Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE

Corso di Laurea in Matematica

REGRESSIONE LINEARE E ANALISI DELLE COMPONENTI PRINCIPALI

Tesi di Laurea in Calcolo Numerico

Relatore:
Chiar.ma Prof.ssa
Valeria Simoncini

Presentata da:
Sofia Cannizzaro

Sessione unica
Anno Accademico 2018/2019

Indice

Introduzione

Questa tesi si prefigge lo scopo di studiare la composizione di due metodi nel campo della statistica multivariata: l'analisi delle componenti principali e la regressione lineare multivariata (MLR). Si tratta di tecniche usate in ogni ambito in cui sia necessario studiare un fenomeno determinato da piu fattori, dall'economia alla medicina.

Nel Capitolo 1 viene presentata l'analisi delle componenti principali (PCA), metodo che analizza e semplifica i dati e attraverso il quale si vuole spiegare la struttura di varianza-covarianza di un insieme di p variabili, date poche combinazioni lineari di esse. Affinché lo studio statistico del campione sia attendibile è necessario avere a disposizione un numero molto elevato, n, di osservazioni per ciascuna variabile. Lo scopo è duplice: ridurre il numero di variabili necessarie a spiegare un fenomeno e renderlo di più facile interpretazione, in modo tale da poter identificare legami fra le variabili ed esprimere le stesse in modo tale da evidenziare le loro similarità e differenze. Questo è uno scopo molto difficile da raggiungere quando le dimensioni del problema sono grandi e quindi non si può sfruttare la rappresentazione grafica. È per questo che la PCA è uno strumento potente. Viene naturale pensare che siano sempre necessarie tutte le p variabili per esprimere la variabilità del sistema, ma spesso esse si possono ridurre ad un numero inferiore di componenti, k, dette componenti principali, quando queste contengono quasi la stessa quantità di informazione. In questo modo l'insieme di dati iniziale, composto da $n \times p$ misure, si può ridurre alle sole misurazioni per k variabili. Ciò comporta un notevole vantaggio sia per i costi sia per la successiva analisi dei dati. Inoltre, lo studio delle componenti principali rivela spesso relazioni tra i dati che non erano evidenti in precedenza, permettendo nuove interpretazioni: illustreremo questo fatto con un esempio

sui dati del mercato azionario (Esempio 2.1). Nonostante una minima perdita di informazioni iniziali sia inevitabile, 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, i vantaggi legati alla semplificazione del problema superano gli svantaggi. L'analisi delle componenti principali è, di solito, uno strumento intermedio per arrivare ad un risultato. Viene usata come passaggio in analisi più complesse, come la regressione multivariata, infatti, o l'analisi dei cluster

Nel Capitolo 2 viene affrontata la regressione lineare multivariata (MLR). Quando vi è un grande numero di variabili che determinano il comportamento di un'altra variabile dipendente (o risposta) è interessante studiare se esiste una relazione lineare che lega queste informazioni. A volte, data la complessità dei dati, questo legame non è evidente e non è possibile esprimerlo sotto forma di una funzione. Allora è necessario raccogliere molti dati per il campione di variabili che si intende studiare, ottenuti con osservazioni ed esperimenti, in modo da capirne il legame attraverso metodi statistici (come la regressione). Insorgono, però, problemi se le variabili iniziali sono dipendenti fra loro: in questi casi su di esse viene sfruttata la PCA e si applica la regressione direttamente sulle componenti principali, trovando modelli equivalenti.

Nel Capitolo 3 mostriamo come combinare l'analisi delle componenti principali e la regressione lineare multivariata. Si nota, infatti, che l'utilizzo della PCA nella regressione comporta dei vantaggi anche in assenza di un problema di collinearità. Utilizzare le componenti principali permette di ridurre la dimensione del problema e, in particolare, di poterlo interpretare meglio e, quindi, determinare da quali variabili, anche non esplicitamente osservabili, dipende effettivamente la risposta. Tramite un esempio si osserverà l'importanza della scelta del numero di componenti principali nell'ambito della regressione.

Richiami

Per lo studio della PCA sono richieste alcune conoscenze di base di statistica e analisi numerica, oltre che una conoscenza di base di algebra lineare. Riporto di seguito alcune definizioni e proposizioni fondamentali.

Definizione 1. *Una matrice $A \in \mathbb{R}^{n\times n}$ si dice simmetrica se è uguale alla sua trasposta, ossia se $A = A'$. Una matrice reale simmetrica si dice definita positiva se $\forall x \in \mathbb{R}^n, x \neq 0$ allora $x'Ax > 0$.*

Definizione 2. *Data una matrice $A \in \mathbb{R}^{n\times n}$ si definisce autovalore di A un numero λ, reale o complesso, per cui esista un vettore $v \neq 0$ tale per cui $Av = \lambda v$. In questo caso chiamiamo v autovettore di A e (λ, v) autocoppia di A.*

Osservazione. Osservo che se la matrice A è simmetrica allora tutti i suoi autovalori sono reali e se, in particolare, è definita positiva allora i suoi autovalori sono positivi.

Teorema 0.0.1. *Data una matrice $A \in \mathbb{R}^{n\times n}$ simmetrica allora esiste una matrice ortogonale $P \in \mathbb{R}^{n\times n}$ tale per cui $A = P\Lambda P'$, dove $\Lambda \in \mathbb{R}^{n\times n}$ è una matrice diagonale avente sulla diagonale principale gli autovalori di A.*

Definizione 3. *Sia $A \in \mathbb{R}^{n\times n}$ allora la sua traccia è definita come la somma degli elementi sulla sua diagonale principale. Indichiamo la traccia di A con $tr(A)$. Se, in particolare, A è simmetrica allora $tr(A) = \lambda_1 + \cdots + \lambda_n$, dove $\lambda_1, \ldots, \lambda_n$ sono gli autovalori di A.*

Definizione 4. *Date due variabili aleatorie $X = [x_1, x_2, \ldots, x_n]$ e $Y = [y_1, y_2, \ldots, y_n]$, siano $\mu_X = [E(X_1), \ldots, E(X_n)]$ e $\mu_Y = [E(Y_1), \ldots, E(Y_n)]$ le medie di X e Y rispettivamente, allora definisco 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 $Var(X) = Cov(X, X)$. Chiamiamo il coefficiente di correlazione fra X e Y il valore:

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

Si dice che due variabili sono non correlate se il loro coefficiente di correlazione, o equivalentemente la loro covarianza, è nullo.

Definizione 5. *La matrice di covarianza Σ del vettore di variabili $X = [X_1, \ldots, X_p]$ è la matrice simmetrica con $\Sigma_{ij} = Cov(X_i, X_j)$, $i, j = 1, \ldots, p$. La matrice di correlazione ρ è rispettivamente $\rho_{ij} = \rho(X_i, X_j)$.*

In questo elaborato verrà trattato lo studio statistico di dati che, solitamente, consistono in un un numero n di osservazioni relative a p variabili. Per esprimere questi dati vengono usati vettori di variabili $x_i = [x_{i1}, \ldots, x_{ip}]$ con $i = 1, \ldots, n$ che rappresentano i valori della i-esima osservazione per ogni variabile. I dati vengono raccolti in una matrice X in cui le n righe consistono nei vettori x_i:

$$X = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Definizione 6. *Dato un set di dati espresso dalla matrice X come definita in precedenza e sia $\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ji}$, con $i = 1, \ldots, p$ la media della variabile i-esima. Allora posso definire la sua matrice di covarianza campionaria S come segue:*

$$S = \begin{bmatrix} s_{11} & s_{12} & \ldots & s_{1p} \\ s_{12} & s_{22} & \ldots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{2p} & \ldots & s_{pp} \end{bmatrix} \qquad \left\{ s_{ik} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \right\}.$$

Posso definire anche la matrice di correlazione campionaria R:

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

Proposizione 1. *Dato un vettore aleatorio* $X = [X_1, \dots, X_p]$ *con matrice di covarianza* Σ_X *ed una matrice* $A \in \mathbb{R}^{p \times p}$*, 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* $Var(\mathbf{c}'X') = \mathbf{c}'\Sigma_X\mathbf{c}$*.*

Proposizione 2. *Sia* $A \in \mathbb{R}^{n \times n}$ *matrice simmetrica definita positiva e sia* λ_{max} *il suo autovalore massimo, allora:*

$$\lambda_{max} = \max_{0 \neq x \in \mathbb{R}^n} \frac{x'Ax}{x'x}. \tag{1}$$

In particolare, sia v *un autovettore di* λ_{max}*, allora* $\frac{v'Av}{v'v} = \frac{\lambda_{max}v'v}{v'v} = \lambda_{max}$*.*

Capitolo 1

Analisi delle componenti principali

1.1 Componenti principali di popolazioni

Siano $X_1, X_2, \ldots, X_p$ variabili aleatorie, l'analisi delle componenti principali permette di ridurre il numero di variabili necessarie, semplificandone lo studio.
Algebricamente le componenti principali consistono in una specifica scelta di combinazioni lineari delle variabili iniziali, presa in modo che venga massimizzata la varianza; geometricamente, invece, se considero $X_1, X_2, \ldots, X_p$ come basi di un sistema di riferimento, il procedimento consiste nel ruotare questo sistema ottenendo nuove coordinate.

