# Aspetti matematici del biplot

Marco Riani
Dipart. di Scienze Economiche e Aziendali, Università di Parma

24 ottobre 2016

Il biplot è una rappresentazione simultanea delle $n$ unità (righe delle matrice dei dati) e delle $p$ variabili (colonne della matrice dei dati) in uno spazio a due dimensioni.

Osservazione: in questa nota faremo sempre riferimento alla matrice $Z$ (di dimensione $n \times p$) degli scostamenti standardizzati. Ovviamente è possibile partire anche dalla matrice $\tilde{X}$ degli scostamenti dalla media. In tal caso i richiami alla matrice di correlazione $R = Z'Z/(n-1)$ devono intendersi riferiti alla matrice di covarianze $S = \tilde{X}'\tilde{X}/(n-1)$.

Si parte dalla scomposizione in valori singolari:

$$Z = U\Gamma V^T$$

Se le prime due componenti principali tengono conto di una quota elevata della varianza totale è possibile sostituire alla matrice $Z$ la sua miglior rappresentazione di rango 2 come segue:

$$Z \approx U_{(2)}\Gamma^*_{(2)}V^T_{(2)}$$

dove $U_{(2)}$ è una matrice di dimensione $n \times 2$ che contiene le prime due colonne della matrice $U$, $V_{(2)}$ è una matrice di dimensione $p \times 2$ che contiene le prime due colonne della matrice $V$. $\Gamma^*_{(2)}$ è la matrice diagonale di dimensione $2 \times 2$ che contiene sulla diagonale principale i primi due valori singolari della matrice $Z$ (ossia la radice quadrata degli autovalori della matrice $Z^TZ = (n-1)R$). Per fare in modo che questa matrice contenga gli autovalori di $Z^TZ/(n-1) = R$, possiamo scrivere la scomposizione in valori singolari come segue:

$$\frac{Z}{\sqrt{n-1}} \approx U_{(2)}\frac{\Gamma^*_{(2)}}{\sqrt{n-1}}V^T_{(2)}$$

$$Z \approx \sqrt{n-1} U_{(2)} \frac{\Gamma^*_{(2)}}{\sqrt{n-1}} V^T_{(2)}$$

$$Z \approx \sqrt{n-1} U_{(2)} \Gamma_{(2)} V^T_{(2)}$$

dove $\Gamma_{(2)} = \Gamma^*_{(2)}/\sqrt{n-1}$.

La matrice $Z$ degli scostamenti standardizzati risulta, quindi, approssimativamente uguale al prodotto di due matrici, la prima ($G$) riferita alle $n$ unità (di dimensione $n \times 2$) e la seconda ($H$) alle $p$ variabili (di dimensione $p \times 2$) come segue:

$$Z_{n\times p} \approx G_{n\times 2} H^T_{2\times p} = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \\ \dots & \dots \\ g_{n1} & g_{n2} \end{pmatrix} \begin{pmatrix} h_{11} & h_{21} & \dots & h_{p1} \\ h_{12} & h_{22} & \dots & h_{p2} \end{pmatrix}$$

A seconda di come vengono specificate $G$ e $H$, possiamo ottenere diverse rappresentazioni per i punti riga ed i punti colonna.

*Rappresentazione 1 dei punti riga e dei punti colonna*

Le $n$ righe della matrice dei dati (di dimensione $p$) possono essere rappresentate tramite la matrice $n \times 2$:

$$G = \sqrt{n-1} U_{(2)} = Z V_{(2)} \Gamma^{-1}_{(2)}$$

Questa matrice contiene gli scores normalizzati delle prime due componenti principali. In altri termini la matrice di varianze e covarianze degli scores normalizzati è la matrice identità di dimensione $2 \times 2$:

$$var(\sqrt{n-1} U_{(2)}) = I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Le $p$ colonne della matrice dei dati possono essere rappresentate tramite $p$ frecce bidimensionali che partono dall'origine degli assi e che presentano coordinate uguali a quelle delle $p$ colonne della matrice che segue:

$$H^T = \Gamma_{(2)} V^T_{(2)} = \begin{pmatrix} \gamma_1 v_{11} & \gamma_1 v_{21} & \dots & \gamma_1 v_{p1} \\ \gamma_2 v_{12} & \gamma_2 v_{22} & \dots & \gamma_2 v_{p2} \end{pmatrix} = \begin{pmatrix} \sqrt{\lambda_1} v_{11} & \sqrt{\lambda_1} v_{21} & \dots & \sqrt{\lambda_1} v_{p1} \\ \sqrt{\lambda_2} v_{12} & \sqrt{\lambda_2} v_{22} & \dots & \sqrt{\lambda_2} v_{p2} \end{pmatrix}$$

Le coordinate della $j$-esima freccia $(\sqrt{\lambda_1}v_{j1}, \sqrt{\lambda_1}v_{j2})$, $(j = 1, 2, \ldots, p)$, non sono altro che i coefficienti di correlazione tra la $j$-esima variabile e le prime due componenti principali. La lunghezza della freccia in questo caso è esattamente uguale alla radice quadrata della comunalità (ossia alla radice quadrata della quota di varianza della $j$-esima variabile spiegata dalle prime due componenti principali).

*Rappresentazione 2 dei punti riga e dei punti colonna*

Le $n$ righe della matrice dei dati (di dimensione $p$) possono essere rappresentate tramite la matrice $n \times 2$:

$$G = \sqrt{n-1}U_{(2)}\Gamma_{(2)} = ZV_{(2)}$$

In questo caso i punti riga sono gli scores non normalizzati cioè:

$$var(\sqrt{n-1}U_{(2)}\Gamma_{(2)}) = var(ZV_{(2)}) = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

In questa rappresentazione i punti colonna (le frecce) non sono altro che le coordinate dei primi due autovettori:

$$H^T = V_{(2)}^T = \begin{pmatrix} v_{11} & v_{21} & \ldots & v_{p1} \\ v_{12} & v_{22} & \ldots & v_{p2} \end{pmatrix}$$

In questa rappresentazione, quindi, la lunghezza delle frecce non è uguale alla radice quadrata della comunilità ma è solo "funzione" della comunalità.

Ci chiediamo: è possibile avere una rappresentazione dinamica dei punti riga e dei punti colonna (frecce) in modo tale che le due reppresentazioni precedenti siano sono dei casi particolari? A questo scopo introduciamo due nuovi parametri $\omega$, e $\alpha$ definiti nell'intervallo $[0, 1]$, e scriviamo la scomposizione in valori singolari come segue:

$$Z \approx \left[(\sqrt{n-1})^{\omega}U_{(2)}\Gamma_{(2)}^{\alpha}\right]\left[\Gamma_{(2)}^{1-\alpha}V_{(2)}^T(\sqrt{n-1})^{1-\omega}\right]$$

In questo caso gli $n$ punti riga possono essere rappresentati tramite la matrice $n \times 2$

$$\left[(\sqrt{n-1})^{\omega}U_{(2)}\Gamma_{(2)}^{\alpha}\right]$$

e i $p$ punti colonna tramite la matrice $2 \times p$

$$\left[\Gamma_{(2)}^{1-\alpha}V_{(2)}^T(\sqrt{n-1})^{1-\omega}\right]$$

La rappresentazione 1 vista in precedenza si ottiene quando $\omega = 1$ e $\alpha = 0$. Al contrario, la rappresentazione 2 si ottiene quando $\omega = 1$ e $\alpha = 1$. Ovviamente, tanto più $\omega$ diminuisce, tanto più la lunghezza delle frecce aumenta e le coordinate dei punti riga si comprimono.