ANALISI DELLE CORRELAZIONI CANONICHE

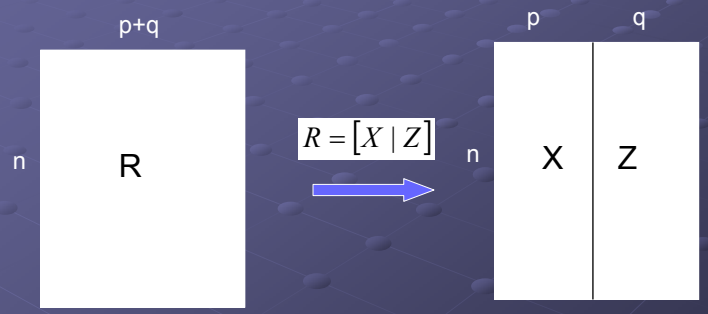
Dott.ssa Agnieszka Stawinoga
Dipartimento di Matematica e Statistica
Università degli Studi di Napoli Federico II
agnieszka.stawinoga@unina.it

- L' Analisi delle Correlazioni Canoniche è stata proposta da H. Hottelling nel 1936.
- Nel 1968 J. D.Carroll ha proposto una generalizzazione dell' ACC per tre o più gruppi di variabili.
- L'ACC è molto importante dal punto di vista teorico e metodologico. Essa può essere considerata come il caso generale delle:
  - Regressione Multipla
  - Analisi delle Corrispondenze
  - Analisi Discriminante

- L' obiettivo dell' ACC è identificare le relazioni lineari esistenti tra due gruppi di variabili quantitative osservate su uno stesso insieme di individui.
- Lo scopo è trovare una combinazione lineare delle variabili del primo gruppo e una combinazione lineare delle variabili del secondo gruppo che siano le più correlate possibile.
- L' ACC opera su una matrice *R* ad *n* righe e *p+q* colonne partizionabile in due sottomatrici *X (n,p)* e *Z(n,q)*.

  Le colonne della matrice *X* sono costituite dalle variabili del primo insieme.
  Le variabili del secondo insieme costituiscono le colonne della matrice *Z*.

- Si supponga che tutte le variabili siano centrate quindi per ogni colonna della matrice R la somma degli elementi è uguale a 0 .
- Dal vettore $(x_{i1}, x_{i2}, \ldots, x_{ip}, z_{i1}, \ldots, z_{iq})$ si identifica il generico individuo *i* della matrice *R*.



- La matrice di varianza-covarianza della matrice R può essere ottenuta da:

$$V(R) = \frac{1}{n} R'R = \frac{1}{n} \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z \end{bmatrix}$$

- Si indica con *a* un vettore di *p* componenti e con *b* un altro a *q* componenti.

  $a = (a_1, a_2, \ldots, a_p)$, $b = (b_1, b_2, \ldots, b_q)$
- Per il generico individuo *i* si definiscono le due combinazioni lineari:

$$a(i) = \sum_{j=1}^{p} a_j x_{ij}, b(i) = \sum_{j=1}^{q} b_j z_{ij}$$

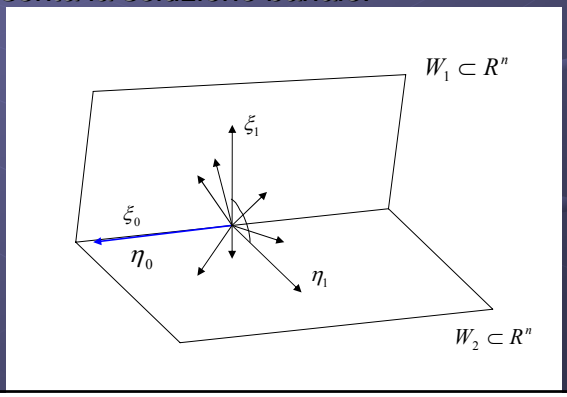
- I valori di *a(i)* e *b(i)* costituiscono le componenti dei vettori:

$$\xi = Xa \quad , \quad \eta = Zb$$

- Lo scopo dell'Analisi delle Correlazioni Canoniche è trovare i coefficienti dei vettori *a* e *b* che massimizzano la correlazione tra ξ ed η.

- Definiamo:
  - Variabili canoniche ⟹ i vettori $\xi \in R^n, \eta \in R^n$
  - Fattori canonici ⟹ i vettori di coefficienti $a \in R^p, b \in R^q$
  - Correlazione canonica ⟹ il coefficiente di correlazione tra ξ ed η
- Il gruppo delle variabili $x_{i1}, x_{i2}, \ldots, x_{ip}$ *costituisce un sottospazio vettoriale* W_1 *di* R^n *chiamato potenziale di previsione del primo gruppo. Ugualmente le variabili del secondo insieme formano un sottospazio* W_2 *di* R^n .