Le componenti principali dipendono solo dalla matrice di covarianza Σ (o da quella di correlazione ρ) e non richiedono l'ipotesi di una distribuzione normale, anche se in questo specifico caso possono essere dedotti ulteriori risultati.
Sia $X = [X_1, X_2, \ldots, X_p]$ un vettore aleatorio con matrice di covarianza Σ, i cui autovalori sono $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$; considero delle generiche combinazioni lineari in cui $a_i' = [a_{i1}, \ldots, a_{ip}]$ con $i = 1, \ldots, p$:

$$\begin{aligned} Y_1 &= a_1' X' = a_{11}X_1 + \cdots + a_{1p}X_p, \\ Y_2 &= a_2' X' = a_{21}X_1 + \cdots + a_{2p}X_p, \\ &\vdots \\ Y_p &= a_p' X' = a_{p1}X_1 + \cdots + a_{pp}X_p. \end{aligned} \tag{1.1}$$

Allora ne consegue, utilizzando la Proposizione 1, che:

$$Var(Y_i) = a_i' \Sigma a_i \quad \text{con } i = 1, \ldots, p, \tag{1.2}$$

$$Cov(Y_i, Y_j) = a_i' \Sigma a_j \quad \text{con } i, j = 1, \ldots, p. \tag{1.3}$$

Le componenti principali consistono nella scelta delle combinazioni lineari, $Y_1, \ldots, Y_p$, che massimizzano la varianza di ciascuna e che siano scorrelate fra loro; poiché la varianza aumenta con l'aumentare in modulo di a_i considero solo i vettori per cui $a_i' a_i = 1$.

- Prima componente principale: combinazione lineare $a_1' X'$ che massimizza $Var(a_1' X')$ con $a_1' a_1 = 1$.
- Seconda componente principale: combinazione lineare $a_2' X'$ che massimizza $Var(a_2' X')$ con $a_2' a_2 = 1$ e per cui $Cov(a_1' X', a_2' X') = 0$.
  $\vdots$
- i-esima componente principale: combinazione lineare $a_i' X'$ che massimizza $Var(a_i' X')$ con $a_i' a_i = 1$ e per cui $Cov(a_j' X', a_i' X') = 0 \ \forall j < i$.

1.2 Alcuni risultati

Lemma 1.2.1. *Sia $X = [X_1, \ldots, X_p]$ un vettore aleatorio con matrice di covarianza Σ. Siano $(\lambda_1, e_1), \ldots, (\lambda_p, e_p)$ 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:*

$$Y_i = e_i' X' = e_{i1} X_1 + \cdots + e_{ip} X_p. \tag{1.4}$$

In particolare:

$$\begin{aligned} Var(Y_i) &= e_i' \Sigma e_i \quad \textit{con } i = 1, \ldots, p, \\ Cov(Y_i, Y_k) &= e_i' \Sigma e_k \quad \textit{con } i \neq k. \end{aligned} \tag{1.5}$$

Se alcuni autovalori non sono distinti allora la scelta dei corrispettivi autovettori, e dunque delle componenti principali, non è unica.

Dimostrazione. Per i risultati della Proposizione 2 sappiamo che $\max_{a\neq 0} \frac{a'\Sigma a}{a'a} = \lambda_1$ e che questo valore è raggiunto con $a = e_1$. Quindi, scegliendo gli autovettori in modo che siano unitari:

$$\lambda_1 = \frac{e_1'\Sigma e_1}{e_1'e_1} = e_1'\Sigma e_1 = Var(Y_1).$$

Posso procedere analogamente ottenendo che, poiché $\max_{a\perp e_1,e_2,\ldots,e_k} \frac{a'\Sigma a}{a'a} = \lambda_{k+1}$ e si ottiene con $a = e_{k+1}$, allora:

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

Osserviamo che, se tutti gli autovalori di Σ sono distinti, allora gli autovettori sono ortogonali e, nel caso non lo siano, posso comunque scegliere gli autovettori associati allo stesso λ_k in maniera che lo siano. Dunque, per ogni $i \neq k$, gli autovettori e_i e e_k sono distinti e ortogonali, con $e_i'e_k = 0$. Si conclude la dimostrazione mostrando che $\forall i \neq k$:

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

Quindi le componenti principali sono scorrelate ed il teorema è dimostrato. □

Lemma 1.2.2. *Sia $X = [X_1, \ldots, X_p]$ un vettore aleatorio con matrice di covarianza Σ. Siano $(\lambda_1, e_1), \ldots, (\lambda_p, e_p)$ le autocoppie di Σ con $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ e $Y_1 = e_1'X', \ldots, Y_p = e_p'X'$ le componenti principali, allora:*

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

Dimostrazione. Per definizione $\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = tr(\Sigma)$.
Sia Λ la matrice diagonale degli autovalori di Σ e sia $P = [e_1, \ldots, e_p]$, allora P è ortogonale $(P'P = I)$ e, in particolare, $\Sigma = P'\Lambda P$. 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. Da questo lemma si può osservare che la varianza totale della popolazione, ossia $\sigma_{11}+\sigma_{22}+\cdots+\sigma_{pp}$, coincide con la traccia di Λ e che, quindi, la porzione di varianza

totale della popolazione dovuta alla k-esima componente principale coincide con $\frac{\lambda_k}{\lambda_1+\cdots+\lambda_p}$, per $k = 1, \ldots, p$. Questo è fondamentale: se la maggior parte della varianza complessiva (per esempio l'80%−90%) è attribuibile alle prime componenti principali allora si possono considerare solo queste senza avere una grossa perdita di informazione. Questo comporta un evidente vantaggio nell'analisi dei dati, avendo ridotto la dimensione del problema iniziale.

Lemma 1.2.3. *Siano $Y_1 = e_1'X', \ldots, Y_p = e_p'X'$ le componenti principali ottenute dalla matrice di covarianza Σ e $(\lambda_1, e_1), \ldots, (\lambda_p, e_p)$ le sue autocoppie, allora:*

$$\rho_{Y_i,X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \quad \textit{con } i,k = 1,\ldots,p \tag{1.6}$$

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

Dimostrazione. Definisco $a_k' = [0, \ldots, 1, \ldots, 0]$ in modo che $X_k = a_k'X'$. In particolare, sfruttando i lemmi precedenti, la Proposizione 1 ed il fatto che $\Sigma e_i = \lambda_i e_i$, si ottiene che:

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

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

Dunque, usando la definizione di coefficienti di correlazione:

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

□

Osservazione. Il coefficiente di correlazione può essere visto come un indicatore dell'importanza che una certa variabile X_k ha nella componente principale Y_i. Tuttavia questo valore non tiene conto del comportamento delle altre variabili. Per questa ragione, nonostante sia un dato utile per l'interpretazione delle componenti, alcuni statistici ritengono sia erroneo usare le correlazioni; viene ritenuto conveniente 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. Tuttavia, nella maggior parte dei casi, si verifica sperimentalmente che queste due stime non sono molto differenti fra loro, ossia che variabili con relativamente grandi correlazioni (in valore assoluto) hanno anche grandi coefficienti. Risulta, in ogni modo, preferibile tenere conto di entrambi i valori per ottenere un analisi più accurata.

1.3 Componenti principali per variabili standardizzate

I dati di partenza possono non essere omogenei, sia perché aventi unità di misura diverse sia perché le dimensioni sono di ordini di grandezza molto differenti; in questi casi le componenti principali ottenute dalla matrice di covarianza possono non essere indicative dell'effettiva relazione fra le variabili. Se una variabile ha valori significativamente più grandi delle altre influenzerà molto di più le componenti principali: per esempio, se viene cambiata l'unità di una misura di lunghezza da Km a cm (incrementando la varianza) allora questa variabile potrebbe passare da avere un piccolo impatto nelle componenti ad avere un ruolo fondamentale. Per ovviare a questo problema si possono standardizzare le variabili. Sia $\mu = [\mu_1, \ldots, \mu_p]$ il vettore media di X:

$$Z_1 = \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}}, \quad Z_2 = \frac{(X_2 - \mu_2)}{\sqrt{\sigma_{22}},} \quad \ldots \quad Z_p = \frac{(X_p - \mu_p)}{\sqrt{\sigma_{pp}}}. \tag{1.7}$$

Sia

$$V^{\frac{1}{2}} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \ldots & 0 \\ 0 & \sqrt{\sigma_{22}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sqrt{\sigma_{pp}} \end{bmatrix}$$

allora in notazione matriciale definiamo le componenti principali $Z = [Z_1, \ldots, Z_p]$ come:

$$Z' = (V^{\frac{1}{2}})^{-1}(X - \mu)'. \tag{1.8}$$

Si nota che $E[Z] = 0$ e che $Cov(Z) = (V^{\frac{1}{2}})^{-1}\Sigma(V^{\frac{1}{2}})^{-1} = \rho$.
Le componenti principali per variabili standardizzate si ottengono, quindi, cercando le autocoppie della matrice di correlazione, in particolare rimangono validi i risultati ottenuti per la matrice di covarianza.

