Alma Mater Studiorum · Università di Bologna

FACOLTA' DI SCIENZE MATEMATICHE, FISICHE E NATURALI
Corso di laurea in Matematica

ANALISI DELLE COMPONENTI PRINCIPALI: ALGORITMI E APPLICAZIONI

Tesi di Laurea in Calcolo Numerico

Relatore:
Chiar.ma Prof.
Valeria Simoncini

Presentato da:
Bruno Farabegoli

♯ **3 Sessione**
Anno Accademico: 2015-2016

Indice

Introduzione

L'analisi delle componenti principali (abbreviato in PCA) è una tecnica utilizzata nel campo della statistica multivariata. Consiste nell'esprimere la struttura di varianza-covarianza di un insieme di variabili attraverso alcune combinazioni lineari di queste ultime.

I suoi obbiettivi sono principalmente 1) la riduzione dei dati originari, e quindi la semplificazione computazionale; 2) la reinterpretazione di tali osservazioni.

Se si considera un insieme di n dati su p variabili, è naturale pensare che siano richieste esattamente p componenti principali per riprodurre la variabilità totale del sistema. In realtà, come verrà mostrato in seguito, gran parte di questa variabilità può essere quantificata da un numero più piccolo di queste cosiddette componenti principali arrivando quindi alla conclusione che k componenti (con $k < p$) possono sostituire le iniziali p variabili. Si ottiene quindi che il set originario dei dati, che consisteva in n misurazioni su p caratteristiche, venga trasformato in un set che consiste in n misurazioni su k componenti principali.

Non solo. La PCA, una volta applicata, spesso rivela relazioni tra osservazioni e nuove variabili che non erano originariamente sospettate e questo comporta una reinterpretazione più profonda dei dati.

Come tutte le applicazioni nell'analisi dei dati, anche la PCA ha un costo: la perdita di alcune informazioni iniziali. D'altra parte è una tecnica largamente utilizzata nei più svariati settori proprio perché confina tale perdita entro limiti accettabili. In altre parole il *trade off* tra la perdita delle informazioni e la semplificazione del problema è quasi sempre a vantaggio di chi decide di utilizzare la PCA, a patto che il numero delle componenti sia scelto in modo "giudizioso".

Capitolo 1

Algebra delle componenti principali

1.1 Richiami di Algebra Lineare e Statistica

Lo studio della PCA richiede la conoscenza degli elementi fondamentali della statistica e dell'algebra lineare. A tale proposito, si vogliono riproporre brevemente le definizioni principali degli strumenti usati in tali campi della matematica. L'oggetto algebrico fondamentale nel lavoro che seguirà, è l'autovalore.

Definizione. *Data una matrice* $A(\in \mathbb{C}^{n*n})$ *si definisce autovalore di A il numero* λ*, reale o complesso, tale per cui esista un vettore v che soddisfi l'equazione*

$$Av = \lambda v$$

In tal caso v verrà chiamato autovettore di A relativo a λ

Per il teorema fondamentale sui sistemi lineari sappiamo che tale equazione ha soluzioni se e solo se

$$\det(A - \lambda I) = P_n(\lambda) = 0$$

dove $P_n(\lambda)$ è detto polinomio caratteristico di A.

Definizione. *La coppia* (λ, v) *con* λ *autovalore e v autovettore a esso associato, è detta autocoppia.*

Inoltre è fatto largo uso del concetto di varianza, ovvero un un indice di variabilità che fornisce una stima su quanto i dati di una popolazione (o di un campione) si discostano dalla sua media. Diamo qui sotto la definizione.

Definizione. *Data una distribuzione di variabile quantitativa* $X = [X_1, ..., X_n]$ *su*

una popolazione di n elementi, la varianza è definita come

$$\sigma_X^2 = \frac{\sum_{i=1}^{n}(X_i - \mu_X)^2}{n}$$

dove $\mu_X = \frac{\sum_{i=1}^{n} X_i}{n}$ *è la media di X.*

La covarianza invece è un indice di variabilità congiunta di due variabili statistiche X, Y

Definizione. *Date due variabili aleatorie* $X = [x_1, ..., x_n]$ *e* $Y = [y_1, ..., y_n]$ *di media rispettivamente* μ_X *e* μ_Y *la covarianza risulta*

$$Cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y)$$

Nel caso in cui si abbia che $Cov(X, Y) = 0$ si che che le variabili X e Y sono non correlate. Da qui andiamo subito a definire il coefficiente di correlazione.

Definizione. *Si definisce coefficiente di correlazione tra due variabili X e Y il numero*

$$\rho_{(X,Y)} = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$$

Ai fini del nostro studio, sono indispensabili anche altri due strumenti, ovvero la matrice di covarianza Σ e la matrice di correlazione ρ.

Definizione. *La matrice di covarianza* Σ *(rispettivamente quella di correlazione* ρ*) di un vettore X*= $[X_1, ..., X_p]$ *è la matrice* $p \times p$ *definita da* $\Sigma_{ij} = Cov(X_i, X_j)$ *(rispettivamente* $\rho_{ij} = \rho(X_i, X_j)$

Osservazione. Sia Σ che ρ sono matrici simmetriche semidefinite positive. Nel caso siano anche non singolari, ovvero abbiano rango massimo, allora sono definite positive.

1.2 Componenti principali delle popolazioni

Da un punto di vista algebrico, date p variabili aleatorie $X_1, ..., X_p$, le componenti principali sono particolari combinazioni lineari di tali variabili con fondamentali proprietà che verranno analizzate di seguito. Tali combinazioni lineari devono comunque descrivere una trasformazione lineare delle variabili che proietta quelle originarie in un nuovo sistema cartesiano nel quale le variabili vengono ordinate in ordine decrescente di varianza. La riduzione della complessità del sistema avviene limitandosi ad analizzare le

principali, per varianza, delle nuove variabili. Le componenti principali dipendono unicamente dalla matrice di covarianza, che indicheremo con Σ o dalla matrice di correlazione, che chiameremo ρ. Notiamo anche che il loro sviluppo non richiede obbligatoriamente l'assunzione di una distribuzione multinormale.

Sia $X^T = [X_1, ..., X_p]$ un vettore aleatorio e sia Σ la matrice di covarianza con autovalori $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$, consideriamo la seguente combinazione lineare

$$\begin{aligned}
&Y_1 = a_1^T X = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p \\
&Y_2 = a_2^T X = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p \\
&\vdots \\
&\vdots \\
&Y_p = a_p^T X = a_{p1}X_1 + a_p 2X_2 + ... + a_{pp}X_p
\end{aligned} \tag{1.1}$$

Da cui segue

$$Var(Y_i) = a_i^T \Sigma a_i \text{ con } i = 1, 2....., p \tag{1.2}$$
$$Cov(Y_i, Y_K) = a_i^T \Sigma a_k \text{ con } i, k = 1, 2....., p \tag{1.3}$$

Le componenti principali sono quelle combinazione lineari $Y_1, ..., Y_k$ *non* correlate le cui varianze abbiano i valori più grandi possibili.
La prima componente sarà la combinazione lineare con massima varianza. Ora, sapendo che $Var(Y_1) = a_1^T \Sigma a_1$, si può pensare che quindi essa possa essere incrementata a piacere moltiplicando a_1 con qualche costante rendendo così impossibile la massimizzazione. Per ovviare a questo problema imponiamo ai coefficienti a_i che abbiano norma uno. perciò definiamo:

