BRIEF ARTICLE

THE AUTHOR

1. Italian stuff used as draft

Si tratta di un metodo di analisi che è stato introdotto nel 1933 da Harold Hotelling [**?**] per analizzare dati di tipo psicometrico. Nella letteratura scientifica in lingua inglese tale metodo prende il nome di *Principal Component Analysis*, il cui acronimo PCA è molto popolare. Il modo più compatto di introdurre la PCA è attraverso una serie di considerazioni di tipo algebrico svolte sulla matrice di varianza-covarianza. Per prima cosa richiamiamo la definizione della matrice di varianza-covarianza e cerchiamo di riscriverla in modo simmetrico:

Tale identità algebrica mette in luce la prima proprietà della matrice di varianza-covarianza, ovvero il fatto che si tratta di una matrice simmetrica – un fatto di cui ci siamo già accorti tramite la definizione dei suoi elementi di matrice, si veda l'Eq. (**??**). Infatti, se introduciamo la matrice $\boldsymbol{A} = \boldsymbol{\chi}/(N_o - 1)$, avremo per la matrice $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \boldsymbol{A}^t \, \boldsymbol{A} \tag{1}$$

Valutando la trasposta di $\boldsymbol{\Sigma}$ otteniamo:

$$\boldsymbol{\Sigma}^t = (\boldsymbol{A}^t \, \boldsymbol{A})^t = \boldsymbol{A}^t \, (\boldsymbol{A}^t)^t = \boldsymbol{A}^t \, \boldsymbol{A} = \boldsymbol{\Sigma}, \tag{2}$$

che dimostra ancora una volta la simmetria della matrice di varianza-covarianza. Un'altra proprietà significativa della matrice di varianza-covarianza è che si tratta di una matrice definita positiva[1], che quindi possiede autovalori ≥ 0 [**?**].

Vediamo ora in che modo si possa introdurre una trasformazione lineare che permetta di passare dagli indici di variabile v a nuovi indici s, che chiameremo indici delle componenti principali. L'idea è quella di sostituire al set di variabili con cui abbiamo operato per raccogliere le osservazioni un nuovo set di variabili "migliori", per due motivi: (1) la loro mutua indipendenza (ortogonalità) e (2) il loro numero, potenzialmente più ridotto del numero di variabili originarie. Tale set di variabili "migliori" è il set delle cosiddette componenti principali, da associare agli indici s. Detto in altri termini, *tramite le componenti principali diventa possibile descrivere le osservazioni sperimentali utilizzando un set ridotto di grandezze tra di loro indipendenti*, con vantaggi evidenti.

[1] Questo risultato è semplice da dimostrare utilizzando la forma $\boldsymbol{\Sigma} = \boldsymbol{A}^t \boldsymbol{A}$. Una matrice semi definita positiva $\boldsymbol{M}$ è tale per cui dato un vettore generico $\boldsymbol{x}$ risulta $\boldsymbol{x}^t \boldsymbol{M} \boldsymbol{x} \geq 0$. Nel nostro caso, sfruttando la definizione $\boldsymbol{\Sigma} = \boldsymbol{A}^t \boldsymbol{A}$ abbiamo $\boldsymbol{x}^t \boldsymbol{\Sigma} \boldsymbol{x} = \boldsymbol{x}^t \boldsymbol{A}^t \boldsymbol{A} \boldsymbol{x} = (\boldsymbol{A}\boldsymbol{x})^t (\boldsymbol{A}\boldsymbol{x}) = |\boldsymbol{A}\boldsymbol{x}|^2$. L'ultimo passaggio dimostra che otteniamo il modulo al quadrato di un generico vettore $\boldsymbol{A}\boldsymbol{x}$; tale modulo al quadrato è certamente una grandezza ≥ 0.

Dal punto di vista tecnico, per implementare quest'idea basta utilizzare la decomposizione spettrale della matrice di covarianza, che è agevole perchè la matrice $\mathbf{\Sigma}$ è simmetrica. A quel punto si cercherà di riportare la decomposizione spettrale in una forma equivalente a quella dell'Eq. 1 in cui tuttavia il primo membro sarà in forma diagonale. Il riconoscimento di una struttura del tipo $\boldsymbol{A}^t\,\boldsymbol{A}$ nel secondo membro fornirà la definizione del dataset nelle nuove variabili semplificatrici, ovvero nelle componenti principali s. È chiaro che, il requisito di *indipendenza* delle componenti principali risulterà soddisfatto se assoceremo a ciascuna di questa uno degli autovettori della matrice di covarianza. Infatti autovettori diversi saranno ortogonali tra di loro (*vide infra*) e forniranno i coefficienti della trasformazione lineare che permette di passare dalle variabili di indice v ad una certa componente principale di indice s.

