Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Matematica

ASPETTI COMPUTAZIONALI NELL'ANALISI DELLE COMPONENTI PRINCIPALI

Tesi di Laurea in Calcolo Numerico

Relatore:
Chiar.ma Prof.ssa
Valeria Simoncini

Presentata da:
Marco Tagliapietra

Sessione unica
Anno Accademico 2018/2019

A mio padre

Indice

Introduzione

Nella seguente tesi verrà trattata la "Principal Components Analysis" (PCA), l'analisi delle componenti principali, argomento relativamente moderno utilizzato nel mondo della statistica multivariata per lo studio di grandi quantità di dati. I campi di applicazione principali possono essere relativi alla medicina (esaminare gli effetti di un medicinale su una parte di popolazione), alla finanza (studiare quali aziende hanno un trend collegato), alla metereologia (classificare zone geografiche analoghe), allo studio socio/economico (trovare una categoria "bersagliabile" con un prodotto commerciale/-politico) e molti altri ambiti.

I motivi del suo utilizzo sono duplici: la riduzione dei dati e la loro interpretazione.
Per quanto concerne la riduzione di dati, se si considera un insieme di p variabili è naturale pensare che siano necessarie p componenti principali per riprodurre correttamente la variabilità del sistema, in vero verrà mostrato che la maggior parte della variabilità (80-90%) può essere spesso descritta da un numero significativamente minore di queste componenti, saranno necessarie quindi k componenti principali (con $k < p$) per descrivere il sistema. Si ottiene quindi che il set originario di osservazioni che consisteva in n misurazioni di p variabili, venga trasformato in un set che consiste in n misurazioni di k componenti principali.
Come detto la PCA è anche un forte strumento di analisi dei dati, infatti la sua applicazione spesso rivela relazioni tra le variabili prima non visibili, raggruppandole in una singola variabile, detta variabile latente, che ne descrive la loro dipendenza. Viceversa, si mettono in risalto variabili non correlate che intuitivamente pensavamo fossero simili.

Nel primo capitolo si analizzerà l'aspetto algebrico della PCA, difatti l'analisi delle componenti principali è basata sulla struttura di varianza-covarianza di un insieme di variabili che, sfruttando combinazioni lineari di quest'ultime, analizza e riduce la mole di osservazioni. Si daranno definizioni formali e utili risultati sulla loro struttura soffermandosi su alcuni aspetti che hanno risvolti interpretativi. Si discuterà inoltre il problema delle unità di misura differenti nei dati introducendo le componenti principali per variabili standardizzate.

Nel secondo capitolo verrà invece analizzata la struttura geometrica della PCA e come interpretare i suoi risultati. Si studierà la struttura con l'ipotesi di una distribuzione normale multivariata, struttura base in questo ambito di applicazione statistica/matematica soffermandosi particolarmente sulla sua interpretazione grafica. In secondo luogo si vedranno due possibili metodi per la scelta delle componenti principali ed anche in questo caso ci sarà l'aiuto grafico della rappresentazione delle componenti principali per interpretare al meglio l'argomento.

Nel terzo capitolo verrà descritto un esempio con l'utilizzo della PCA passo per passo, si tratterà dell'analisi di un set di dati riguardanti le prestazioni atletiche di 55 nazioni su 9 tipologie di gare. Da queste nove variabili ci si ridurrà a soltanto due e attraverso la teoria descritta nei primi due capitoli si analizzerano i risultati ottenuti.

Infine viene descritto l'argoritmo di un comando computazionale cruciale nella risoluzione di problemi attraverso la PCA, il comando *eig*. Si vedrà quindi in particolare l'iterazione e la fattorizzazione QR.

Nozioni Preliminari

Al fine di comprendere l'analisi delle componenti principali vengono riportate le seguenti definizioni, proposizioni e strutture base riguardanti la statistica e l'algebra lineare su cui si baseranno tutti i ragionamenti successivi:

Notazione 0.0.1. ***(Vettore Colonna e Vettore Riga)***
Nella trattazione verrà considerato un generico vettore x come un vettore colonna e il suo trasposto x' come un vettore riga.

Definizione 0.0.2. ***(Covarianza e Varianza)***
Siano date due variabili aleatorie $X' = [X_1, X_2, \ldots, X_n]$ e $Y' = [Y_1, Y_2, \ldots, Y_n]$, e siano $\mu'_X = [E(X_1), E(X_2), \ldots, E(X_n)]$ e $\mu'_Y = [E(Y_1), E(Y_2) \ldots, E(Y_n)]$ il valore atteso rispettivamente di X e Y, allora si definisce la covarianza fra X e Y come:

$$Cov(X, Y) = \frac{1}{n}[(X - \mu_X)'(Y - \mu_Y)].$$

La varianza di X è definita come $Var(X) = Cov(X, X)$.

Definizione 0.0.3. ***(Coefficiente di Correlazione)***
Si denota il coefficiente di correlazione fra X e Y il valore:

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}.$$

Definizione 0.0.4. ***(Matrice di Covarianza e di Correlazione)***
La matrice di covarianza Σ del vettore di variabili $X' = [X_1, X_2, \ldots, X_p]$ è la matrice simmetrica con $\Sigma_{ij} = Cov(X_i, X_j)$ al variare di $i, j = 1, 2, \ldots, p$.
La matrice di correlazione ρ è data da $\rho_{ij} = \rho(X_i, X_j)$.

Proposizione 0.0.5. ***(Varianza Di Una Combinazione Lineare)***
Sia data una matrice $A \in \mathbb{R}^{p\times p}$ *e un vettore aleatorio* $X' = [X_1, X_2, \ldots, X_p]$ *con matrice di covarianza* Σ_X *e sia* $Y = AX$.
Allora la matrice di covarianza di Y *è* $\Sigma_Y = A\Sigma_X A'$.
In particolare se $\mathbf{c} \in \mathbb{R}^p$ *allora* $Var(\mathbf{c}'X) = \mathbf{c}'\Sigma_X\mathbf{c}$.

Sono riportati altri generici risultati senza prova, che verranno richiamati nel corso della trattazione.

Proposizione 0.0.6. ***(Autovalore Massimo)***
Sia $A \in \mathbb{R}^{p\times p}$ *matrice simmetrica definita positiva e sia* λ_{max} *il suo autovalore massimo, allora:* $\lambda_{max} = \max_{0\neq x\in\mathbb{R}^p} \frac{x'Ax}{x'x}$.
In particolare, sia v *un autovettore di* λ_{max}, *allora:* $\frac{v'Av}{v'v} = \frac{\lambda_{max}v'v}{v'v} = \lambda_{max}$.

Teorema 0.0.7. ***(Teorema di Courant-Fischer)***
Sia B *una matrice* $p\times p$ *definita positiva con autovalori* $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$ *associati agli autovettori* $e_1, e_2, \ldots, e_p$, *allora:*

$$\max_{x\neq 0} \frac{x'Bx}{x'x} = \lambda_1 \quad (x = e_1)$$

$$\max_{x\neq 0} \frac{x'Bx}{x'x} = \lambda_p \quad (x = e_p)$$

inoltre

$$\max_{x\perp e_1, e_2, \ldots, e_k} \frac{x'Bx}{x'x} = \lambda_{k+1} \tag{1}$$

ottenuto quando $x = e_{k+1}$, $k = 1, 2, \ldots, p-1$.

Proposizione 0.0.8. ***(Decomposizione Spettrale)***
La decomposizione spettrale di una matrice A *simmetrica di dimensione* $p \times p$ *è data da:*

$$A = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \cdots + \lambda_p e_p e_p'$$

dove $\lambda_1, \lambda_2, \ldots, \lambda_p$ *sono gli autovalori di* A *e* $e_1, e_2, \ldots, e_p$ *sono gli autovettori associati. Dunque* $e_i'e_i = 1$ *per* $i = 1, 2, \ldots, p$ *e* $e_i'e_j = 0$ *per* $i \neq j$.

Viene data ora l'importante definizione di distribuzione normale e una sua generalizzazione a più dimensioni che gioca un ruolo fondamentale nell'analisi multivariata in quanto si assume che la maggior parte dei dati venga generata da una distribuzione normale multivariata.

