Capitolo 3
Analisi Statistica Multivariata delle Componenti Principali

3.1 Analisi Statistica Preliminare

La statistica fornisce strumenti e metodi per organizzare, riassumere e rappresentare in modo significativo i dati raccolti, per evidenziare gli aspetti rilevanti ivi contenuti e descrivere quindi le caratteristiche della popolazione.

La statistica descrittiva si occupa dell'analisi dei dati osservati, prescindendo da qualsiasi modello probabilistico che descriva il fenomeno in esame e dal fatto che l'insieme dei dati sia un campione estratto da una popolazione più vasta o sia invece l'intera popolazione. Lo scopo basilare della statistica descrittiva è di ridurre il volume dei dati osservati, esprimendo l'informazione rilevante contenuta in tali dati per mezzo di grafici e indicatori numerici che li descrivono; inoltre, possono essere fatte indagini di tipo comparativo e si può verificare l'adattarsi dei dati sperimentali a un certo modello teorico.

Quando si raccolgono dei dati su una popolazione o su un campione, i valori ottenuti si presentano inizialmente come un insieme di dati disordinati; i dati che non sono stati organizzati, sintetizzati o elaborati in alcun modo sono chiamati dati grezzi. Mediante l'analisi statistica descrittiva è possibile organizzare e sintetizzare i dati in modo da poter evidenziare le loro caratteristiche importanti e individuare le informazioni da essi fornite. Si va quindi ad operare su una "matrice di dati" che raccoglie le osservazioni effettuate su *n* unità statistiche con riferimento a *p* variabili.

Per riassumere l'impatto di un determinato fenomeno su una generica popolazione è utile riuscire a sintetizzare le numerose informazioni a disposizione con misure tali da compendiare le principali caratteristiche e fornire un'indicazione più facilmente utilizzabile, propriamente dette *indici statistici*.

3.1.1 Parametri statistici principali

La Media (μ) è l'indice statistico più utilizzato per la facilità di calcolo e per le proprietà di cui gode. Essa esprime la posizione globale di una distribuzione di frequenza.

Si definisce media aritmetica di N dati $x_1, x_2, \ldots x_N$ la quantità

$$\mu = \frac{1}{N}\sum_{i=1}^{N} \chi_i$$

Nel nostro caso andremo a considerare uno stimatore corretto della media μ di una popolazione, che chiameremo **Media Campionaria** $\overline{X}$, il cui valore numerico rappresenta una stima puntuale di μ.

Uno stimatore puntuale è uno stimatore corretto per il parametro, quando il suo valore medio coincide con il valore del parametro da stimare per qualsiasi suo valore, ovvero

$$E(\overline{X}) = \mu$$

dove

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} \chi_i$$

La **Moda** di una distribuzione statistica di frequenza rappresenta il valore che compare con la massima frequenza. È un indice che sintetizza le modalità di un carattere qualitativo sconnesso rilevato su n unità statistiche.

La **Mediana** si definisce come la modalità associata all'unità statistica che si trova nella posizione centrale della distribuzione, quando le unità statistiche sono ordinate. Per il calcolo della mediana è quindi necessaria l'operazione preliminare di ordinamento delle n unità statistiche, successivamente si procede con il calcolo dell'unità statistica che si trova al centro della distribuzione. Se la taglia N del campione è pari la mediana è

$$m = \frac{1}{2}\left(\chi_{\frac{N}{2}} + \chi_{\frac{N+1}{2}}\right)$$

mentre se è dispari la formula da utilizzare sarà

$$m = \chi_{\frac{N+1}{2}}$$

La mediana è una misura della tendenza centrale per una distribuzione più robusta rispetto ad altri indici, in quanto non è influenzata da nessuna modalità estrema che potrebbe rappresentare un caso di anomalia. Dalle definizioni è chiaro che la media, la moda e la mediana consentono di valutare l'ordine di grandezza della variabile aleatoria e aiutano a localizzare la distribuzione, ovvero ad individuare attorno a quale valore si incentra la distribuzione stessa.

