

Analisi in componenti principali

Carla Rampichini

rampichini@ds.unifi.it
Dipartimento di Statistica "G. Parenti"
Università di Firenze

Indice

- 1 Definizione delle componenti principali
- 2 Proprietà delle CP
- 3 Riduzione di dimensionalità
- 4 Interpretazione delle CP campionarie
- 5 Biplot
 - Esempio biplot
 - Scomposizione in valori singolari
 - Costruzione biplot

Analisi in componenti principali

- Date p v.c. quantitative $\mathbf{x}' = [X_1, X_2, \dots, X_p]$, con vettore delle medie μ e matrice di covarianza Σ , si sostituisce alle variabili originali *correlate* un nuovo insieme di variabili, dette *componenti principali* che hanno le seguenti proprietà:
 - sono **incorrelate** (ortogonali)
 - sono elencate in ordine decrescente rispetto alla loro **varianza**
- La prima componente principale Y_1 è la *combinazione lineare* delle p variabili di partenza avente la *massima varianza*.
- La seconda componente principale Y_2 è la *combinazione lineare* delle p variabili con la varianza immediatamente inferiore alla varianza di Y_1 e ad essa *incorrelata*.
- ecc. fino alla p -esima componente.
- Se le p variabili originali sono molto correlate, un numero $q < p$ tiene conto di una quota elevata di varianza totale per cui le prime q componenti forniscono una buona descrizione della struttura dei dati.

Come si determina la prima CP?

Sarà necessario trovare il vettore di dimensione p dei coefficienti della combinazione lineare che rende massima la varianza di Y_1 .

- $Y_1 = \mathbf{a}'_1 \mathbf{x}$, $\mathbf{a}'_1 = [a_{11}, \dots, a_{1p}]$
- Y_1 deve avere varianza massima, cioè si dovrà trovare $\mathbf{a}_1$:
 $\text{var}(Y_1) = \text{var}(\mathbf{a}'_1 \mathbf{x}) = \mathbf{a}'_1 \boldsymbol{\Sigma} \mathbf{a}_1$ sia massima.
- Le soluzioni di questo problema di massimo sono infinite proporzionali perchè la combinazione lineare contiene un fattore di scala arbitrario.
- Si impone il **vincolo** che il vettore dei coefficienti abbia lunghezza unitaria: $\mathbf{a}'_1 \mathbf{a}_1 = 1$
- Il problema di massimo vincolato si può risolvere attraverso il metodo dei moltiplicatori di Lagrange
 - si esprime il vincolo come eguaglianza a zero $\mathbf{a}'_1 \mathbf{a}_1 - 1 = 0$
 - si calcolano le derivate prime parziali di $\mathbf{a}'_1 \boldsymbol{\Sigma} \mathbf{a}_1 - \lambda(\mathbf{a}'_1 \mathbf{a}_1 - 1)$
 - si eguagliano a zero le derivate prime.

Determinazione del massimo vincolato

$$\max_{\mathbf{a}_1} : \mathbf{a}_1' \mathbf{\Sigma} \mathbf{a}_1, \text{ s.t. } \mathbf{a}_1' \mathbf{a}_1 = 1$$

$$\frac{\partial [\mathbf{a}_1' \mathbf{\Sigma} \mathbf{a}_1 - \lambda (\mathbf{a}_1' \mathbf{a}_1 - 1)]}{\partial \mathbf{a}_1} = \mathbf{0}$$

$$2\mathbf{\Sigma} \mathbf{a}_1 - 2\lambda \mathbf{a}_1 = 2[\mathbf{\Sigma} - \lambda \mathbf{I}] \mathbf{a}_1 = \mathbf{0}$$

$$(\mathbf{\Sigma} - \lambda \mathbf{I}) \mathbf{a}_1 = \mathbf{0}$$

- Questo è un sistema lineare che ammette soluzioni non tutte nulle se il suo determinante è diverso da zero
- equazione caratteristica di $\mathbf{\Sigma}$: $|\mathbf{\Sigma} - \lambda \mathbf{I}| = 0$
- è un polinomio di grado p il λ , che ammette p soluzioni (non necessariamente distinte): $\lambda_1 \geq \dots \geq \lambda_p$
- poichè la matrice di covarianza è definita semipositiva gli autovalori sono tutti non negativi.

