UNIVERSITÀ DEGLI STUDI DELLA BASILICATA
FACOLTÀ DI INGEGNERIA
CORSO DI LAUREA IN INGEGNERIA MECCANICA

TESINA
IN
COMPLEMENTI DI PROBABILITÀ E STATISTICA
3 crediti

Analisi delle componenti principali

DOCENTE:

Prof.: ELVIRA DI NARDO

STUDENTE:

D'ANDRIA PATRIZIA **22673**

ANNO ACCADEMICO 2005/2006

Indice

Analisi delle componenti principali

ANALISI DELLE COMPONENTI PRINCIPALI

Introduzione

L'analisi delle componenti principali (detta pure *PCA* oppure *CPA*) è una tecnica utilizzata nell'ambito della statistica multivariata per la semplificazione dei dati d'origine.

Lo scopo primario di questa tecnica è la riduzione di un numero più o meno elevato di variabili (rappresentanti altrettante caratteristiche del fenomeno analizzato) in alcune variabili latenti. Ciò avviene tramite una trasformazione lineare delle variabili che proietta quelle originarie in un nuovo sistema cartesiano nel quale le variabili vengono ordinate in ordine decrescente di varianza: pertanto, la variabile con maggiore varianza viene proiettata sul primo asse, la seconda sul secondo asse e così via. La riduzione della complessità avviene limitandosi ad analizzare le principali (per varianza) tra le nuove variabili.

Diversamente da altre trasformazioni (lineari) di variabili praticate nell'ambito della statistica, in questa tecnica sono gli stessi dati che determinano i vettori di trasformazione.

La PCA è una tecnica statistica adoperata in molti ambiti: nell'astronomia, nella medicina, in campo agro-alimentare, ecc... fino anche alla compressione di immagini; questo perché quando ci si trova a semplificare un problema, riducendo la dimensione dello spazio di rappresentazione, si ha allo stesso

tempo una perdita dell'informazione contenuta nei dati originali. La PCA consente di controllare egregiamente il "trade-off" tra la perdita di informazioni e la semplificazione del problema (basta scegliere il numero appropriato di autovettori).

Il presente elaborato mira a descrivere tale metodologia dal punto di vista sia matematico che qualitativo.

1. *Descrizione del metodo*

L'analisi delle componenti principali con riferimento a p variabili, $X_1, X_2, ..., X_i, ..., X_p \quad con \quad i = 1,2,...,p$ (vettore casuale multivariato), consente di individuare altrettante p variabili (diverse dalle prime), $Y_1, Y_2, ..., Y_i, ..., Y_p \quad con \quad i = 1,2,...,p$ (vettore multivariato), ognuna combinazione lineare delle p variabili di partenza.

L'obiettivo della PCA consiste nell'individuare opportune trasformazioni lineari Y_i delle variabili osservate facilmente interpretabili e capaci di evidenziare e sintetizzare l'informazione insita nella matrice iniziale $\vec{X}$. Tale strumento risulta utile soprattutto allorquando si ha a che fare con un numero di variabili considerevole da cui si vogliono estrarre le maggiori informazioni possibili pur lavorando con un set più ristretto di variabili.

I dati di partenza vengono organizzati in una matrice, indicata con $\vec{X}$**:**

$$\vec{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pp} \end{pmatrix} \quad con \quad i = 1,2,...,p \quad e \quad j = 1,2,...,p$$

dove:

- le colonne rappresentano le p osservazioni effettuate;
- le righe sono le p variabili considerate per il fenomeno in analisi.

Si può notare come la matrice dei dati d'origine viene sinteticamente rappresentata con un vettore casuale multivariato ($\vec{X} = (X_1 \quad X_2 \quad \dots \quad X_p)^T$).

Data la matrice $\vec{X}$, che contiene p variabili correlate tra loro, si vuole ottenere una matrice di nuovi dati $\vec{Y}$, composta da p variabili incorrelate tra loro, che risultano essere combinazione lineare delle prime. E quindi si ha:

$$\vec{Y} = \overline{\overline{L}}\ \vec{X} \tag{1.1}$$

$$(p \times p) = (p \times p)\ (p \times p)$$

in forma estesa è:

$$\vec{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \\ Y_p \end{pmatrix} = \begin{pmatrix} Y_{11} & Y_{12} & \cdots & Y_{1p} \\ Y_{21} & Y_{22} & \cdots & Y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{p1} & Y_{p2} & \cdots & Y_{pp} \end{pmatrix} = \begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1p} \\ l_{21} & l_{22} & \cdots & l_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ l_{p1} & l_{p2} & \cdots & l_{pp} \end{pmatrix} \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pp} \end{pmatrix} \tag{1.2}$$

Una generica componente di $\vec{Y}$, ad esempio la prima, si esprimerà come:

$$Y_1 = (Y_{11}, Y_{12}, \ldots, Y_{1p}) = \left(\sum_{i=1}^{p} l_{1i} X_{i1}, \sum_{i=1}^{p} l_{1i} X_{i2}, \ldots, \sum_{i=1}^{p} l_{1i} X_{ip} \right) = l_1^T \vec{X} \tag{1.3}$$

In sintesi, si ha che l'i-esima componente di $\vec{Y}$ è data da:

$$Y_i = l_i^T \vec{X} \tag{1.4}$$

a cui corrisponde una varianza pari a:

$$Var(Y_i) = \vec{l}_i^{\,T} \Sigma \vec{l}_i \tag{1.5}$$

e una covarianza di:

$$Cov(Y_i, Y_j) = \vec{l}_i^{\,T} \Sigma \vec{l}_j \tag{1.6}$$

$\overline{\overline{L}}$ è la matrice caratteristica della trasformazione lineare, mentre le Y_i sono dette *componenti principali*.

Il vettore multivariato $\vec{Y} = (Y_1 \quad Y_2 \quad \dots \quad Y_p)^T$ è tale che il primo elemento Y_1 comprenda la maggiore variabilità possibile (e quindi maggiori informazioni) delle variabili originarie, e che Y_2 rappresenti la maggiore variabilità delle X_i dopo la prima componente, e così fino a Y_p che tiene conto della più piccola frazione dell'originaria varianza. Perciò le componenti principali sono quelle combinazioni lineari delle variabili aleatorie X_i a norma unitaria che ne rendono massima la varianza e che sono incorrelate.

1.1 *La varianza delle componenti principali*

Si è definita la varianza delle componenti principali secondo l'espressione (1.5), ossia:

$$Var(Y_i) = \vec{l}_i^{\,T} \Sigma \vec{l}_i$$

Occorre però porre un vincolo sul vettore dei coefficienti. Supponendo di aver trovato un vettore $\vec{l}_1$ che massimizzi la varianza di Y_1, tale varianza potrà essere ulteriormente incrementata utilizzando anziché il vettore $\vec{l}_1$ appena trovato, un nuovo vettore $c\,\vec{l}_1$, con $c > 1$. Con tale ragionamento si otterranno un'infinità di soluzioni, note a meno di un fattore di proporzionalità c.

