Canonical Correlation Analysis

(1) Background:

Recall : regressional analysis concerns with relationship between a single response and a set of predictors.

What if we have more than one responses?

i) Multivariate Regression : $Y_k = \sum_{1}^{p} \beta_{ki} X_i + \varepsilon_k$. $1 \le k \le 2$

ii) Apply PCA on Y_k. iii) CCA.

e.g. Variables related to arithmetic power and variables related to reading power.

$\Rightarrow$ The advantage of CCA is that:

It seeks to identify and quantify linear associations between two sets of variables by find L.F. of variables maximally correlated. exact the valuable information about correlation.

(2) Population CCA:

① Data: 1^{st} group $X^{(1)} = (X_{11} \cdots X_{1p})^T$ suppose $p \le q$.

2^{nd} group $X^{(2)} = (X_{21} \cdots X_{2q})^T$

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}. \quad E(X) = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}. \quad Var(X) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Rmk: pq elements Σ_{12} measures the association between two sets. Commonly, we require two variables are homogeneous

Set $u = a^T X^{(1)}$, $V = b^T X^{(2)}$. Then we obtain:

$Var(u) = a^T \Sigma_{11} a$; $Var(V) = b^T \Sigma_{22} b$.

$cov(u, V) = a^T \Sigma_{12} b$. We seek a, b s.t.

maximizes: $corr^2(u, V) = \dfrac{(a^T \Sigma_{12} b)^2}{(a^T \Sigma_{11} a)(b^T \Sigma_{22} b)}$ $(*)$

i) The 1^{st} pair (u_1, V_1) maximizes $(*)$, which's constrainted by: $cov(u_1) = cov(V_1) = 1$.

ii) The 2^{nd} pair (u_2, V_2) need to exact the max information of correlation which are uncorrelated with (u_1, V_1), i.e. maximize $(*)$, and s.t. $cov(u_2) = cov(V_2) = 1$, $cov(u_2, u_1) = cov(u_2, V_1) = cov(V_2, u_1) = cov(V_2, V_1) = 0$.

iii) The k^{th} pair (u_k, V_k) maximizes $(*)$, s.t. have unit Var. uncorrelated with the first $k-1$ pairs, maximizes $(*)$.

Rmk: Compare to PCA:

CCA	PCA
two sets	one set
2 proj.	1 proj.
max $corr^2$.	max Var

Thm. Denote $\mathrm{Corr}^2(U_k, V_k) = \ell_k^{*2}$. $\ell_1^{*2} \geq \ell_2^{*2} \cdots \geq \ell_p^{*2}$.

Then $a_k^T = e_k^T \Sigma_{11}^{-\frac{1}{2}}$. $b_k^T = f_k^T \Sigma_{22}^{-\frac{1}{2}}$ where (ℓ_k^{*2}, e_k)

is eigen-pair of $\Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-\frac{1}{2}}$. (ℓ_k^{*2}, f_k)

is eigen-pair of $\Sigma_{22}^{-\frac{1}{2}} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}}$.

Pf: 1°) By Schwartz Inequi. fix b:

$$\mathrm{corr}^2(a^T X_{(1)}, b^T X_{(2)}) \leq \frac{a^T \Sigma_{11} a \, (b^T \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} b)}{(a^T \Sigma_{11} a)(b^T \Sigma_{22} b)}$$

$$= \frac{b^T \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} b}{b^T \Sigma_{22} b} \leq \ell_1^{*2} \text{ which is maximum}$$

eigenvalue of $\Sigma_{22}^{-\frac{1}{2}} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}}$.

Besides $\tilde{a}_1 = c \Sigma_{11}^{-1} \Sigma_{12} \tilde{b}_2$. 1^{st} "=" holds.

Normalize: $a_1 = \Sigma_{11}^{-\frac{1}{2}} e_1 = \frac{1}{\sqrt{\lambda_1}} \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}} f_1$

$b_1 = \Sigma_{22}^{-\frac{1}{2}} f_1$. $\lambda_1 = \ell_1^{*2}$.

2°) By induction: suppose $1 \leq k \leq n-1$ holds.

Set $u_k = \Sigma_{22}^{\frac{1}{2}} b_k$. $1 \leq k \leq p$. $u = \Sigma_{22}^{\frac{1}{2}} b$.

From above: $\ell^2(U_n, V_n) \leq \tilde{u}^T T^T T \tilde{u}$. $\tilde{u}^T \tilde{u} = 1$.

and $T = \Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}}$. $\tilde{u} \in \mathrm{span}^{\perp}(u_1, \cdots u_{n-1})$.