La prima componente principale

- Si sceglie come prima CP la combinazione lineare $Y_1 = \mathbf{e}_1' \mathbf{x}$, dove $\mathbf{e}_1$ è l'autovettore associato al primo autovalore λ_1 di Σ
- Ricordando che vale la seguente eguaglianza $\Sigma \mathbf{e}_1 = \lambda_1 \mathbf{e}_1$, si ha:

$$\text{var}(Y_1) = \mathbf{e}_1' \Sigma \mathbf{e}_1 = \mathbf{e}_1' \lambda_1 \mathbf{e}_1 = \lambda_1 \mathbf{e}_1' \mathbf{e}_1 = \lambda_1$$

La j -ma componente principale

- La j -ma componente principale del vettore casuale $\mathbf{x} = [X_1, \dots, X_p]$, la cui matrice di covarianza Σ ha autovalori-autovettori $(\lambda_j, \mathbf{e}_j)$, $j = 1, \dots, p$, è data da:

$$Y_j = \mathbf{e}_j' \mathbf{x} = e_{j1} X_1 + \dots e_{jp} X_p,$$
- $\text{Var}(Y_j) = \mathbf{e}_j' \Sigma \mathbf{e}_j = \lambda_j$ (ricordare: $\Sigma \mathbf{e}_j = \lambda_j \mathbf{e}_j$)
- $\text{Covar}(Y_j, Y_k) = \mathbf{e}_j' \Sigma \mathbf{e}_k = \mathbf{e}_j' \lambda_k \mathbf{e}_k = \lambda_k \mathbf{e}_j' \mathbf{e}_k = 0$

Varianza totale

La varianza totale delle p componenti principali è uguale alla varianza totale delle $X_1, \dots, X_p$:

$$\sum_{j=1}^p \text{Var}(X_j) = \sum_{j=1}^p \sigma_{jj} = \sum_{j=1}^p \lambda_j = \sum_{j=1}^p \text{Var}(Y_j)$$

Infatti:

- $\sum_{j=1}^p \sigma_{jj} = \text{tr}(\mathbf{\Sigma})$
- poichè $\mathbf{\Sigma} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}'$, con $\mathbf{\Lambda} = \text{diag}(\lambda_j)$ e $\mathbf{E} = [\mathbf{e}_1, \dots, \mathbf{e}_p]$, si ha
- $\text{tr}(\mathbf{\Sigma}) = \text{tr}(\mathbf{E}\mathbf{\Lambda}\mathbf{E}')$
- ricordare $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$, posto $\mathbf{A} = \mathbf{E}$ e $\mathbf{B} = \mathbf{\Lambda}\mathbf{E}'$, si ha infine
- $\text{tr}(\mathbf{\Sigma}) = \text{tr}(\mathbf{E}\mathbf{\Lambda}\mathbf{E}') = \text{tr}(\mathbf{\Lambda}\mathbf{E}'\mathbf{E}) = \text{tr}(\mathbf{\Lambda}\mathbf{I}) = \text{tr}(\mathbf{\Lambda})$

Correlazione tra CP e X

La correlazione tra le CP e le X originali è data da

$$\rho(Y_j, X_k) = \frac{\mathbf{e}_{jk} \sqrt{\lambda_j}}{\sqrt{\sigma_{kk}}}$$

Infatti:

- posto $\mathbf{a}'_k = [0, \dots, 0, 1, 0, \dots]$, $X_k = \mathbf{a}'_k \mathbf{x}$
- $\text{Cov}(X_k, Y_j) = \text{Cov}(\mathbf{a}'_k \mathbf{x}, \mathbf{e}'_j \mathbf{x}) = \mathbf{a}'_k \Sigma \mathbf{e}_j$
- ricordando $\Sigma \mathbf{e}_j = \lambda_j \mathbf{e}_j$ si ha
- $\text{Cov}(X_k, Y_j) = \mathbf{a}'_k \lambda_j \mathbf{e}_j = \lambda_j \mathbf{a}'_k \mathbf{e}_j = \lambda_j \mathbf{e}_{jk}$
- poichè $\text{Var}(X_k) = \sigma_{kk}$ e $\text{Var}(Y_j) = \lambda_j$ si ottiene
- $\rho(Y_j, X_k) = \frac{\text{Cov}(X_k, Y_j)}{\sqrt{\text{Var}(X_k) \text{Var}(Y_j)}} = \frac{\lambda_j \mathbf{e}_{jk}}{\sqrt{\sigma_{kk} \lambda_j}}$