Pertanto per avere un'unica soluzione è necessario porre un vincolo sugli elementi del vettore $\vec{l}_1$, espresso nella seguente condizione:

$$\vec{l}_1^{\,T} \vec{l}_1 = 1$$

ovvero il vettore $\vec{l}_1$ deve avere norma unitaria.

Per individuare la prima componente principale bisognerà risolvere il seguente problema di massimo vincolato:

$$Var(Y_1) = \vec{l}_1^{\,T}\Sigma\vec{l}_1 \equiv \max \qquad con \quad \vec{l}_1^{\,T}\vec{l}_1 = 1 \tag{1.7}$$

1.2 *Procedura di estrazione delle componenti principali della matrice di covarianza*

Data la matrice di covarianza Σ, dovendo perseguire l'obiettivo stabilito dalla (1.7), si definisce come funzione obiettivo da massimizzare la funzione di Langrange:

$$P = \vec{l}_i^{\,T}\Sigma\vec{l}_i - \lambda(\vec{l}_i^{\,T}\vec{l}_i - 1)$$

dove λ è il moltiplicatore di Lagrange.

Massimizzare la funzione obiettivo rispetto a $\vec{l}_i$ significa trovare l'opportuno vettore di pesi da assegnare alle variabili presenti nella matrice $\vec{X}$ in modo tale che la nuova variabile ottenuta, Y_i, spieghi la massima quota possibile della variabilità totale, Σ.

Trattandosi di un problema di massimo vincolato, la soluzione si trova uguagliando a zero la derivata , rispetto al vettore $\vec{l}_i$, della funzione Lagrangiana:

$$\frac{\partial P}{\partial \vec{l}_i} = 2\Sigma\vec{l}_i - 2\lambda\vec{l}_i = 2(\Sigma - \lambda I)\vec{l}_i = 0 \tag{1.8}$$

dove I è la matrice identità.

Dal teorema di Rouchè-Capelli, l'equazione (1.8) individua un sistema lineare omogeneo che ammette soluzioni se e solo se la matrice $(\Sigma - \lambda I)$ è singolare, ovvero:

$$\det(\Sigma - \lambda I) = 0 \tag{1.9}$$

Le soluzioni della (1.9) sono gli autovalori della matrice Σ, per cui la risoluzione della (1.9) comporta la ricerca del rango della matrice $(\Sigma - \lambda I)$. Poiché Σ ha dimensione $(p \times p)$, si avranno al massimo p soluzioni.

Ordinando le soluzioni λ_i in senso decrescente, si ha:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$$

Presa la massima soluzione λ_1 della (1.9), si troverà il vettore $\vec{l}_1$ corrispondente risolvendo il seguente sistema:

$$(\Sigma - \lambda I)\vec{l}_1 = 0 \tag{1.10}$$

e quindi:

$$\Sigma \vec{l}_1 = \lambda_1 \vec{l}_1 \tag{1.11}$$

Ne deriva che il problema del massimo vincolato si tradurrà in un problema di autovalori e autovettori, in quanto il vettore $\vec{l}_1$ non è altro che l'autovettore di norma unitaria della matrice Σ associata all'autovalore λ_1.

Moltiplicando entrambi i membri della (1.11) per $\vec{l}_1^{\,T}$, si ha:

$$\vec{l}_1^{\,T} \Sigma \vec{l}_1 = \lambda_1 \vec{l}_1^{\,T} \vec{l}_1$$

essendo il vettore $\vec{l}_1$ di norma unitaria, si ottiene:

$$\vec{l}_1^{\,T}\Sigma\vec{l}_1 = \lambda_1 = Var(Y_1)$$

La varianza della prima componente principale sarà dunque massimizzata in quanto si è scelta per λ_1 il più grande degli autovalori di Σ.

Per trovare le componenti successive alla prima si deve seguire un procedimento analogo che dovrà tener conto delle componenti già valutate.

Tutto questo discorso viene sintetizzato nel seguente teorema:

- **<u>TEOREMA:</u>** Sia Σ la matrice di covarianza associata al vettore casuale $(X_1 \quad X_2 \quad \dots \quad X_p)$. Indicati con $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ i p autovalori della matrice Σ e con $(\vec{e}_1 \quad \vec{e}_2 \quad \dots \quad \vec{e}_p)$ i rispettivi autovettori, la i-esima componente principale Y_i è data da:

  $$Y_i = \vec{e}_i^{\,T}\vec{X} \qquad con \;\; i = 1,2,\dots,p$$

  con questa scelta, risulta $Var(Y_i) = \lambda_i$ per $i = 1,2,\dots,p$ e $Cov(Y_i, Y_j) = 0$ per $i \neq j$.

- **<u>COROLLARIO 1:</u>** $\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_P^2 = \sum_{I=1}^{P} Var(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{I=1}^{P} Var(Y_i)$

  dove: → $\sum_{I=1}^{P} Var(X_i)$ è la varianza totale della popolazione;

  → $\sum_{I=1}^{P} Var(Y_i)$ è la varianza totale delle componenti principali.

Tale corollario sta ad indicare che la varianza dei dati di partenza si ridistribuisce in quella delle componenti principali.

1.3 *Proprietà delle componenti principali*

a) Per ogni $i \neq j$, si ha che $Y_i \perp Y_j$, il che equivale alla condizione $\vec{l}_i^T \vec{l}_j = 0$ (essendo Σ simmetrica, gli autovettori saranno a due a due ortogonali). Quindi la matrice di covarianza di $\vec{Y}$ è una matrice diagonale, pertanto si ha $Cov(Y_i, Y_j) = 0$, per $i \neq j$.

b) La somma degli autovalori è uguale alla traccia di Σ

$$\sum_{i=1}^{p} \lambda_i = tr(\Sigma) = \sum_{i=1}^{p} \sigma_i^2 .$$

Va evidenziato che l'estrazione delle componenti principali può essere effettuata anche dalla matrice di correlazione R (con la medesima procedura vista per Σ). In tal caso, si avrà:

$$\sum_{i=1}^{p} \lambda_i = tr(R) = p .$$

c) Il prodotto degli autovalori è uguale al determinante della matrice Σ:

$$\prod_{i=1}^{p} \lambda_i = \det(\Sigma)$$

oppure

$\prod_{i=1}^{p} \lambda_i = \det(R)$, se si lavora con la matrice di correlazione.

d) È sempre possibile scomporre una matrice di covarianza o una matrice di correlazione in un numero di componenti principali mai superiore al numero di variabili osservate.

e) Le componenti principali non sono indipendenti dall'unità di misura delle variabili. Se si moltiplica una variabile osservata per un valore costante, la matrice di covarianza cambia determinando una corrispondente variazione delle componenti principali.

f) Le componenti principali non variano per variabili standardizzate, mentre non è necessario standardizzare valori percentuali o rapporti tra grandezze che variano in intervalli limitati.

g) Essendo Σ una matrice simmetrica, gli autovalori λ_i associati sono reali.

h) Il rango della matrice Σ coincide con il numero di autovalori λ_i non nulli.