Esaminiamo quindi i dettagli algebrici in modo dettagliato. Cominciamo con l'equazione agli autovalori per la matrice di covarianza $\mathbf{\Sigma}$, in cui associamo ogni coppia autovalore/autovettore ad un indice di componente principale s:

$$\mathbf{\Sigma}_{vv}\boldsymbol{L}_{vs} = \boldsymbol{L}_{vs}\boldsymbol{\sigma}_{ss} \tag{3}$$

Nell'equazione (3) $\boldsymbol{\sigma}_{ss}$ è la matrice diagonale degli autovalori di $\mathbf{\Sigma}_{vv}$. $\boldsymbol{L}_{vs}$ è la matrice degli autovettori di $\mathbf{\Sigma}_{vv}$. Dato che $\mathbf{\Sigma}_{vv}$ è simmetrica, $\boldsymbol{L}_{vs}$ è una matrice ortogonale, per la quale valgono le relazioni:

$$\begin{aligned}\boldsymbol{L}_{vs}\boldsymbol{L}_{sv} &= \mathbf{1}_{vv}\\ \boldsymbol{L}_{sv}\boldsymbol{L}_{vs} &= \mathbf{1}_{ss}.\end{aligned} \tag{4}$$

Moltiplicando da sinistra l'Eq. (3) per la matrice $\boldsymbol{L}_{sv}$, e sfruttando le sue proprietà di ortogonalità, si ricava la *decomposizione spettrale* della matrice di varianza-covarianza:

$$\boldsymbol{L}_{sv}\mathbf{\Sigma}_{vv}\boldsymbol{L}_{vs} = \boldsymbol{\sigma}_{ss} \tag{5}$$

Introduciamo ora nel secondo membro dell'Eq. (5) la definizione di $\mathbf{\Sigma}_{vv} = \boldsymbol{\chi}_{vo}\boldsymbol{\chi}_{ov}/(N_o-1)$ (cfr. Eq. **??**):

$$\boldsymbol{\sigma}_{ss} = \frac{1}{N_o-1}\,\boldsymbol{L}_{sv}\boldsymbol{\chi}_{vo}\,\boldsymbol{\chi}_{ov}\boldsymbol{L}_{vs} \tag{6}$$

Con un piccolo sforzo possiamo riconoscere al secondo membro una struttura data dal prodotto di una nuova matrice $\boldsymbol{S}$ e della sua trasposta:

$$\boldsymbol{\sigma}_{ss} = \left[\frac{1}{\sqrt{N_o-1}}\boldsymbol{L}_{sv}\boldsymbol{\chi}_{vo}\right]\left[\boldsymbol{\chi}_{ov}\boldsymbol{L}_{vs}\frac{1}{\sqrt{N_o-1}}\right] = \boldsymbol{S}_{so}\boldsymbol{S}_{os} = \boldsymbol{S}^t\boldsymbol{S}. \tag{7}$$

Le righe di tale matrice, detta matrice degli *scores* ($\boldsymbol{S} = \boldsymbol{S}_{os}$), definiscono le osservazioni o in funzione delle componenti principali s:

$$\boldsymbol{S}_{os} \equiv \left[\frac{1}{\sqrt{N_o-1}}\boldsymbol{\chi}_{ov}\boldsymbol{L}_{vs}\right] \tag{8}$$

La diagonalizzazione della matrice di covarianza fornisce l'insieme degli autovalori delle componenti principali, rappresentati in quello che viene solitamente chiamato *screeplot*. Tale grafico fornisce l'andamento decrescente delle varianze principali (σ_s) in funzione

dell'indice s della componente principale. Da questo grafico si può rapidamente giudicare l'importanza relativa delle componenti principali nel descrivere la varianza totale dei dati del dataset.

La matrice $\boldsymbol{L}_{vs}$ degli autovettori della matrice di covarianza $\boldsymbol{\Sigma}_{vv}$ è detta matrice dei *loadings*. Essa rappresenta il legame tra le variabili (v) e le componenti principali (s). Nel caso di dataset formati da spettri, la rappresentazione dei *loadings* di una componente principale in funzione delle variabili (e.g., lunghezze d'onda), è affine a quella di uno spettro. I picchi (positivi o negativi) nella rappresentazione di uno dei *loadings* mostrano in quali regioni spettrali sono osservate le maggiori variazioni (crescita/decrescita) del segnale all'interno del dataset.

Per quanto riguarda invece la matrice $\boldsymbol{S}$ degli *scores*, il grafico cartesiano delle prime due colonne della matrice $\boldsymbol{S}_{os}$ (pensate come coordinate x, y) fornisce la posizione delle osservazioni del dataset rispetto al sistema di riferimento ortogonale fornito dalle componenti principali s_1, s_2. Tale grafico è talvolta definito *scatterplot* (grafico di dispersione dei dati). È utile osservare che talvolta si utilizza una matrice di *scores* $\boldsymbol{S}'_{os}$ normalizzata rispetto alle deviazioni standard principali:

$$\boldsymbol{S}'_{os} \equiv \boldsymbol{S}_{os}\boldsymbol{\sigma}_{ss}^{-1/2} \tag{9}$$

In questo modo nella presentazione dello *scatterplot* ci si può ricondurre ad un sistema di riferimento adimensionale (standardizzato). Infatti, le varianze calcolate rispetto alla matrice $\boldsymbol{S}'_{os}$ risultano date da un prodotto uguale alla matrice identità:

$$(\boldsymbol{S}')^t\boldsymbol{S}' = \left(\boldsymbol{\sigma}_{ss}^{-1/2}\boldsymbol{S}_{so}\right)\left(\boldsymbol{S}_{os}\boldsymbol{\sigma}_{ss}^{-1/2}\right) = \boldsymbol{\sigma}_{ss}^{-1/2}\boldsymbol{\sigma}_{ss}\boldsymbol{\sigma}_{ss}^{-1/2} = \mathbf{1}. \tag{10}$$