Componenti principali campionarie

- date n osservazioni indipendenti $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ da una popolazione p -dimensionale con centroide μ e matrice di covarianza Σ
 - $\bar{\mathbf{x}}$ vettore delle medie campionarie
 - $\mathbf{S}$ e $\mathbf{R}$ matrici di covarianza e correlazione campionarie
- le *componenti principali* campionarie sono definite come le *combinazioni lineari* delle $\mathbf{x}$ con la varianza campionaria massima.

obiettivo: trovare $k < p$ combinazioni lineari incorrelate che tengano conto il più possibile della variabilità presente nei dati.

Combinazioni lineari

Dati n valori campionari, per una qualsiasi combinazione lineare delle p variabili $\mathbf{x}$:

$$\mathbf{a}'_1 \mathbf{x} = \sum_{j=1}^p a_{1j} x_{ij}, \quad i = 1, \dots, n$$

- la media campionaria della combinazione lineare è $\mathbf{a}'_1 \bar{\mathbf{x}}$
- la varianza campionaria della combinazione lineare è $\mathbf{a}'_1 \mathbf{S} \mathbf{a}_1$
- le due combinazioni lineari $(\mathbf{a}'_1 \mathbf{x}, \mathbf{a}'_2 \mathbf{x})$ hanno covarianza $\mathbf{a}'_1 \mathbf{S} \mathbf{a}_2$

Come trovare le CP campionarie

- Matrice di covarianza campionaria $\mathbf{S}$, con autovalori e autovettori $(\hat{\lambda}_j, \hat{\mathbf{e}}_j)$, $j = 1, \dots, p$
- $\hat{\lambda}_1 \geq \hat{\lambda}_2 \dots \geq \hat{\lambda}_p \geq 0$
- la j -ma CP è

$$\hat{y}_j = \hat{\mathbf{e}}_j' \mathbf{x} = \hat{e}_{j1} x_1 + \dots + \hat{e}_{jp} x_p$$

- $\text{var}(\hat{y}_j) = \hat{\lambda}_j$
- $\text{covar}(\hat{y}_j, \hat{y}_k) = 0$
- varianza campionaria totale $\sum_{j=1}^p s_{jj} = \sum_{j=1}^p \hat{\lambda}_j$
- $\text{corr}(\hat{y}_j, \mathbf{x}_k) = \frac{\hat{e}_{jk} \sqrt{\hat{\lambda}_j}}{\sqrt{s_{kk}}}$

Componenti principali campionarie: osservazioni

- Le CP campionarie possono essere ricavate da $\mathbf{S}$ o da $\mathbf{R}$
- Le CP ricavate da $\mathbf{S}$ sono diverse dalle CP ricavate da $\mathbf{R}$
- Di solito le $\mathbf{x}$ vengono *centrate*: $\mathbf{x} - \bar{\mathbf{x}}$
- La *centratura* non modifica $\mathbf{S}$ e quindi nemmeno autovalori e autovettori
- CP *centrate* $\hat{y}_j = \mathbf{e}_j'(\mathbf{x} - \bar{\mathbf{x}})$
- In tal modo la media campionaria delle CP è zero. Infatti, considerando le n osservazioni campionarie $\mathbf{x}_i$ (vettore $p \times 1$ i -ma oss.), $i = 1, \dots, n$,

$$\bar{\hat{y}}_j = \frac{1}{n} \mathbf{e}_j' \left(\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) \right) = \frac{1}{n} \mathbf{e}_j' \mathbf{0} = 0$$

CP campionarie nel caso di distribuzione normale

Se $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- le CP campionarie $\hat{y}_j = \hat{\mathbf{e}}_j'(\mathbf{x} - \bar{\mathbf{x}})$ sono *realizzazioni* delle CP della popolazione $Y_j = \mathbf{e}_j'(\mathbf{x} - \boldsymbol{\mu})$
- $\mathbf{Y} \sim N_p(\mathbf{0}, \boldsymbol{\Lambda})$, $\boldsymbol{\Lambda} = \text{diag}\{\lambda_j\}$, $j = 1, \dots, p$
- Il profilo che si ottiene considerando per tutti i vettori $\mathbf{x}$ di dimensione $p \times 1$ che soddisfano