Lo **Scarto Quadratico Medio** o **Deviazione Standard** ha simbolo σ nel caso della popolazione ed ***s*** nel caso di un campione e rappresenta una misura dello scostamento dalla media, fornisce quindi un'idea della dispersione della variabile casuale intorno alla media.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(\chi_i - \overline{\chi})^2}{N-1}}$$

In statistica l'**Errore Standard** di una misura è definito come la stima della deviazione standard dello stimatore, cioè una misura della sua imprecisione. Se lo stimatore è la media campionaria di N variabili indipendenti con medesima distribuzione statistica, lo *standard error* è:

$$se = \frac{\sigma}{\sqrt{N}}$$

La **Varianza (σ^2)** è un indicatore di dispersione, in quanto è nulla solo nei casi in cui tutti i valori sono uguali tra di loro (e pertanto uguali alla loro media) e cresce con il crescere delle differenze reciproche dei valori.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (\chi_i - \overline{\chi})^2}{N - 1}$$

Anche qui andremo a considerare uno stimatore puntuale corretto del campione $X_1, X_2, \ldots X_N$.
La **Varianza Campionaria** sarà quindi calcolabile con la seguente formula

$$S^2 = \frac{\sum_{i=1}^{N} (\chi_i - \overline{\chi})^2}{N - 1}$$

il cui valor medio

$$E(S^2) = \sigma^2$$

La **Curtosi (γ_2)** misura il "grado di appiattimento" di una distribuzione rispetto alla curva normale, cioè il grado di addensamento dei valori attorno alla media.
Una *curtosi positiva* ($\gamma_2 > 0$) indica che ci sono più valori agli estremi della distribuzione di quanto aspettato, quindi avremo distribuzioni piatte con code ampie, mentre nel caso di *curtosi negativa* ($\gamma_2 < 0$) saremo in presenza di curve appuntite con code piccole, cioè meno valori di quelli attesi agli estremi. L'indice γ_2 è uguale a zero nel caso di distribuzione normale o gaussiana.

$$\gamma_2 = \frac{\sum \frac{(\chi - \mu)^4}{\sigma^4}}{N} - 3 \qquad \text{(Formula di Fisher)}$$

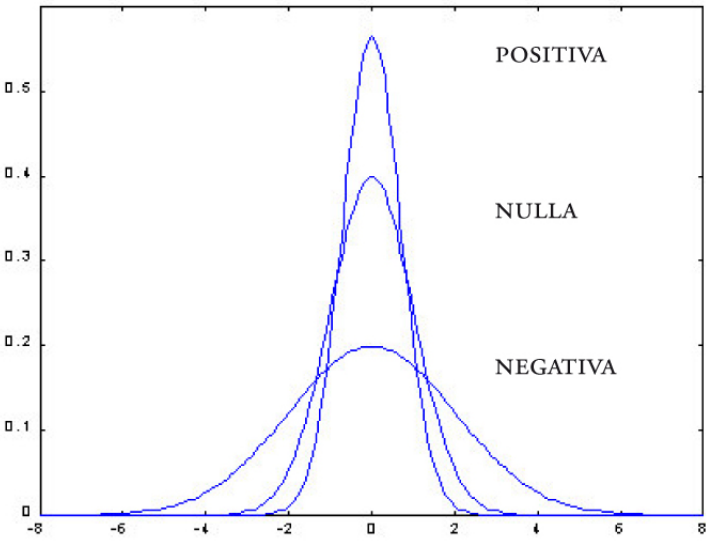

Figura **3.1**: Distribuzioni con diversi gradi di curtosi.

La curtosi misura, insieme all'asimmetria, quanto una distribuzione sia simile o no alla distribuzione normale presa a modello da molte tecniche statistiche.

L'**Asimmetria** è un valore caratteristico che misura la simmetria di una distribuzione rispetto alla media e di conseguenza ne indica il grado di asimmetria intorno ad essa.