$\Rightarrow \tilde{u} = \sum_n^p c_k u_k = \sum_n^p c_k f_k$. $\sum c_k^2 = 1$.

$\therefore \ell^2(U_n, V_n) \leq \sum_n^p c_k^2 \lambda_k \leq \lambda_n = \ell_n^{*2}$.

where "=" holds when $\tilde{u} = f_n = u_n$.

Rmk: 1) Nonzero eigenvalues of $T^T T$ is identical with $T T^T$. Besides, $e_k = \frac{1}{\sqrt{\lambda_k}} \Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}} f_k$.

$\lambda_k = \ell_k^{*2}$.

ii) Note the matrix is symmetric. So $e_i \perp e_j$, $f_i \perp f_j$, $i \neq j$. But $b_i \not\perp b_j$, $a_i \not\perp a_j$ may hold!

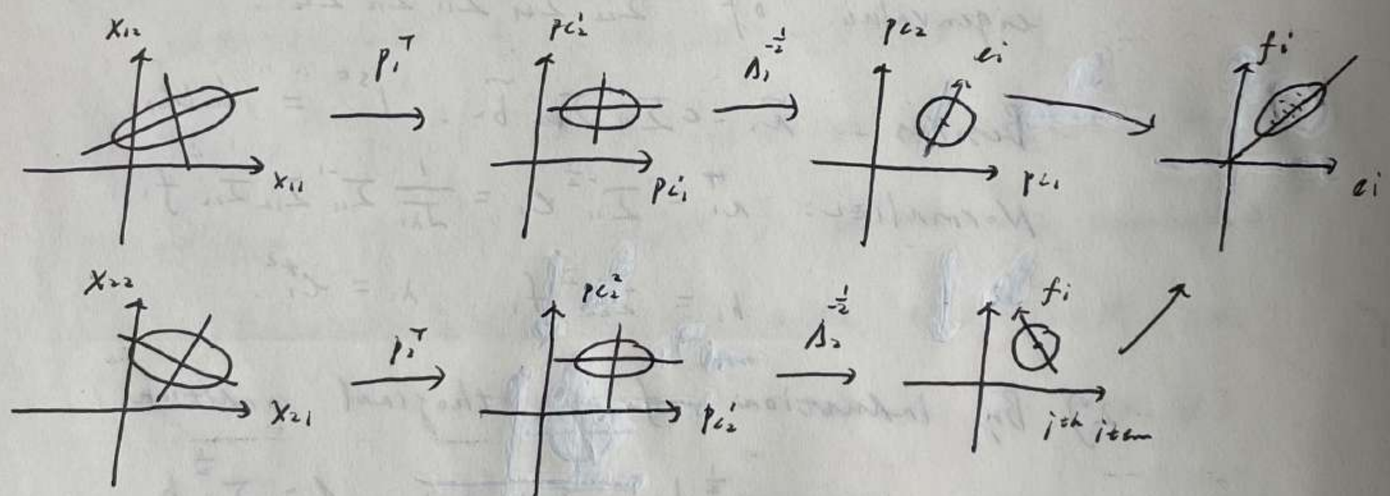
Geometric Integration:

Set $A = (a_1 \cdots a_p)_{p\times p}$, $B = (b_1 \cdots b_p)_{q\times p}$

$\Rightarrow U = A^T X_{(1)}$, $V = B^T X_{(2)}$, $A^T = E^T \Sigma_{11}^{-\frac{1}{2}}$, $E = (e_1 \cdots e_p)$

Decompose $\Sigma_{11}^{-\frac{1}{2}} = P_1 \Lambda_1^{-\frac{1}{2}} P_1^T$. $\therefore U = E^T P_1 \Lambda_1^{-\frac{1}{2}} P_1^T X_{(1)}$

Note that: $P_1^T X_{(1)}$ is PC (analogous to $X_{(2)}$)



Rmk: The last $E^T P_1$ is just rotation to select a direction to projection, which guarantee large correlation.

The direction may do nothing with PCs.

② Properties of Canonical variables:

i) $Cov(U, X_{(1)}) = A^T \Sigma_{11}$, $Cov(V, X_{(2)}) = B^T \Sigma_{22}$

$Cov(U, X_{(2)}) = A^T \Sigma_{12}$, $Cov(V, X_{(1)}) = B^T \Sigma_{21}$.

