio test statistic with respect the hypothesis testing is

$$\Lambda = \frac{\max_{under H_o} L(\boldsymbol{\mu}, \Sigma)}{\max_{\boldsymbol{\mu},\Sigma} L(\boldsymbol{\mu}, \Sigma)} = \left(\frac{|\boldsymbol{S}_{11}| \times |\boldsymbol{S}_{22}|}{|\boldsymbol{S}|}\right)^{-n/2}$$

Note that often $|\boldsymbol{S}|$ is used instead of $|\hat{\Sigma}| = |\boldsymbol{S}_n| = |\frac{n-1}{n}\boldsymbol{S}|$ because the cancelation of common factors by the ratio.

The common test statistic for the likelihood ratio test is

$$-2\ln\Lambda = n \ln \frac{|\boldsymbol{S}_{11}| \times |\boldsymbol{S}_{22}|}{|\boldsymbol{S}|}$$

The idea is to reject the null $H_0 : \Sigma_{12} = \boldsymbol{O}_{p\times q}$ when the test statistic is too large.

In the following, we derive a computationally convenient form for the test statistics.

Consider the case $p \leq q$. Using the Schur complement for block matrix, the determinant of the sample covariance matrix $\boldsymbol{S}$ can be expressed as

$$|\boldsymbol{S}| = |\boldsymbol{S}_{22}| \times |\boldsymbol{S}_{11} - \boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21}| = |\boldsymbol{S}_{22}| \times |\boldsymbol{S}_{11}| \times |I_p - \boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21}|$$

assuming the non-degenerate case so that all inverse matrices exist. Note that

$$|\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21}| = |\boldsymbol{S}_{11}^{-1/2}\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1/2}| = |\hat{\boldsymbol{C}}\hat{\boldsymbol{C}}'| = \prod_{i=1}^{p}(\hat{\rho}_i^*)^2$$

where we used the fact that if λ_i are the eigenvalues of a matrix M, then the determinant $|M| = \prod_i \lambda_i$. In addition, the matrix $I - M$ has corresponding eigenvalues $1 - \lambda_i$. The likelihood ratio test statistic $-2\ln\Lambda$ can be written in terms of the eigenvalues $(\hat{\rho}_i^*)^2$'s,

$$-2\ln\Lambda = n\ln\frac{|\boldsymbol{S}_{11}| \times |\boldsymbol{S}_{22}|}{|\boldsymbol{S}|} = n\ln\frac{1}{|I_p - \boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21}|} = -n\sum_{i=1}^{r}\ln\left[1 - (\hat{\rho}_i^*)^2\right]$$

where $r = \min\{p, q\}$.

Asymptotic tests (Results only)

Under the null $H_0 : \Sigma_{12} = \boldsymbol{O}_{p\times q}$, the test statistic $-2\ln\Lambda$ is asymptotically of χ^2 distribution, for large n. The degrees of freedom of the χ^2 is the difference of the number of parameters under H_a and that under H_0.

The number of parameters under H_0 is

$$\frac{1}{2}p(p+1) + \frac{1}{2}q(q+1)$$

The number of parameters under H_a is

$$\frac{1}{2}(p+q)(p+q+1) = \frac{1}{2}p(p+1) + \frac{1}{2}q(q+1) + pq$$

Therefore, approximately, the test statistic

$$-2\ln\Lambda \sim \chi^2_{pq} \qquad under\ H_0 \quad for\ large\ n.$$

Again, Bartlett suggested the modification of using the test statistic

$$-\left(n - 1 - \frac{1}{2}(p+q+1)\right)\sum_{i=1}^{r}\ln\left[1 - (\hat{\rho}_i^*)^2\right] \qquad \sim \chi^2_{pq} \qquad under\ H_0 \quad for\ large\ n.$$

In the case of

$$\rho_1^* \geq \cdots \geq \rho_k^* \ > \ \rho_{k+1}^* = \cdots = \rho_r^* = 0,$$

it is possible to use the above idea to test sequentially. The hypotheses are

$$\begin{cases} H_o^k : & \rho_k^* \neq 0,\ \rho_{k+1}^* = \cdots = \rho_r^* = 0 \\ H_a : & \rho_{k+1}^* \neq 0 \end{cases}$$

The Bartlett statistic is used:

$$-\left(n - 1 - \frac{1}{2}(p+q+1)\right)\sum_{i=k+1}^{r}\ln\left[1 - (\hat{\rho}_i^*)^2\right] \qquad \sim \chi^2_{(p-k)(q-k)} \qquad under\ H_0^k \quad for\ large\ n.$$

The tests can also be applied in testing for co-integration in multivariate time series analysis (Johansen, 1988).

4 Extension of CCA*

Non-linear situations are abound in practice, especially in large data era. Applications call for more generail ways to define canonical correlations.

The classical CCA is to maximize

$$\max_{a,b} corr(a'X, b'Y) = \max_{a,b}\frac{cov(a'X, b'Y)}{\sqrt{Var(a'X)}\sqrt{Var(b'Y)}}$$

under constraints $Var(a'X) = 1, Var(b'Y) = 1$.

In many machine learning application, various extensions of canonical correlations are used.

For example, Functional CCA is defined as

$$\max_{f\in\mathcal{H}_x, g\in\mathcal{H}_y} corr(f(X), g(Y)) = \max_{f\in\mathcal{H}_x, g\in\mathcal{H}_y}\frac{cov(f(X), g(Y))}{\sqrt{Var(f(X))}\sqrt{Var(g(Y)}}$$

where $\mathcal{H}_x, \mathcal{H}_y$ are reproducible kernel Hilbert spaces (RKHS).

Other generalizations include Kernel CCA (KCCA), Deep CCA, ...

Some research work in deep learning has used CCA and its extensions. For example,

Deep Variational Canonical Correlation Analysis (2017) (https://arxiv.org/abs/1610.03454)

Large-Scale Approximate Kernel Canonical Correlation Analysis (2016) (https://arxiv.org/abs/1511.04773)

both by K. Livescu et al.

Note Relevant chapter in the book by Johnson and Wichern: Chapter 10.