Lemma 1.3.1. *Siano $(\lambda_1, e_1), \ldots, (\lambda_p, e_p)$ 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 $Z = [Z_1, \ldots, Z_p]$ con matrice di covarianza ρ è:*

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

Inoltre

$$\sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} Var(Z_i) = \rho \tag{1.9}$$

e

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

Osservazione. Si nota che, considerando ρ come matrice di covarianza, allora la varianza complessiva è p poiché gli elementi sulla diagonale sono unitari. Se voglio considerare la porzione di varianza totale che è dovuta alla k-esima componente questa risulterà essere

$$\frac{\lambda_k}{p}.$$

Osservazione. Infine è importante osservare che, generalmente, le componenti principali ottenute da ρ o da Σ, ossia standardizzando o meno le variabili, sono differenti. La scelta di quale matrice utilizzare dipende dal tipo di dati di cui si dispone e dal tipo di analisi che si intende effettuare.

1.4 Variazione di un campione tramite le componenti principali

In questa sezione si intendono studiare le componenti principali di un campione di dati, ossia avremo n misurazioni di ciascuna delle p variabili che componevano il vettore $X = [X_1,\ldots,X_p]$. Siamo interessati ad esprimere, con poche combinazioni lineari opportunamente scelte, la variazione di queste misurazioni sulle p variabili. Lo scopo è costruire delle combinazioni scorrelate e che massimizzino la variabilità che ognuna delle p variabili ha nel campione considerato; chiameremo queste combinazioni le componenti principali del campione.

Facendo riferimento alla notazione definita nei Richiami per lo studio di un campione

considero la matrice

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Per uno specifico campione di dati, utilizzando la Definizione 6, si possono trovare una media del campione di $\bar{x} = [\bar{x}_1, \dots, \bar{x}_p]$, una matrice di covarianza del campione S e una matrice di correlazione del campione R.
Dato $a_1' = [a_{11}, \dots, a_{1p}]$ e siano

$$a_1' x_j' = a_{11}x_{j1} + +a_{1p}x_{jp} \quad \text{con } j = 1, \dots, n$$

n generiche combinazioni lineari, allora la media campionaria è $a_1'\bar{x}'$ mentre la varianza del campione è $a_1' S a_1$ (sfrutto la Proposizione 1). Inoltre per una coppia di combinazioni lineari $(a_1'x_j', a_2'x_j')$ la covarianza campionaria risulta essere $a_1' S a_2$.
Nel definire le componenti principali, per evitare problemi di indeterminazione, mi restringo alle combinazioni per cui $a_i' a_i = 1$:

- Prima componente principale: combinazione lineare $a_1'x_j'$ che massimizza la varianza campionaria di $a_1'x_j'$ con $a_1'a_1 = 1$.
- Seconda componente principale: combinazione lineare $a_2'x_j'$ che massimizza la varianza campionaria di $a_2'x_j'$ con $a_2'a_2 = 1$ e per cui le coppie $(a_1'x_j', a_2'x_j')$ abbiano covarianza campionaria nulla.

  $\vdots$ $\vdots$

- i-esima componente principale: combinazione lineare $a_i'x_j'$ che massimizza la varianza campionaria di $a_i'x_j'$ con $a_i'a_i = 1$ e per cui le coppie $(a_i'x_j', a_k'x_j')$ abbiano covarianza campionaria nulla $\forall k < i$.

Affermare che a_1 venga scelto in modo che massimizzi la covarianza campionaria significa che rende massimo $a_1' S a_1$ e quindi $\frac{a_1' S a_1}{a_1' a_1}$. Per la Proposizione 2 questo massimo è $\hat{\lambda}_1$, ossia il massimo autovalore di S, e il valore è ottenuto con $a_1 = \hat{e}_1$ ($\hat{e}_1$ autovettore di S relativo a $\hat{\lambda}_1$). Procedendo similmente posso ottenere risultati analoghi a quelli ottenuti per le componenti principali di popolazioni.

Lemma 1.4.1. *Sia $S = \{s_{ik}\}$ matrice di covarianza campionaria con autocoppie $(\hat{\lambda}_1, \hat{e}_1), \ldots, (\hat{\lambda}_p, \hat{e}_p)$ con $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$. Sia $x = [x_1, \ldots, x_p]$ una generica osservazione sulle variabili $X_1, \ldots, X_p$ allora la i-esima componente principale è:*

$$\hat{y}_i = \hat{e}_i' x' = \hat{e}_{i1}x_1 + \hat{e}_{i2}x_2 + \cdots + \hat{e}_{ip}x_p \quad \textit{con } i = 1, \ldots, p. \tag{1.10}$$

Inoltre:

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

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

In particolare la varianza totale del campione è $\sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \cdots + \hat{\lambda}_p$ 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. Denoto con $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_p$ le componenti principali del campione indipendentemente dal fatto che siano ottenute dalla matrice di covarianza S o di correlazione R, risulterà chiaro dal contesto quale matrice utilizzare.

Spesso nello studio delle componenti principali le osservazioni x sono centrate in modo che la media campionaria di ogni componente sia nulla. Infatti se considero:

$$\hat{y}_i = \hat{e}_i'(x - \bar{x})' \quad \text{con } i = 1, \ldots, p, \tag{1.11}$$

allora:

$$\bar{\hat{y}}_i = \frac{1}{n}\sum_{j=1}^{n} \hat{e}_i'(x_j - \bar{x})' = \frac{1}{n}\hat{e}_i' \sum_{j=1}^{n} (x_j - \bar{x})' = \frac{1}{n}\hat{e}_i'\mathbf{0} = 0. \tag{1.12}$$

1.5 Selezione delle componenti principali

Lo scopo della PCA è quello di sintetizzare p variabili $(X_1, X_2, \ldots, X_p)$ in un numero k di variabili, con $k < p$, affinché sia possibile analizzare un numero di dati inferiore a quello di partenza. Risulta fondamentale, per non avere un'eccessiva perdita di informazione, la scelta di quante componenti principali considerare. Non vi è una soluzione definitiva: occorre considerare sia la quantità di varianza totale che dipende da quelle variabili, sia l'interpretazione e l'importanza che queste hanno per lo specifico modello che si sta analizzando. Si possono riassumere tre metodi principali per la selezione delle componenti principali:

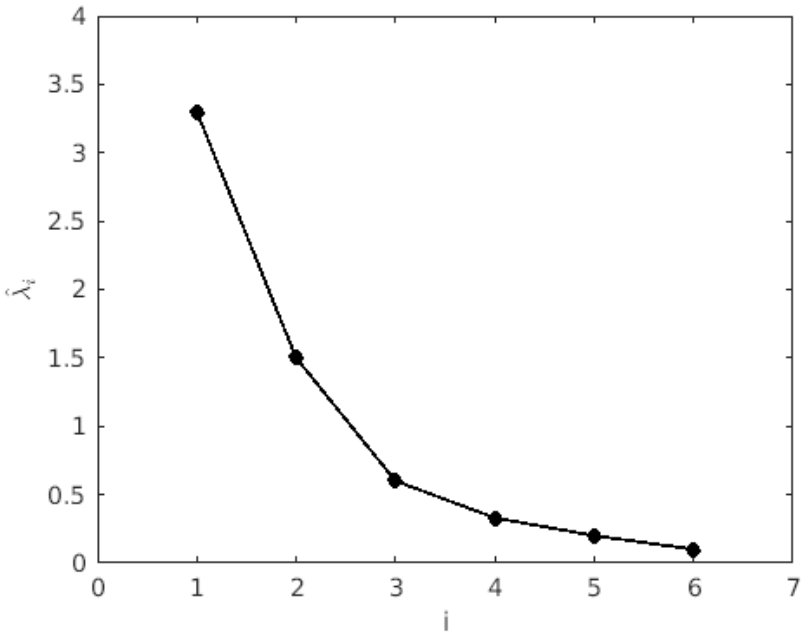
- **Percentuale di varianza totale**: Considero la proporzione di varianza complessiva dovuta alle prime k componenti principali e mi fermo quando questa esaurisce l' $80\% - 90\%$ della varianza complessiva, ossia quando:

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p} \approx 80\% - 90\%.$$

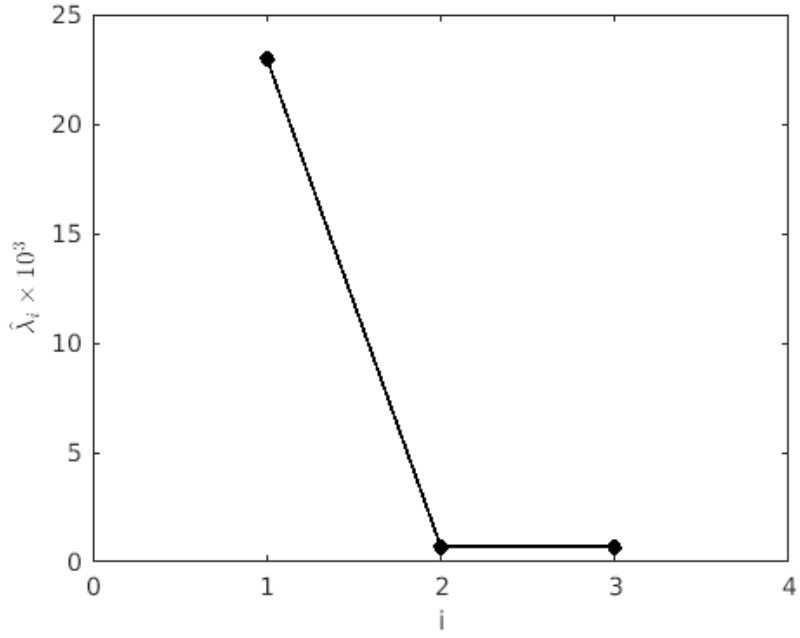
- **Dimensione degli autovalori**: Consiste nel selezionare solo quelle componenti la cui varianza (ossia il corrispettivo autovalore λ_k) è maggiore dell'unità.

- **Scree plot**: Consiste nel tracciare un grafico con il numero di componenti nell'asse delle ascisse ed il valore dei relativi autovalori, dal più grande al più piccolo, nell'asse delle ordinate. Si tratta di un metodo visuale: per decidere il numero di componenti da analizzare si deve guardare nel grafico dove si trova una curva a gomito, ossia dove il valore degli autovalori comincia a livellarsi, e considerare solo le variabili prima di questo punto.

  Riporto in Figura 1.1 due esempi di *scree plot*. Nel primo grafico si può notare un "gomito" a $i = 3$, gli autovalori successivi a λ_2 sono relativamente piccoli e posso, quindi, considerare le prime due (o al massimo 3) componenti principali. Nel secondo grafico la situazione è ancora più chiara, la prima componente è dominante sulle altre e risolve la maggior parte della varianza totale.



(a) *Scree plot 1*



(b) *Scree plot 2*

Figura 1.1: Esempi di scree plot

1.6 Standardizzazione delle componenti principali per un campione

In caso di disomogeneità fra i dati, come per lo studio delle componenti principali di una popolazione, si possono standardizzare le variabili per l'analisi delle componenti principali di un campione. In questo caso si definisce:

$$D^{\frac{1}{2}} = \begin{bmatrix} \sqrt{s_{11}} & 0 & \dots & 0 \\ 0 & \sqrt{s_{22}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{s_{pp}} \end{bmatrix},$$

e si procede costruendo le nuove variabili:

$$z_j' = D^{-\frac{1}{2}}(x_j - \bar{x})' = \begin{bmatrix} \frac{x_{j1}-\bar{x}_1}{\sqrt{s_{11}}} \\ \frac{x_{j2}-\bar{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \frac{x_{jp}-\bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} \quad j = 1, 2, \dots, n. \tag{1.13}$$

Si ottiene, dunque, la matrice standardizzata delle osservazioni:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z_1} \\ \mathbf{z_2} \\ \vdots \\ \mathbf{z_n} \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & \dots & z_{1p} \\ z_{21} & z_{22} & \dots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \dots & z_{np} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{x_{11}-\bar{x}_1}{\sqrt{s_{11}}} & \frac{x_{12}-\bar{x}_1}{\sqrt{s_{11}}} & \dots & \frac{x_{1p}-\bar{x}_1}{\sqrt{s_{11}}} \\ \frac{x_{21}-\bar{x}_2}{\sqrt{s_{22}}} & \frac{x_{22}-\bar{x}_2}{\sqrt{s_{22}}} & \dots & \frac{x_{2p}-\bar{x}_2}{\sqrt{s_{22}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{n1}-\bar{x}_n}{\sqrt{s_{pp}}} & \frac{x_{n2}-\bar{x}_n}{\sqrt{s_{pp}}} & \dots & \frac{x_{np}-\bar{x}_n}{\sqrt{s_{pp}}} \end{bmatrix}.$$

Da queste nuove variabili posso ottenere un nuovo vettore di media campionaria:

$$\bar{\mathbf{z}}' = \frac{1}{n}(\mathbf{1}'\mathbf{Z})' = \frac{1}{n}(\mathbf{Z}'\mathbf{1}) = \frac{1}{n}\begin{bmatrix} \sum_{j=1}^{n} \frac{x_{j1}-\bar{x}_1}{\sqrt{s_{11}}} \\ \sum_{j=1}^{n} \frac{x_{j2}-\bar{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \sum_{j=1}^{n} \frac{x_{jp}-\bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} = 0. \tag{1.14}$$

Sfruttando questo risultato posso dimostrare che la matrice di covarianza campionaria data dalle nuove variabili coincide con la matrice di correlazione R delle variabili iniziali:

$$\begin{aligned} S_z &= \frac{1}{n-1}(\mathbf{Z} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{Z})'(\mathbf{Z} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{Z}) \\ &= \frac{1}{n-1}(\mathbf{Z} - \mathbf{1}\bar{\mathbf{z}})'(\mathbf{Z} - \mathbf{1}\bar{\mathbf{z}}) = \frac{1}{n-1}\mathbf{Z}'\mathbf{Z} \\ &= \frac{1}{n-1}\begin{bmatrix} \frac{(n-1)s_{11}}{s_{11}} & \frac{(n-1)s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \cdots & \frac{(n-1)s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} \\ \frac{(n-1)s_{21}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \frac{(n-1)s_{22}}{s_{22}} & \cdots & \frac{(n-1)s_{2p}}{\sqrt{s_{22}}\sqrt{s_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{(n-1)s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} & \frac{(n-1)s_{2p}}{\sqrt{s_{22}}\sqrt{s_{pp}}} & \cdots & \frac{(n-1)s_{pp}}{s_{pp}} \end{bmatrix} = R. \end{aligned} \tag{1.15}$$

Per ottenere le componenti principali dalle variabili standardizzate è necessario considerare la matrice di correlazione R. Inoltre, poiché le variabili sono già centrate, non ho bisogno di scriverle come in (1.11). Posso ottenere i medesimi risultati ottenuti per la PCA di popolazioni.

Lemma 1.6.1. *Sia $z = [z_1, z_2, \ldots, z_p]$ un vettore di variabili standardizzate con matrice di covarianza campionaria R e siano $(\hat{\lambda}_i, \hat{e}_i)$ le autocoppie di R con $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$. La i-esima componente principale delle variabili standardizzate per un campione è:*

$$\hat{y}_i = \hat{e}_i' z' = \hat{e}_{i1}z_1 + \hat{e}_{i2}z_2 + \cdots + \hat{e}_{ip}z_p \quad \textit{con } i = 1, \ldots, p. \tag{1.16}$$

Inoltre:

$$\textit{Varianza campionaria}(\hat{y}_k) = \hat{\lambda}_k \quad \textit{con } k = 1, \ldots, p,$$

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

In particolare la varianza standardizzata totale del campione è uguale a

$$tr(R) = p = \hat{\lambda}_1 + \cdots + \hat{\lambda}_p$$

e quindi :

$$\begin{pmatrix} proporzione\ di\ varianza\ totale \\ spiegata\ dalla\ k\text{-}esima\ variabile \end{pmatrix} = \frac{\hat{\lambda}_k}{\hat{\lambda}_1 + \cdots + \hat{\lambda}_p}.$$

Inoltre:

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

Esempio 1.1. Considero i tassi di rendimento settimanali determinati dal Gennaio 1975 al Dicembre 1976 di 5 azioni (Allied Chemical, du Pont, Union Carbide, Exxon e Texaco) inseriti nel mercato azionario di New York. Queste osservazioni sembrano essere indipendentemente distribuite, ma i tassi di rendimento di azioni diverse nello stesso periodo risultano essere correlate: tendono a muoversi insieme in risposta alle condizioni economiche.
Se le variabili $x_1, x_2, \ldots, x_5$ rappresentano i tassi di rendimento di Allied Chemical, du Pont, Union Carbide, Exxon e Texaco rispettivamente allora:

$$\bar{x} = [0.0054, 0.0048, 0.0057, 0.0063, 0.0037].$$

La matrice di covarianza delle osservazioni standardizzate è

$$R = \begin{bmatrix} 1.000 & 0.577 & 0.509 & 0.387 & 0.462 \\ 0.577 & 1.000 & 0.599 & 0.389 & 0.462 \\ 0.509 & 0.599 & 1.000 & 0.436 & 0.426 \\ 0.387 & 0.389 & 0.436 & 1.000 & 0.523 \\ 0.462 & 0.322 & 0.436 & 0.523 & 1.000 \end{bmatrix}.$$