1.4 *Il rango della matrice di covarianza*

Bisogna a questo punto aprire una parentesi sul rango della matrice Σ. La matrice di covarianza Σ ha lo stesso rango della matrice dei dati $\vec{X}$ ed è simmetrica, per cui il suo rango sarà pari al numero di autovalori non nulli. Inoltre si dimostra che la matrice Σ è semi-definita positiva; ciò comporta che i suoi autovalori sono sempre o positivi o nulli. Quindi:

- se le righe della matrice dei dati sono linearmente indipendenti (possibile solo se il numero di osservazioni, ad esempio n, è maggiore di p), Σ avrà rango p, ed i suoi autovalori saranno tutti positivi.

- Se qualche autovalore risulta nullo (poniamo $p-k$ autovalori nulli), allora Σ avrà rango k, e si potranno determinare i k autovalori positivi.

In conclusione, si può asserire che se Σ ha rango p (ovvero se è definita positiva) si otterranno p autovalori positivi, p autovettori corrispondenti e quindi p componenti principali.

Se Σ è semi-definita positiva e di rango k, si determineranno k autovalori non nulli e k componenti principali.

1.5 *La scelta del numero delle componenti principali*

Si è partiti da p variabili ($X_1 \quad X_2 \quad \dots \quad X_p$), con l'obiettivo di sintetizzarle in un numero inferiore di variabili "artificiali". A seconda del rango della matrice Σ, si potranno trovare fino a p componenti principali. Il compito della PCA è quello di analizzare un numero di dati inferiore a quello di partenza, a tale scopo vengono elencati, di seguito, i criteri adoperati per ridurre il numero delle componenti principali da p a k, con $p \geq k$.

I criteri adoperati per la scelta del numero di componenti sono tre (*Criteri Euristici*), e sono:

1. Prendere solo quelle componenti che rappresentano l' 80-90% della variabilità complessiva, ovvero:

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \approx 80 - 90\% \tag{1.12}$$

dove il numeratore rappresenta la varianza delle prime k componenti principali, mentre il denominatore rappresenta la varianza di tutte le componenti principali.

2. Seguire la "*Regola di Kaiser*": prendere solo quelle componenti che hanno un autovalore maggiore o uguale ad uno, oppure, equivalentemente, le componenti che hanno varianza maggiore di quella media (ottenuta come media delle λ_i);

3. La scelta del numero di componenti (sufficienti a riprodurre con una buona approssimazione i dati di partenza) può essere fatta attraverso il grafico degli autovalori o "Screen Plot". All'interno del grafico si sceglie il numero di componenti corrispondente al punto di "gomito" della spezzata.

   Lo Screen Plot è costruito ponendo sull'asse delle ascisse i numeri d'ordine degli autovalori $(1,2,\ldots,k)$ e in ordinata gli autovalori ad essi corrispondenti $(\lambda_1,\lambda_2,\ldots,\lambda_k)$. I punti di coordinate (j,λ_j), con $j=1,2,\ldots,k$, vengono uniti con segmenti. Il numero di componenti principali da utilizzare sarà dato dal più piccolo k tale che a sinistra di k l'andamento dei λ_j sia fortemente decrescente, mentre a destra l'andamento deve essere pressoché costante, o comunque debolmente decrescente.

1.6 *Standardizzazione delle variabili d'origine*

Spesso può accadere che i dati d'origine, che si hanno a disposizione, siano caratterizzati da unità di misura non paragonabili tra loro oppure da ampiezze dei sottocampioni molto diverse. In tali condizioni, non è possibile lavorare con il campione noto ma è necessario standardizzare le variabili aleatorie.

A partire dal campione d'origine, $(X_1 \quad X_2 \quad \dots \quad X_p)$, e note la media e la deviazione standard delle popolazioni di appartenenza delle variabili X_i, il processo di standardizzazione permette di normalizzare le variabili:

$$\left(\frac{X_1 - \mu_1}{\sigma_1}, \frac{X_2 - \mu_2}{\sigma_2}, \dots, \frac{X_p - \mu_p}{\sigma_p} \right)$$

ed ottenere così un nuovo campione, $(Z_1 \quad Z_2 \quad \dots \quad Z_p)$, che avrà media nulla e varianza unitaria.

Per questo nuovo campione, si avrà che la matrice di covarianza delle Z_i coincide con la matrice di correlazione del campione d'origine:

$$\Sigma_Z = R_X \tag{1.13}$$

Pertanto, quando si ha un campione X_i "non omogeneo" si adopera la matrice di correlazione, R.

Si deve sottolineare che gli autovalori ottenuti con la matrice di correlazione R sono diversi da quelli della matrice di covarianza Σ relativa al campione d'origine.

1.7 *Interpretazione delle componenti principali*

L'interpretazione delle componenti principali è una fase del metodo assai delicata. Attribuire un significato "semantico" alle Y_i è spesso legato alla capacità, all'esperienza e alla sensibilità del ricercatore. Da questo punto di vista non è possibile formalizzare statisticamente tali caratteristiche.

Esiste, d'altro canto, un modo per individuare il significato insito nelle variabili "latenti" (dette così in quanto non è possibile effettuare una diretta misurazione di tali "parametri nascosti").

Si è detto che ogni componente principale si può esprimere nel seguente modo:

$$Y_i = l_{i1}X_1 + l_{i2}X_2 + \cdots + l_{ip}X_p \qquad con \quad i = 1,2,\ldots,p$$

Pertanto il generico coefficiente l_{ij} rappresenta il peso che la variabile X_j ha nella determinazione della componente principale Y_i (con $i = 1,2,\ldots,p$). Quanto più grande è l_{ij} (in valore assoluto), tanto maggiore sarà il peso che i valori X_j ($j = 1,2,\ldots,p$) hanno nel determinare la componente principale i-esima. Ciò significa che la componente principale Y_i sarà maggiormente caratterizzata dalle variabili X_j a cui corrispondono i coefficienti l_{ij} più grandi in valore assoluto. In tal modo sono proprio i coefficienti l_{ij} a conferire un significato alla componente principale Y_i.