Definizione 0.0.9. ***(Distribuzione Normale Reale)***
La Distribuzione Normale Reale $\mathcal{N}(\mu, \sigma^2)$ *di parametri* $\mu \in \mathbb{R}$ *e* $\sigma>0$ *è la distribuzione che ha come funzione di densità:*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\sigma}{\mu})^2}$$

Proposizione 0.0.10. ***(Valore Atteso e Varianza di*** $X \sim \mathcal{N}(\mu, \sigma^2)$***)***
Sia X *una variabile aleatoria di distribuzione normale reale,* $X \sim \mathcal{N}(\mu, \sigma^2)$ *con* $\mu \in \mathbb{R}$ *e* $\sigma > 0$*, allora* $\mu = E[X]$ *e* $\sigma^2 = Var(X)$*.*

Definizione 0.0.11. ***(Distribuzione Normale Multivariata)***
Sia $X' = [X_1, X_2, \ldots, X_p]$ *dove ogni variabile aleatoria* X_i *è indipendente con distribuzione normale* $\mathcal{N}(\mu_i, \sigma^2)$*. Si dice allora che* $X \sim \mathcal{N}_p(\mu, \Sigma)$ *ha distribuzione normale multivariata con funzione di densità*

$$f(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{\frac{-(X-\mu)'\Sigma^{-1}(X-\mu)}{2}}$$

con $\mu' = [\mu_1, \mu_2, \ldots, \mu_p]$ *e* Σ *matrice di covarianza di* X*.*

Capitolo 1

L'aspetto Algebrico della PCA

1.1 Componenti principali delle popolazioni

Le componenti principali sono specifiche combinazioni lineari di p variabili aleatorie $X_1, X_2, \ldots, X_p$, prese in modo che venga massimizzata la varianza.
In questo capitolo viene presentata e analizzata la loro struttura algebrica. Si consideri quindi $X' = [X_1, X_2, \ldots, X_p]$ un vettore aleatorio con matrice di covarianza Σ, i cui autovalori sono $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.
Si prenda la combinazione lineare $Y = AX$:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1p} \\ a_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{p1} & \ldots & \ldots & a_{pp} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \text{ quindi: } \begin{cases} Y_1 = a_1'X = a_{11}X_1 + \cdots + a_{1p}X_p, \\ Y_2 = a_2'X = a_{21}X_1 + \cdots + a_{2p}X_p, \\ \vdots \\ Y_p = a_p'X = a_{p1}X_1 + \cdots + a_{pp}X_p. \end{cases}$$

Ne consegue, utilizzando la Proposizione 0.0.5, che:

$$Var(Y_i) = a_i'\Sigma a_i \qquad con \quad i = 1, \ldots, p. \tag{1.1}$$

$$Cov(Y_i, Y_j) = a_i'\Sigma a_j \qquad con \quad i, j = 1, \ldots, p. \tag{1.2}$$

Le componenti principali sono quelle combinazioni lineari non correlate $Y_1, Y_2, \ldots, Y_p$ che massimizzano la varianza $Var(Y_i)$.

Si nota quindi che dipendono solo dalla matrice di covarianza Σ di $X_1, X_2, \ldots, X_p$.
La prima componente principale è la combinazione lineare con la varianza maggiore; cioè che massimizza $Var(Y_1) = a_1'\Sigma a_1$.

Definizione 1.1.1. ***(Componenti Principali)***
Si definiscono le componenti principali nel seguente modo:
La prima componente principale è la combinazione lineare $Y_1 = a_1'X$ *che massimizza* $Var(a_1'X)$ *con* $a_1'a_1 = 1$.
La seconda componente principale è la combinazione lineare $Y_2 = a_2'X$ *che massimizza* $Var(a_2'X)$ *con* $a_2'a_2 = 1$ *e tali che* $Cov(a_1'X, a_2'X) = 0$.
In generale:
la i-esima componente principale è la combinazione lineare $Y_i = a_i'X$ *che massimizza* $Var(a_i'X)$ *con* $a_i'a_i = 1$ *e tali che* $Cov(a_k'X, a_i'X) = 0$ *per* $k < i$.

Si osserva che è sufficiente moltiplicare il vettore a_1 per una costante per aumentare il valore della varianza; per ovviare a questo fatto viene aggiunta una restrizione al vettore dei coefficienti, rendendolo di norma unitaria.

Proposizione 1.1.2. ***(Componenti Principali non Correlate)***
Sia Σ *la matrice di covarianza associata al vettore aleatorio* $X' = [X_1, X_2, \ldots, X_p]$.
Siano (λ_i, e_i) *con* $i = 1, 2, \ldots, p$ *le autocoppie di* Σ *con* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.
Allora la i-esima componente principale è data da:

$$Y_i = e_i'X = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p. \quad con \quad i = 1, 2, \ldots, p.$$

In particolare, con questa scelta:

$$Var(Y_i) = e_i'\Sigma e_i = \lambda_i \quad con \quad i = 1, 2, \ldots, p.$$

$$Cov(Y_i, Y_k) = e_i'\Sigma e_k = 0 \quad con \quad i \neq k.$$

Se alcuni degli autovalori λ_i *sono uguali allora la scelta dei corrispettivi autovettori* e_i*, e di conseguenza delle componenti principali* Y_i*, non è unica.*

Dimostrazione. Per la Proposizione 0.0.6 si ottiene che $\max_{a \neq 0} \frac{a'\Sigma a}{a'a} = \lambda_1$ e che questo valore è raggiunto con $a = e_1$.

Quindi, essendo l'autovettore e_1 di norma unitaria ($e_1'e_1 = 1$) e ricordando la (1.1), si ha:

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

Si ha inoltre per il Teorema 0.0.7 (in particolare per (1)) che $\max_{a\perp e_1,e_2,\ldots,e_k} \frac{a'\Sigma a}{a'a} = \lambda_{k+1}$. Considerando $a = e_{k+1}$, anch'esso unitario, e osservando che $e_{k+1}'e_i = 0$ per $i = 1, 2, \ldots, k$ e $k = 1, 2, \ldots, p-1$, si ha:

$$\frac{e_{k+1}'\Sigma e_{k+1}}{e_{k+1}'e_{k+1}} = e_{k+1}'\Sigma e_{k+1} = Var(Y_{k+1}).$$

Ma $e_{k+1}'(\Sigma e_{k+1}) = \lambda_{k+1}e_{k+1}'e_{k+1} = \lambda_{k+1}$ quindi $Var(Y_{k+1}) = \lambda_{k+1}$
Rimane da mostrare che gli autovettori e_i perpendicolari a e_k(ovvero tali che che $e_i'e_k = 0$ per $i \neq k$), danno la matrice di covarianza nulla, cioè $Cov(Y_i, Y_k) = 0$.
Si osserva che gli autovalori di Σ sono ortogonali se tutti gli autovettori sono distinti, nel caso non lo fossero però, gli autovettori corrispondenti allo stesso λ_k possono essere scelti in maniera che siano ortogonali. Dunque, per ogni due autovettori e_i e e_k, con $i \neq k$, ricordando che $\Sigma e_k = \lambda e_k$ e la (1.2), si ha che:

$$Cov(Y_i, Y_k) = e_i'\Sigma e_k = \lambda_k e_i'e_k = 0.$$

Quindi le componenti principali non sono correlate e la proposizione è provata. □

Proposizione 1.1.3. ***(Varianza Totale)***
Sia $X' = [X_1, X_2, \ldots, X_p]$ un vettore aleatorio con matrice di covarianza Σ le cui autocoppie sono (λ_i, e_i) con $i = 1, 2, \ldots, p$ e con $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.
Siano $Y_1 = e_1'X$, $Y_2 = e_2'X$, $\ldots$, $Y_p = e_p'X$ le componenti principali, allora:

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} Var(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} Var(Y_i).$$

Dimostrazione. Per la definizione di traccia si ha: $\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = tr(\Sigma)$. Si può scrivere $\Sigma = P'\Lambda P$ con Λ la matrice diagonale degli autovalori di Σ e $P = [e_1, e_2, \ldots, e_p]$, in particolare P è ortogonale ($P'P = PP' = I$). Quindi:

$$\sum_{i=1}^{p} Var(X_i) = tr(\Sigma) = tr(P'\Lambda P) = tr(\Lambda PP') = tr(\Lambda) = \lambda_1 + \cdots + \lambda_p = \sum_{i=1}^{p} Var(Y_i).$$

□

Osservazione 1.1.4. *La Proposizione 1.1.3 è di estrema importanza per la scelta del numero di componenti principali, in quanto afferma che la varianza totale della popolazione coincide con la somma degli autovalori $\lambda_1, \lambda_2, \ldots, \lambda_p$ e di conseguenza la porzione di varianza totale della popolazione dovuta alla k-esima componente principale è:*

$$\begin{pmatrix} \textit{porzione di varianza} \\ \textit{totale della} \\ \textit{popolazione dovuta} \\ \textit{alla k-esima} \\ \textit{componente principale} \end{pmatrix} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \qquad k = 1, 2, \ldots, p$$

Se la maggior parte della varianza complessiva (indicativamente 80-90%) è dovuta alle prime tre componenti principali per esempio, si possono considerare solo queste tre senza avere una perdita significativa di informazioni; è evidente come questo procedimento riduca le quantità di dati, riducendo quindi la dimensione del problema, uno dei due obbiettivi prefissati all'inizio.

Proposizione 1.1.5. ***(Coefficiente di Correlazione fra la Componente** Y_i **e la Variabile** X_k**)***
Se $Y_1 = e_1'X$, $Y_2 = e_2'X$, $\ldots$, $Y_p = e_p'X$ sono le componenti principali ottenute dalla matrice di covarianza Σ e (λ_i, e_i) con $i = 1, 2, \ldots, p$ le sue autocoppie, allora:

$$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad con \quad i, k = 1, 2, \ldots, p$$

sono i coefficienti di correlazione fra la componente Y_i e la variabile X_k.