Si determinano gli autovalori e autovettori di R:

$$\begin{aligned}
\hat{\lambda}_1 &= 2.857, \quad \hat{e}_1' = \Big[0.464, \quad 0.457, \quad 0.470, \quad 0.421, \quad 0.421\Big], \\
\hat{\lambda}_2 &= 0.809, \quad \hat{e}_2' = \Big[0.240, \quad 0.509, \quad 0.260, \quad -0.526, \quad -0.582\Big], \\
\hat{\lambda}_3 &= 0.540, \quad \hat{e}_3' = \Big[-0.612, \quad 0.178, \quad 0.335, \quad 0.541, \quad -0.435\Big], \\
\hat{\lambda}_4 &= 0.452, \quad \hat{e}_4' = \Big[0.387, \quad 0.206, \quad -0.662, \quad 0.472, \quad -0.382\Big], \\
\hat{\lambda}_5 &= 0.343, \quad \hat{e}_5' = \Big[-0.451, \quad 0.676, \quad -0.400, \quad -0.176, \quad 0.385\Big].
\end{aligned}$$

Considerando le variabili standardizzate si ottengono le prime due componenti principali per il campione:

$$\begin{aligned}
\hat{y}_1 &= \hat{\mathbf{e}}_{\mathbf{1}}'\mathbf{z}' = 0.464z_1 + 0.457z_2 + 0.470z_3 + 0.421z_4 + 0.421z_5, \\
\hat{y}_2 &= \hat{\mathbf{e}}_{\mathbf{2}}'\mathbf{z}' = 0.240z_1 + 0.509z_2 + 0.260z_3 - 0.526z_4 - 0.582z_5.
\end{aligned}$$

Queste componenti spiegano il:

$$\left(\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{p}\right) 100\% = \left(\frac{2.857 + 0.809}{5}\right) 100\% = 73\%$$

della varianza totale.

La prima componente è, approssimativamente, una somma con lo stesso peso delle cinque variabili, si può chiamare *componente generale per il mercato azionario* o semplicemente una *componente di mercato*. La seconda, invece, contrasta una somma positiva per le prime tre variabili (che indicano le industrie chimiche) ad una negativa per le ultime due (industrie petrolifere): posso chiamarla *componente dell'industria*. Si tratta di un esempio in cui una componente il cui autovalore è minore di 1 risulta essere rilevante nell'analisi. Le restanti componenti non sono di facile interpretazione ma non vengono considerate poiché non risolvono una parte significativa della variabilità totale.

Capitolo 2

Il modello di regressione lineare

2.1 Regressione lineare multivariata

La regressione lineare è un metodo statistico che ha come scopo principale la previsione: mira alla costruzione di un modello attraverso cui prevedere i valori di una variabile dipendente (risposta) a partire dai valori di una (regressione lineare semplice), o più (regressione lineare multipla), variabili indipendenti. Questo metodo cerca di spiegare una connessione lineare di casualità fra i "predittori" e la risposta: quest'ultima è conseguenza di alcune caratteristiche della popolazione e nella regressione viene fatta dipendere da un numero r di variabili dipendenti, che indicheremo con $z_1, z_2, \ldots, z_r$.
Il modello per la regressione lineare classica afferma che y è composto da una media, che dipende in maniera lineare dagli z_j e da un errore aleatorio ϵ. Risulta necessario considerare un errore poiché le misurazioni possono essere imprecise e altre concause, oltre a quelle considerate nella regressione, potrebbero essere presenti nella determinazione di y. Questo errore viene visto come un vettore aleatorio che segue un comportamento caratterizzato da delle ipotesi sulla sua distribuzione. Specificatamente il modello è:

$$y = \beta_0 + \beta_1 z_1 + \cdots + \beta_r z_r + \epsilon \tag{2.1}$$
$$[\text{Risposta}] = [\text{media (dipendente da } z_1, \ldots, z_r)] + [\text{errore}].$$

Se considero n osservazioni indipendenti di y e i relativi valori di $z_j = [z_{j1}, \ldots, z_{jr}]$, con $j = 1, \ldots, n$, posso dedurre il modello completo:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 z_{11} + \cdots + \beta_r z_{1r} + \epsilon_1, \\ y_2 &= \beta_0 + \beta_1 z_{21} + \cdots + \beta_r z_{2r} + \epsilon_2, \\ &\vdots \\ y_n &= \beta_0 + \beta_1 z_{n1} + \cdots + \beta_r z_{nr} + \epsilon_n. \end{aligned} \tag{2.2}$$

Assumiamo che i termini degli errori abbiano le seguenti proprietà:

1. $E(\epsilon_j) = 0$
2. $Var(\epsilon_j) = \sigma^2$
3. $Cov(\epsilon_j, \epsilon_i) = 0 \quad \text{con } i \neq j.$

Il modello in notazione matriciale si può esprimere nel seguente modo:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & z_{11} & z_{12} & \ldots & z_{1r} \\ 1 & z_{21} & z_{22} & \ldots & z_{2r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n1} & z_{n2} & \ldots & z_{nr} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

oppure

$$\begin{underset{n \times 1}{\mathbf{y}} = \underset{(n \times (r+1))((r+1) \times 1)}{\mathbf{Z}} \underset{}{\boldsymbol{\beta}} + \underset{(n \times 1).}{\boldsymbol{\epsilon}}} \tag{2.3}$$

con:

1. $E(\epsilon) = 0$
2. $Cov(\epsilon) = E(\epsilon\epsilon') = \sigma^2 \mathbf{I}.$

Osservazione. Si nota che la prima colonna della matrice Z viene moltiplicata per la costante β_0, risulta quindi ottimale costruire la variabile artificiale $z_{j0} = 1$ in modo che Z abbia come j-esima riga $[z_{j0}, z_{j1}, \ldots, z_{jr}]$.

2.2 Il metodo dei minimi quadrati

Lo scopo della regressione lineare è sviluppare un'equazione che permetta di predire una risposta date le variabili indipendenti da cui la si vuole far dipendere. Occorre determinare il coefficiente di regressione β e la varianza dell'errore σ^2 in modo che siano valori consistenti con i dati disponibili. Voglio che, dato b un valore di prova per β, la differenza, $y_i - \hat{y}_i = y_i - b_o - b_1 z_{i1} - \cdots - b_r z_{ir}$, fra l'effettivo valore della risposta y_i e il valore atteso sia la più piccola possibile. Questo valore non sarà mai nullo poiché le risposte fluttuano attorno al loro valore atteso a seconda delle ipotesi che facciamo sull'errore. Si può, tuttavia, utilizzare il metodo dei minimi quadrati per minimizzare la somma dei quadrati delle differenze:

$$S(\mathbf{b}) = \sum_{i=1}^{n} (y_i - b_o - b_1 z_{i1} - \cdots - b_r z_{ir})^2 = (\mathbf{y\text{-}Zb})'(\mathbf{y\text{-}Zb}). \tag{2.4}$$

I valori ottenuti per b sono chiamati stime dei minimi quadrati per i coefficienti di regressione. Le informazioni per calcolare σ^2 sono nel residuo $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}$.

Teorema 2.2.1. *Sia $\boldsymbol{Z}$ una matrice di rango massimo: $r+1 \leq n$. La stima con i minimi quadrati di β è:*

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}.$$

Se $\hat{\mathbf{y}} = \boldsymbol{Z}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$ denota il valore previsto per $\mathbf{y}$, con $\mathbf{H} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$, allora i residui:

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = [\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z'Z})^{-1}\boldsymbol{Z'}]\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

soddisfano $\boldsymbol{Z'}\hat{\boldsymbol{\epsilon}} = \mathbf{0}$ e $\hat{\boldsymbol{y}}'\hat{\boldsymbol{\epsilon}} = \mathbf{0}$. Inoltre:

$$\begin{aligned} \textit{somma dei quadrati degli errori} &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 z_{i1} - \cdots - \hat{\beta}_r z_{ir})^2 = \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} \\ &= \boldsymbol{y}'[\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z'Z})^{-1}\boldsymbol{Z'}]\boldsymbol{y} = \boldsymbol{y'y} - \boldsymbol{y'Z}\hat{\boldsymbol{\beta}}. \end{aligned}$$

Dimostrazione. Suppongo che $\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$. Allora:

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}} = [\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z'Z})^{-1}\boldsymbol{Z'}]\boldsymbol{y}.$$

Poichè $[\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z'Z})^{-1}\boldsymbol{Z'}]$ è una matrice di proiezione ortogonale, soddisfa le seguenti proprietà:

1. $[\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}']' = [\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}']$ (simmetrica),
2. $[\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'][\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'] = \boldsymbol{I} - 2\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' +$
   $+\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' = \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$ (idempotente),
3. $\boldsymbol{Z}'[\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'] = \boldsymbol{Z}' - \boldsymbol{Z}' = \boldsymbol{0}$.