> Prima componente principale = la combinazione lineare $a_1^T X$ che massimizza $Var(a_1^T X)$ e per cui valga: $a_1^T a_1 = 1$
>
> Seconda componente principale = la combinazione lineare $a_1^T X$ che massimizza $Var(a_2^T X)$ e per cui valga: $a_2^T a_2 = 1$ e $Cov(a_1^T X, a_2^T X) = 0$

All'i-esimo passo avremo perciò

> I-esima componente principale = la combinazione lineare $a_i^T X$ che massimizza $Var(a_i^T X)$ e per cui valga: $a_i^T a_i = 1$ e $Cov(a_i^T X, a_k^T X) = 0$ per $k < i$

Osservazione. Per una generica nuova variabile y_i si può riformulare il problema come la ricerca del massimo del quoziente di Rayleigh

$$\max_{a \neq 0} \frac{a^T \Sigma a}{a^T a}$$

1.3 Risultati fondamentali

Lemma 1. *Data la matrice di covarianza Σ associata al vettore aleatorio $X = [X_1, ...X_p]$ e date le autocoppie $(\lambda_1, e_1)....(\lambda_p, e_p)$ con $\lambda_1 \geq \lambda_2 \geq ..... \geq \lambda_p$ allora la i-esima componente principale è data dalla seguente formula*

$$Y_i = e_i^T X = e_{i1}X_1 + ... + e_{ip}X_p \qquad con\ i = 1, 2, ..., p$$

Inoltre

$$Var(Y_i) = e_i^T \Sigma e_i = \lambda_i \qquad con\ i = 1, 2, ..., p \tag{1.5}$$

E

$$Cov(Y_i, Y_k) = e_i^T \Sigma e_k = 0 \quad per\ \ i \neq k \tag{1.6}$$

Se alcuni λ_i sono uguali le scelte dei corrispondenti autovettori e_i (e perciò delle componenti principali) è molteplice.

Dimostrazione. Per il Lemma di Massimizzazione sappiamo che

$$\max_{a \neq 0} \frac{a^T \Sigma a}{a^T a} = \lambda_1$$

con $a = e_1$ e che gli autovettori sono unitari

$$\max_{a \neq 0} \frac{a^T \Sigma a}{a^T a} = \lambda_1 = \frac{e_1^T \Sigma e_1}{e_1^T e_1} = e_1^T \Sigma e_1 = Var(Y_1)$$

in modo simile otteniamo anche

$$\max_{a \perp e_1 ... e_k} \frac{a^T \Sigma a}{a^T a} = \lambda_{k+1}$$

con $k = 1, 2, ..., p-1$
Scegliendo $a = e_{k+1}$ con $e_{k+1}^T e_i = 0 \qquad per\ i = 1, 2, ..., k \quad$ e $\ k = 1, 2, ..., p-1$

$$\frac{e_{k+1}^T \Sigma e_{k+1}}{e_{k+1}^T e_{k+1}} = e_{k+1}^T \Sigma e_{k+1} = Var(Y_{k+1})$$

Ma $e_{k+1}(\Sigma e_{k+1}) = \lambda_{k+1} e_{k+1}^T e_{k+1} = \lambda_{k+1}$ e quindi $Var(Y_{k+1}) = \lambda_{k+1}$
Rimane da dimostrare che gli autovettori e_i perpendicolari a e_k(ovvero tali per cui $e_i^T e_k = 0, i \neq k$) danno la matrice di covarianza nulla, cioè $Cov(Y_i Y_k) = 0$
Ora, gli autovettori di Σ sono ortogonali se tutti gli autovalori di siffatta matrice sono distinti. Se gli autovalori non lo sono, gli autovettori corrispondenti agli autovalori in comune possono essere scelti ortogonali. Perciò, per ogni due autovettori e_i, e_k, si ha

$e_i^T e_k = 0, i \neq k$. Poiché $\Sigma e_k = \lambda e_k$ moltiplicando per e_i otteniamo

$$Cov(Y_i, Y_k) = e_i^T \Sigma e_k = e_i^T \lambda e_k = \lambda e_i^T e_k = 0$$

per ogni $i \neq k$, e con questo la dimostrazione è completata.

□

Lemma 2. *Dato un vettore aleatorio $X = [X_1, ...X_p]$, sia Σ la matrice di covarianza associata con le autocoppie $(\lambda_1, e_1), ...(\lambda_p, e_p)$ tali che $\lambda_1 \geq \lambda_2 \geq ..... \geq \lambda_p \geq 0$. Siano $Y_1 = e_1^T X, ..., Y_p = e_p^T X$ le componenti principali. Allora*

$$\sigma_{11} + ... + \sigma_{pp} = \sum_{i=1}^{p} Var(X_i) = \lambda_1 + ... + \lambda_p = \sum_{i=1} Var(Y_i)$$

Dimostrazione. E' noto che $\sigma_{11} + ... + \sigma_{pp} = tr(\Sigma)$ e inoltre che è possibile scrivere $\Sigma = \Gamma\Lambda\Gamma^T$ dove Λ è la matrice diagonale degli autovalori e $\Gamma = [e_1, ..., e_p]$ e quindi $\Gamma\Gamma^T = \Gamma^T\Gamma = I$. Si ottiene allora

$$tr(\Sigma) = tr(\Gamma\Lambda\Gamma^T) = tr(\Lambda\Gamma^T\Gamma) = tr(\Lambda) = \lambda_1 + ... + \lambda_p$$

dunque

$$\sum_{i=1}^{p} Var(X_i) = tr(\Sigma) = tr(\Lambda) = \sum_{i=1}^{p} Var(Y_i)$$

e con ciò la dimostrazione è conclusa.

□

Osservazione. Il precedente risultato ci dice che la varianza totale della popolazione, dove per varianza totale si intende la somma di tutte le varianze relative alle variabili originarie, equivale alla seguente somma

$$\sigma_{11} + ... + \sigma_{pp} = \lambda_1 + ... + \lambda_p \qquad (1.7)$$

di conseguenza il rapporta tra la varianza totale e quella relativa alla k-esima componente principale è

$$\frac{\lambda_k}{\lambda_1 + ... + \lambda_p} \qquad (1.8)$$

con $k = 1, 2, ..., p$. Questa proporzione è essenziale per riuscire a ridurre le variabili del sistema. Se infatti, per esempio, immaginiamo un sistema con p variabili dove p è preso abbastanza grande, e dove la prima, la seconda e la terza componente principale esprimano solo loro l'80% o il 90% della varianza totale della popolazione, allora queste tre componenti possono *sostituire* le originali p variabili senza una grossa perdita delle informazioni che fornivano i dati iniziali. Risulta quindi evidente la forte semplificazione

apportata al sistema al fronte di una minima dispersione dei dati. Non solo. Merita attenzione anche ogni componente del vettore dei coefficienti $e_i^T = [e_{i1}, ...e_{ik}, ...e_{ip}]$ Infatti la grandezza di e_{ik} misura l'importanza, o per meglio dire il peso, che la k-esima variabile apporta alla i-esima componente principale, a prescindere dall'influenza delle altre variabili. In particolare ogni e_{ik} è proporzionale al coefficiente della correlazione tra Y_i e X_k.