$\mathrm{corr}(u, X^{(1)}) = \mathrm{Cov}(u, V_{11}^{-\frac{1}{2}} X_{(1)}) = A^T \Sigma_{11} V_{11}^{-\frac{1}{2}}$

$\mathrm{corr}(V, X_{(2)}) = \mathrm{Cov}(V, V_{22}^{-\frac{1}{2}} X_{(2)}) = B^T \Sigma_{22} V_{22}^{-\frac{1}{2}}$

where $V_{11} = \mathrm{diag}\, \Sigma_{11}$. $V_{22} = \mathrm{diag}\, \Sigma_{22}$.

ii) <u>Invariant:</u>

Consider model: $X_{(1)}^{*} = P X_{(1)} + c_1$. $X_{(2)}^{*} = Q X_{(2)} + c_2$.

Then CCA on $(X_{(1)}, X_{(2)})$ is essentially same to $(X_{(1)}^{*}, X_{(2)}^{*})$. and $a_i^{*} = P^{-1} a_i$. $b_i^{*} = Q^{-1} b_i$. $1 \le i \le p$.

iii) To ease the computation burden:

Calculate: $\Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}\, a = \ell^{*2} a$. $\Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}\, b = \ell^{*2} b$.

i.e $|\Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} - \ell_{*}^{2} I| = 0$.

iv) <u>Some Intepretion:</u>

- Compute correlation between canonical variables and original variables can be a way to determine the relative important of origin and canonical
- Canonical correlation generalizes the correlation between 2 variables to 2 groups variables.

  $|\mathrm{Corr}^2(X_i^{(1)}, X_k^{(2)})| = |\mathrm{Corr}^2(e_i^T X^{(1)}, e_k^T X^{(2)})| \le \ell_1^{*2}$
- In multiple correlation coefficient intepretion: ℓ_k^{*2} is the proportion of Var of $u_k = a_k^T X^{(1)}$ explained by the linear combination of data $X^{(2)}$. It's also explain V_k by $X^{(1)}$.

③ CCA for Standardilization:

· Consider $Z^{(1)} = (Z_1^{(1)}, \cdots Z_p^{(1)})^T$. $Z^{(2)} = (Z_1^{(2)} \cdots Z_q^{(2)})^T$

$$\Rightarrow \begin{cases} U_k = a_k^T Z^{(1)} = e_k^T \rho_{11}^{-\frac{1}{2}} Z^{(1)} \\ V_k = b_k^T Z^{(2)} = f_k^T \rho_{22}^{-\frac{1}{2}} Z^{(2)} \end{cases} \quad \text{where}$$

(e_k, ρ_k^{*2}), (f_k, ρ_k^{*2}) are eigenpairs of $\rho_{11}^{-\frac{1}{2}} \rho_{12} \rho_{22}^{-1} \rho_{21} \rho_{11}^{-\frac{1}{2}}$, $\rho_{22}^{-\frac{1}{2}} \rho_{21} \rho_{11}^{-1} \rho_{12} \rho_{22}^{-\frac{1}{2}}$. $e_k = \frac{1}{\sqrt{\rho_k^{*2}}} \cdot \rho_{11}^{-\frac{1}{2}} \rho_{12} \rho_{22}^{-\frac{1}{2}} f_k$.

Rmk: CCA is unchanged under standardilization.

i.e. $\Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-\frac{1}{2}}$ and $\rho_{11}^{-\frac{1}{2}} \rho_{12} \rho_{22}^{-1} \rho_{21} \rho_{11}^{-\frac{1}{2}}$ will have same eigenvalues. Suppose e_k, e_k^* are correspond eigenvectors. $a_k = \Sigma_{11}^{-\frac{1}{2}} e_k$, $a_k^* = \rho_{11}^{-\frac{1}{2}} e_k^*$

$$\Rightarrow a_k^* = V_{11}^{\frac{1}{2}} a_k. \quad V_{11} = \operatorname{diag} \Sigma_{11}.$$

(3) CCA on Samples:

Data: $X = (X^{(1)} \mid X^{(2)}) = \begin{array}{c} \quad Var_1^{(1)} \; Var_2^{(1)} \cdots Var_p^{(1)} \; Var_1^{(2)} \cdots Var_q^{(2)} \\ \text{item} \rightarrow \left(\begin{array}{cccc|ccc} X_{11}^{(1)} & X_{12}^{(1)} & \cdots & X_{1p}^{(1)} & X_{11}^{(2)} & \cdots & X_{1q}^{(2)} \\ \vdots & & & \vdots & \vdots & & \vdots \\ X_{n1}^{(1)} & \cdots & \cdots & X_{np}^{(1)} & X_{n1}^{(2)} & \cdots & X_{nq}^{(2)} \end{array} \right) \end{array}$

$$= \left(\begin{array}{c|c} X_1^{(1)T} & X_1^{(2)T} \\ \vdots & \vdots \\ X_n^{(1)T} & X_n^{(2)T} \end{array} \right)$$

Find CCA: i) Replace population dist by empirical dist.

ii) Replace Σ by S. ρ by R.