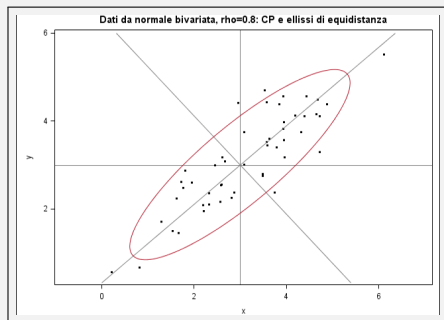
$$(\mathbf{x} - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}}) = c^2$$

è una stima del profilo a densità costante $(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2$ della densità normale sottostante.

- I profili $(\mathbf{x} - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}}) = c^2$ sono ellissoidi centrati in $\bar{\mathbf{x}}$ e con assi pari agli autovettori di $\mathbf{S}^{-1}$ (equivalenti a quelli di $\mathbf{S}$)
- la lunghezza degli assi è proporzionale a $\sqrt{\hat{\lambda}_j}$, ($\hat{\lambda}_j$ autovalori di $\mathbf{S}$)

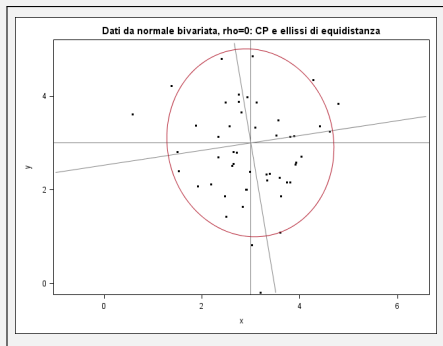
CP campionarie e distribuzione normale

- I profili $(\mathbf{x} - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}}) = c^2$ possono essere sovrapposti allo scatterplot della nuvola di punti per evidenziare la distribuzione normale generatrice dei dati
- le CP campionarie $\hat{\mathbf{y}}_j$ possono essere costruite anche quando la distribuzione delle $\mathbf{X}$ non è normale
- Le CP sono le proiezioni ortogonali delle $\mathbf{x}$ centrate sugli autovettori $\hat{\mathbf{e}}_j \Rightarrow$ (i) *traslazione* dell'origine degli assi in $\bar{\mathbf{x}}$; (ii) *rotazione* degli assi fino a passare per la nuvola di punti seguendo la direzione di massima varianza (gli assi dell'ellissoide)



CP campionarie per variabili incorrelate

- Se $\hat{\lambda}_1 \doteq \hat{\lambda}_2$ la direzione degli assi dell'ellissi non è univocamente determinata e gli assi possono giacere in qualsiasi direzione, compresa quella degli assi originali.
- quando il profilo di equidistanza è approx circolare, la variabilità campionaria è omogenea lungo qualsiasi direzione
- non è possibile rappresentare i dati in un spazio di dimensione ridotta



(dati e grafici generati con cp_bivariata.sas, SAS v9.2)

Standardizzazione delle variabili

In generale le CP campionarie *non sono invarianti* per trasformazioni di scala.

- La standardizzazione delle variabili si ottiene come

$$\mathbf{z}_i = \mathbf{D}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}}) = \left[\frac{x_{i1} - \bar{x}_1}{\sqrt{s_{11}}} \cdots \frac{x_{ip} - \bar{x}_p}{\sqrt{s_{pp}}} \right]'$$

$$i = 1, \dots, n, \mathbf{D}^{-1/2} = \text{diag}\left\{\frac{1}{\sqrt{s_{jj}}}\right\}, j = 1, \dots, p.$$

- La matrice dei dati standardizzati è data da $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_n]'$
- con vettore delle medie

$$\bar{\mathbf{z}} = \frac{1}{n}(\mathbf{1}'\mathbf{Z})' = \frac{1}{n}\mathbf{Z}'\mathbf{1} = \left[\sum_{i=1}^n \frac{x_{i1} - \bar{x}_1}{\sqrt{s_{11}}} \cdots \sum_{i=1}^n \frac{x_{ip} - \bar{x}_p}{\sqrt{s_{pp}}} \right]'$$
- e matrice di covarianza $\frac{1}{n-1}\mathbf{Z}'\mathbf{Z} = \mathbf{R}$