Lemma 3. *Siano* $Y_1 = e_1^T X, ...Y_p = e_p^T X$ *le componenti principali ottenute dalla matrice di covarianza* Σ*. Allora*

$$\rho_{Y_i,X_k} = \frac{e_{i,k}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad (1.9)$$

sono i coefficienti delle correlazioni tra le componenti Y_i *e le variabili* X_k*, dove* $(\lambda_1, e_1), ..., (\lambda_p, e_p)$ *sono le autocoppie di* Σ*.*

Dimostrazione. Sia $\mathbf{a}_k' = [0, ..., 0, 1, 0, ...0]$ tale che $X_k = \mathbf{a}_k^T X$ e $Cov(X_k, Y_k) = Cov(\mathbf{a}_k^T, \mathbf{e}_j^T X) = \mathbf{a}_k^T \Sigma \mathbf{e}_j$. poiché $\Sigma \mathbf{e}_i = \lambda_i \mathbf{e}_i, Cov(X_k, Y_i) = \mathbf{a}_k^T \lambda_i \mathbf{e}_i = \lambda_i e_{ik}$. Allora $Var(Y_i) = \lambda_i$ (come vuole la proposizione 1)e $Var(X_k) = \sigma_{kk}$ E perciò

$$\rho_{(Y_i,X_k)} = \frac{Cov(Y_i, X_k)}{\sqrt{Var(Y_i)}\sqrt{Var(X_k)}} = \frac{\lambda_i e_{ik}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$$

per $i, k = 1, ..., p$

□

Osservazione. Sebbene le correlazioni tra le variabili e le componenti principali aiutino spesso a interpretare le componenti, esse misurano però solo il contributo di una singola X su una componente Y. Ad esempio nulla dicono sul peso di una variabile X rispetto a una componente Y in relazione a un'altra variabile X^T. Proprio per questo molti statisti raccomandano che solo i coefficienti e_{ik} debbano essere usati per interpretare le componenti senza contare le correlazioni. Detto ciò, è esperienza molto frequente quella di osservare che variabili con coefficienti relativamente grandi in valore assoluto, tendono ad avere anche ampie correlazioni, e che quindi entrambe le misure sui pesi, una multivariata mentre l'altra univariata danno spesso risultati simili. Si può quindi concludere che tenere in considerazione sia i coefficienti che le correlazioni renda lo studio sulle componenti principali più preciso.

1.4 Componenti principali per variabili standardizzate

Accade spesso che i dati originari a disposizione siano caratterizzati da unità di misura diverse tra loro e risultino quindi non confrontabili. In tali situazioni si procede alla standardizzazione delle variabili aleatorie.

Sia $X = [X_1, ..., X_p]$ il vettore aleatorio originario. Standardizziamo le sue variabili ottenendo:

$$Z_1 = \frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}} \cdot \cdot \cdot \cdot \cdot \cdot Z_p = \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}} \qquad (1.10)$$

Dove $\mu_1, ..., \mu_p$ sono le componenti del vettore della media di X.
In notazione matriciale:

$$Z = (V^{\frac{1}{2}})^{-1}(X - \mu) \qquad (1.11)$$

dove $V^{\frac{1}{2}}$ è la matrice diagonale di deviazione standard. Risulta evidente che $E(Z) = 0$ e che

$$Cov(Z) = (V^{\frac{1}{2}})^{-1}\Sigma(V^{\frac{1}{2}})^{-1} = \rho$$

con ρ matrice di correlazione di X.

Le componenti principali di Z possono essere ottenuti proprio dalla matrice di correlazione ρ. Tutti i precedenti risultati rimangono validi anche in questo caso con qualche ulteriore semplificazione dovuta al fatto che la varianza di ogni Z_i è uguale a 1. Continuiamo perciò a usare la notazione Y_i per riferirci alla i-esima componente principale e $(\lambda_i, \mathbf{e}_i)$ per riferirci alle autocoppie di ρ o Σ tenendo però presente che in generale le autocoppie ottenute da una non sono le stesse ottenute dall'altra.

Lemma 4. *Siano* $(\lambda_1, e_1), ...(\lambda_p, e_p)$ *sono le autocoppie della matrice* ρ *con* $\lambda_1 \geq \lambda_2 \geq ..... \geq \lambda_p \geq 0$. *La i-esima componente principale di variabili standardizzate* $Z' = [Z_1, ...Z_p]$ *con* $Cov(Z) = \rho$ *è data dalla seguente formula*

$$Y_i = \boldsymbol{e}_i^T Z = \boldsymbol{e}_i^T (V^{\frac{1}{2}})^{-1}(X - \mu) \qquad con\ i = 1, 2, ...p$$

Inoltre

$$\sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} Var(Z_i) = p \qquad (1.12)$$

E

$$\rho_{(Y_i, Z_k)} = e_{ik}\sqrt{\lambda_i} \qquad con\ i, k = 1, 2, ..., p$$

Dimostrazione. Discende immediatamente dalle precedenti dimostrazioni sostituendo $X_1, ..., X_p$ con $Z_1, ..., Z_p$ e Σ con ρ.

□

Osservazione. Dal fatto che $\sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} Var(Z_i) = p$ evinciamo che è possibile esprimere il rapporto tra la varianza totale e quella della k-esima componente principale con

$$\frac{\lambda_k}{p} \qquad (1.13)$$

dove λ_k con $k = 1, 2, ..., p$ sono gli autovalori della matrice di correlazione ρ

1.5 Variazione del campione attraverso le componenti principali

Abbiamo a questo punto la struttura algebrica necessaria per studiare la variazione di un sistema con n misurazioni su p variabili per mezzo di alcune combinazioni lineari *opportunamente* scelte. Supponiamo che i seguenti dati $x_1, ..., x_n$ rappresentino n studi indipendenti di alcune popolazioni p dimensionali. Sia μ il vettore media e Σ la matrice di covarianza. Restringiamoci allo studio su un campione di tali popolazioni, quindi consideriamo il vettore media del campione $\bar{x}$, la matrice di covarianza del campione e la matrice di correlazione del campione, rispettivamente indicati con S, R. Il nostro obiettivo in questa sezione sarà quello di costruire delle combinazioni lineari di nuove variabili non correlate che esprimano gran parte della variabilità all'interno del campione. Le combinazioni lineari con più ampia varianza saranno definite come *componenti principali del campione*. Ricordiamo che n valori di una combinazione lineare

$$a_1^T x_j = a_{11} x_{j1} + ... + a_{1p} x_{jp}$$

con $j = 1, 2, ..., n$ hanno media campionaria $a_1^T \bar{x}$ e varianza campionaria $a_1^T S a_1$. Inoltre le coppie di valori $(a_1^T x_j, a_2^T x_j)$ per due combinazioni lineari, hanno covarianza campionaria $a_1^T S a_2$. Le componenti principali sono definite come quelle combinazioni aventi la massima varianza campionaria. Come fatto precedentemente per le popolazioni, imponiamo ai vettori di a_i dei coefficienti di avere norma uno. Quindi in pratica

> Prima componente principale del campione = combinazione lineare $a_1^T x_j$ che massimizza la varianza campionaria di $a_1^T x_j$ e tale che $a_1^T a_1 = 1$
>
> seconda componente principale del campione = combinazione lineare $a_2^T x_j$ che massimizza la varianza di $a_2^T x_j$ e tale che $a_2^T a_2 = 1$ e $Cov(a_1^T x_j, a_2^T x_j) = 0$

All'i-esimo passo

I-esima componente principale= combinazione lineare $a_i^T x_j$ che massimizza la varianza campionaria di $a_i^T x_j$ tale che $a_j^T a_j = 1$ e $Cov(a_j^T x_j, a_k^T x_j) = 0$ per $k < i$