Dimostrazione. Sia $a_k' = [0, \ldots, 0, 1, 0, \ldots, 0]$ tale che $X_k = a_k'X$ e $Cov(X_k, Y_i) = Cov(a_k'X, e_i'X) = a_k'\Sigma e_i$. In particolare, sfruttando le proposizioni precedenti, il fatto che $\Sigma e_i = \lambda_i e_i$ e quanto appena detto, si ha che:

$$0 = Cov(X_k, Y_i) = Cov(a_k'X, e_i'X) = a_k'\Sigma e_i = a_k'\lambda_i e_i = \lambda a_k' e_i = \lambda_i e_{ik},$$

Inoltre:

$$Var(Y_i) = \lambda_i, \qquad Var(X_k) = \sigma_{kk}.$$

Dunque, usando la definizione di coefficienti di correlazione:

$$\rho_{Y_i,X_k} = \frac{Cov(X_k, Y_i)}{\sqrt{Var(X_k)Var(Y_i)}} = \frac{\lambda_i e_{ik}}{\sqrt{\lambda_i \sigma_{kk}}} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad con \quad i,k = 1,2,\ldots,p$$

□

É importante considerare anche un altro aspetto, si consideri ogni componente del vettore dei coefficienti $e_i' = [e_{i1}, \ldots, e_{ik}, \ldots, e_{ip}]$. La grandezza di e_{ik} misura il peso della k-esima variabile sull'i-esima componente principale, a prescindere dall'influenza delle altre variabili. In particolare ogni e_{ik} è proporzionale al coefficiente di correlazione tra Y_i e X_k. Si può perciò dire che il coefficiente di correlazione è un indicatore del peso che una certa variabile X_k ha nella componente principale Y_i. Bisogna notare però che questo valore non tiene conto del comportamento delle altre variabili. Infatti sebbene le correlazione tra le variabili e le componenti principali aiutino spesso a interpretare le componenti, esse misurano solo il contributo di una singola X su una componente Y. Per questa ragione alcuni statistici ritengono sia sbagliato usare le correlazioni; viene invece ritenuto più corretto e completo usare i coefficienti e_{ik} (la k-esima componente dell'autovettore e_i) che indicano comunque il peso della k-esima variabile in Y_i e sono proporzionali al coefficiente di correlazione. Sebbene queste due scelte portino a pesi diversi da parte della stessa variabile per una componente, è prassi molto frequente osservare che variabili con coeffcienti relativamente grandi in valore assoluto, tendono ad avere anche ampie correlazioni, e che quindi entrambe le misure sui pesi 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.

1.2 Componenti principali per variabili standardizzate

I dati del problema possono avere unità di misura differenti o per esempio essere di ordini di grandezza diversi; per questi casi si avrebbe un problema procedendo con le componenti principali ottenute dalla matrice di covarianza poichè quest'ultime possono descrivere la relazione fra le variabili in maniera non funzionale. Infatti se una variabile ha ordini di grandezza significativamente più grandi inciderà molto di più sulle componenti principali: ragionando su una stessa variabile, per esempio, se viene cambiata la sua unità di misura da Km a cm allora questa variabile potrebbe passare da avere un peso irrelevante nelle componenti ad avere un ruolo fondamentale. Per ovviare a questo problema si introducono e si studiano le variabili standardizzate:
$Z = (V^{\frac{1}{2}})^{-1}(X - \mu)$, segue la sua forma matriciale,

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_p \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{\sigma_{11}}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & \frac{1}{\sqrt{\sigma_{pp}}} \end{pmatrix} \begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_p - \mu_p \end{pmatrix} \qquad \text{cioè} \qquad \begin{cases} Z_1 = \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}} \\ Z_2 = \frac{(X_2 - \mu_2)}{\sqrt{\sigma_{22}}} \\ \vdots \qquad \vdots \\ Z_p = \frac{(X_p - \mu_p)}{\sqrt{\sigma_{pp}}} \end{cases}$$

dove $\mu' = [\mu_1, \mu_2, \dots, \mu_p]$ è il valore atteso del vettore $X' = [X_1, X_2, \dots, X_p]$ e V è la matrice che ha sulla diagoanale le varianze σ_{kk} di X_k, con $k = 1, 2, \dots, p$.
Si ha che $E(Z) = 0$ ed in particolare, con ρ e Σ rispettivamente la matrice di correlazione e quella di covarianza di X si ha anche che:

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

Le componenti principali per variabili standardizzate si ottengono dalla matrice di correlazione ρ di X, i risultati visti in precedenza continuano a valere poichè la varianza di ogni Z_i è unitaria infatti:

$$\begin{aligned} Var(Z_i) &= E[(Z_i - E[Z_i])^2] = E[(\frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}} - \frac{\mu_i - \mu_i}{\sqrt{\sigma_{ii}}})^2] = \\ &= E[\frac{(X_i - \mu_i)^2}{\sigma_{ii}}] = \frac{E[(X_i - \mu_i)^2]}{\sigma_{ii}} = \frac{Var(X_i)}{\sigma_{ii}} = \\ &= \frac{\sigma_{ii}}{\sigma_{ii}} = 1. \end{aligned}$$

Notazione 1.2.1. *Si può continuare a usare la notazione Y_i per riferirsi alla i-esima componente principale e (λ_i, e_i) per l'autocoppia ottenuta da ρ o Σ, anche se, in generale, le autocoppie derivate da Σ non sono le stesse generate da ρ.*

Proposizione 1.2.2. ***(Componenti Principali Per Variabili Standardizzate)*** *Siano $(\lambda_1, e_1), (\lambda_2, e_2), \ldots, (\lambda_p, e_p)$ le autocoppie di ρ con $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.*
La i-esima componente principale delle variabili standardizzate $Z' = [Z_1, Z_2, \ldots, Z_p]$ con $Cov(Z) = \rho$ è data da:

$$Y_i = e_i' Z = e_i' (V^{\frac{1}{2}})^{-1} (X - \mu) \qquad con \quad i = 1, 2, \ldots, p.$$

Inoltre

$$\sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} Var(Z_i) = p \tag{1.3}$$

e

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

Dimostrazione. La dimostrazione è del tutto analoga a quelle precedenti (Proposizione 1.1.2, Proposizione 1.1.3 e Proposizione 1.1.5) con $Z_1, Z_2, \ldots, Z_p$ al posto di $X_1, X_2, \ldots, X_p$ e ρ al posto di Σ. □

Osservazione 1.2.3. *Come si è osservato nella situazione generale delle componenti principali si ha che anche la varianza complessiva delle variabili standardizzate è p, la somma degli elementi della diagonale della matrice ρ (1.3). Si nota quindi che:*

$$\begin{pmatrix} \text{porzione di varianza} \\ \text{totale della} \\ \text{popolazione (standardizzata)} \\ \text{dovuta alla k-esima} \\ \text{componente principale} \end{pmatrix} = \frac{\lambda_k}{p} \qquad k = 1, 2, \ldots, p$$

con λ_k autovalore di ρ.

1.3 Componenti principali di un campione

In questo paragrafo si studieranno le componenti principali di un campione di dati, si analizzano cioè p variabili delle quali si hanno n misurazioni; si considera quindi un vettore aleatorio $X' = [X_1, X_2, \ldots, X_p]$ e il suo corrispettivo set di dati che indicheremo con $x' = [x_1, x_2, \ldots, x_p]$ con x_i di dimensione n che rappresenta tutte le osservazioni che si hanno sulla i-esima variabile.

Notazione 1.3.1. *Si denota con $\tilde{x}_i'$ il vettore contenente i valori delle p variabili aleatorie di un singolo caso (l'i-esimo) degli n considerati.*

Facendo quindi attentione alla differenza di significato tra x_i e $\tilde{x}_i$ e ricordando la notazione per cui x_i è un vettore colonna e $\tilde{x}_i'$ è un vettore riga si scrive esplicitamente il set di dati x nella seguente maniera:

$$x = \begin{pmatrix} (& \tilde{x}_1' &) \\ (& \tilde{x}_2' &) \\ (& \vdots &) \\ (& \tilde{x}_n' &) \end{pmatrix} = \left(\begin{pmatrix} x_1 \end{pmatrix} \begin{pmatrix} x_2 \end{pmatrix} \cdots \begin{pmatrix} x_p \end{pmatrix} \right) = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

Si definiscono prima di proseguire alcuni utili strumenti matematici:

Definizione 1.3.2. ***(Media Campionaria)***
Dato x come in precedenza si denota il vettore media dell'i-esima componente (o campionario) nel seguente modo: $\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ji}$, con $i = 1, 2, \ldots, p$.