Informazioni aggiuntive sono fornite dai coefficienti di correlazione tra le variabili X_j e l'i-esima componente principale Y_i. Si dimostra che:

$$r_{Y_i X_j} = Corr(Y_i, X_j) = e_{ij}\frac{\sqrt{\lambda_i}}{\sigma_j} \tag{1.14}$$

dove la (1.14) rappresenta il coefficiente di correlazione, e e_{ij} è un autovettore.

È chiaro che più il valore di tale coefficiente è elevato tanto maggiore sarà il legame tra X_j e Y_i. Ciò significa che a determinare il significato delle componenti principali saranno le variabili X_j con cui è maggiormente correlata. Inoltre, questo tipo di analisi può essere compiuta anche graficamente, ovvero: se si considera nel piano delle prime due componenti principali un cerchio di raggio unitario, è possibile valutare il coefficiente di correlazione tra le X_j e Y_1, e tra X_j e Y_2. Ogni variabile X_j verrà plottata all'interno del cerchio, con le seguenti coordinate: $(Corr(Y_1, X_j), Corr(Y_2, X_j))$. In questo modo, si avrà un'indicazione grafica di quali variabili determinino maggiormente l'una, l'altra o entrambe le componenti principali; di quali siano correlate positivamente e quali negativamente e così via.

Pertanto l'interpretazione delle componenti principali individuate viene di solito effettuata sulla base dell'osservazione della matrice di correlazione tra le variabili originarie e le componenti stesse nonché degli autovettori di ciascuna componente. Probabilmente uno dei maggiori punti deboli di tale strumento statistico è proprio questo: l'interpretazione dell'output risulta estremamente soggettiva poiché determinati valori dei coefficienti di correlazione possono risultare significativi per alcuni, non significativi per altri.

- **NOTA:** Se si lavora con le variabili standardizzate $(Z_1 \quad Z_2 \quad \dots \quad Z_p)$, si avrà un diverso valore del coefficiente di correlazione (in quanto gli autovalori della matrice Σ non coincidono con gli autovalori della matrice R).

1.8 *Interpretazione geometrica delle componenti principali*

Dal punto di vista geometrico, la matrice dei dati $\vec{X}$ è rappresentabile come p punti nello spazio dimensionale $\boldsymbol{R}^p$. Si è ampiamente detto che la PCA mira a ridurre il numero di variabili da analizzare, ciò si traduce, da un punto di vista geometrico, nel proiettare i p punti in un sottospazio $\boldsymbol{R}^k$, individuato in modo tale che la nuvola dei punti p in $\boldsymbol{R}^p$ sia deformata il meno possibile.

Le componenti principali individuano un nuovo sistema di coordinate che è tale da avere sul primo asse (Y_1) la massima variabilità del sistema, sul secondo si ha una varianza inferiore alla prima ma massima rispetto alle altre, e così via. Pertanto, si avrà che:

→ Y_1 spiega la massima varianza su riduzione uni-dimensionale;

→ $\{Y_1, Y_2\}$ spiegano la massima varianza su riduzione bi-dimensionale;

……

→ $\{Y_1, Y_2, \ldots, Y_p\}$ spiegano la varianza totale.

Le immagini che seguono permettono di comprendere al meglio il significato di quanto detto.

IMMAGINE 1: I Dati di partenza sono plottati nel piano R^3, ed i tre segmenti indicati sono i tre autovettori della matrice di covarianza.

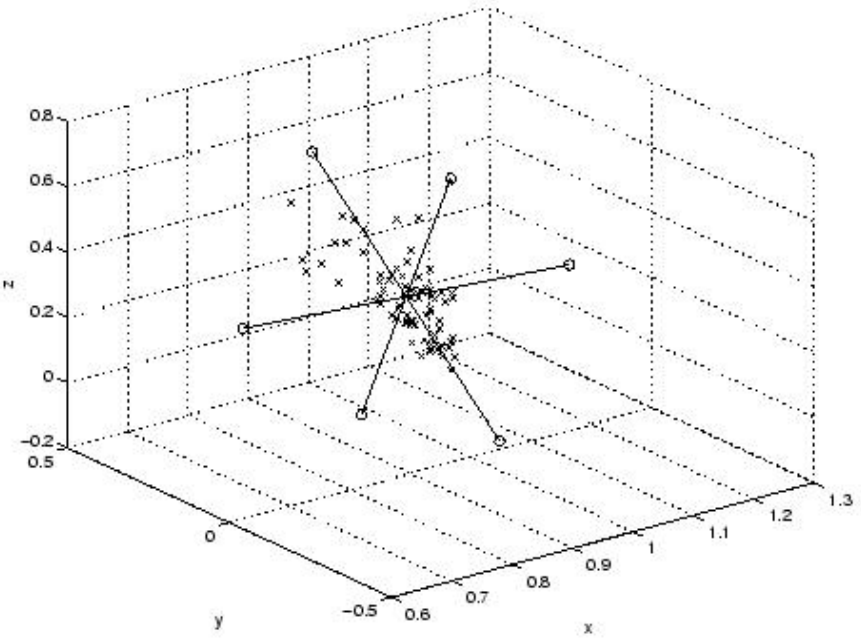


IMMAGINE 2: Si riporta il piano delle due componenti principali scelte.

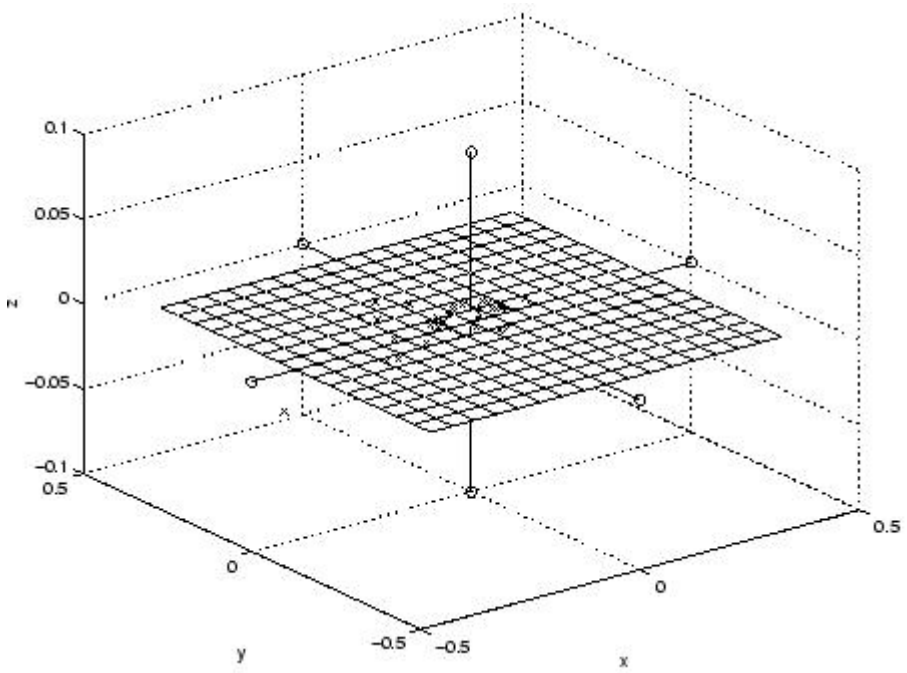


IMMAGINE 3: Rappresentazione visiva della perdita di informazione relativa alla prima componente principale.