La prima componente principale, equivalentemente, è

$$\max_{a=\|1\| a \neq 0} = \frac{a_1^T S a_1}{a_1^T a_1} \qquad (1.14)$$

Sappiamo che tale massimo è il più grande autovalore $\hat{\lambda}_1$ ottenuto per la scelta di $a_1 = \hat{e}_1$ con $\hat{e}_1$ autovettore di S. I successivi a_i scelti massimizzeranno a loro volta le varianza di $a_i^T x_j$ e saranno perpendicolari agli autovettori $\hat{e}_k$ per ogni $k < i$.
Diamo uno schema dei risultati trovati:

Se S è la matrice di covarianza del campione con autocoppie $(\hat{\lambda}_1, \hat{e}_1), ..., (\hat{\lambda}_p, \hat{e}_p)$ la i-esima componente principale del campione è data da

$$\hat{y}_i = \hat{e}_i^T X = \hat{e}_{i1} x_1 + ... + \hat{e}_{ip} x_p$$

con $i = 1, 2, ..., p$ e dove $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ..... \geq \hat{\lambda}_p \geq 0$ e x è un'osservazione sulle variabili $X_1, ...X_p$.
Inoltre

$$Var(\hat{y}_k) = \hat{\lambda}_k \qquad con\ k = 1, 2, ..., p.$$
$$Cov(\hat{y}_i, \hat{y}_k) = 0 \qquad per\ i \neq k.$$

Infine

$$\text{Varianza totale del campione} = \sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + ... + \hat{\lambda}_p \qquad (1.15)$$

E

$$r(\hat{y}_i, x_k) = \frac{\hat{e}_{ik}}{\sqrt{\hat{\lambda}_i}} \sqrt{s_{kk}} \qquad con\ i, k = 1, 2, ..., p$$

Osservazione. Ricordiamo ancora una volta che abbiamo denotato le componenti principali con $\hat{\lambda}_1, ..., \hat{\lambda}_p$ a prescindere dal fatto se esse siano stati ottenuti da S o R. Anche se in generale le componenti ottenute da una o dall'altra non sono le stesse, risulterà chiaro dal contesto quale matrice viene utilizzata, d'altra parte adottare una singola notazione riesce molto più comodo. Per questo motivo manteniamo singole anche le notazioni per i vettori dei coefficienti $\hat{e}_i$ e le varianze $\hat{\lambda}_i$.

Le osservazioni x_j sono spesso "centrate" in $\bar{x}$. Questo non ha effetti rilevanti sulla matrice di covarianza S e per ogni vettore dell'osservazione ci viene data la i-esima componente principale seguente

$$\hat{y}_i = \hat{e}_i(x - \bar{x}) \qquad per\ i = 1, 2, ..., p \tag{1.16}$$

Se consideriamo il valore della i-esima componente invece

$$\hat{y}_{ji} = \hat{e}_i(x_j - \bar{x}) \qquad con\ j = 1, 2, ..., n \tag{1.17}$$

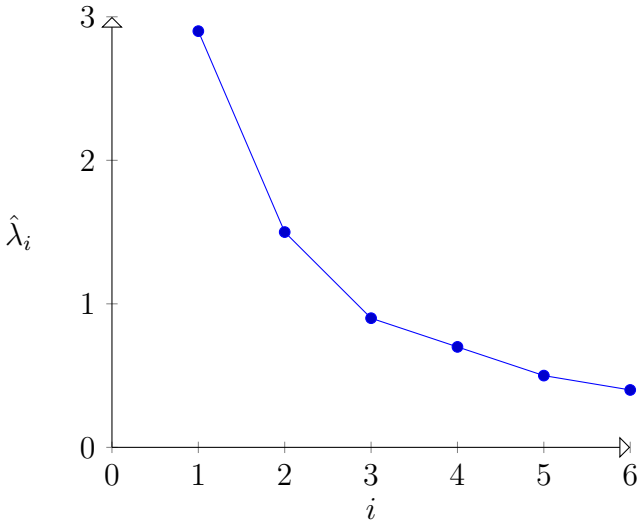
Risulta evidente quindi che

$$\hat{y}_i = \frac{1}{n}\sum_{j=1}^{n} \hat{e}_i^T(x_j - \bar{x}) = \frac{1}{n}\hat{e}_i^T(\sum_{j=1}^{n}(x_j - \bar{x})) = \frac{1}{n}\hat{e}_i^T 0 = 0 \tag{1.18}$$

Cioè, la media campionaria di ogni singola componente principale è zero.

1.6 Il numero delle componenti principali

Un problema che spesso coinvolge chi usa la PCA è quello di decidere il numero delle componenti principali da utilizzare. Non esiste una soluzione definitiva a questo problema. Ci sono molteplici fattori da tenere in considerazione, come la quantità della varianza totale del campione e le dimensioni degli autovalori. Noi ci limitiamo ad esporre un aiuto "visivo" che in alcuni casi (non tutti) può essere efficace nella determinazione del numero delle componenti.
Riportiamo un esempio per spiegare la metodologia:



Questo grafico,relativo a un generico sistema di dati a 6 variabili, posto sul piano Cartesiano $i \times \hat{\lambda}_i$ ordina gli autovalori dal più grande al più piccolo. Per individuare il numero delle componenti si cerca il cosiddetto gomito della curva che appare in figura. Infatti in corrispondenza del "gomito" si trova il punto per cui da lì in avanti i rimanenti autovalori si fanno abbastanza piccoli, e tutti dello stesso ordine di grandezza, da poter essere trascurati. Nella fattispecie della nostra figura, potremmo indicare 2 o al massimo 3 componenti principali necessari per riscrivere il sistema di origine, e non 6. Notiamo quindi una notevole riduzione delle variabili, che si traduce in un'utile semplificazione.

1.7 Componenti principali per campioni standardizzati

Anche per campioni, le componenti principali non sono invarianti rispetto ai cambiamenti di unità di misura. Quindi, come già detto per le popolazioni, anche le componenti principali di campioni che raccolgono dati misurati con diverse unità, o anche sulla stessa scala ma su range di valori molto differenti, vengono standardizzate. La procedura è la seguente

$$z_j = D^{-\frac{1}{2}}(x_j - \bar{x}) = \begin{bmatrix} \frac{x_{j1}-\bar{x}}{\sqrt{s_{11}}} \\ . \\ . \\ . \\ . \\ \frac{x_{jp}-\bar{x}}{\sqrt{s_{pp}}} \end{bmatrix} \qquad j = 1, 2, ..., p \tag{1.19}$$

La matrice $n \times p$ dei dati standardizzati sarà

$$Z = \begin{bmatrix} Z_1^T \\ . \\ . \\ . \\ Z_n^T \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{p1} & z_{p2} & \cdots & z_{pp} \end{bmatrix} =$$

$$\begin{bmatrix} \frac{x_{11}-\bar{x}_1}{\sqrt{s_{11}}} & \frac{x_{12}-\bar{x}_1}{\sqrt{s_{12}}} & \cdots & \frac{x_{1p}-\bar{x}_1}{\sqrt{s_{1p}}} \\ \frac{x_{21}-\bar{x}_2}{\sqrt{s_{12}}} & \frac{x_{22}-\bar{x}_2}{\sqrt{s_{22}}} & \cdots & \frac{x_{2p}-\bar{x}_2}{\sqrt{s_{2p}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{n1}-\bar{x}_p}{\sqrt{s_{1p}}} & \frac{x_{n2}-\bar{x}_p}{\sqrt{s_{2p}}} & \cdots & \frac{x_{np}-\bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} \tag{1.20}$$