Definizione 1.3.3. ***(Matrice di Covarianza e Correlazione Campionarie)***
Si definisce la matrice di covarianza campionaria S come segue:

$$S = \begin{pmatrix} s_{11} & s_{12} & \ldots & s_{1p} \\ s_{12} & s_{22} & \ldots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{2p} & \ldots & s_{pp} \end{pmatrix} \qquad s_{ik} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \quad con \quad i, k < p$$

Inoltre si definisce anche la matrice di correlazione campionaria R:

$$R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1p} \\ r_{12} & r_{22} & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \dots & r_{pp} \end{pmatrix} \qquad r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}s_{kk}}}.$$

L'obbiettivo è costruire delle combinazioni non correlate e che massimizzino la variabilità che ognuna delle p variabili ha nel campione considerato, queste combinazioni vengono chiamate le componenti principali del campione.
Considerando le definizioni precedenti e il vettore $a_1' = [a_{11}, a_{12}, \dots, a_{1p}]$, si ottengono n generiche combinazioni lineari:

$$a_1'\tilde{x}_j = a_{11}x_{j1} + a_{12}x_{j2} + \cdots + a_{1p}x_{jp} \qquad con \quad j = 1, 2, \dots, n$$

con la media campionaria che è $a_1'\bar{x}$ mentre la varianza del campione è $a_1'Sa_1$. Inoltre per una coppia di combinazioni lineari $(a_1'\tilde{x}_j, a_2'\tilde{x}_j)$ la covarianza campionaria risulta essere $a_1'Sa_2$.
Le componenti principali sono definite come quelle combinazioni aventi massima varianza campionaria. Come fatto precedentemente per le popolazioni, per evitare problemi di indeterminazione si pone $a_i'a_i = 1$, cioè che il vettore abbia norma unitaria, quindi:

Definizione 1.3.4. ***(Componenti Principali del Campione)***
La prima componente principale del campione è la combinazione lineare $a_1'\tilde{x}_j$ che massimizza la varianza campionaria di $a_1'\tilde{x}_j$ tale che $a_1'a_1 = 1$.
La seconda componente principale del campione è la combinazione lineare $a_2'\tilde{x}_j$ che massimizza la varianza campionaria di $a_2'\tilde{x}_j$ tale che $a_2'a_2 = 1$ e $Cov(a_1'\tilde{x}_j, a_2'\tilde{x}_j) = 0$.
In generale:
La i-esima componente principale del campione è la combinazione lineare $a_i'\tilde{x}_j$ che massimizza la varianza campionaria di $a_i'\tilde{x}_j$ tale che $a_i'a_i = 1$ e $Cov(a_k'\tilde{x}_j, a_i', \tilde{x}_j) = 0$ per $k < i$.

Affermare che a_1 venga scelto in modo che massimizzi la varianza campionaria significa che rende massimo $a_1' S a_1$ e quindi $\max_{a=||1||, a\neq 0} \frac{a_1' S a_1}{a_1' a_1}$. Per la Proposizione 0.0.6 si è osservato 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 la varianza di $a_i' \tilde{x}_j$ e saranno perpendicolari agli autovettori $\hat{e}_k$ per ogni $k < i$.
In modo analogo a quanto fatto per le componenti principali delle popolazioni si ottengono risultanti anche per quelle del campione:

Proposizione 1.3.5. ***(Risultati sulle Componenti Principali del Campione)*** *Sia S la matrice $p \times p$ di covarianza campionaria con autocoppie $(\hat{\lambda}_1, \hat{e}_1), (\hat{\lambda}_2, \hat{e}_2), \ldots, (\hat{\lambda}_p, \hat{e}_p)$ con $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$.*
Allora la i-esima componente principale è:

$$\hat{y}_i = \hat{e}_i' x = \hat{e}_{i1} x_1 + \hat{e}_{i2} x_2 + \cdots + \hat{e}_{ip} x_p \quad \textit{con } i=1,2,\ldots,p.$$

con $x = [x_1, x_2, \ldots, x_p]$ un generico set di dati.
Inoltre:

$$\textit{Varianza campionaria}(\hat{y}_i) = \hat{\lambda}_i, \qquad \textit{con} \quad i = 1, 2, \ldots, p.$$

$$\textit{Covarianza campionaria}(\hat{y}_k, \hat{y}_i) = 0, \quad \textit{con } i \neq k.$$

$$\textit{Varianza totale del campione} = \sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p$$

e

$$r_{\hat{y}_i, x_k} = \frac{\hat{e}_{ik} \sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}} \quad \textit{con } i, k = 1, \ldots, p.$$

Osservazione 1.3.6. *Si faccia attenzione anche in questo caso alla denotazione delle componenti principali del campione con $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_p$ a prescindere dal fatto che siano ottenute dalla matrice di covarianza S o di correlazione R, infatti, di solito, le componenti costruite da S e da R non sono uguali. In generale però sarà chiaro dal contesto quale matrice viene utilizzata, d'altra parte adottare una singola notazione riesce molto più comodo. Per questo motivo si mantengono singole anche le notazioni per i vettori dei coefficienti $\hat{e}_i$ e le varianze $\hat{\lambda}_i$.*

Spesso nello studio delle componenti principali le osservazioni $\tilde{x}_j$ sono centrate in $\hat{x}$ in modo che la media campionaria di ogni componente sia nulla, questo non ha effetti rilevanti sulla matrice di covarianza S. Infatti si ha per ogni vettore dell'osservazioni x:

$$\hat{y}_i = \hat{e}_i'(x - \bar{x}) \quad \text{con } i = 1, 2, \ldots, p, \tag{1.4}$$

allora, considerando il valore della i-esima componente:

$$\bar{\hat{y}}_i = \frac{1}{n}\sum_{j=1}^{n} \hat{e}_i'(x_j - \bar{x}) = \frac{1}{n}\hat{e}_i' \sum_{j=1}^{n} (x_j - \bar{x}) = \frac{1}{n}\hat{e}_i'\mathbf{0} = 0.$$

1.4 Standardizzazione delle componenti principali per un campione

Anche nel caso di campioni, nelle componenti principali bisogna fare attenzione alle unità di misura, quindi, in caso di grandi differenze fra i dati, come si è visto per lo studio delle componenti principali di una popolazione, si devono standardizzare le variabili per l'analisi delle componenti principali di un campione. Il procedimento è del tutto analogo, la variabile z_j sarà ottenuta nel seguente modo:

$$z_j = D^{-\frac{1}{2}}(x_j - \bar{x}) = \begin{pmatrix} \frac{x_{j1}-\bar{x}_1}{\sqrt{s_{11}}} \\ \frac{x_{j2}-\bar{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \frac{x_{jp}-\bar{x}_p}{\sqrt{s_{pp}}} \end{pmatrix} \quad j = 1, 2, \ldots, n.$$

con D definita come la seguente matrice diagonale:

$$D = \begin{pmatrix} s_{11} & 0 & \ldots & 0 \\ 0 & s_{22} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & s_{pp} \end{pmatrix},$$

Si scrive per intero la matrice (di dimensione n × p) standardizzata delle osservazioni Z:

$$Z = \begin{pmatrix} z_1' \\ z_2' \\ \vdots \\ z_n' \end{pmatrix} = \begin{pmatrix} z_{11} & z_{12} & \ldots & z_{1p} \\ z_{21} & z_{22} & \ldots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \ldots & z_{np} \end{pmatrix} = \begin{pmatrix} \frac{x_{11}-\bar{x}_1}{\sqrt{s_{11}}} & \frac{x_{12}-\bar{x}_1}{\sqrt{s_{22}}} & \ldots & \frac{x_{1p}-\bar{x}_1}{\sqrt{s_{pp}}} \\ \frac{x_{21}-\bar{x}_2}{\sqrt{s_{11}}} & \frac{x_{22}-\bar{x}_2}{\sqrt{s_{22}}} & \ldots & \frac{x_{2p}-\bar{x}_2}{\sqrt{s_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{n1}-\bar{x}_n}{\sqrt{s_{11}}} & \frac{x_{n2}-\bar{x}_n}{\sqrt{s_{22}}} & \ldots & \frac{x_{np}-\bar{x}_n}{\sqrt{s_{pp}}} \end{pmatrix}.$$

Dunque il vettore della media campionaria ottenuto da queste nuove variabili è:

$$\bar{z} = \frac{1}{n}(\mathbf{1}'Z)' = \frac{1}{n}Z'\mathbf{1} = \frac{1}{n}\begin{pmatrix} \sum_{j=1}^{n} \frac{x_{j1}-\bar{x}_1}{\sqrt{s_{11}}} \\ \sum_{j=1}^{n} \frac{x_{j2}-\bar{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \sum_{j=1}^{n} \frac{x_{jp}-\bar{x}_p}{\sqrt{s_{pp}}} \end{pmatrix} = 0.$$

Mentre la matrice di covarianza del campione, che si dimostra essere uguale alla matrice di correlazione R delle variabili iniziali, è:

$$\begin{aligned} S_z &= \frac{1}{n-1}(Z - \frac{1}{n}\mathbf{1}\mathbf{1}'Z)'(Z - \frac{1}{n}\mathbf{1}\mathbf{1}'Z) = \\ &= \frac{1}{n-1}(Z - \mathbf{1}\bar{z}')'(Z - \mathbf{1}\bar{z}') = \\ &= \frac{1}{n-1}Z'Z = \\ &= \frac{1}{n-1}\begin{pmatrix} \frac{(n-1)s_{11}}{s_{11}} & \frac{(n-1)s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \cdots & \frac{(n-1)s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} \\ \frac{(n-1)s_{21}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \frac{(n-1)s_{22}}{s_{22}} & \cdots & \frac{(n-1)s_{2p}}{\sqrt{s_{22}}\sqrt{s_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{(n-1)s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} & \frac{(n-1)s_{2p}}{\sqrt{s_{22}}\sqrt{s_{pp}}} & \cdots & \frac{(n-1)s_{pp}}{s_{pp}} \end{pmatrix} = R. \end{aligned}$$

É evidente quindi che per ottenere le componenti principali dalle variabili standardizzate bisogna considerare la matrice di correlazione R. Inoltre, essendo per costruzione le osservazioni già centrate, non c'è bisogno di scriverle come in (1.4).
Si riassume il tutto, analogamente ai paragrafi precedenti, con una proposizione:

Proposizione 1.4.1. ***(Componenti Principali per Campioni Standardizzati)*** *Sia $z' = [z_1, z_2, \ldots, z_p]$ un vettore di dati standardizzati con matrice di covarianza campionaria R e siano $(\hat{\lambda}_i, \hat{e}_i)$ le autocoppie di R con $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$.*
Allora la i-esima componente principale delle variabili standardizzate per un campione è:

$$\hat{y}_i = \hat{e}_i' z = \hat{e}_{i1}z_1 + \hat{e}_{i2}z_2 + \cdots + \hat{e}_{ip}z_p \qquad con \quad i = 1, \ldots, p$$

Inoltre:

$$Varianza\ campionaria(\hat{y}_i) = \hat{\lambda}_i \quad con\ i = 1, 2, \ldots, p,$$

$$Covarianza\ campionaria(\hat{y}_i, \hat{y}_k) = 0 \quad con\ i \neq k.$$

$$Varianza\ totale\ del\ campione\ standardizzato = tr(R) = p = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p \quad (1.5)$$

e

$$r_{\hat{y}_i, x_k} = \hat{e}_{ik}\sqrt{\hat{\lambda}_i} \quad con\ i, k = 1, 2, \ldots, p.$$

quindi, usando la (1.5):

$$\begin{pmatrix} \textit{porzione di varianza totale} \\ \textit{del campione standardizzatao} \\ \textit{dovuta alla k-esima} \\ \textit{componente principale} \end{pmatrix} = \frac{\hat{\lambda}_k}{\hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p} = \frac{\hat{\lambda}_k}{p} \qquad k = 1, 2, \ldots, p$$

Capitolo 2

Aspetti Geometrici della PCA

2.1 Componenti principali e distribuzione multivariata

Si è mostrato che algebricamente le componenti principali sono combinazioni lineari con certe proprietà, si studia adesso il loro aspetto geometrico. Geometricamente, queste combinazioni lineari rappresentano la scelta di nuove coordinate di riferimento ottenute dalla rotazioni di quelle originali che avevano come base $X_1, X_2, \ldots, X_p$. Il nuovo sistema rappresenta le direzioni con la massima variabilità e fornisce una descrizione più semplice e più facilmente utilizzabile della covarianza.
La loro struttura non necessita l'ipotesi di una distribuzione normale multivariata, d'altra parte però nel caso specifico di componenti principali derivate da una distribuzione normale multivariata si ottengono utili informazioni nel caso di ellisoidi a densità costante, inoltre in questo caso possono essere fatte inferenze sul campione , come sarà mostrato successivamente. Si supponga quindi di avere un vettore aleatorio di distribuzione multivariata $X \sim \mathcal{N}_p(\mu, \Sigma)$.

Proposizione 2.1.1. ***(Grafico Distribuzione Normale p-Dimensionale)***
I grafici della distribuzione normale p-dimensionale $\mathcal{N}_p(\mu, \Sigma)$ con densità costante sono ellissoidi definiti da $x \in \mathbb{R}^p$ tale che:

$$(x - \mu)'\Sigma^{-1}(x - \mu) = c^2$$

questi ellissoidi sono centrati in μ *e con assi rispettivamente* $c\sqrt{\lambda_i}e_i$ *e* $-c\sqrt{\lambda_i}e_i$ *dove* (λ_i, e_i)*, con* $i = 1, 2, \ldots, p$*, sono le autocoppie di* Σ*.*

Dimostrazione. Si ponga la funzione di densità della distribuzione uguale a costante:

$$f(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} e^{\frac{-(X-\mu)'\Sigma^{-1}(X-\mu)}{2}} = c_1$$

allora:

$$e^{\frac{-(X-\mu)'\Sigma^{-1}(X-\mu)}{2}} = \sqrt{(2\pi)^p|\Sigma|}c_1$$

$$(X-\mu)'\Sigma^{-1}(X-\mu) = -2\ln\sqrt{(2\pi)^p|\Sigma|}c_1,$$

denotando $-2\ln\sqrt{(2\pi)^p|\Sigma|}c_1 = c^2$, considerando $\Sigma = Q\Lambda Q'$ la decomposizione spettrale di Σ con Q ortogonale e $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$ dove λ_i sono gli autovalori di Σ, con il cambio di variabili $\hat{\mu} = Q'\mu$ e $\hat{x} = Q'x$ si ottiene:

$$(x-\mu)'Q\Lambda^{-1}Q'(x-\mu) = c^2$$

$$(\hat{x}-\hat{\mu})'\Lambda^{-1}(\hat{x}-\hat{\mu}) = c^2$$

che è un'ellissoide della seguente forma:

$$\frac{(\hat{x}_1-\hat{\mu}_1)^2}{\lambda_1} + \frac{(\hat{x}_2-\hat{\mu}_2)^2}{\lambda_2} + \cdots + \frac{(\hat{x}_p-\hat{\mu}_p)^2}{\lambda_p} = c^2$$

□

Un punto giacente nell'i-esimo asse dell'ellissoide avrà le coordinate proporzionali al vettore $e_i' = [e_{i1}, e_{i2}, \ldots, e_{ip}]$ nel sistema che ha origine in μ e gli assi paralleli alle direzioni $x_1, x_2, \ldots, x_p$ del sistema iniziale. Senza perdita di generalità si impone $\mu = 0$, infatti un vettore X con distribuzione normale può essere sempre traslato in un vettore $W = X - \mu$ sempre con distribuzione normale e con $Cov(W) = 0$, d'altronde $Cov(X) = Cov(W)$.
Per quanto detto nella Proposizione 0.0.8, con $A = \Sigma^{-1}$, si può scrivere:

$$c^2 = x'\Sigma^{-1}x = \frac{1}{\lambda_1}(e_1'x)^2 + \frac{1}{\lambda_2}(e_2'x)^2 + \ldots, \frac{1}{\lambda_p}(e_p'x)^2$$

dove $e_1'x$, $e_2'x$, $\ldots$, $e_p'x$ sono identificati con le componenti principali di X.
Poste quindi le componenti principali $y_1 = e_1'x$, $y_2 = e_2'x$, $\ldots$, $y_p = e_p'x$ si scrive:

$$c^2 = \frac{1}{\lambda_1}y_1^2 + \frac{1}{\lambda_2}y_2^2 + \cdots + \frac{1}{\lambda_p}y_p^2$$

Quest'ultima equazione definisce un ellissoide (poichè si è supposto $\lambda_1, \lambda_2, \ldots, \lambda_p$ positivi) con assi $y_1, y_2, \ldots, y_p$ corrispondenti rispettivamente alle direzioni di $e_1, e_2, \ldots, e_p$. Se λ_1 è l'autovalore maggiore, allora l'asse maggiore dell'ellissoide giace sulla direzione di e_1; i rimanenti assi minori invece giacciono nelle direzioni definite da $e_2, \ldots, e_p$.
Si è fatto quindi notare, per riassumere, che le componenti principali $y_1 = e_1'x$, $y_2 = e_2'x$, $\ldots$, $y_p = e_p'x$ giacciono sugli assi dell'ellissoide a densità constante. Perciò, preso un qualsiasi punto nell'i-esimo asse dell'elissoide, esso ha le coordinate di X proporzionali a $e_i' = [e_{i1}, e_{i2}, \ldots, e_{ip}]$ e necessariamente le coordinate della componente principale sono nella forma $[0, \ldots, 0, y_i, 0, \ldots, 0]$.

Esempio 2.1. (Ellissoide)
Un elissoide a densità costante e le componenti principali per una variabile aleatoria X di distribuzione normale bidimensionale con $\mu = 0$ sono mostrate in Figura 2.1. Si evince che le componenti principali sono ottenute ruotando le coordinate iniziali di un angolo θ che le porta negli assi dell'ellissoide. Questo risultato continua a valere per $p > 2$ dimensioni.

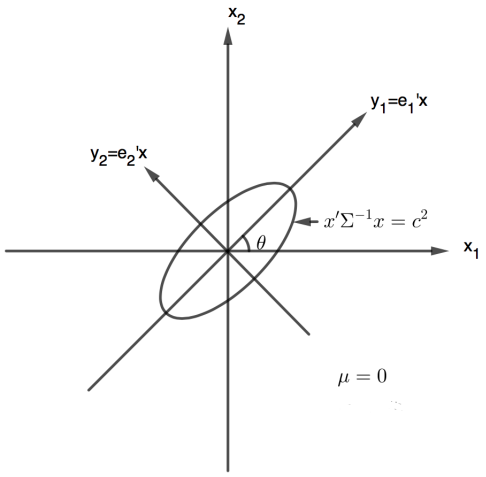


Figura 2.1: É rappresentato l'ellissoide di densità costante, $x'\Sigma^{-1}x = c^2$, e le componenti principali y_1, y_2 per una variabile aleatoria X bidimensionale con valore atteso $\mu = 0$