CP campionarie sulle variabili standardizzate

- La varianza totale delle variabili standardizzate è

$$tr(\mathbf{R}) = p = \sum_{j=1}^p \hat{\lambda}_j$$
- Autovalori e autovettori di $\mathbf{R}$ sono diversi da quelli di $\mathbf{S}$
- $corr(\mathbf{z}_j, \hat{\mathbf{y}}_k) = \hat{e}_{jk} \sqrt{\hat{\lambda}_k}$
- regola empirica: considerare solo le CP che spiegano una proporzione pari almeno a $1/p$ della varianza totale, cioè con

$$var(\hat{y}_j) = \hat{\lambda}_j > 1$$

Rappresentazione dei dati multivariati

- Quando si hanno più di due variabili
 - la posizione di ciascuna **unità** rispetto alle altre può essere rappresentata nel diagramma relativo alle **prime due CP**
 - l'importanza relativa di ciascuna **variabile** e la loro relazione può essere rappresentata nella **matrice degli scatterplot**
- ma come possiamo rappresentare queste due informazioni in un unico grafico?
- per $p = 2$ lo scatterplot rappresenta entrambe queste informazioni:
 - la posizione di ciascuna unità rispetto alle altre
 - l'importanza relativa di ciascuna variabile rispetto alla posizione delle unità
- ma per $p > 2$ come possiamo fare?

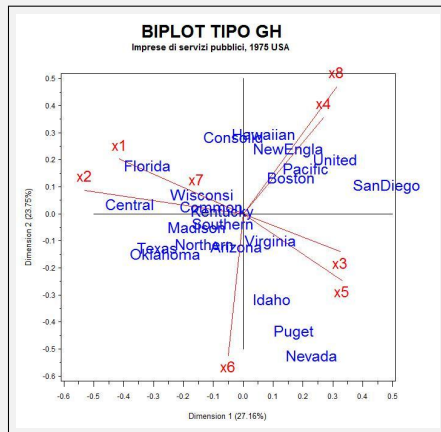
Biplot

- Il **biplot** è una rappresentazione grafica dell'informazione contenuta in una matrice **X** di dimensione $n \times p$
- L'idea sottostante al biplot consiste nell'aggiungere l'informazione sulla relazione tra variabili al grafico delle componenti principali.
- Il suffisso *bi* indica le due informazioni contenute in **X** e rappresentate nel grafico
 - le righe di **X** rappresentano le *osservazioni* campionarie
 - le colonne di **X** rappresentano le *variabili*.
- quindi come si costruisce il biplot: la costruzione del biplot si basa sulla *scomposizione in valori singolari* della matrice dei dati (centrati) **X** (o della matrice dei dati standardizzati **Z**)
- vedremo ora un esempio

Esempio biplot

Il biplot mostra quali compagnie sono tra loro più vicine e quali sono le variabili che contribuiscono a questo raggruppamento.

- X_4 e X_8 sono responsabili del gruppo in alto a destra, che comprende le compagnie della costa;
- mentre X_1 e X_2 contribuiscono alla vicinanza delle compagnie della Florida e della Louisiana.



Dati raccolti su 22 imprese che hanno erogato servizi pubblici nel 1975 negli USA
(JW Tab 12.12, `utility_biplot.sas`)

Scomposizione in valori singolari

Singular Value Decomposition (SVD)

- Si dimostra che una matrice $\mathbf{X}$ di dimensione $n \times p$ può essere scomposta nel modo seguente

$$\underset{n \times p}{\mathbf{X}} = \underset{n \times p}{\mathbf{U}} \underset{p \times p}{\mathbf{\Gamma}} \underset{p \times p}{\mathbf{A}'}'$$

- sotto il vincolo che $\mathbf{U}'\mathbf{U} = \mathbf{A}'\mathbf{A} = \mathbf{I}$, dove $\mathbf{I}$ è la matrice identità
- La matrice ortonormale $\mathbf{A}$ è la matrice degli autovettori di $\mathbf{X}'\mathbf{X} = (n-1)\mathbf{S}$
- $\mathbf{\Gamma} = \text{diag}\{\lambda_j^{1/2}\}$, λ_j autovettori di $\mathbf{X}'\mathbf{X} = (n-1)\mathbf{S}$