Dunque il vettore della media campionaria è

$$\bar{\mathbf{z}} = \frac{1}{n}(1^T Z)^T = \frac{1}{n} Z^T 1 = \frac{1}{n} \begin{bmatrix} \sum_{j=1}^{n} \frac{x_{j1}-\bar{x}_1}{\sqrt{s_{11}}} \\ \vdots \\ \sum_{j=1}^{n} \frac{x_{jp}-\bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} = 0 \qquad (1.21)$$

Mentre la matrice di covarianza del campione

$$S_z = \frac{1}{n-1}(Z - \frac{1}{n}11^T Z)^T (Z - \frac{1}{n}11^T Z) =$$

$$\frac{1}{n-1}(Z - 1\bar{Z^T})^T (Z - 1\bar{Z^T}) = \frac{1}{n-1} Z^T Z =$$

$$= \frac{1}{n-1} \begin{bmatrix} \frac{(n-1)s_{11}}{s_{11}} & \frac{(n-1)s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \cdots & \frac{(n-1)s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} \\ \frac{(n-1)s_{1p2}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \frac{(n-1)s_{22}}{\sqrt{s_{22}}} & \cdots & \frac{(n-1)s_{2p}}{\sqrt{s_{22}}\sqrt{s_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{(n-1)s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} & \frac{(n-1)s_{2p}}{\sqrt{s_{22}}\sqrt{s_{pp}}} & \cdots & \frac{(n-1)s_{pp}}{\sqrt{s_{pp}}} \end{bmatrix} = R \qquad (1.22)$$

I problemi di massimo che individuano le componenti principali del campione di osservazioni standardizzate sono già state date in (1.15) con la matrice R al posto di S. Poiché le osservazioni sono già "centrate" per costruzione, non abbiamo bisogno di scrivere le componenti nella forma (1.16).
Riassumendo:
Se $z_1, ..., z_n$ sono dati standardizzati con matrice di covarianza R, la i-esima componente principale è

$$\hat{y}_i = \hat{e}_i^T z = \hat{e}_{i1} z_1 + ... + \hat{e}_{ip} z_p \quad i = 1, 2, ..., p$$

Dove $(\hat{\lambda}_i, \hat{e}_i)$ è la i-esima autocoppia di R con $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ..... \geq \hat{\lambda}_p \geq 0$
Inoltre

$$\text{Varianza campionaria di } \hat{y}_i = \hat{\lambda}_i \; i = 1, 2, ..., p$$

$$\text{Covarianza del campione di } (\hat{y}_i, \hat{y}_k) = 0 \quad i \neq k$$

In ultimo

$$\text{Varianza totale del campione } = tr(R) = p = \hat{\lambda}_1 + ... + \hat{\lambda}_p \qquad (1.22)$$

$$r_{(\hat{\lambda}_i, z_k)} = \hat{e}_{ik}\sqrt{\hat{\lambda}_i} \quad i, k = 1, 2, ..., p$$

Usando la (1.23) notiamo che il rapporto tra la varianza totale e quella relativa alla i-esima componente è

$$\frac{\hat{\lambda}_i}{p} \quad i = 1, 2, ..., p \qquad (1.24)$$

In prima istanza si suggerisce di mantenere solo quelle componenti le cui varianze sono più grandi di 1, oppure, equivalentemente, solo quelle componenti che, da sole, esprimano almeno, in proporzione, $\frac{1}{p}$ della varianza totale. A onor del vero, questa regola non ha alle spalle una teoria consolidata e completa e perciò non deve essere applicata ciecamente in ogni situazione. Il metodo del grafico decrescente prima accennato è forse più affidabile.

Capitolo 2

Alcune applicazioni della PCA

2.1 Esempio di PCA per variabili non standardizzate

Cominciamo con un esercizio abbastanza semplice, dove non viene usata la standardizzazione.

Supponiamo di possedere un campione di un'indagine di censimento su alcune popolazioni, che per 14 aree geografiche diverse abbia raccolto dati relativi al numero degli abitanti, la media degli anni scolastici, l'occupazione totale, l'occupazione in campo sanitario, e il valore medio della casa.
Riportiamo i risultati su questa tabella:

Area	numero di abitanti (in migliaia)	media anni scolastici	occupazione totale (in migliaia)	occupazione in campo sanitario (in centinaia)	valore medio della casa (in decine di migliaia)
1	5.935	14.2	2.265	2.27	2.91
2	1.523	13.1	0.597	0.75	2.62
3	2.599	12.7	1.237	1.11	1.07
4	4.009	15.2	1.649	0.81	3.02
5	4.687	14.7	2.312	2.50	2.22
6	8.044	15.6	3.641	4.51	2.36
7	2.766	13.3	1.244	1.03	1.97
8	6.538	17.0	2.618	2.39	1.85
9	6.451	12.9	3.147	5.52	2.01
10	3.314	12.2	1.606	2.18	1.82
11	3.777	13.0	2.119	2.83	1.80
12	1.530	13.8	0.080	0.84	4.25
13	2.768	13.6	1.336	1.75	2.64
14	6.585	14.9	2.763	1.91	3.17

A questo punto ci poniamo i seguenti obiettivi:

1) Costruire la matrice di covarianza dei dati e il vettore della media del campione.
2) Trovare le autocoppie relative alle prime due componenti principali del campione.
3) Trovare le correlazioni ed esprimere il rapporto tra la varianza delle prime due componenti e quella totale. Dare infine una interpretazione delle nuove variabili se possibile.

Utilizziamo il comando di matlab per trovare la matrice di covarianza. Poniamo quindi A come la matrice dei dati, ovvero la matrice 14×5 che ha come righe le osservazioni rispetto alle 14 aree geografiche. Diamo in pasto a matlab "Cov(A)", e otteniamo la matrice di covarianza di A:

$$\begin{bmatrix} 4.308 & 1.683 & 1.803 & 2.155 & -0.253 \\ 1.683 & 1.768 & 0.588 & 0.177 & 0.176 \\ 1.803 & 0.588 & 0.801 & 1.065 & -0.158 \\ 2.155 & 0.177 & 1.065 & 1.970 & -0.357 \\ -0.253 & 0.176 & -0.158 & -0.357 & 0.504 \end{bmatrix}$$

che chiameremo B.
Mentre per calcolare il vettore media usiamo Mean(A) e Matlab ci restituirà il vettore

$$\bar{X} = (4.32, 14.01, 1.95, 2.17, 2.45)$$

Ora, per calcolare le autocoppie applichiamo il seguente algoritmo $[E, V] = eig(B)$
E otteniamo

$$V = \begin{bmatrix} 0.302 & -0.541 & -0.004 & -0.070 & 0.781 \\ 0.009 & 0.545 & 0.162 & -0.764 & 0.306 \\ -0.937 & -0.051 & -0.015 & 0.083 & 0.334 \\ 0.172 & 0.636 & -0.220 & 0.579 & 0.426 \\ -0.024 & -0.051 & -0.962 & -0.262 & -0.054 \end{bmatrix}$$

E anche

$$D = \begin{bmatrix} 0.014 & 0 & 0 & 0 & 0 \\ 0 & 0.230 & 0 & 0 & 0 \\ 0 & 0 & 0.390 & 0 & 0 \\ 0 & 0 & 0 & 1.786 & 0 \\ 0 & 0 & 0 & 0 & 6.931 \end{bmatrix}$$