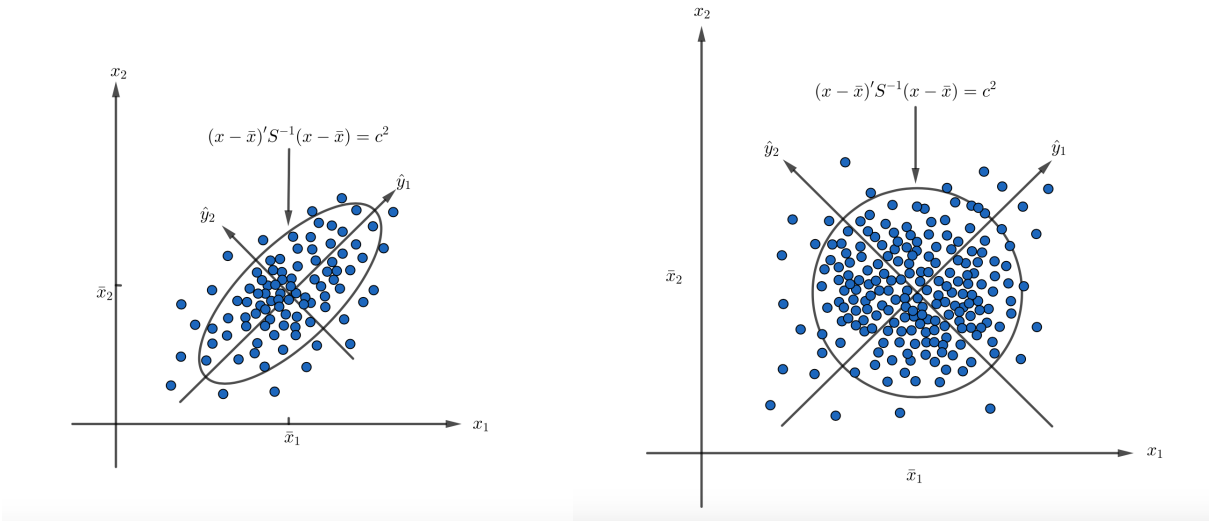
2.2 Interpretazione delle componenti principali del campione

Lo studio dei grafici dipendenti dai dati serve a rendere più tangibile la parte teorica e per localizzare casi singolari o al contrario identificare la zona con maggior concentrazione di dati che per esempio in ambito economico può rappresentare il target pubblicitario al quale siamo interessati. Osservare i grafici può aiutare anche a capire meglio la teoria e il perchè in alcuni casi bisognerà avere particolari accorgimenti.
Le componenti principali del campione hanno diverse interpretazioni, per prima cosa si supponga che X abbia distribuzione normale multivariata $\mathcal{N}_p(\mu, \Sigma)$, ipotesi sperimentalmente non restrittiva. Le componenti principali del campione $\hat{y}_i = \hat{e}_i'(x - \bar{x})$ sono ottenute dalle componenti principali della popolazione $Y_i = e_i'(X - \mu)$ le quali hanno distribuzione $\mathcal{N}_p(\mathbf{0}, \Lambda)$ con Λ matrice diagonale di elementi $\lambda_1, \lambda_2, \ldots, \lambda_p$, autovalori delle autocoppie (λ_i, e_i) di Σ infatti il valore atteso di Y_i, con $i = 1, 2, \ldots, p$, è uguale a 0 poichè $E[X - \mu] = \mu - \mu = 0$.
Inoltre, dalle variabili campionarie x_j, si può sostituire μ con $\bar{x}$ e Σ con S, infatti se S è definita positiva il bordo del grafico si ottiene con vettori x di dimensione p $\times$ 1 che soddisfano:

$$(x - \bar{x})' S^{-1} (x - \bar{x}) = c^2 \tag{2.1}$$

cioè che approssimano il grafico della densità normale, $(x - \mu)'\Sigma^{-1}(x - \mu) = c^2$. Questa approssimazione può essere vista graficamente attraverso uno "scatter plot" che indica la distribuzione normale che genera i dati.
Geometricamente, le osservazioni possono essere inserite come n punti in un grafico p-dimensionale per poi essere espresse nelle nuove coordinate che coincidono con gli assi dell'ellissoide descritto da (2.1), esso sarà centrato in $\bar{x}$ e i suoi assi, la cui lunghezza è proporzionale a $\sqrt{\hat{\lambda}_i}$ *con* $i = 1, 2, \ldots, p$ ($\hat{\lambda}_i$ autovalori in ordine decrescente di S), coincidono con gli autovettori di S.
Poichè $\hat{e}_i$ ha lunghezza unitaria, il valore assoluto della i-esima componente principale, $|\hat{y}_i| = |\hat{e}_i'(x - \bar{x})|$, indica la lunghezza della proiezione del vettore $(x - \bar{x})$ nel vettore $\hat{e}_i$.
Riassumendo, le componenti principali $\hat{y}_i = \hat{e}_i'(x - \bar{x})$ con $i = 1, 2, \ldots, p$ giacciono lungo gli assi dell'ellissoide e il loro valore assoluto è dato dalla lunghezza della proiezione di

$(x - \bar{x})$ in direzione di $\hat{e}_i$. Ne consegue quindi che le componenti principali del campione possono essere viste come il risultato di una traslazione dalle coordinate iniziali a quelle centrate in $\bar{x}$ e di una rotazione che porta gli assi iniziali negli assi dell'ellissoide.



(a) $\hat{\lambda}_1 > \hat{\lambda}_2$ (b) $|\hat{\lambda}_1| = |\hat{\lambda}_2|$

Figura 2.2: Componenti Principali del Campione e Ellissi a Densità Costante

L'interpretazione geometrica delle componenti principali del campione è illustrata in Figura 2.2

In particolare si ha nel caso della Figura 2.2(a) un'ellisse di densità costante centrata in $\bar{x} = (\bar{x}_1, \bar{x}_2)$ con $\hat{\lambda}_1 > \hat{\lambda}_2$. Le componenti principali sono ben determinate e giacciono lungo gli assi dell'ellisse (nelle direzioni perpendicolari del massimo della varianza campionaria).

In Figura 2.2(b) viene mostrato il caso in cui i due autovalori sono uguali, $\hat{\lambda}_1 = \hat{\lambda}_2$. Si è ora in un caso specifico di ellisse, la circonferenza, nel quale giungono problemi: gli assi non sono unicamente determinati, possono essere qualsiasi due vettori perpendicolari del piano, compresi quelli iniziali, di conseguenza le stesse componenti principali del campione che derivano direttamente dagli assi non sono uniche. Quando il bordo è approssimabile ad una circonferenza, quindi quando il valore degli autovalori di S è vicino ad essere uguale, le componenti principali del campione sono omogene in tutte le direzioni quindi non è possibile esprimere i dati in meno di p dimensioni.

Al contrario se gli ultimi autovalori sono sufficientemente piccoli rispetto agli altri autovalori allora si possono ignorare e considerare i dati adeguatamente approssimati nello spazio generato dalle altre componenti principali del campione corrispondenti a quegli autovalori più significativi.

2.3 Il numero delle componenti principali

Rimane da discutere, per completare il procedimento, la scelta del numero delle componenti principali da considerare, infatti uno degli scopi della PCA è quello di sintetizzare p variabili $(X_1, X_2, \ldots, X_p)$ in un numero k di variabili, con $k < p$. Questo passaggio è di fondamentale importanza, con una scelta opporta si può avere una minima perdita di informazione con una grossa riduzione di dati. Non vi è una procedura standard poichè ci sono molteplici fattori da tenere in considerazione: bisogna per esempio considerare la quantità di varianza totale che dipende da quelle variabili o l'interpretazione e la coerenza che queste hanno per lo specifico modello che si voleva studiare.
I metodi più comuni per la selezione delle componenti principali possono essere riassunti in due categorie (delle quali in questo paragrafo si approfondirà con un esempio la seconda):

1) Percentuale di Varianza Totale:
Ricordando quanto detto nell'Osservazione 1.1.4 e nell'Osservazione 1.2.3 si calcola la porzione di varianza complessiva dovute alle prime k componenti principali, nel caso questa descriva più dell'80%, si può considerare il problema ottimamente descritto.

2) Scree plot:
Un utile aiuto grafico per determinare un appropriato numero di componenti principali è lo "scree plot".

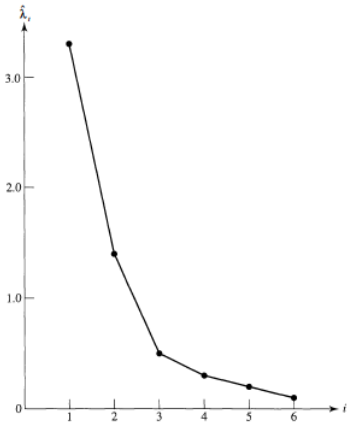


Figura 2.3: In Figura è riportato un esempio di scree plot in un piano cartesiano $i \times \hat{\lambda}_i$.

Si considerano gli autovalori relativi alle componenti principali ordinati dal maggiore al minore e si posizionano nell'asse delle ordinate mentre nell'asse delle ascisse vengono segnati i numeri delle componenti principali e infine si traccia il grafico.
Al fine di decidere il corretto numero di componenti da considerare bisogna guardare nel grafico dove si trova una curva a "gomito", ossia dove il valore degli autovalori non differenzia di tanto da quello precedente, e considerare unicamente le variabili prima di quel valore, in quanto 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 , o inferiore, ordine di grandezza, da poter essere trascurati.
In Figura 2.3 vi è la presenza di un gomito in prossimità di $i = 3$, questo significa che dopo $\hat{\lambda}_2$, gli autovalori sono relativamente piccoli e non differiscono di tanto dall'autovalore precedente. Nella fattispecie, quindi, si possono indicare 2 (o al massimo 3) componenti principali necessarie per riscrivere il sistema di origine di 6 componenti. Notiamo quindi una notevole riduzione delle variabili nonostante il semplice esempio.