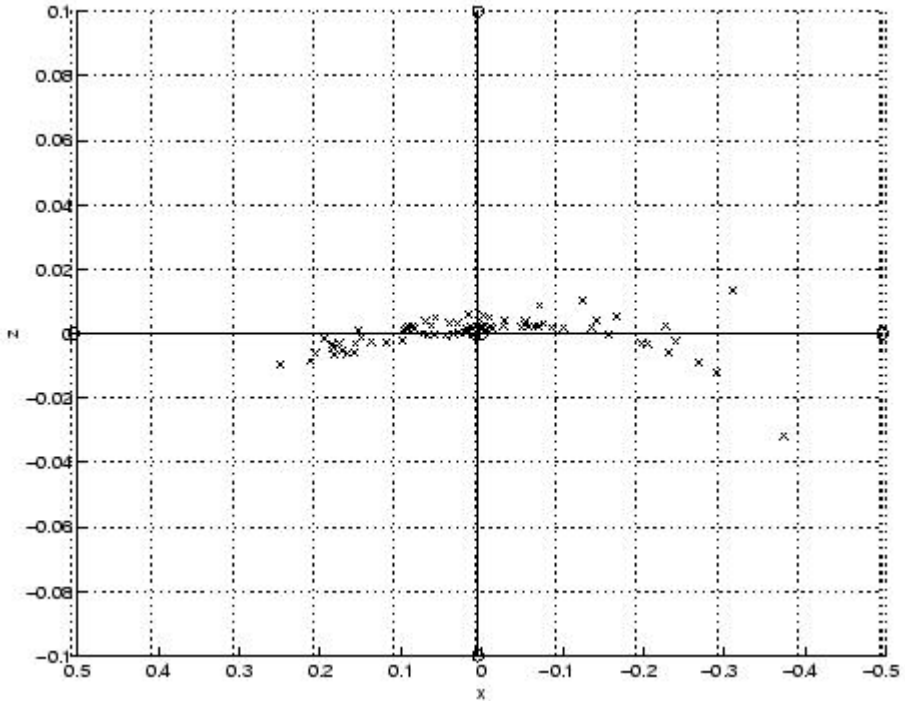
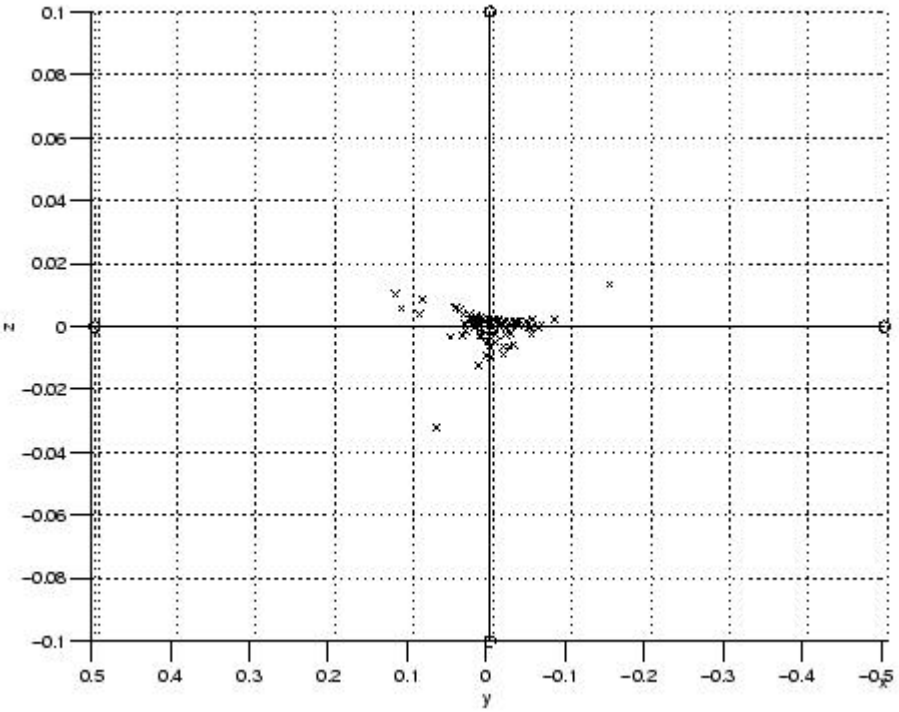


IMMAGINE 4: Rappresentazione visiva della perdita di informazione relativa alla seconda componente principale.



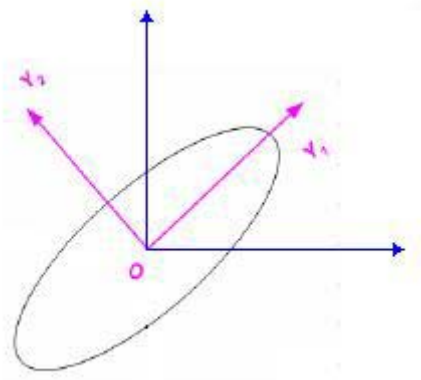
1.9 *Le componenti principali nel caso di campione multivariato gaussiano*

Se il campione di partenza $(X_1 \quad X_2 \quad \dots \quad X_p)$ è distribuito secondo una variabile aleatoria normale $N(0,\Sigma)$, allora la matrice di covarianza è definita positiva ed è dunque invertibile. Allora $\vec{x}^T\Sigma^{-1}\vec{x} = c^2$ descrive un ellissoide di centro nell'origine. Essendo $\Sigma^{-1} = \sum_{i=1}^{p} \frac{1}{\lambda_i}\vec{e}_i\vec{e}_i{}^T$, si ha che:

$$c^2 = \sum_{i=1}^{p} \frac{\left(\vec{e}_i{}^T x\right)^2}{\lambda_i} = \sum_{i=1}^{p} \frac{y_i{}^2}{\lambda_i} \tag{1.15}$$

La (1.15) fornisce un ellissoide nel sistema di coordinate con assi y_i e direzioni $\vec{e}_i$. Inoltre, le componenti principali sono indipendenti.