Ne consegue che $\boldsymbol{Z}'\hat{\boldsymbol{\epsilon}} = \boldsymbol{Z}'(\boldsymbol{y} - \hat{\boldsymbol{y}}) = \boldsymbol{Z}'[\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}']\boldsymbol{y} = \boldsymbol{0}$ e quindi $\hat{\boldsymbol{y}}'\hat{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\beta}}'\boldsymbol{Z}'\hat{\boldsymbol{\epsilon}} = \boldsymbol{0}$. In particolare $\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} = \boldsymbol{y}'[\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'][\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}']\boldsymbol{y} = \boldsymbol{y}'[\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}']\boldsymbol{y}$ $= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{Z}\hat{\boldsymbol{\beta}}$. Occorre verificare l'espressione di $\hat{\boldsymbol{\beta}}$, per farlo considero l'espressione:

$$\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{b} = \boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{b} = \boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}).$$

Quindi, ricordando che $(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}})'\boldsymbol{Z} = \hat{\boldsymbol{\epsilon}}'\boldsymbol{Z} = \boldsymbol{0}$,

$$\begin{aligned}\boldsymbol{S}(\boldsymbol{b}) &= (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{b}) \\ &= (\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{b})'\boldsymbol{Z}'\boldsymbol{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}) \\ &+2(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}})'\boldsymbol{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}) \\ &= (\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{b})'\boldsymbol{Z}'\boldsymbol{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{b})\ .\end{aligned}$$

Il primo termine di $\mathbf{S(b)}$ non dipende da $\mathbf{b}$ mentre il secondo è la lunghezza al quadrato di $\boldsymbol{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{b})$. Poichè abbiamo assunto che $\boldsymbol{Z}$ abbia rango massimo allora $\boldsymbol{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}) \neq \boldsymbol{0}$ se $\hat{\boldsymbol{\beta}} \neq \boldsymbol{b}$, quindi il valore minimo di $\mathbf{S(b)}$ è unico ed è ottenuto con $\boldsymbol{b} = \hat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}$. In particolare osservo che se $\mathbf{Z}$ non avesse rango massimo, allora non si potrebbe definire $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ poiché in quel caso anche $(\mathbf{Z}'\mathbf{Z})$ non avrebbe rango massimo. □

Questo risultato è importante perché mostra come il metodo dei minimi quadrati permetta di stimare il parametro $\hat{\boldsymbol{\beta}}$ ed il residuo $\hat{\boldsymbol{\epsilon}}$ partendo solo dalla matrice $\boldsymbol{Z}$ e dalla risposta $\mathbf{y}$ con semplici operazioni matriciali.

Osservazione. Applicando il precedente risultato (in particolare $\hat{\boldsymbol{y}}'\hat{\boldsymbol{\epsilon}} = \boldsymbol{0}$) posso decomporre la somma dei quadrati della risposta totale nel seguente modo:

$$\boldsymbol{y}'\boldsymbol{y} = (\hat{\boldsymbol{y}} + \boldsymbol{y} - \hat{\boldsymbol{y}})'(\hat{\boldsymbol{y}} + \boldsymbol{y} - \hat{\boldsymbol{y}}) = (\hat{\boldsymbol{y}} + \hat{\boldsymbol{\epsilon}})'(\hat{\boldsymbol{y}} + \hat{\boldsymbol{\epsilon}}) = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} + \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}. \tag{2.5}$$

Inoltre, poiché la prima colonna di $\boldsymbol{Z}$ è $\mathbf{1}$, la condizione $\boldsymbol{Z}'\hat{\boldsymbol{\epsilon}} = 0$ comporta:

$$\boldsymbol{0} = \mathbf{1}'\hat{\boldsymbol{\epsilon}} = \sum_{j=1}^{n} \hat{\boldsymbol{\epsilon}}_j = \sum_{j=1}^{n} \boldsymbol{y}_j - \sum_{j=1}^{n} \hat{\boldsymbol{y}}_j,$$

ossia $\bar{y} = \bar{\hat{y}}$. Sottraendo $n\bar{y}^2 = n(\bar{\hat{y}})^2$ all'equazione 2.5 ottengo una decomposizione della somma dei quadrati:

$$\boldsymbol{y'y - n\bar{y}^2 = \hat{y}'\hat{y} - n(\bar{\hat{y}})^2 + \hat{\epsilon}'\hat{\epsilon}}.$$

Ossia:

$$\sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} \boldsymbol{(y_j - \bar{y})^2} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} \boldsymbol{(\hat{y}_j - \bar{\hat{y}})^2} + \sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} \boldsymbol{\hat{\epsilon}_j^2} \tag{2.6}$$

$$\begin{pmatrix}\text{somma totale}\\ \text{dei quadrati}\end{pmatrix} = \begin{pmatrix}\text{somma dei quadrati}\\ \text{di regressione}\end{pmatrix} + \begin{pmatrix}\text{somma dei quadrati}\\ \text{degli errori}\end{pmatrix}.$$

Da questa decomposizione posso osservare che la qualità del modello può essere stimata dal *coefficiente di determinazione*:

$$R^2 = 1 - \frac{\sum_{j=1}^n \hat{\epsilon}_j^2}{\sum_{j=1}^n (y_j - \bar{y})^2} = \frac{\sum_{j=1}^n (\hat{y}_j - \bar{\hat{y}})^2}{\sum_{j=1}^n (y_j - \bar{y})^2}.$$

Questo fattore indica la quantità della totale varianza degli y_i dovuta alle variabili $z_1, z_2, \ldots, z_r$. In particolare $R^2 = 1$ se il modello spiega completamente la varianza delle y, ossia l'equazione interpola tutti i dati e gli errori sono sempre nulli; mentre $R^2 = 0$ se $\hat{\beta}_0 = \bar{y}$ e $\hat{\beta}_1, \ldots, \hat{\beta}_r = 0$, ossia le variabili considerate non hanno alcuna influenza nella risposta.

2.3 Selezione delle variabili indipendenti

Nella ricerca di un modello per la regressione lineare che meglio approssimi l'effettiva relazione fra dati e risposta, occorre scegliere quali variabili predittive considerare: se i fattori che incidono nella risposta sono molti è necessario farne una selezione. Un modello con un numero inferiore di variabili indipendenti è più efficiente e di più facile interpretazione. Cerco una "strategia" per trovare il minimo numero di variabili utili per la valutazione della variabile dipendente e che, inoltre, mantenga la possibilità di ottenere un risultato preciso.

Un metodo può essere provare tutti i sottoinsiemi possibili, prima ogni variabile presa

singolarmente, poi a coppie e così via, per poi selezionare il miglior sottoinsieme sfruttando un criterio, come per esempio R^2. Tuttavia, poiché R^2 aumenta sempre con l'aggiunta di variabili predittive, è più preciso considerare il valore modificato:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-r-1}.$$

Un altro valore che si può prendere in considerazione è il criterio di stima di Mallow C_p:

$$C_p = \left(\frac{\begin{array}{c}\text{(somma dei quadrati degli errori del modello}\\ \text{ottenuto dal sottoinsieme con p parametri)}\end{array}}{\text{(varianza degli errori per il modello completo)}} \right) - n + 2p.$$

Per decidere con quali variabili si ottenga il miglior modello si considera un grafico delle coppie (p, C_p) ed i valori ottimali sono vicini alla retta a 45°.
Se il numero di possibili predittori è elevato vi sono delle limitazioni sul numero possibile di combinazioni che si possono provare. Esistono, dunque, altri metodi che selezionano un sottoinsieme di variabili senza considerare ogni possibilità, ma incrementando o decrescendo a ogni passo il numero delle variabili indipendenti.
Un esempio che segue questo procedimento è la regressione *stepwise*:

Step 1 Vengono considerate tutte le variabili prese singolarmente: la prima ad entrare nella regressione sarà quella che spiega la più significativa porzione di varianza totale.

Step 2 Per determinare la seconda variabile che verrà considerata nel modello si utilizza un test F, questo test ha lo scopo di controllare se la variabile dipendente è effettivamente correlata a quella indipendente considerata, ci si chiede se il valore del corrispondente coefficiente di regressione è significativamente diverso da zero oppure no. Si determina un valore di tolleranza per il test F sopra al quale viene aggiunta la variabile considerata.

Step 3 Una volta aggiunta una nuova variabile nell'equazione viene ricontrollato il contributo delle altre variabili alla somma dei quadrati della regressione. Si utilizza

nuovamente il test F, se il valore è minore del criterio di tolleranza la variabile viene eliminata.

Step 4 Questi passaggi sono ripetuti, eseguendo il test F su ogni nuova variabile inserita nella regressione e ricontrollando, eventualmente eliminando, quelle già presenti. Il procedimento ha termine quando l'aggiunta di una nuova variabile non è più considerata significativa e nessuna deve essere eliminata in base al test F.

Questo metodo non garantisce la selezione del miglior sottoinsieme di variabili ed, inoltre, essendo una selezione automatica, non indica quando un cambiamento di variabili potrebbe essere utile.