- *Indichiamo con $(\xi_1,\xi_2,\ldots,\xi_k)$, $(\eta_1,\eta_2,\ldots,\eta_k)$ due basi ortonormali rispettivamente di W_1 e W_2 tali che le coppie (ξ_i, η_i) i=1,...,k siano le più correlate possibile.*
- *Poiché le variabili sono centrate l' obiettivo è cercare le coppie (ξ_i, η_i) i=1,...,k tali che massimizzano il coseno dell' angolo formato da loro.*
- *Il vettore comune di due sottospazi (la prima bisettrice) rappresenta la soluzione banale.*



- Il valore del coseno dell' angolo formato tra due variabili canoniche ξ ed η *è uguale al valore del loro coefficiente di correlazione.*

$$\cos(\xi,\eta) = \frac{a' X' Zb}{\sqrt{(a' X' Xa)(b' Z' Zb)}}$$

- *L' angolo tra due vettori non dipende dalla loro norma quindi possiamo porre:*

$$\|\xi\| = a' X' Xa = 1, \|\eta\| = b' Z' Zb = 1$$

- *Usando la funzione di Lagrange si possono ottenere sotto le condizioni di normalizzazione i vettori a e b che massimizzano la quantità a'X'Zb.*

$$L = a' X' Zb - \lambda(a' X' Xa - 1) - \mu(b' Z' Zb - 1)$$

- Derivando l'equazione di Lagrange rispetto ad *a* e *b* e ponendo i risultati uguali a 0 otteniamo:

(1)
$$\begin{aligned} X'Zb - 2\lambda X'Xa = 0 \\ Z'Xa - 2\mu Z'Zb = 0 \end{aligned}$$

- Sotto le condizioni di normalizzazione moltiplichiamo le equazioni riportate sopra rispettivamente per *a'* e *b':*

$$\begin{aligned} a'X'Zb = 2\lambda \\ b'Z'Xa = 2\mu \end{aligned}$$

- Ricordando che il trasposto di uno scalare è lo scalare stesso otteniamo la quantità:

$$\beta = 2\lambda = a'X'Zb$$

Questa quantità è il coefficiente di correlazione massimo che abbiamo trovato.

- Il nostro sistema da risolvere lo possiamo scrivere nella forma:

(2)
$$\begin{aligned} X'Zb = \beta X'Xa \\ Z'Xa = \beta Z'Zb \end{aligned}$$

- Per poter risolvere questo sistema le matrici *X'X* e *Z'Z* devono essere non singolari (det(X'X) ≠0, det(Z'Z) ≠0)

- Ricavando *a* dalla prima equazione del sistema (2)e *b* dalla seconda otteniamo:

(3) $a = \frac{1}{\beta}(X'X)^{-1}X'Zb$ (4) $b = \frac{1}{\beta}(Z'Z)^{-1}Z'Xa$

- Adesso sostituendo *a* nella equazione (4) si ottiene:

$$Z'X(X'X)^{-1}X'Zb = \beta^2 Z'Zb$$

- Il vettore *b* soluzione del sistema è quindi l'autovettore della matrice:

$$(Z'Z)^{-1}Z'X(X'X)^{-1}X'Z$$

associato al più grande autovalore β^2 .

- β^2 rappresenta il quadrato del coefficiente di correlazione tra ξ ed η.
- Il vettore *a* può essere calcolato dalla equazione (3) o come autovettore associato al più grande autovalore della matrice:

$$(X'X)^{-1}X'Z(Z'Z)^{-1}Z'X$$

- Per poter ottenere le variabili canoniche ξ *ed* η moltiplichiamo le equazioni (3) e (4) rispettivamente per *X* e *Z*.

- Si ottiene:

(5) $$\xi = Xa = \frac{1}{\beta}\underbrace{X(X'X)^{-1}X'}_{P_1}Zb$$

(6) $$\eta = Zb = \frac{1}{\beta}\underbrace{Z(Z'Z)^{-1}Z'}_{P_2}Xa$$

- Dalle equazioni (5) e (6) deriva un risultato importante. Le matrici:

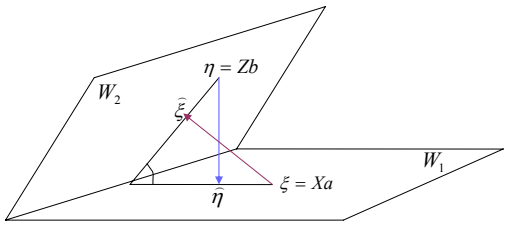
$$P_1 = X(X'X)^{-1}X' \qquad P_2 = Z(Z'Z)^{-1}Z'$$

- Le matrici P_1 e P_2 sono simmetriche ed idempotenti. Possiamo considerarle come operatori di proiezione di R^n sui sottospazi W_1 W_2 generati dalle colonne delle matrici *X* e *Z*.