① Matrixs of Error Approxi.:

$\hat{u} = \hat{A}^T x^{(1)}$. $\hat{v} = \hat{B}^T x^{(2)}$. $\Rightarrow x^{(1)} = (\hat{A}^T)^{-1} \hat{u}$. $x^{(2)} = (\hat{B}^T)^{-1} \hat{v}$.

Note that: $\mathrm{cov}(\hat{u}, \hat{v}) = \left(\begin{array}{ccc|c} \hat{e}_1^* & & & \\ & \ddots & & 0 \\ & & \hat{e}_p^* & \end{array} \right) = \hat{A}^T S_{12} \hat{B}$

<u>Denote</u>: $(\hat{A}^T)^{-1} = (\hat{a}^{(1)} \cdots \hat{a}^{(p)})$

$(\hat{B}^T)^{-1} = (\hat{b}^{(1)} \cdots \hat{b}^{(q)})$

$\Rightarrow S_{12} = \sum_1^p \hat{e}_i^* \hat{a}^{(i)} \hat{b}^{(i)T}$. Similarly. $\mathrm{cov}(\hat{u}) = \mathrm{cov}(\hat{v}) = I$

$\Rightarrow S_{11} = \sum_1^p \hat{a}^{(i)} \hat{a}^{(i)T}$, $S_{22} = \sum_1^q \hat{b}^{(i)} \hat{b}^{(i)T}$.

<u>Rmk</u>: $X^{(1)} = (\hat{A}^T)^{-1} \hat{u} = \sum_{i=1}^p \hat{u}_i \hat{a}^{(i)}$. $\Rightarrow \mathrm{cov}(X^{(1)}, \hat{u}_i) = \hat{a}^{(i)}$

$\Rightarrow$ The first r columns contain samples cov of

$\hat{u}_1 \cdots \hat{u}_r$ and $X_1^{(1)} \cdots X_p^{(1)}$.

Similar for $(\hat{B}^T)^{-1}$ and $\hat{v}_1 \cdots \hat{v}_r$.

If only the first r canonical pairs are used:

$$\tilde{X}^{(1)} = (\hat{a}^{(1)} \cdots \hat{a}^{(r)}) \begin{pmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_r \end{pmatrix} \in \mathbb{R}^p.$$

$\Rightarrow S_{12}$ is approxi

$$\tilde{X}^{(2)} = (\hat{b}^{(1)} \cdots \hat{b}^{(r)}) \begin{pmatrix} \hat{v}_1 \\ \vdots \\ \hat{v}_r \end{pmatrix} \in \mathbb{R}^q$$

by $\mathrm{cov}(\tilde{X}^{(1)}, \tilde{X}^{(2)})$

<u>Residues</u>:
$$\begin{cases} S_{11} - \sum_1^r \hat{a}^{(i)} \hat{a}^{(i)T} = \sum_{r+1}^p \hat{a}^{(i)} \hat{a}^{(i)T} \\ S_{22} - \sum_1^r \hat{b}^{(i)} \hat{b}^{(i)T} = \sum_{r+1}^q \hat{b}^{(i)} \hat{b}^{(i)T} \\ S_{12} - \sum_1^r \hat{e}_i^* \hat{a}^{(i)} \hat{b}^{(i)T} = \sum_{r+1} \hat{e}_i^* \hat{a}^{(i)} \hat{b}^{(i)T} \end{cases}$$

<u>Rmk</u>: i) Large entries of residual matrix indicate a poor fit for correspond variables.

ii) S_{12} is better fit than S_{11}. S_{22}. Since We can select r. s.t. $\hat{e}_k^*$ is small when $k \geq r+1$. But for S_{11}, S_{22}. the approxi. may not as good as S_{12}. Since it can't be controlled

② Proportions of Explained Sample Var:

Suppose the observations are standardilized.