2.4 Il problema della collinearità

Nella regressione lineare multipla una delle più grandi difficoltà è il problema della collinearità, ossia quando $\mathbf{Z}$ non ha rango massimo ed esiste una qualche combinazione lineare per cui $\boldsymbol{Za} = \mathbf{0}$. Nella pratica è raro che esista esattamente tale combinazione, ma è possibile che $\mathbf{Za} = \mathbf{0}$ si avvicini drasticamente a 0. In quel caso il calcolo di $(\mathbf{Z'Z})^{-1}$ risulta essere numericamente instabile ed i valori sulla diagonale di $(\mathbf{Z'Z})^{-1}$ diventano grandi, generando un aumento della varianza per le stime $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ dei coefficienti della regressione e rendendo difficile scegliere quali siano significativi ai fini dell'analisi.

Questo problema può essere risolto considerando una selezione delle variabili iniziali, eliminando, ossia, quelle fortemente correlate. In alternativa si può considerare la relazione fra la risposta e le componenti principali delle variabili indipendenti, ottenute considerando $\boldsymbol{Z}$ come la matrice dei dati relativi ad un campione.
Sia il modello per la regressione lineare standard:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}. \tag{2.7}$$

Sia $\mathbf{A}$ la matrice che abbia come colonne gli autovettori di $\mathbf{X}$, allora i valori delle componenti principali di $\boldsymbol{X}$ sono:

$$\boldsymbol{Z} = \boldsymbol{XA}.$$

Essendo $\boldsymbol{A}$ ortogonale, si ottiene:

$$\boldsymbol{X\beta} = \boldsymbol{XAA'\beta} = \boldsymbol{Z\gamma},$$

dove $\boldsymbol{\gamma} = \boldsymbol{A'\beta}$.
Posso, dunque, riscrivere il modello della regressione, applicandola a $\boldsymbol{Z}$, nel seguente modo:

$$\boldsymbol{y} = \boldsymbol{Z\gamma} + \boldsymbol{\epsilon}. \tag{2.8}$$

Se decido di considerare solo le prime k componenti principali ottengo un modello ridotto:

$$\boldsymbol{y_k} = \boldsymbol{Z_k\gamma_k} + \boldsymbol{\epsilon_k}.$$

Chiaramente usare il metodo dei minimi quadrati per trovare $\hat{\boldsymbol{\gamma}}$ dall'equazione (2.8) e, di conseguenza, $\hat{\boldsymbol{\beta}} = \boldsymbol{A}\hat{\boldsymbol{\gamma}}$ è equivalente ad ottenere direttamente $\hat{\boldsymbol{\beta}}$ dall'equazione (2.7). Inoltre, essendo le colonne di $\boldsymbol{Z}$ ortogonali, il procedimento per trovare $\boldsymbol{\gamma}$ risulta essere più efficiente.
In particolare, poiché le componenti principali sono scorrelate fra loro, l'importanza di ciascuna di esse nell'equazione non è influenzata dalla presenza o meno delle altre; nella regressione standard, invece, il contributo di una variabile può cambiare drasticamente se ne viene inserita, o tolta, un'altra.

Capitolo 3

L'uso della PCA nell'ambito della regressione lineare

In questo capitolo si intende combinare la regressione lineare multivariata e l'analisi delle componenti principali. Lo scopo della regressione lineare è trovare un modello che regoli la relazione fra r variabili ed una risposta attraverso un metodo statistico: sono necessarie un numero n alto di osservazioni per ogni variabile e per la risposta, al fine di avere risultati attendibili. Tuttavia, se la risposta dipende da un numero elevato di variabili, il metodo standard può avere un alto costo computazionale e la necessità di numerose osservazioni per ogni variabile. L'utilizzo della PCA nell'ambito della regressione, dunque, può portare a notevoli vantaggi, riducendo la dimensione del problema ad numero inferiore k di componenti principali senza una significativa perdita di informazioni. Tuttavia, diversamente da quanto accade nell'analisi delle componenti principali, la quantità di varianza complessiva spiegata da una componente può non essere sufficiente per decidere se considerarla o meno al momento della regressione. Può succedere che una variabile che spiega una percentuale minima della varianza (ad esempio che abbia un autovalore quasi nullo) sia, invece, strettamente correlata con la risposta nella regressione e non convenga eliminarla. Chiamiamo questa combinazione "regressione delle componenti principali".

Il principale vantaggio nella regressione delle componenti principali è l'interpretazione. Le componenti principali sono di più facile comprensione e possono mostrare legami fra

i dati che non erano precedentemente sospettabili; come si può vedere nell'Esempio 1.1 dove sono state evidenziate due componenti principali che spiegavano da sole quasi tutta la varianza iniziale: una componente di mercato e una dell'industria. Nell'ambito della regressione questo tipo di analisi può permettere di capire meglio quali aspetti condizionano la risposta e di darne una migliore interpretazione.

Infine si può osservare che, nella regressione lineare multivariata, si assume che non vi siano errori nelle misurazioni delle variabili, ma solo in quelle delle risposte. Si tratta di un'ipotesi forte e, solitamente, non vera. Utilizzando la PCA, invece, si presuppone che siano le componenti principali a non avere errori di misurazioni e non le variabili iniziali. Tuttavia errori di misurazioni delle variabili iniziali comportano principalmente un aumento della varianza più che del valore atteso delle componenti, rendendo, di conseguenza, la regressione su queste componenti più attendibile.

3.1 Un esempio

Considero i dati relativi allo studio delle batterie argento-zinco usate in alcune applicazioni per i satelliti. La tabella 3.1 riporta i dati raccolti per caratterizzare le prestazioni della batteria durante il suo ciclo di vita.

Si nota che la risposta y, ossia il numero di cicli della batteria, dipende da un vettore di 5 variabili $x = [x_1, x_2, \ldots, x_5]$ in cui:

$$x_1 = \text{Tasso di carica}, \quad x_2 = \text{Tasso di scarica}, \quad x_3 = \text{Intensità di scarica},$$
$$x_4 = \text{Temperatura}, \quad x_5 = \text{Tensione finale di scarica}.$$

Lo scopo è quello di confrontare una regressione lineare sulla totalità delle variabili ed una invece su una selezione delle componenti principali di queste variabili; per determinare quale sia più attendibile si utilizzerà il coefficiente di determinazione R^2 o il suo valore corretto $\bar{R}^2$. Osservo che, in entrambi casi, sarà trovata una stima del modello di regressione per $ln(Y)$.

Tasso di carica	Tasso di scarica	Intensità di scarica	Temperatura	Tensione finale di carica	Numero di cicli
0.38	3.13	60.00	40.00	2.00	101.00
1.00	3.13	76.80	30.00	1.99	141.00
1.00	3.13	60.00	20.00	2.00	96.00
1.00	3.13	60.00	20.00	1.98	125.00
1.62	3.13	43.20	10.00	2.01	43.00
1.62	3.13	60.00	20.00	2.00	16.00
1.62	3.13	60.00	20.00	2.02	188.00
0.38	5.00	76.80	10.00	2.01	10.00
1.00	5.00	43.20	10.00	1.99	3.00
1.00	5.00	43.20	30.00	2.01	386.00
1.00	5.00	100.00	20.00	2.00	45.00
1.62	5.00	76.80	10.00	1.99	2.00
0.38	1.25	76.80	10.00	2.01	76.00
1.00	1.25	43.20	10.00	1.99	78.00
1.00	1.25	76.80	30.00	2.00	160.00
1.00	1.25	60.00	0.00	2.00	3.00
1.62	1.25	43.20	30.00	1.99	216.00
1.62	1.25	60.00	20.00	2.00	73.00
0.38	3.13	76.80	30.00	1.99	314.00
0.38	3.13	60.00	20.00	2.00	170.00

Tabella 3.1: Dati

Costruisco la matrice delle variabili standardizzate:

$$Z = \begin{bmatrix} 1 & -1.38 & \dots & 0.10 \\ 1 & -0.06 & \dots & -0.92 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & -1.38 & \dots & 0.10 \end{bmatrix}$$

da cui posso ottenere le stime per i coefficienti di regressione, tramite il Teorema 2.1.1, dal vettore:

$$\hat{\beta}_1 = [4.00, -0.22, -0.47, -0.17, 1.16, 0.33].$$

Il valore del coefficiente di determinazione è:

$$R_1^2 = 0.6633.$$

Per studiare le componenti principali delle variabili standardizzate $z_1, z_2, \dots, z_5$ determi-

no autovalori e autovettori della matrice di correlazione campionaria:

$$\begin{aligned}
\hat{\lambda}_1 &= 1.447, \quad \hat{e}_1' = \Big[-0.6064, \quad 0.3901, \quad 0.6357, \quad 0.2755, \quad 0.0045\Big], \\
\hat{\lambda}_2 &= 1.144, \quad \hat{e}_2' = \Big[0.1089, \quad 0.4003, \quad 0.1125, \quad -0.5975, \quad 0.6769\Big], \\
\hat{\lambda}_3 &= 0.894, \quad \hat{e}_3' = \Big[0.3644, \quad 0.7184, \quad -0.0388, \quad -0.1163, \quad -0.5797\Big], \\
\hat{\lambda}_4 &= 0.855, \quad \hat{e}_4' = \Big[0.2618, \quad 0.3359, \quad -0.2772, \quad 0.7329, \quad 0.4523\Big], \\
\hat{\lambda}_5 &= 0.661, \quad \hat{e}_5' = \Big[0.6474, \quad -0.2421, \quad 0.7105, \quad 0.1277, \quad 0.0336\Big].
\end{aligned}$$

Osservando lo scree plot in Figura 3.1 posso pensare di considerare le prime tre componenti principali che spiegano il:

$$\left(\frac{\hat{\lambda}_1 + \hat{\lambda}_2 + \hat{\lambda}_3}{\hat{\lambda}_1 + \cdots + \hat{\lambda}_5}\right) 100\% = \left(\frac{1.447 + 1.144 + 0.894}{5}\right) 100\% = 70\%$$

della varianza totale.