Capitolo 3

Applicazioni

3.1 PCA di un esempio in ambito sportivo

Viene riportata ora la risoluzione di un problema attraverso i passi fondamentali della PCA per poi analizzare quanto ottenuto; i dati presi in considerazione sono le osservazioni delle prestazione da parte di 55 nazioni nelle olimpiadi nell'ambito dell'atletica leggera che comprende 8 categorie (che rappresentano le 8 variabili del problema): i 100m, i 200m e i 400m con i tempi calcolati in secondi e gli 800m, i 1500 metri, i 5000 metri, i 10000m e la maratona calcolati in minuti.
Segue la tabella dei dati:

Nazione	**100m(s)**	**200m(s)**	**400m(s)**	**800m(min)**	**1500m(min)**	**5000m(min)**	**10000m(min)**	**Maratona(min)**
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
Birmania	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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Svizzera	10.37	20.46	45.78	1.78	3.55	13.22	27.91	131.20
Taiwan	10.59	21.29	46.80	1.79	3.77	14.07	30.07	139.27
Thailandia	10.39	21.09	47.91	1.83	3.84	15.23	32.56	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

L'obbiettivo è quello di ridurre la mole di dati e di analizzare eventuali analogie tra stati o tra variabili; per prima cosa si standardizza il set di dati, poi si calcola la sua matrice di correlazione, per poi procedere nel calcolo degli autovettori e autovalori e infine calcolarsi le nuove componenti principali, come studiato nella parte algebrica nel capitolo 1.

```
Z=zscore(X);      % Z  indica la matrice 55x8 dei dati standardizzati
R=corr(Z);        % R  indica la matrice 8x8 di correlazione di Z
[V,D]=eig(R);     % V  indica la matrice 8x8 con gli autovettori di R
                  % D  indica la matrice 8x8 che sulla diagonale ha
                  % i relativi autovalori
p=trace(D);       % p  indica la traccia di D, serve per calcolare la
                  % percentuale della varianza totale per la scelta
                  % del numero delle componenti principali
%si nota che le prime due componenti esprimono il 93,75%
v1=V(:,1);        %  v1 indica il primo autovettore
v2=V(:,2);        %  v2 indica il secondo autovettore
Y1=v1'*(Z');      %  viene calcolata la prima componente principale
Y2=v2'*(Z');      %  viene calcolata la seconda componente principale
```

Nel seguente codice Matlab con X si è indicato il set di dati di partenza, attraverso il comando *zscore* si ottiene invece il set di dati standardizzato che dato come argomento alla funzione *corr* restituisce la matrice di correlazione.
Per trovare gli autovettori e gli autovalori si utilizza il comando *eig*, esso restituisce la matrice diagonale D e la matrice degli autovettori V così fatte:

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

$$V = \begin{pmatrix} -0.318 & 0.567 & -0.332 & -0.128 & 0.105 & 0.263 & 0.594 & 0.136 \\ -0.337 & 0.462 & -0.361 & 0.259 & -0.096 & -0.154 & -0.6561 & -0.113 \\ -0.356 & 0.248 & 0.560 & -0.652 & -0.0001 & -0.218 & -0.157 & -0.003 \\ -0.369 & 0.012 & 0.532 & 0.480 & -0.0380 & 0.540 & 0.0147 & -0.238 \\ -0.373 & -0.140 & 0.153 & 0.405 & 0.139 & -0.488 & 0.158 & 0.610 \\ -0.364 & -0.312 & -0.190 & -0.0296 & 0.547 & -0.254 & 0.141 & -0.591 \\ -0.367 & -0.307 & -0.182 & -0.080 & -0.797 & -0.133 & 0.219 & -0.177 \\ -0.342 & -0.439 & -0.263 & -0.300 & 0.158 & 0.498 & -0.315 & 0.390 \end{pmatrix}$$

Con il metodo degli autovalori descritto nella sezione 2.2 e nell'Osservazione 1.2.3 si nota che i primi due autovalori hanno una varianza che è pari al 93.75% infatti, ricordando che la traccia è uguale a p (1.5):

$$\frac{\lambda_1 + \lambda_2}{p} = \frac{6.6221 + 0.8776}{8.0000} = 93,75\%$$

Si può quindi ridurre il numero di variabili in due componenti principali che nel codice sono state calcolate nel seguente modo:

$$Y_1 = v_1' Z'$$

$$Y_2 = v_2' Z'$$

con v_1, v_2 gli autovettori relativi ai primi due autovalori, sono le prime due colonne della matrice V.

Ci si sofferma ora ad analizzare nel dettaglio questi due vettori per capire meglio il significato delle due componenti principali:

Per quanto riguarda la prima componente studiando il vettore a lei relativo, ovvero

$$v_1' = \Big(-0.318, \quad -0.337, \quad -0.356, \quad -0.369, \quad 0.373, \quad -0.364, \quad -0.367, \quad -0.342\Big)$$

si nota che i pesi dei coefficienti sono quasi uguali quindi incidono nella stessa maniera sulla componente principale. Si può dunque considerare la prima componente principale derivante da questo vettore come un indicatore generale della qualità di una nazione complessiva in tutte le categorie. Questa interpretazione è del tutto ragionevole poichè la prima componente principale tende a rappresentare la maggior parte delle informazioni,

mentre le componenti successive tentano di ottimizzare l'informazione restante.
Infatti, rispetto alla seconda componente principale, considerando il suo autovettore

$$v_2' = \Big(0.567, \quad 0.462, \quad 0.248, \quad 0.012, \quad -0.140, \quad -0.312, \quad -0.307, \quad -0.439\Big)$$

si osserva che i primi quattro valori sono positivi mentre gli altri quattro sono negativi. Ricordando che le prime quattro variabili (100m, 200m, 400m e gli 800m) possono essere considerate prove atletiche sulla velocità intesa come sprint mentre le ultime quattro (1500m, 5000m, 10000m e la maratona) sono prove atletiche sulla resistenza, la seconda componente principale sembra descrivere il contrasto che ha una nazione tra lo sprint e la resistenza. Più nel dettaglio si vede come il primo valore e l'ultimo sono in modulo i maggiori poichè possono essere considerati come i rappresentanti delle due categorie, i 100m per gli sprint e la maratona per la resistenza.
Vengono riportati di seguito il nuovo set di dati con le nazioni numerate e la sua rappresentazione attraverso un plot per poi fare le ultime osservazioni sull'esercizio:

Nazione	Y_1="qualità generale"	Y_2="contrasto sprint-resistenza"
1) Argentina	-0.2619	-0.3449
2) Australia	2.4464	-0.2162
3) Austria	0.8076	0.4869
4) Belgio	2.0413	0.2619
5) Bermuda	-0.7392	-1.7669
6) Brasile	1.5583	-0.6412
⋮	⋮	⋮
50) Taipei	-0.9505	0.0420
51) Thailandia	-2.7618	-1.6698
52) Turchia	-0.266	1.3830
53) Usa	3.4305	-1.1101
54) Russia	2.6269	-0.7570
55) Samoa	-7.2312	-1.9020

Il codice Matlab riportato successivamente mostra il plot delle nazioni in un grafico (Figura 3.1) che ha sulle ascisse la prima componente principale Y_1 e nelle ordinate la seconda componente principale Y_2.

```
A=[Y1;Y2];        % Si crea una matrice che ha come colonne le due
                  % componenti principali per visualizzare meglio i
                  % dati
labels=cellstr(num2str([1:55]')); % Questa "stringa" serve per
                                  % numerare le nazioni nel Plot
figure
plot(Y1,Y2,'b.')% Restituisce il grafico con Y1 nelle ascisse e Y2
                  % nelle ordinate
text(Y1,Y2, labels) % Si numerano i punti per identifarcli con le
                    % corrispettive nazioni
hold on           % Si evidenzia l'Usa con un asterisco rosso
plot(3.43055603805251,-1.11019112023025,'r*')
hold on           % Si evidenziano le Isole cook con un asterisco
                  % rosso
plot(-10.5556255526907,1.50876823465337,'r*')
hold on           % Si evidenzia la Costa Rica con una croce rossa
plot(-2.29664651830155,1.67064120240189,'rx')
hold on           % Si evidenzia la Repubblica Domenicana con una
                  % croce rossa
plot(-1.71488638867262,-2.44900165039329,'rx')
```

Si nota che la nazione che ha il valore più grande nelle ascisse è l'Usa rappresentata dal numero 53 (è evidenziata con un asterisco rosso), si può quindi evincere, per come si è interpretata la prima componente, che essa è la nazione che a livello generale è migliore, con ottimi risultati in tutte le competizioni, a seguire dopo l'Usa per esempio ci sono la Gran Bretagna (21), la Russia (54) e l'Italia (29); mentre invece la nazione peggiore a livello generale sono le Isole Cok (12), anch'esse evidenziate con un asterisco rosso.
Altro discorso più complicato è l'analisi rispetto alla seconda componente, si considerino le due nazioni con i casi estremi, la Costa Rica (13) e la Repubblica Domenicana (16),