Datogli tale ordine, Matlab individua la matrice V, le cui colonne non sono altro che gli autovettori dei rispettivi autovalori, e la matrice diagonale D, sulla cui diagonale giacciono gli autovalori stessi.
Risulta quasi immediato che sono sufficienti due componenti principali per descrivere il sistema. Infatti, relativamente all'autovalore 6.931 abbiamo che il rapporto

$$\frac{6.931}{(6.931 + 0.014 + 0.230 + 0.390 + 1.786)} = 74.1$$

esprime già il 74 per cento della varianza totale del sistema. Aggiungiamo anche l'autovalore 1.785 e arriviamo addirittura al 93 per cento. Infatti

$$\frac{6.931 + 1.786}{(6.931 + 0.014 + 0.230 + 0.390 + 1.786)} = 93.2$$

Accostiamo a tali autovalori i rispettivi autovettori.

$$(\lambda_1, e_1) = (6.931, [0.781(\mathbf{0.99}), 0.306(\mathbf{0.61}), 0.334(\mathbf{0.98}), 0.426(\mathbf{0.80}), -0.054(\mathbf{\text{-}0.20})])$$
$$(\lambda_2, e_2) = (1.786, [-0.071(\mathbf{\text{-}0.04}), -0.764(\mathbf{\text{-}0.76}), 0.083(\mathbf{0.12}), 0.0579(\mathbf{0.55}), -0.262(\mathbf{\text{-}0.49})]$$

Osservazione. Abbiamo riportato (in neretto tra parentesi) anche i coefficienti delle correlazioni, per mostrare come (almeno in questo caso) forniscano di fatto le stesse indicazioni dei vettori dei coefficienti.

Otteniamo dunque le componenti principali attraverso la formula già riportata:

$$Y_i = e_i^T(X - \bar{X})$$

per ogni vettore x delle 14 osservazioni.

Interpretiamo la *prima componente principale*. Il suo autovettore associato ci indica che tale variabile apporta un peso rilevante rispetto alle prime 4 caratteristiche facendo quasi una media ponderata ma con una leggera preferenza verso la variabile del numero degli abitanti.
Per quanto riguarda la *seconda componente principale*, sempre considerando i coefficienti dell'autovettore, vediamo che essa influisce prevalentemente sulla seconda, quarta e quinta caratteristica. In particolare mette a confronto l'occupazione in campo sanitario con la media degli anni scolastici e il valore della casa. Riscriviamo adesso la tabella dati con le componenti principali.

Area	Y1	Y2
1	1.4377	-0.2926
2	-3.5349	-0.0826
3	-2.4002	0.6444
4	-0.5953	-1.8459
5	0.7667	-0.2679
6	4.9575	0.0449
7	-2.1317	0.0629
8	2.9913	-2.0973
9	3.1718	2.8563
10	-1.4207	1.6001
11	-0.3649	1.3806
12	-3.2985	-0.9767
13	-1.7374	0.0827
14	2.1585	-1.1089

2.2 A proposito dell'algoritmo Eig

Data l'importanza che hanno le autocoppie per il calcolo delle componenti principali, vale la pena descrivere l'algoritmo di Matlab che le individua, ovvero il comando *eig*. Tale algoritmo è basato sulla iterazione QR. il metodo iterativo QR prende a sua volta il nome dalla fattorizzazione QR, che viene utilizzata nel procedimento. Esponiamo innanzitutto questa fattorizzazione.

Definizione. *Data una matrice rettangolare $A \in \mathbb{R}^{n\times m}$, una fattorizzazione QR di A è una decomposizione di A per cui $A = QR$ con Q matrice ortogonale e R una matrice triangolare superiore*

Diamo anche la definizione di matrici simili:

Definizione. *Due matrici quadrate A e B si dicono simili se esiste una matrice M invertibile tale che $A = M^{-1}BM$*

Il metodo QR per la ricerca degli autovalori di una matrice genera una successione di matrici simili. Lo schema dell'algoritmo è particolarmente semplice. Nella generica iterazione k si effettua la fattorizzazione QR della matrice A_k e si calcola la nuova iterata rimoltiplicando i fattori in ordine inverso:
Si pone inizialmente

$$A = A_1$$

e

$$A_1 = A = Q_1 R_1$$

con Q_1 unitaria ed R_1 triangolare superiore. poi

$$A_2 = R_1 Q_1$$

e ottenendo:

$$A_2 = R_1 Q_1 = (Q_1^T Q)(R_1 Q_1) = Q_1^T A_1 Q_1$$

e si ha che A_2 è simile ad A_1.

Alla k-esima interazione

$$A_k = Q_k R_k \qquad A_{k+1} = R_k Q_k \qquad k = 1, 2, ...$$

e

$$A_{k+1} = Q_k^T A_k Q_k$$

Osservazione. Tutte le matrici della forma A_K sono simili, e come tali hanno gli stessi autovalori.

Si dimostra che per matrici simmetriche l'iterazione appena descritta converge a una matrice diagonale D i cui autovalori, ovvero gli elementi sulla diagonale, saranno perciò gli stessi della matrice A di partenza. In Matlab facendo D=eig(A) ci viene trovata la matrice diagonale degli autovalori attraverso questa procedura, scrivendo invece [V,D]=eig(A) ci viene fornita anche la matrice V degli autovettori.

2.3 PCA per campioni standardizzati

Svolgiamo ora il seguente esercizio, che richiede anzitutto una standardizzazione delle variabili. Riportiamo in una tabella i vari record nazionali dell'atletica leggera.

Paese	100m(s)	200m(s)	400m(s)	800m(m)	1500m(m)	5000m(m)	10000m(m)	42200(m)
Argentina	10.39	20.81	46.84	1.81	3.70	14.04	29.36	137.72
Australia	10.31	20.06	44.84	1.74	3.57	13.28	27.66	128.30
Austria	10.44	20.81	46.82	1.79	3.60	13.26	27.72	135.90
Belgio	10.34	20.68	45.04	1.73	3.60	13.22	27.45	129.95
Bermuda	10.28	20.58	45.91	1.80	3.75	14.68	30.55	146.62
Brasile	10.22	20.43	45.21	1.73	3.66	13.62	28.62	133.13
Burma	10.64	21.52	48.30	1.80	3.85	14.45	30.28	139.95
Canada	10.17	20.22	45.68	1.76	3.63	13.55	28.09	130.15
Cile	10.34	20.80	46.20	1.79	3.71	13.61	29.30	134.03
Cina	10.51	21.04	47.30	1.81	3.73	13.90	29.13	133.53
Colombia	10.43	21.05	46.10	1.82	3.74	13.49	27.88	131.35
Isole cok	12.18	23.20	52.94	2.02	4.24	16.70	35.38	164.70
Costa Rica	10.94	21.90	48.66	1.87	3.84	14.03	28.81	136.58
Slovacchia	10.35	20.65	45.64	1.76	3.58	13.42	28.19	134.32
Danimarca	10.56	20.52	45.89	1.78	3.61	13.50	28.11	130.78
R.D.	10.14	20.65	46.80	1.82	3.82	14.91	31.45	154.12
Finlandia	10.43	20.69	45.49	1.74	3.61	13.27	27.52	130.87
Francia	10.11	20.38	45.28	1.73	3.57	13.34	27.97	132.30
Germania	10.12	20.33	44.87	1.73	3.56	13.17	27.42	129.92
Germ.ovest	10.16	20.37	44.50	1.73	3.53	13.21	27.61	132.23
Bretagna	10.11	20.21	44.93	1.70	3.51	13.01	27.51	129.13
Grecia	10.22	20.71	46.56	1.78	3.64	14.59	28.45	134.60
Guatemala	10.98	21.82	48.40	1.89	3.80	14.16	30.11	139.33
Ungheria	10.26	20.62	46.02	1.77	3.62	13.49	28.44	132.58
India	10.60	21.42	45.73	1.76	3.73	13.77	28.81	131.98
Indonesia	10.59	21.49	47.80	1.84	3.92	14.73	30.79	148.83
Irlanda	10.61	20.96	46.30	1.79	3.56	13.32	27.81	132.35
Israele	10.71	21.00	47.80	1.77	3.72	13.66	28.93	137.55
Italia	10.01	19.72	45.26	1.73	3.60	13.23	27.52	131.08
Giappone	10.34	20.81	45.86	1.79	3.64	13.41	27.72	128.63
Kenya	10.46	20.66	44.92	1.73	3.55	13.10	27.38	129.75
Korea	10.34	20.89	46.90	1.79	3.77	13.96	29.23	136.25
Korea N.	10.91	21.94	47.30	1.85	3.77	14.13	29.67	130.87
Lussemb.	10.35	20.77	47.40	1.82	3.67	13.64	29.08	141.27
Malesia	10.40	20.92	46.30	1.82	3.80	14.64	31.01	154.10
Mauritius	11.19	22.45	47.70	1.88	3.83	15.06	31.77	152.23
Messico	10.42	21.30	46.10	1.80	3.65	13.46	27.95	129.20
Olanda	10.52	20.95	45.10	1.74	3.62	13.36	27.61	129.02