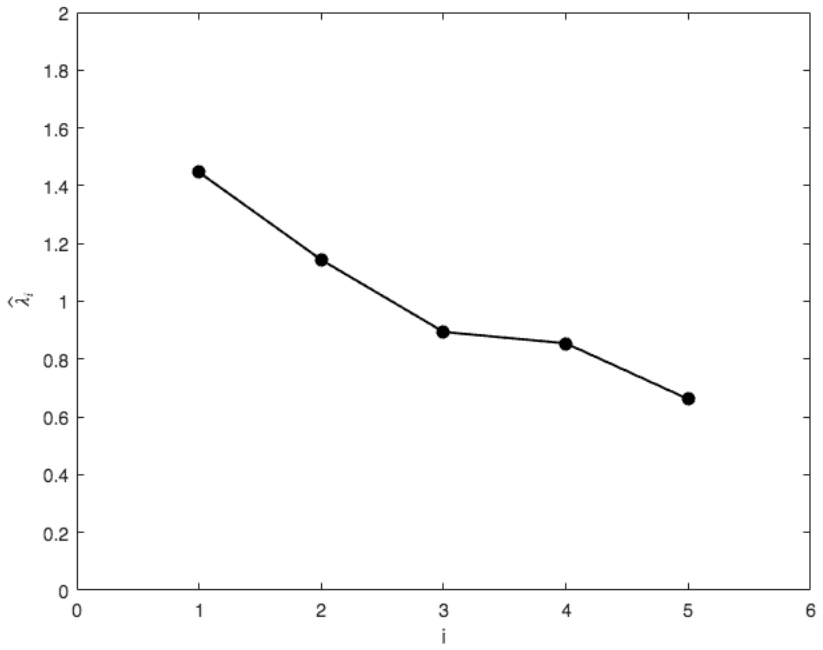


Figura 3.1: Scree plot

Queste componenti sono:

$$\hat{y}_1 = \hat{\mathbf{e}}_\mathbf{1}'\mathbf{x}' = -0.6064z_1 + 0.3901z_2 + 0.6357z_3 + 0.2755z_4 + 0.0045z_5,$$
$$\hat{y}_2 = \hat{\mathbf{e}}_\mathbf{2}'\mathbf{x}' = 0.1089z_1 + 0.4003z_2 + 0.1125z_3 + -0.5975z_4 + 0.6769z_5,$$
$$\hat{y}_3 = \hat{\mathbf{e}}_\mathbf{3}'\mathbf{x}' = 0.3644z_1 + 0.7184z_2 - 0.0388z_3 - 0.1163z_4 - 0.5797z_5.$$

In questo caso i coefficienti risultano essere:

$$\hat{\beta}_2 = [4.00, 0.16, -0.70, -0.73].$$

Il coefficiente di determinazione è:

$$R_2^2 = 0.4284.$$

Bisogna osservare che il coefficiente R^2, tuttavia, aumenta sempre con l'aumentare delle variabili considerate ed è quindi opportuno fare affidamento al coefficiente corretto $\bar{R}^2$:

$$\bar{R}_1^2 = 0.5430, \quad \bar{R}_2^2 = 0.3212.$$

Il coefficiente della regressione effettuata su tutte le variabili risulta essere maggiore. In effetti le prime tre componenti risolvono solo il 70% della varianza complessiva e può, quindi, essere utile considerare anche la quarta componente principale:

$$\hat{y}_4 = \hat{\mathbf{e}}_\mathbf{4}'\mathbf{x}' = 0.2618z_1 + 0.3359z_2 - 0.2772z_3 + 0.7329z_4 + 0.4523z_5.$$

In questo modo le componenti spiegano l'87% della varianza complessiva ed i coefficienti della regressione sono:

$$\hat{\beta}_3 = [4.00, 0.16, -0.70, -0.73, 0.82].$$

In questo caso il coefficiente di determinazione corretto è:

$$\bar{R}_3^2 = 0.5734.$$

Dunque questo modello può ritenersi più attendibile del primo (che considerava tutte le 5 variabili iniziali) nonostante vi siano un numero inferiore di variabili predittive.
Per quanto riguarda l'interpretazione delle componenti principali si può osservare che la

prima componente è ben correlata sia alla prima che alla terza variabile; si può osservare sia dai coefficienti $e_{11} = -0.6064$ e $e_{13} = 0.6357$ sia dai coefficienti di correlazione:

$$r_{\hat{y}_1,z_1} = -0.7293, \qquad r_{\hat{y}_1,z_3} = 0.7646.$$

Le altre componenti invece sono ben correlate solo ad una delle variabili iniziali, sono indicati in seguito i relativi coefficienti di correlazione:

$$r_{\hat{y}_2,z_5} = 0.7238, \qquad r_{\hat{y}_3,z_2} = 0.6793, \qquad r_{\hat{y}_4,z_4} = 0.6775, \qquad r_{\hat{y}_5,z_3} = 0.5779.$$

Si può concludere che la prima variabile può essere interpretata come una variabile di dipendenza della capacita di ricarica dal calore mentre le altre sono associate a una delle variabili iniziali.

In questo esempio si nota quanto la scelta del numero di componenti principali da considerare sia importante nell'ambito della regressione. Se quasi ogni variabile è importante nel determinare la risposta, l'uso della PCA può non essere così vantaggioso; inoltre è difficile determinare a priori il numero giusto di componenti da considerare poiché anche se una di esse risolve una minima parte della varianza complessiva può essere strettamente correlata alla risposta e quindi fondamentale nella regressione.
Tuttavia per quanto riguarda l'interpretazione del problema l'analisi delle componenti principali rimane uno strumento potente. Inoltre, se le dimensioni del problema sono nettamente maggiori rispetto a questo esempio, una riduzione delle variabili dovuta all'uso della PCA può comportare notevoli vantaggi a livello di costo computazionale.

Conclusioni

Abbiamo analizzato separatamente due metodi di analisi dei dati quali l'analisi delle componenti principali e la regressione lineare multivariata e, infine, li abbiamo combinati nel metodo della regressione delle componenti principali. Per farlo si sono precedentemente date le nozioni di base per la comprensione delle strutture matematiche utilizzate.

Nel Capitolo 1 è stata esaminata la procedura per determinare le componenti principali sia per una popolazione sia per un campione, in entrambi i casi evidenziando la possibilità di standardizzare le variabili per avere risultati più omogenei. I due principali vantaggi nell'uso della PCA (riduzione della dimensione del problema e migliore interpretazione dei dati) sono discussi nel capitolo, anche tramite esempi.
In seguito è stata discussa la regressione lineare multivariata, presentandone il modello e l'utilizzo del metodo dei minimi quadrati. In particolare si è sottolineata l'importanza di scegliere il numero minore di variabili, che, tuttavia, mantenga la possibilità di un risultato attendibile; per questo sono state presentate alcune strategie per la selezione di variabili (come la regressione *stepwise*). Infine si è presentato il problema della riduzione del numero di variabili predittive mediante l'uso dell'analisi delle componenti principali.

Tuttavia bisogna prestare molta attenzione, come si vede nell'esempio sulla batteria argento-zinco, al numero di componenti da selezionare: una componente che risolve una minima parte della varianza complessiva può essere fondamentale nella determinazione della risposta nella regressione. In conclusione l'uso delle componenti principali nella regressione è vantaggiosa, sia permettendo una migliore interpretazione nel contesto in cui si opera sia riducendo le dimensioni.

Bibliografia

[1] R. Johnson, D. Wichern, *Applied Multivariate Statistical Analysis*, Prentice Hall, 2002.

[2] I.T.Joliffe, *Principal Component Analysis*, Springer, 2002, pp.167-195.

[3] R.Ricci, *Appunti di Statistica*, Università di Firenze, 2003.

[4] F.Ruini, *Regressione lineare multipla*, Università di Modena.

[5] Lexi V. Perez, *Principal Component Analysis to Address Multicollinearity*, Withman College, 2017.