Passare dal campione casuale iniziale multivariato a quello delle componenti principali, equivale a ruotare gli assi coordinati fino a quando essi coincidono con gli assi dell'ellissoide di concentrazione costante. Per cui, per un campione multivariato gaussiano si ha un'interpretazione grafica differente.



1.10 *Sintesi delle caratteristiche delle componenti principali*

Le componenti principali (CP) saranno:

- tra loro incorrelate $Corr(Y_i, Y_j) = 0$, per $i \neq j$;
- ordinate in ragione della variabilità complessiva che esse possono sintetizzare: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$;
- la variabilità dei due sistemi di variabili, è tale che: $\sum_{I=1}^{P} Var(X_i) = \sum_{I=1}^{P} Var(Y_i)$;
- Inoltre, se si ha $\lambda_i \neq \lambda_j$, allora $\vec{e}_i \perp \vec{e}_j$. Se $\lambda_i = \lambda_j$, allora è sempre possibile scegliere $\vec{e}_i$ e $\vec{e}_j$ tali che $\vec{e}_i \perp \vec{e}_j$.

2. *Esempi*

2.1 Esempio n° 1

16 scrittori di romanzi sono stati valutati da un campione di lettori i quali hanno espresso opinioni sul tipo di contenuto e sul modo in cui sono state scritte delle opere dei scrittori indicati in base ai parametri riportati (vedi tab. 1): X1: giudizio, X2: leggibilità, X3: politica, X4: fantasia, X5: rilettura, X6: thriller, X7: attualità. Si vuole condurre un'analisi delle componenti principali.

Campione casuale multivariato X							
	X1	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**
	giud	**leggib**	**politic**	**fantast**	**rilettur**	**thrill**	**attual**
Fallaci	8.90	9.40	7.50	2.30	7.40	1.50	10.50
Tolkien	7.99	10.30	0.40	9.70	8.90	6.50	1.30
Rowling	8.01	10.56	1.60	10.02	7.59	7.01	1.68
Hardt	9.06	6.25	7.09	2.30	4.50	1.33	8.45
Cornwel	6.75	7.88	6.22	8.42	6.04	6.99	6.05
Camilleri	6.99	7.56	6.78	3.01	7.85	3.14	4.25
Mazzantini	8.33	8.01	2.52	3.50	6.55	2.47	6.70
Allende	5.95	7.73	3.50	5.80	6.45	3.20	2.40
Nasar	7.23	8.21	3.56	1.50	3.24	4.01	5.63
Eco	8.68	9.04	4.56	4.36	8.25	2.32	6.78
Gadda	9.12	7.56	6.89	1.23	5.89	2.32	7.88
Grass	7.56	6.12	6.52	4.56	6.01	1.25	6.12
King	6.22	8.44	1.01	2.14	7.13	4.22	1.23
Magris	8.14	7.06	2.10	6.55	4.12	0.88	2.26
Mishima	7.99	7.12	5.98	5.45	5.01	1.08	4.16
Lucarelli	6.04	6.42	3.01	5.02	3.01	6.89	3.01

tab. 1: Dati di partenza.

Mediante il comando "correlazione" di Excel, si ottiene la seguente matrice di correlazione:

	Matrice della correlazione di X						
	X1	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**
X1	1.0000	0.1676	0.3377	-0.1691	0.1699	-0.5003	0.6112
X2	0.1676	1.0000	-0.4485	0.4085	0.6898	0.4430	-0.1985
X3	0.3377	-0.4485	1.0000	-0.4554	-0.1115	-0.4862	0.7873
X4	-0.1691	0.4085	-0.4554	1.0000	0.2764	0.5708	-0.5688
X5	0.1699	0.6898	-0.1115	0.2764	1.0000	0.1289	-0.0710
X6	-0.5003	0.4430	-0.4862	0.5708	0.1289	1.0000	-0.4891
X7	0.6112	-0.1985	0.7873	-0.5688	-0.0710	-0.4891	1.0000

tab. 2: Matrice di Correlazione.

Attraverso il comando di Matlab "*svd*" è possibile determinare gli autovalori e gli autovettori associati alla matrice di correlazione, che sono riportati di seguito:

Autovettori della matrice di correlazione di X						
e1	**e2**	**e3**	**e4**	**e5**	**e6**	**e7**
-0.2793	-0.5153	0.3395	-0.5104	0.1871	-0.3805	-0.3176
0.3254	-0.5439	0.0048	0.1376	0.3763	0.6406	-0.1655
-0.4508	-0.0771	-0.5998	-0.0562	-0.3287	0.2409	-0.5118
0.4167	-0.0975	-0.1483	-0.6995	-0.4436	0.1876	0.2714
0.1894	-0.5910	-0.1050	0.4623	-0.4765	-0.3859	0.1188
0.4353	0.0666	-0.6019	-0.0963	0.4491	-0.4489	-0.1772
-0.4625	-0.2635	-0.3601	-0.0717	0.2999	0.0341	0.7005

tab. 3: Matrice degli autovettori associati alla matrice di correlazione.

Autovalori della matrice di correlazione						
3.2778	0	0	0	0	0	0
0	1.7679	0	0	0	0	0
0	0	0.6649	0	0	0	0
0	0	0	0.5978	0	0	0
0	0	0	0	0.4911	0	0
0	0	0	0	0	0.1142	0
0	0	0	0	0	0	0.0863

tab. 4: Matrice degli autovalori associati alla matrice di correlazione.

A questo punto, bisogna adoperare uno dei tre criteri euristici per la determinazione del numero di componenti principali che dovranno rappresentare il campione iniziale.

- **1° CRITERIO:** In base al primo criterio (esposto nel §1.5), si sceglie un numero di componenti principali pari al numero di autovalori che riescono a ricoprire l'80-90% della variabilità totale. Pertanto di seguito si propone una tabella in cui sono indicate le percentuali di ciascuna λ_i, e anche le cumulate:

		percent	cumulata
λ1	3.2778	46.83%	46.83%
λ2	1.7679	25.26%	72.08%
λ3	0.6649	9.50%	81.58%
λ4	0.5978	8.54%	90.12%
λ5	0.4911	7.02%	97.14%
λ6	0.1142	1.63%	98.77%
λ7	0.0863	1.23%	100.00%
tot	**7.0000**	**100%**	

tab. 5: Percentuali degli autovalori.

Si può notare come gli autovalori λ_i siano ordinati in maniera decrescente ed inoltre essi corrispondano alle stime delle varianze campionarie delle Y_i.

Dalla tabella 5, si evince che i primi due valori delle λ_i ricoprono il 72% della varianza totale, mentre se si considera anche il terzo autovalore si arriva fino all'81,58%.