Paese	100m(s)	200m(s)	400m(s)	800m(m)	1500m(m)	5000m(m)	10000m(m)	42200(m)
N.Z	10.51	20.88	46.10	1.74	3.54	13.21	27.70	128.98
Norvegia	10.55	21.16	46.71	1.76	3.62	13.34	27.69	131.48
Guinea	10.96	21.78	47.90	1.90	4.01	14.72	31.36	148.22
Filippine	10.78	21.64	46.24	1.81	3.83	14.74	30.64	145.27
Polonia	10.16	20.24	45.36	1.76	3.60	13.29	27.89	131.58
Portogallo	10.53	21.17	46.70	1.79	3.62	13.13	27.38	128.65
Romania	10.41	20.98	45.87	1.76	3.64	13.25	27.77	132.50
Singapore	10.38	21.28	47.40	1.88	3.89	15.11	31.32	157.77
Spagna	10.42	20.77	45.98	1.76	3.55	13.31	27.73	131.57
Svezia	10.25	20.61	45.63	1.77	3.61	13.29	27.94	130.63
Svizzera	10.37	20.46	45.78	1.78	3.55	13.22	27.91	131.20
Taipei	10.59	21.29	46.80	1.79	3.77	14.07	30.07	139.27
Thailanda	10.39	21.09	47.91	1.83	3.84	15.23	32.65	149.90
Turchia	10.71	21.43	47.60	1.79	3.67	13.56	28.58	131.50
Usa	9.93	19.75	43.86	1.73	3.53	13.20	27.43	128.22
Russia	10.07	20.00	44.60	1.75	3.59	13.20	27.53	130.55
Samoa	10.82	21.86	49.00	2.02	4.24	16.28	34.71	161.83

Per studiare e semplificare la mole consistente di questi dati procediamo nel modo seguente: come prima cosa, determiniamo la matrice di correlazione campionaria dei dati standardizzati e le sue autocoppie. Quindi chiamando A la matrice 55×8 delle osservazioni riportate in tabella

$$B = zscore(A)$$

$$R = corr(B)$$

E poi

$$[V, D] = eig(R)$$

Otteniamo la matrice diagonale D degli autovalori della matrice di correlazione R

$$D = \begin{bmatrix} 6.6221 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.8776 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.1593 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.1240 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0226 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.0799 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.0680 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0464 \end{bmatrix}$$

e la matrice degli autovettori V.

$$V = \begin{bmatrix} -0.3176 & 0.5669 & -0.3323 & -0.1276 & +0.1055 & 0.2626 & -0.5937 & -0.1362 \\ -0.3370 & 0.4616 & -0.3607 & 0.2591 & -0.0961 & -0.1540 & 0.6561 & 0.1126 \\ -0.3556 & 0.2483 & 0.5605 & -0.6523 & -0.0001 & -0.2183 & -.1566 & 0.0029 \\ -0.3687 & 0.0124 & 0.5325 & 0.4800 & -0.0382 & 0.5401 & -0.0147 & 0.2380 \\ -0.3728 & -0.1398 & 0.1534 & 0.4045 & +0.1393 & -0.4877 & -0.1578 & -0.6100 \\ -0.3644 & -0.3120 & -0.1898 & -0.0296 & 0.5467 & -0.2540 & -0.1413 & 0.5913 \\ -0.3668 & -0.3069 & -0.1818 & -0.0801 & -0.7968 & -0.1332 & -0.2190 & 0.1769 \\ -0.3419 & -0.4390 & -0.2632 & -0.2995 & 0.1582 & 0.4979 & 0.3153 & -0.3988 \end{bmatrix}$$

I primi due autovalori hanno una varianza che è pari al 93,75 per cento di quella totale. infatti

$$\frac{\lambda_1 + \lambda_2}{Tr(D)} = \frac{6.221 + 0.0776}{8.0000} = 93,75$$

Lavoriamo allora solo sulle prime due componenti principali:
Rispetto alla prima componente guardando l'autovettore a lei relativo, ovvero il vettore

$$v_1 = [-0.3176, -0.3370, -0.3556, -0.3687, -0.3728, -0.3644, -0.3668, -0.3419]$$

Notiamo che i pesi dei coefficienti sono essenziali simili, quasi una media ponderata dei risultati ottenuti su tutte le distanze. Potremo quindi pensare la prima componente principale come un indicatore della qualità della prestazione di una nazione a livello globale. Questa interpretazione è giustificata anche dal fatto che la prima componente è quella che cerca di estrarre la massima quantità di informazione, mentre le seguenti tentano di ottimizzare l'informazione residua. Per quanto riguarda la seconda componente principale sempre osservando l'autovettore associato

$$v_2 = [0.5669,\ 0.4616,\ 0.2483,\ 0.0124,\ -0.1398,\ -0.3120,\ -0.3069,\ -0.4390]$$