- Ciascun vettore η (oppure ξ) è collineare con la proiezione ortogonale dell' altro.

  Quindi se η_1 è un vettore unitario di W_2 , il vettore di W_1 che forma l' angolo minimo con η_1 è il vettore $\hat{\eta}_1$ proiezione ortogonale di η_1 su W_1 .

Collinearità dei vettori $\hat{\eta}$ e ξ



- Il legame dell' ACC con la regressione multipla

In questo caso la matrice *Z* è formata da una sola colonna(q =*1*). La matrice *Z* è costituita dalla variabile da spiegare z e *X* è costituita dalle *p* variabili esplicative $x_1, x_2, \ldots, x_p$. Il vettore *b* ha una sola componente ed è quindi uno scalare, lo stesso il prodotto *Z'Z*. Possiamo scrivere:

$$\beta^2 = \frac{Z'X(X'X)^{-1}X'Z}{Z'Z}$$

La quantità β^2 costituisce il coefficiente di correlazione multipla tra la variabile da spiegare e le variabili esplicative. Dalla (3) otteniamo:

$$a = \frac{b}{\beta}(X'X)^{-1}X'Z$$

Il vettore *a* è proporzionale al vettore dei coefficienti della regressione multipla con $x_1, x_2, \ldots, x_p$ variabili esplicative e *z* la variabile dipendente. Dalla condizione di normalizzazione si ricava:

$$b = \frac{1}{\sqrt{Z'Z}}$$

Il legame con l' Analisi delle Corrispondenze

In questo caso si considerano due matrici Z_1 e Z_2 di dimensioni rispettivamente (n, q_1) e (n, q_2).
L' Analisi delle Corrispondenze di una tabella di contingenza può essere vista come l' Analisi delle Correlazioni Canoniche della matrice $[Z_1 Z_2]$.
Nella matrice Z_1 vengono riportati i valori in codifica disgiuntiva completa delle q_1 modalità di una variabile X.
La matrice Z_2 è costituita dalle q_2 modalità di una variabile Y.
Tutte due osservate su n unità. Dal prodotto $Z_1'Z_2$ otteniamo la tabella di contingenza F di dimensioni (q_1, q_2).
Con D_1 e D_2 indichiamo le matrici diagonali rispettivamente dei marginali di riga e di colonna.
Otteniamo :

$$\varphi_1 = \frac{1}{\sqrt{\lambda}} D_1^{-1} Z_1' Z_2 \varphi_2 \qquad \varphi_2 = \frac{1}{\sqrt{\lambda}} D_2^{-1} Z_2' Z_1 \varphi_1$$

Se moltiplichiamo per Z_1 e Z_2 otteniamo:

$$Z_1\varphi_1 = \frac{1}{\sqrt{\lambda}} Z_1 D_1^{-1} Z_1' Z_2 \varphi_2 \qquad Z_2\varphi_2 = \frac{1}{\sqrt{\lambda}} Z_2 D_2^{-1} Z_2' Z_1 \varphi_1$$

La matrice $Z = [Z_1 Z_2]$ ha $q_1 + q_2$ colonne con corrispondenti $q_1 + q_2$ punti nello spazio R^n . Ogni sottomatrice Z_i (i = 1,2) genera in R^n un sottospazio lineare W_i a q_i dimensioni. Le componenti del vettore φ_1 $(q_1,1)$ costituiscono le coordinate di un punto m_1 nel sottospazio W_1.
Le coordinate di m_1 in R^n : $m_1 = Z_1\varphi_1$

Con P_1 e P_2 abbiamo indicato gli operatori di proiezione su W_1 e W_2 .

$$P_1 = Z_1(Z_1'Z_1)^{-1}Z_1' \qquad P_2 = Z_2(Z_2'Z_2)^{-1}Z_2'$$

Adesso i nostri vettori m_1 e m_2 possono essere ottenute come: $m_1 = \frac{1}{\sqrt{\lambda}} P_1 m_2$ $m_2 = \frac{1}{\sqrt{\lambda}} P_2 m_1$

La proiezione di m_2 su W_1 è collineare a m_1 .
L' analisi delle corrispondenze di una tabella di contingenza può essere vista come lo studio della posizione relativa dei sottospazi W_1 e W_2 e quindi come l' Analisi delle Correlazioni Canoniche della matrice $[Z_1 Z_2]$.

- Riferimenti bibliografici

- Bouroche J.-M., Saporta G. (1983)- *L'Analisi dei Dati*, (a cura di) D'Alfonso G., CLU, Napoli
- Gherghi M.,Lauro N.C. (2004)-*Appunti di Analisi dei Dati Multidimensionali,* RCEedizioni, Napoli
- Lebart L., Morineau A., Warwick K.M. (1984)- *Multivariate Descriptive Statistical Analysis*, Wiley & Sons