Si decide di adoperare solo le prime due varianze, λ_1 e λ_2, per cui le componenti principali si esprimeranno nel seguente modo:

$$Y_1 = -0.2793(giud) + 0.3254(leggib) - 0.4508(politic) + 0.4167(fantas) + 0.1894(rilett) + 0.4353(thrill) - 0.4625(attual)$$

$$Y_2 = 0.5153(giud) - 0.5439(leggib) - 0.0771(politic) - 0.0975(fantas) - 0.5910(rilett) + 0.0666thrill) - 0.2635(attual)$$

(2.1)

I coefficienti delle combinazioni lineari coincidono con le componenti degli autovettori corrispondenti ai due autovalori scelti.

In base ai pesi riportati per queste combinazioni lineari, si ha:

- **Componente Y_1 :** tiene conto soprattutto del contenuto indicando una certa preferenza per i seguenti temi: politica, fantasia, thriller e attualità.
- **Componente Y_2 :** tiene conto principalmente del modo in cui è stato scritto il romanzo, esprimendo la preferenza per: il giudizio, la leggibilità e la rilettura.

- **2° CRITERIO:** Secondo la regola di Kaiser, andrebbero presi i primi due autovalori in quanto il loro valore risulta essere maggiore di 1.

λ1	3.2778
λ2	1.7679
λ3	0.6649
λ4	0.5978
λ5	0.4911
λ6	0.1142
λ7	0.0863

tab. 6: Gli autovalori.

- **3° CRITERIO:** Dallo Screen Plot, si evince che il numero di autovalori deve essere di 3, in quanto in corrispondenza di quel valore si ha il brusco cambiamento di pendenza.

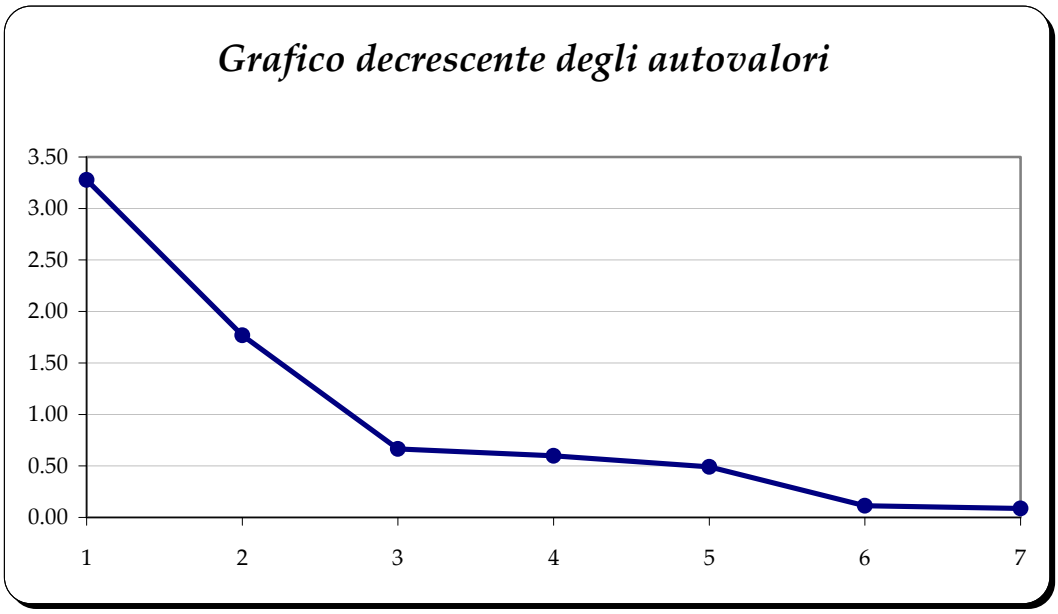


fig. 1: Screen Plot degli autovalori.

Poiché due metodi su tre restituiscono lo stesso risultato, si decide che le componenti principali sono solo due, e sono quelle indicate dalle (2.1).

2.2 *Esempio n° 2*

E' stata impostata una prova di confronto tra erbicidi per il diserbo chimico della barbabietola da zucchero. L'efficacia di ogni p.a. è stata misurata attraverso la percentuale di ricoprimento di sei specie infestanti (Polygonum lapathyfolium, Chenopodium polyspermum, Echinochloa crus-galli, Amaranthus retroflexus, Xanthium strumarium e Polygonum aviculare). Si vuole esprimere un giudizio di merito tra le diverse soluzioni erbicide, considerando l'insieme delle specie infestanti rilevate. I dati ottenuti sono riportati in tabella:

Tabella - Flora infestante rilevata con diversi prodotti diserbanti.

Erbicida	Code	POLLA	CHEPO	ECHCH	AMARE	XANST	POLAV
Tryflusulfuron methyl + olio	A	0.1	33	11	0	0.1	0.1
Tryflusulfuron methyl + phenmedipham + olio	B	0.1	3	3	0	0.1	0
Quinmerac + chloridazon + phenmedipham	C	7	19	19	4	7	1
Phenmedipham + etophumesate	D	18	3	28	19	12	6
Phenmedipham + etophumesate + chloridazon	E	5	7	28	3	10	1
Phenmedipham + etophumesate + metamitron	F	11	9	33	7	10	6
Phenmedipham + desmedipham + ethofumesate	G	8	13	33	6	15	15
Phenmedipham + desmedipham + ethofumesate + chloridazon	H	18	5	33	4	19	12
Phenmedipham + desmedipham + ethofumesate + metamitron	I	6	6	38	3	10	6

tab. 7: Dati di partenza.

Mediante Excel, si valuta la matrice di correlazione:

	Matrice della correlazione di X					
	POLLA	**CHEPO**	**ECHCH**	**AMARE**	**XANST**	**POLAV**
POLLA	1.00000	-0.47143	0.63094	0.74998	0.82359	0.61526
CHEPO	-0.47143	1.00000	-0.36033	-0.37682	-0.45998	-0.28664
ECHCH	0.63094	-0.36033	1.00000	0.39646	0.84664	0.69440
AMARE	0.74998	-0.37682	0.39646	1.00000	0.44580	0.31777
XANST	0.82359	-0.45998	0.84664	0.44580	1.00000	0.84416
POLAV	0.61526	-0.28664	0.69440	0.31777	0.84416	1.00000

tab. 8: Matrice di Correlazione.