Canonical coefficients are $\hat{A}_z$. $\hat{B}_z$.

$$\Rightarrow \begin{cases} Cov(Z^{(1)}, \hat{U}_z) = (\hat{A}_z^T)^{-1} = e^2(Z^{(1)}, \hat{U}_z) \\ Cov(Z^{(2)}, \hat{V}_z) = (\hat{B}_z^T)^{-1} = e^2(Z^{(2)}, \hat{V}_z) \end{cases}$$

$$\Rightarrow \begin{cases} tr(R_{11}) = tr(\sum_1^p \hat{a}_z^{(i)} \hat{a}_z^{(i)T}) = p = \sum_1^r \sum_1^p \hat{r}^2_{\hat{U}_{zi}, \hat{z}_k^{(1)}} \\ tr(R_{22}) = tr(\sum_1^2 \hat{b}_z^{(i)} \hat{b}_z^{(i)T}) = q = \sum_1^2 \sum_1^p \hat{r}_{\hat{V}_{zi}, \hat{z}_k^{(2)}} \end{cases}$$

Rmk: We can calculate the proportion of total sample Variances explained by first r variables.

$$R^2_{z^{(1)}|\hat{u}_{z1}\cdots\hat{u}_{zr}} = tr(\sum_1^r \hat{a}_z^{(i)} \hat{a}_z^{(i)T}) / tr(R_{11}) = \frac{\sum_1^r \sum_1^p \hat{r}^2_{\hat{u}_{zi}, \hat{z}_k^{(1)}}}{p}$$

It indicate how well the r canonical variates represent the original sets $Z^{(1)}$.

And the matrix of error can be interpreta by $1 - R^2_{z^{(1)}|\hat{u}_{z1}\cdots\hat{u}_{zr}}$. Similar for $Z^{(2)}$, $\hat{V}_z$.

(4) Large Sample Test:

Assume: $\begin{pmatrix} X_j^{(1)} \\ X_j^{(2)} \end{pmatrix} \overset{i.i.d}{\sim} N_{p+q}(\mu, \Sigma)$. $1 \le j \le n$.

Test $H_0: \Sigma_{12} = 0$ v.s. $H_1: \Sigma_{12} \neq 0$. By MLE test:

$$-2\ln\Lambda = n\ln\left(\frac{|S_{11}||S_{22}|}{|S|}\right) = n\ln|I - S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}|$$

By $\begin{vmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{vmatrix} = \left|\begin{pmatrix} I & 0 \\ -S_{21}S_{11}^{-1} & I \end{pmatrix}\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}\begin{pmatrix} I & -S_{11}^{-1}S_{12} \\ 0 & I \end{pmatrix}\right|$

$$= |S_{22}||S_{11} - S_{12}S_{22}^{-1}S_{21}| \hat{=} |S_{22}||S_{11}||I - \hat{M}|$$

$\therefore \Lambda = \prod_1^p (1 - \hat{\ell}_i^{*2})$. $T = -n\sum_1^p \ln(1 - \hat{\ell}_i^{*2})$

$T \sim \chi^2_{pq}$ when $n \to \infty$.

Note: $T \uparrow$ if $\exists\ \hat{\ell}_i^{*2} \to 1$. $\therefore R = \{T > \chi^2_{pq}(\alpha)\}$.

<u>Rmk</u>: i) $H_0: \Sigma_{12} = 0 \iff H_0: \ell_1^2 = \cdots \ell_p^2 = 0$

ii) Bartlett suggest: $-(n - 1 - \frac{1}{2}(p+q+1))\ln\Lambda$

<u>Continuation</u>: Test: $H_0^k: \ell_1^* \cdots \ell_k^* \neq 0$. $\ell_{k+1}^* = \cdots \ell_p^* = 0$

v.s $H_1^k: \exists i \ge k+1$. $\ell_i^* \neq 0$

$$R = \{-(n-1-\tfrac{1}{2}(p+q+1))\ln\prod_{k+1}^{p}(1 - \hat{\ell}_i^{*2}) > \chi^2_{(p-k)(q-k)}(\alpha)\}.$$

<u>Rmk</u>: It can reduce the numbers of canonical varinbles. (c.f. (3). ①)

(5) <u>Procedure</u>:

i) From samples of X, Y $\Rightarrow$ Calculate R

ii) Calculate canonical coefficients and varinbles

iii) Test $\Sigma_{XY} = 0$?   iv) Apply CCA base on iii)