SVD e CP

Consideriamo la matrice dei dati centrata $\mathbf{X}$

- ACP fattorizza $\mathbf{X}'\mathbf{X}$ (cioè la matrice di covarianza $\mathbf{S}$)
 - $\mathbf{X}'\mathbf{X} = (n - 1)\mathbf{S} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}'$.
 - le CP si ottengono come combinazione lineare $\mathbf{Y} = \mathbf{X}\mathbf{A}$
- SVD fattorizza la matrice dei dati
 $\mathbf{X} = \mathbf{U}\mathbf{\Gamma}\mathbf{A}'$
 - postoltiplicando per $\mathbf{A}$ si ottiene $\mathbf{X}\mathbf{A} = \mathbf{U}\mathbf{\Gamma}$
 - e infine postmoltiplicando per $\mathbf{\Gamma}^{-1}$ si ha

$$\begin{array}{c} \Downarrow \\ \mathbf{X}\mathbf{A}\mathbf{\Gamma}^{-1} = \mathbf{U} \end{array}$$

- Le p colonne di $\mathbf{U}$ contengono i punteggi normalizzati ottenuti con l'ACP $\mathbf{U} = [\frac{1}{\sqrt{\lambda_1}}\mathbf{y}_1, \dots, \frac{1}{\sqrt{\lambda_p}}\mathbf{y}_p], \mathbf{y}_j = \mathbf{X}\mathbf{a}_j$
- $\Rightarrow$ ogni riga di $\mathbf{U}$ rappresenta una unità in base ai punteggi normalizzati $\mathbf{u}_i = [y_{i1}/\sqrt{\lambda_1}, \dots, y_{ip}/\sqrt{\lambda_p}]$

SVD e approssimazione di $\mathbf{X}$

Se le prime $q < p$ CP tengono conto di una % elevata di varianza totale, allora la matrice dei dati può essere approssimata dai primi q autovalori λ_j e autovettori $\mathbf{a}_j, j = 1, \dots, q$:

$$\mathbf{X} \approx \underset{n \times q}{\mathbf{U}} \underset{q \times q}{\mathbf{\Gamma}} \underset{q \times p}{\mathbf{A}'}'$$

- per $q = 2$:

$$\mathbf{X} \approx \underset{n \times 2}{\mathbf{U}} \underset{2 \times 2}{\mathbf{\Gamma}} \underset{2 \times p}{\mathbf{A}'}'$$

- $\mathbf{U} = [\mathbf{u}'_1, \dots, \mathbf{u}'_n], \mathbf{u}'_i = [y_{i1}/\sqrt{\lambda_1}, y_{i2}/\sqrt{\lambda_2}], i = 1, \dots, n$
- $\mathbf{\Gamma} = \begin{pmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{pmatrix}$
- $\mathbf{A}' = \begin{pmatrix} a_{11} & \dots & a_{1p} \\ a_{21} & \dots & a_{2p} \end{pmatrix}$

Scomposizione **GH** di **X**

X matrice dei dati centrata, di dimensione $n \times p$.

SVD di **X**

$$\mathbf{X} = \mathbf{U}\mathbf{\Gamma}\mathbf{A}' = \underbrace{(\mathbf{U}\mathbf{\Gamma}^\alpha)}_{\mathbf{G}} \underbrace{(\mathbf{\Gamma}^{1-\alpha}\mathbf{A}')}_{\mathbf{H}} = \underset{n \times p}{\mathbf{G}} \underset{p \times p}{\mathbf{H}'}$$

- $\mathbf{\Gamma}^\alpha = \text{diag}\{\lambda_j^{\alpha/2}\}$
- $0 \leq \alpha \leq 1$

Approssimazione **GH** di **X**

Considerando le prime $q = 2$ colonne di **G** e le prime $q = 2$ righe di **H**

$$\mathbf{X} = \underset{n \times 2}{\mathbf{G}} \underset{2 \times p}{\mathbf{H}'}$$