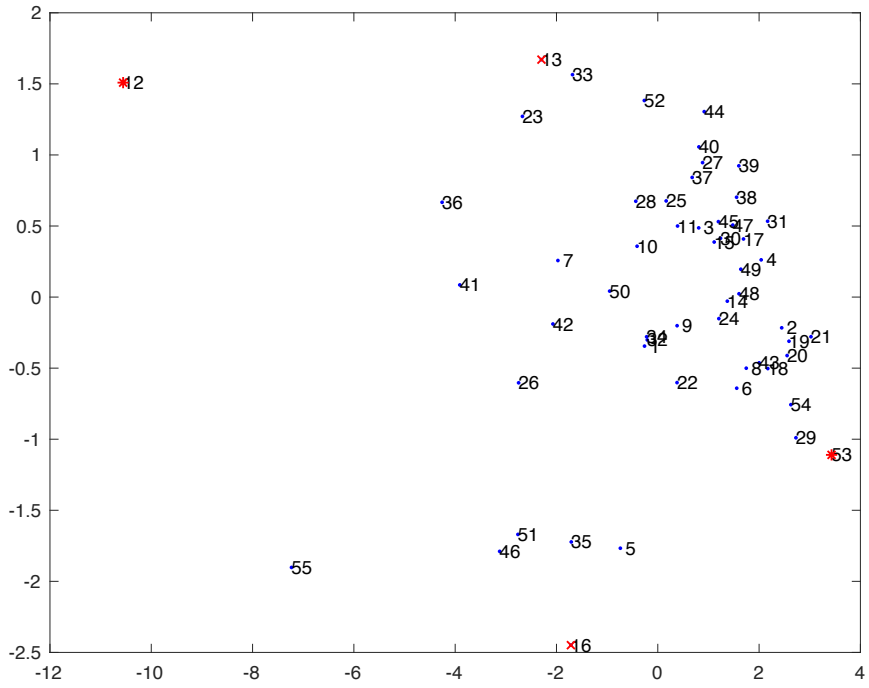


Figura 3.1: Plot delle nazioni attraverso le due componenti principali

evidenziati con una croce rossa in Figura 3.1 : hanno entrambi valori alti in valore assoluto per cui si deduce che c'è una significativa differenza tra i tempi ottenuti nelle gare di sprint rispetto a quelli ottenuti nelle gare di resistenza. Il segno positivo nel caso della Costa Rica indica che i suoi atleti corrono le lunghe distanze in meno tempo di quello che ci si potrebbe aspettare se si considerano i loro tempi nelle brevi distanze, al contrario il segno negativo nel caso della Repubblica Domenicana ci suggerisce che i suoi atleti hanno ottenuto dei tempi molto bassi nelle gare di resistenza rispetto ai tempi delle gare di sprint che comunque sono discreti e non ottimi infatti a livello generale, si guardi la prima componente, questa nazione non eccelle.
Grazie alla rappresentazione in un piano cartesiano si possono osservare due gruppi, uno in alto a destra ed uno in basso, e un *outlaier* che sono le Isole Cok (12); questo è possibile fino a tre componenti principali, si vedono anche in questo caso i vantaggi che si ottengono con la riduzione delle variabili del problema.

Si può concludere notando che se la prima componente è forse facilmente intuibile anche ad occhio umano, sicuramente non si può dire lo stesso della seconda, specialmente con una grossa quantità di dati, essa descrive un aspetto intrinseco dei dati che solamente la PCA può far notare in così pochi passaggi e in così poco tempo.

Appendice A

Algoritmo Eig

Nel capitolo 3 è stata usata la funzione *eig*, essa è indispensabile per il calcolo degli autovalori e dei corrispettivi autovettori. Tale algoritmo si basa sull'iterazione QR che a sua volta si basa sulla fattorizzazione QR.
La fattorizzazione QR dice che data una matrice rettangolare $A \in \mathbb{R}^{n\times m}$ con $n \leq m$, può essere scritta come prodotto di due matrici nel seguente modo:

$$A = QR = \Big(Q_1, Q_2\Big) \begin{pmatrix} R_1 \\ \mathbf{0} \end{pmatrix},$$

dove $R_1 \in \mathbb{R}^{m\times m}$ e Q è una matrice $n \times n$ reale ortogonale, con $Q_1 \in \mathbb{R}^{n\times m}$ e $Q_2 \in \mathbb{R}^{n\times(n-m)}$ e avente le colonne ortonormali; inoltre R_1 è una matrice triangolare superiore.
Nel caso dell'iterazione QR si suppone che A sia una matrice quadrata per poterne calcolare gli autovalori e inoltre, nel caso questa sia anche normale, i corrispettivi autovettori.

Definizione A.0.1. ***(Matrice Normale)***
Una matrice $A \in \mathbb{C}^{n\times n}$ si dice normale se $AA^H = A^HA$.

In particolare si dimostra che A è normale se e solo se è diagonalizzabile mediante una matrice unitaria, cioè $A = Q\Lambda Q^H$ con $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$.
Tornando all'iterazione QR, si ha che l'algoritmo determina una successioni di matrici $(T_k)_{k\in\mathbb{N}}$ tale che:

$$T_k = U_k^H A U_k \to T, \qquad per \quad k \to +\infty, \tag{A.1}$$

con T matrice diagonale con elementi gli autovalori di A (perchè A normale, altrimenti T sarebbe stata triangolare superiore).
Il primo passaggio per determinare la successione è porre $T_0 = A$, successivamente si inzia un ciclo for con $k = 0, 1, \ldots,$ in cui nella prima operazione viene calcolta la fattorizzazione QR della matrice quadrata T_k, $T_k = Q_k R_k$, nella seconda operazione invece viene definita la matrice T_{k+1} come prodotto dei due fattori R_k, Q_k in ordine scambiato, $T_{k+1} = R_k Q_k$. In questo modo, dato che dalla prima operazione segue che $Q^H T_k = R_k$, si ha, riscrivendo la seconda operazione, $T_{k+1} = Q_k^H T_k Q_k$.
Essendo le matrici Q_k unitarie per ogni k, le matrici T_k sono simili tra loro, inoltre si ha che:

$$T_{k+1} = Q_k^H Q_{k-1}^H T_{k-1} Q_{k-1} Q_k = \cdots = Q_k^H Q_{k-1}^H \ldots Q_0^H T_0 Q_0 \ldots Q_{k-1} Q_k = U_k^H A U_k$$

con U_k unitaria, quindi gli autovalori di A vengono mantenuti.
Si conclude il discorso con il seguente teorema che prova la (A.1).

Teorema A.0.2. *($T_k \to T$)*
Sia $A \in \mathbb{C}^{n \times n}$ con $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0$ dove λ_i con $i = 1, 2, \ldots, n$ indica un autovalore di A, allora:

$$\lim_{k \to +\infty} T_k = T,$$

dove T è triangolare superiore (diagonale se A è normale) avente come elementi gli autovalori di A.

Riguardo quindi all'utilizzio di eig si ha che $D = eig(A)$ restituisce la matrice D diagonale degli autovalori e $[V, D] = eig(A)$ restituisce anche la matrice V dei corrispettivi autovettori.

Conclusioni

In questo elaborato è stato analizzato il metodo della PCA, sia soffermadonsi sui suoi aspetti principali, quello algebrico, descritto nel Capitolo 1 e quello geometrico, esposto nel Capitolo 2, sia presentando un'applicazione di quanto appena esposto in ambito sportivo. Inizialmente sono presenti alcune nozioni matematico-statistiche necessarie per la comprensione dell'argomento.

Nel Capitolo 1 sono state introdotte le componenti principali, studiando separatamente il caso per una popolazione e per un campione. Inoltre sono stati evidenziati i problemi dovuti alla presenza di dati non omogenei, presentando, quindi, un modello efficiente per la standardizzazione delle variabili.

In seguito è stata presa in considerazione la rappresentazione geometrica delle componenti principali nel caso di variabili aleatorie con distribuzioni normali multivariate, studiando l'ellissoide descritto da queste. Nel capitolo 2 si è, inoltre, analizzato il problema della scelta del numero di componenti, evidenziandone l'importanza e descrivendo alcune delle metodologie utilizzate. Il tutto è stato supportortato con figure e grafici che accompagno la parte discorsiva.

Infine sono stati utilizzati i risultati dei capitoli precedenti per analizzare un set di dati riguardante i tempi in 8 categorie delle Olimpiadi e sono stati riportati i passaggi computazionali necessari. Questo esempio è stato fondamentale per evidenziare come l'analisi delle componenti principali abbia effettivamente ridotto la mole di dati (nel caso specifico la dimensione del problema si è ridotta di tre quarti dei dati) e abbia rilevato

aspetti latenti non visibili altrimenti.

Nonostante una minima perdita di informazioni iniziali sia inevitabile, presente in tutte le applicazioni sull'analisi dei dati, la PCA è una tecnica largamente utilizzata proprio perché limita tale perdita entro limiti accettabili. Infatti, se la scelta del numero di componenti da analizzare è fatta in maniera giudiziosa, il $trade\ off$ tra la perdita delle informazioni e la semplificazione del problema è quasi sempre a favore dell'utilizzo di questa tecnica.

Bibliografia

[1] Richard Johnson, Dean Wichern, *Applied Multivariate Statistical Analysis*, Pearson New International Edition, 2014.

[2] Valeria Simoncini, *Dispense del corso di Calcolo Numerico*, Università di Bologna, 2019.

[3] Andrea Pascucci, *Teoria della Probabilità: Variabili aleatorie e distribuzioni*, Università di Bologna, 2020.