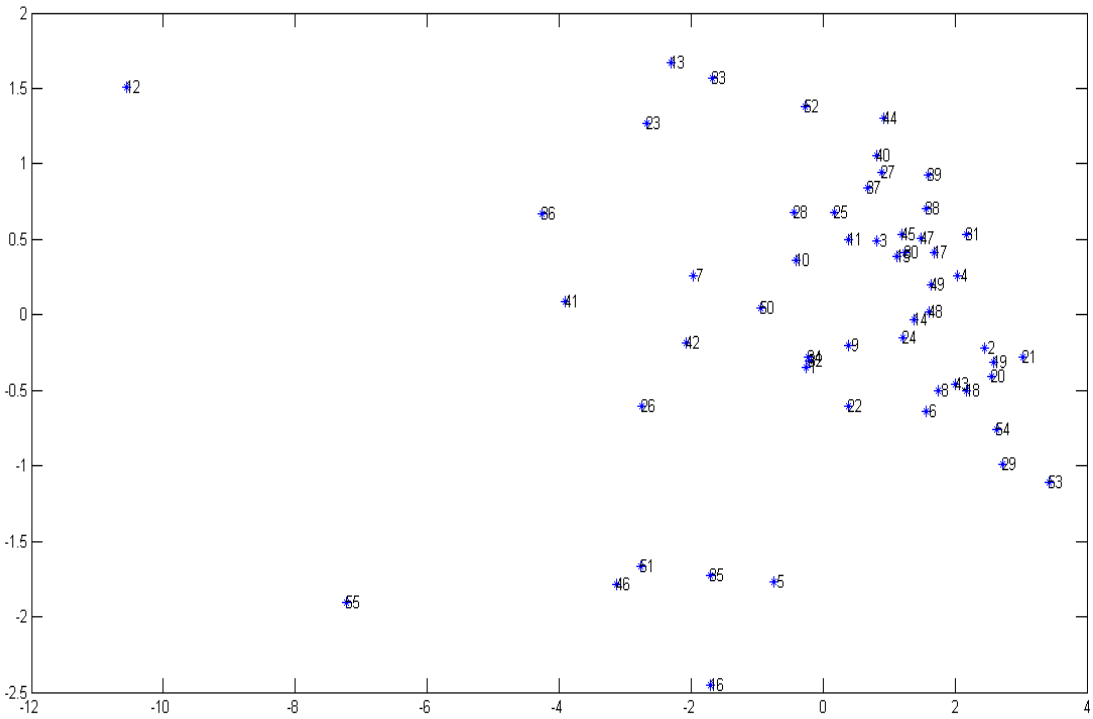
Si nota che i primi 4 coefficienti sono positivi mentre gli ultimi 4 sono negativi. Questo contrasto potrebbe suggerire che tale variabile cerchi di misurare lo "scarto prestazionale" di una nazione tra le gare di velocità, quindi fino agli 800 metri, e quelle di resistenza, con un condizionamento leggermente maggiore da una parte verso la velocità più pura (infatti nei cento metri il peso è 0.5669, superiore a tutti gli altri 3 coefficienti dello stesso segno) e dall'altra parte verso il fondo a scapito del mezzofondo (infatti la maratona ha un peso di 0,4390, più grande degli altri 3 coefficienti del medesimo segno).

Riportiamo adesso in tabella il nuovo set di dati.

numero nazione	**Y1**	**Y2**
1	-0.2619	-0.3449
2	2.4464	-0.2162
3	0.8076	0.4869
4	2.0413	0.2619
5	-0.7393	-1.7669
6	1.5583	-0.6412
7	-1.9719	0.2572
8	1.7464	-0.5003
9	0.3811	-0.2014
10	-0.4090	0.3578
11	0.3901	0.5000
12	-10.5556	1.5088
13	-2.2966	1.6706
14	1.3726	-0.0288
15	1.1132	0.3873
16	-1.7149	-2.4490
17	1.6920	0.4088
18	2.1719	-0.5029
19	2.5901	-0.3107
20	2.5527	-0.4114
21	3.0242	-0.2789
22	0.3796	-0.6018
23	-2.6724	1.2712
24	1.2052	-0.1515
25	0.1652	0.6765
26	-2.7478	-0.6032
27	0.8842	0.9460
28	-0.4346	0.6743
29	2.7269	-0.9899
30	1.2379	0.4136
31	2.1683	0.5337
32	-0.2075	-0.3015
33	-1.6837	1.5647

numero nazione	**Y1**	**Y2**
34	-0.2205	-0.2793
35	-1.7083	-1.7227
36	-4.2587	0.6670
37	0.6785	0.8418
38	1.5554	0.7024
39	1.5997	0.9234
40	0.8115	1.0566
41	-3.9092	0.0855
42	-2.0704	-0.1893
43	2.0006	-0.4626
44	0.9164	1.3047
45	1.1965	0.5308
46	-3.1221	-1.7890
47	1.4806	0.5067
48	1.6032	0.0232
49	1.6390	0.1959
50	-0.9505	0.0420
51	-2.7618	-1.6698
52	-0.2661	1.3830
53	3.4306	-1.1102
54	2.6269	-0.7570
55	-7.2312	-1.9021

Di seguito viene mostrato anche il plot delle componenti, numerate:



Vediamo subito che gli Stati Uniti (numero 53) hanno la prima componente più grande di tutti gli altri, e infatti globalmente gli atleti più validi sono americani. Lo si vede guardando anche la tabella originaria dei dati: detengono il record dei cento metri che è di 9.93, (ricordiamo che non viene rappresentata la Giamaica in questa classifica), e sebbene nelle distanze più lunghe non siano i migliori mantengono risultati di tutto rispetto. Valutare invece le nazioni relativamente alla seconda componente è un po' più complicato. Paesi che riportano un valore molto alto non significa che necessariamente abbiano atleti di fondo eccellenti. Ricordiamo che la seconda componente riporta la forza relativa di una nazione nella corsa di fondo rispetto a quella di velocità. Si prenda come esempio il paese con valore più alto: la Costa Rica (13). A livello globale i suoi atleti hanno raggiunto risultati mediocri (si veda la prima componente). Eppure, all'aumentare della distanza, detenendo comunque valori abbastanza bassi in senso assoluto, essi corrono le lunghe distanze in meno tempo di quello che ci si potrebbe aspettare se si considerano i loro tempi nelle distanze brevi. Confrontiamo anche i Samoani (55) con gli atleti delle Isole Cook (12). Entrambi hanno tempi mediocri, in particolare le isole Cook. Le isole Cook hanno seconda componente principale alta, Le Samoa invece molto bassa. Significa di fatto che in rapporto ai tempi conseguiti nelle distanze brevi, gli atleti delle isole Cook

si comportano meglio nel fondo rispetto ai Samoani. Infatti se facciamo riferimento alla tabella iniziale i tempi delle prime distanze dei samoani sono molto più brevi rispetto a quelli delle isole Cook, pur essendo scadenti. Ma nelle ultime distanze la differenza si assottiglia molto: sono 67 secondi nei 10000 metri e 3 minuti in maratona ad esempio. Per quanto riguarda invece le nazioni più forti, che formano la nuvola di punti nel grafico vicino agli USA, risulta ad esempio che i keniani (31) pur essendo meno validi degli americani sulle gare di velocità hanno la seconda componente più alta di loro. Ciò è in linea con la conoscenza comune dell'eccellente valore nelle corse di resistenza di atleti di certe zone dell'Africa. Infine citiamo la Gran Bretagna (21), che riporta buoni valori per entrambe le componenti ma senza spiccare in nessuna delle due, e infatti come si può dedurre dal set di origine, è un paese "equilibrato", dotato di buoni atleti, alcuni anche ottimi, per ogni distanza.

2.4 Conclusione

Riassumendo il lavoro svolto, possiamo dire di aver descritto il metodo della PCA nei suoi aspetti più caratteristici. Si sono date le nozioni matematico-statistiche necessarie per esporre la struttura matematica su cui è fondato tale metodo. Dopo aver dimostrato i lemmi fondamentali abbiamo riportato le formule che individuano le componenti principali. Infine si sono mostrate alcune applicazioni del metodo nell'Analisi Dati. Si è anche evidenziato, trasversalmente attraverso gli esempi, l'importanza di reinterpretare correttamente le nuove variabili ottenute dal procedimento suddetto, ovvero di ridefinirle in modo consono al contesto in cui si opera.

Bibliografia

[1] Richard A. Johnson, Dean W. Wichern *Applied Multivariate Statistical Analysis.* Prentice hall, 2005.

[2] Alfio Quarteroni, Fausto Saleri, Paola Gervasio *Calcolo scientifico.* Springer, Quinta edizione.