- le righe di **G** sono i *punti unità*
 $\mathbf{g}'_i = [g_{i1}, g_{i2}], i = 1, \dots, n$
- le colonne di **H** sono i *punti variabile*
 $\mathbf{h}'_j = [h_{1j}, h_{2j}], j = 1, \dots, p$
- $\Rightarrow$ ogni elemento x_{ij} di **X** può essere scomposto nel prodotto tra l' i -esimo vettore riga di **G** e il j -mo vettore colonna $\mathbf{h}_j$ di **H**

$$x_{ij} = \mathbf{g}'_i \mathbf{h}_j$$

Costruzione biplot

- Punto di partenza per costruire il biplot è la matrice dei dati. Si considera
 - la matrice dei dati centrata $\mathbf{X}$ se le componenti sono state estratte dalla matrice di covarianza
 - la matrice $\mathbf{Z}$ dei dati standardizzati se le CP sono estratte dalla matrice di correlazione.
- Si procede quindi alla *scomposizione in valori singolari* della matrice, p.e. con riferimento a $\mathbf{X}$

$$\mathbf{X} \approx (\mathbf{U}\mathbf{\Gamma}^\alpha)(\mathbf{\Gamma}^{1-\alpha}\mathbf{A}') = \underset{n \times 2}{\mathbf{G}} \quad \underset{2 \times p}{\mathbf{H}'}$$

- questa scomposizione contiene le informazioni per costruire il **biplot**

Costruzione biplot (2)

$$\mathbf{X} \approx (\mathbf{U}\mathbf{\Gamma}^\alpha)(\mathbf{\Gamma}^{1-\alpha}\mathbf{A}') = \underset{n \times 2}{\mathbf{G}} \quad \underset{2 \times p}{\mathbf{H}'}$$

Nel biplot

- le n righe della matrice $\mathbf{G}$ $n \times 2$ sono rappresentate come *punti-unità*, corrispondenti alle *osservazioni*.
- le p righe della matrice $\mathbf{H}$ $p \times 2$ sono rappresentate come *vettori*, corrispondenti alle *variabili*.
- Il biplot consente una rappresentazione simultanea nel piano cartesiano dei *punti-unità* e dei *punti-variabile*

Costruzione biplot: scelta di α

$$\mathbf{X} \approx (\mathbf{U}\mathbf{\Gamma}^\alpha)(\mathbf{\Gamma}^{1-\alpha}\mathbf{A}') = \underset{n \times 2}{\mathbf{G}} \underset{2 \times p}{\mathbf{H}'}$$

- In generale, è impossibile rappresentare in maniera accurata sia le osservazioni che le variabili in due sole dimensioni.
- La scelta di α determina la scala utilizzata per rappresentare le osservazioni e l'interpretazione dei vettori nel biplot.
- L'interpretazione del biplot dipende dal valore di α
 - si possono scegliere i valori di α che consentono di conservare alcune proprietà dei dati multivariati.
 - Valori di α comunemente scelti: 0, 1/2, 1.

Biplot per $\alpha = 0$

$$\mathbf{X} = \underset{n \times p}{\mathbf{G}} \quad \underset{p \times p}{\mathbf{H}'}$$

- $\mathbf{G} = \mathbf{U}\mathbf{\Gamma}^\alpha = \mathbf{U} \Rightarrow$ punti-unità: **scores normalizzati** (valori dei punteggi divisi per $\sqrt{\lambda_j}$)
- $\mathbf{H} = \mathbf{\Gamma}^{1-\alpha}\mathbf{A}' = \mathbf{\Gamma}\mathbf{A}' \Rightarrow$ punti-variabile: **coefficienti di correlazione** tra variabili originali e le CP: $r(Y_j, X_k) = a_{kj}\sqrt{\lambda_j}$, $j = 1, 2$, $k = 1, \dots, p$

Informazioni fornite dal Biplot (per $\alpha = 0$)

- Lunghezza dei vettori: $r_{1k}^2 + r_{2k}^2$ quota di varianza della variabile X_k spiegata dalle prime 2 CP (comunalità)
- Angolo di ciascun vettore con gli assi: $r(Y_j, X_k) = a_{kj}\sqrt{\lambda_j}$ correlazione tra variabile e CP
- Angoli tra coppie di vettori: $r(X_k, X_l)$ correlazione tra due variabili

Biplot per $\alpha = 1$

$$\mathbf{X} = \underset{n \times 2}{\mathbf{G}} \quad \underset{2 \times p}{\mathbf{H}'}$$