Con l'ausilio del software Matlab, si valutano gli autovalori e gli autovettori:

Autovettori della matrice di correlazione di X					
e1	**e2**	**e3**	**e4**	**e5**	**e6**
-0.4601	-0.2212	0.2498	-0.1658	0.6411	0.4884
0.2937	0.4726	0.8208	0.0553	0.1058	-0.0492
-0.4288	0.3177	-0.0436	0.7830	-0.1780	0.2617
-0.3400	-0.5991	0.5084	0.0570	-0.4600	-0.2285
-0.4818	0.2519	-0.0596	-0.0369	0.3377	-0.7651
-0.4128	0.4521	0.0016	-0.5931	-0.4695	0.2300

tab. 9: Matrice degli autovettori.

Autovalori della matrice di correlazione					
3.858	0	0	0	0	0
0	0.9366	0	0	0	0
0	0	0.675	0	0	0
0	0	0	0.3039	0	0
0	0	0	0	0.192	0
0	0	0	0	0	0.0344

tab. 10: Matrice degli autovalori.

Si applicano, ora, i tre criteri di scelta del numero delle componenti principali:

- **1° CRITERIO:** Secondo questo metodo, il numero di autovalori da scegliere è pari a due, in quanto i primi due autovalori riescono ad esprimere il 79,91% della varianza totale.

		percent	cumulata
λ1	3.8580	64.30%	64.30%
λ2	0.9366	15.61%	79.91%
λ3	0.6750	11.25%	91.16%
λ4	0.3039	5.07%	96.23%
λ5	0.1920	3.20%	99.43%
λ6	0.0344	0.57%	100.00%
tot	**6.000**	**100%**	

tab. 11: Percentuale degli autovalori.

Per cui, le componenti principali sono:

$$Y_1 = 0.4601(polla) + 0.2937(chepo) - 0.4288(echch) - 0.34(amare) - 0.4818(xanst) - 0.4128(polav)$$

$$Y_2 = -0.2212(polla) + 0.4726(chepo) + 0.3177(echch) - 0.5991(amare) + 0.2519(xanst) + 0.4521(polav)$$

(2.2)

In base ai pesi riportati per queste combinazioni lineari, si ha:

– **Componente Y_1 :** tiene conto soprattutto delle seguenti specie infestanti: polla, di echch, di xanst e polav.

– **Componente Y_2 :** tiene conto principalmente delle seguenti specie infestanti: chepo e amare.

- **2° CRITERIO:** Secondo la regola di Kaiser, andrebbe preso solo il primo autovalore in quanto risulta essere maggiore di 1.

λ1	3.8580
λ2	0.9366
λ3	0.6750
λ4	0.3039
λ5	0.1920
λ6	0.0344
tot	**6.000**

tab. 12: Gli autovalori.

- **3° CRITERIO:** Dallo Screen Plot, si evince che il numero di autovalori deve essere di 2, in quanto in corrispondenza di quel valore si ha il brusco cambiamento di pendenza.

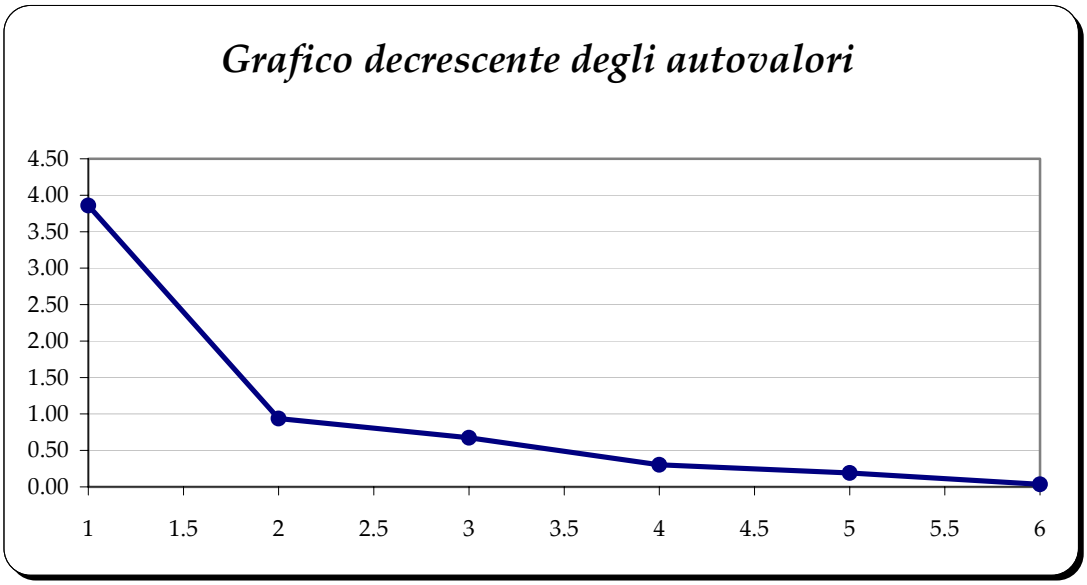


fig. 2: Screen Plot degli autovalori.

Poiché due metodi su tre restituiscono lo stesso risultato, si decide che le componenti principali sono solo due, e sono quelle indicate nella (2.2).

Appendice – Alcune definizioni

COVARIANZA: La covarianza tra variabili aleatorie X_i e X_j è la quantità:

$$\text{cov}(X_i, X_j) = \sigma_{X_i X_j} = E\left[\left(X_i - \mu_{X_i}\right)\left(X_j - \mu_{X_j}\right)\right] = E\left[X_i X_j\right] - \mu_{X_i}\mu_{X_j}$$

La covarianza è una misura della relazione lineare tra due variabili aleatorie.

MATRICE DI COVARIANZA: Misura il grado di correlazione tra due variabili.

$$\begin{pmatrix} \sigma_1^2 & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_p) \\ \text{cov}(X_2, X_1) & \sigma_2^2 & \cdots & \text{cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_p, X_1) & \text{cov}(X_p, Y_2) & \cdots & \sigma_p^2 \end{pmatrix}$$

Se la matrice di covarianza viene stimata, allora il singolo elemento della stessa è così definito:

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ik} - \bar{x}_{i\bullet})(x_{jk} - \bar{x}_{j\bullet})$$

dove il pedice i è relativo alle righe, mentre j alle colonne.

CORRELAZIONE: La correlazione tra variabili aleatorie X_i e X_j è la quantità:

$$\rho = \frac{\text{cov}(X_i, X_j)}{\sqrt{Var(X_i)Var(X_j)}} = \frac{\sigma_{X_i X_j}}{\sigma_{X_i}\sigma_{X_j}}$$

MATRICE DI CORRELAZIONE: Misura il grado di correlazione lineare tra due variabili.

$$\begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix}$$

Se la matrice di covarianza viene stimata, allora il singolo elemento della stessa è così definito:

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii} s_{jj}}}$$

dove il pedice i è relativo alle righe, mentre j alle colonne.