- $\mathbf{G} = \mathbf{U}\mathbf{\Gamma}^\alpha = \mathbf{U}\mathbf{\Gamma} \Rightarrow$ i punti-unità sono gli scores non normalizzati (punteggi delle CP con varianza pari a λ_j)
- $\mathbf{H} = \mathbf{\Gamma}^{1-\alpha}\mathbf{A}' = \mathbf{A}' \Rightarrow$ i punti-variabile sono gli autovettori di $\mathbf{X}'\mathbf{X}$

Tipologie Biplot in SAS

Con la macro `biplot.sas` possiamo costruire 4 tipi diversi di biplot

- **GH'** Questa fattorizzazione corrisponde a $\alpha = 0$. Questo biplot cerca di conservare la relazione tra le variabili e ha due proprietà molto utili
 - la lunghezza di ogni vettore (righe di **H**) è proporzionale alla varianza della variabile corrispondente.
 - la distanza euclidea tra la i -ma e la l -ma riga di **G** (punto-unità) è proporzionale alla distanza di Mahalanobis tra la i -ma e la l -ma osservazione sui dati originali.
- **JK'** Questa fattorizzazione corrisponde a $\alpha = 1$. Questo biplot cerca di conservare la distanza tra le osservazioni, e ha le due proprietà seguenti
 - la posizione di punti-unità nel biplot è identica ai punteggi delle prime due CP ($\Rightarrow$ stessa nuvola di punti che si ottiene rappresentando $PRIN1*PRIN2$)
 - La distanza Euclidea tra la i -ma e la l -ma riga di **G** (punto-unità) è pari alla distanza euclidea tra la i -ma e la l -ma osservazione sui dati originali.

Tipologie Biplot in SAS (2)

- **COV** Questa fattorizzazione corrisponde a $\alpha = 0$, con **G** moltiplicata per $\sqrt{n-1}$ e **H** divisa per $\sqrt{n-1}$. Questo biplot ha due utili proprietà:
 - la lunghezza di un vettore (una riga di **H**) è pari alla varianza della variabile corrispondente.
 - La distanza Euclidea tra la i -ma e la l -ma riga di **G** (punto-unità) è pari alla distanza di Mahalanobis tra la i -ma e la l -ma osservazione sui dati originali.
- **SYM** Questa fattorizzazione corrisponde a $\alpha = 1/2$. Questo biplot tratta le osservazioni e le variabili in maniera simmetrica, cercando di conservare i valori delle osservazioni.

Costruzione Biplot in SAS

- Gli assi in basso e a sinistra del biplot definiscono le coordinate delle osservazioni.
- Gli assi in alto e a destra del biplot sono gli assi che definiscono le coordinate dei vettori.
- Se la matrice dei dati $\mathbf{X}$ non è ben approssimata da una matrice di rango 2, l'informazione contenuta nel biplot non è una buona rappresentazione dei dati
⇒ in tal caso, non cercare di interpretare il biplot!
- se $\mathbf{X}$ è ben approssimata dalle prime due CP, allora il biplot ha seguente interpretazione

Interpretazione Biplot in SAS

- il coseno dell'angolo tra un vettore e un asse principale è proporzionale al contributo della variabile corrispondente alla determinazione dell'asse.
- Il coseno dell'angolo tra due vettori indica la correlazione tra le due variabili corrispondenti. Variabili altamente correlate puntano nella stessa direzione, variabili poco correlate sono ad angolo retto.
- punti vicini nel biplot rappresentano osservazioni con valori simili.

Esempio Biplot in SAS

dati baseball (`baseball.sas`), biplot generato con macro `biplot.sas`
(`baseball_biplot.sas`)

- Le prime 2 CP raccolgono il 74% della varianza totale dei dati originali.
- $\Rightarrow$ il biplot rappresenta abbastanza bene (ma non bene) i dati.
- Abbiamo costruito il biplot
 - a partire dalla matrice dei dati standardizzati **Z**
 - di tipo **COV**, cioè con $\alpha = 0$, **G** moltiplicata per $\sqrt{n-1}$ e **H** divisa per $\sqrt{n-1}$.
- Questo tipo di biplot ha due utili proprietà:
 - la lunghezza di un vettore (una riga di **H**) è pari alla varianza della variabile corrispondente.
 - La distanza Euclidea tra la i -ma e la l -ma riga di **G** (punto-unità) è pari alla distanza di Mahalanobis tra la i -ma e la l -ma osservazione sui dati originali.