Può essere misurata confrontando gli indici di posizione più comuni, ad esempio la media e la mediana. Se la Mediana è minore della media la gran parte delle osservazioni si posiziona su valori bassi, ma alcuni valori particolarmente alti spostano la media verso destra: si parla in tal caso di *asimmetria positiva* (la distribuzione presenta una "coda" verso il semiasse positivo delle ascisse). Una distribuzione con *asimmetria negativa* ha invece una coda più lunga a sinistra del valore centrale. Valori assoluti di asimmetria maggiori di 1 indicano distribuzioni molto diverse da una distribuzione normale. Per indice di asimmetria nullo avremo che gli scarti negativi sono bilanciati da quelli positivi e quindi saremo in presenza di una distribuzione simmetrica, come nel caso normale o gaussiano.

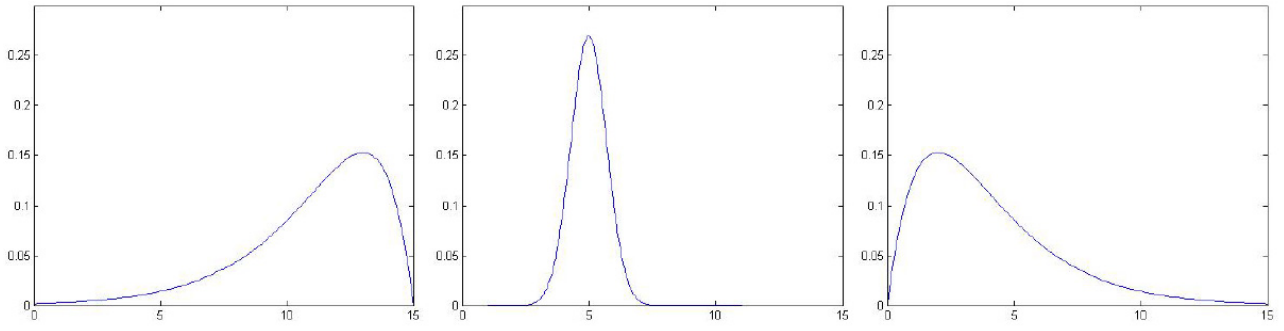

Figura **3.2**: Distribuzioni con differenti coefficienti di asimmetria.

3.2 Analisi Statistica Multivariata o Multidimensionale

Quando in statistica succede di dover analizzare fenomeni descritti da tre o più variabili, l'analisi descrittiva diventa estremamente difficoltosa. L'Analisi Statistica Multivariata considera i fenomeni (biologici, clinici, fisici, economici, sociali, ecc.) nella loro interezza, tenendo quindi conto delle diverse caratteristiche che meglio servono a descrivere i fenomeni stessi: da tre o quattro variabili sino, in casi particolari, a diverse centinaia di variabili.

Per tale vocazione la Statistica Multivariata è orientata a fornire rappresentazioni più che a valutare test di ipotesi, anche se questi non mancano. Oltre a tale premessa non esiste una precisa definizione di statistica multivariata, né questa sarebbe condivisa unanimemente dagli addetti al lavoro.

L'analisi multivariata raccoglie una grande famiglia di tecniche matematico-statistiche che,
come dice il nome stesso, consentono l'esame simultaneo di tutte le variabili caratteristiche di un set di dati. Grazie all'esame simultaneo di tutte le variabili, l'analisi multivariata consente una completa classificazione dei dati.

Per molti aspetti, l'analisi multivariata può essere considerata un'estensione sia della statistica monovariata che di quella bivariata (per alcuni sono invece queste ultime due da considerarsi riduzione della realtà multivariata). Per questo fatto, molti concetti e molte spiegazioni prendono spunto da esempi di situazioni bivariate, col vantaggio di ragionare su semplici grafici 2D.

L'analisi multivariata non ha soltanto scopi di classificazione, poiché sono possibili anche utilizzazioni di tipo modellistico, analoghe per finalità all'analisi di regressione multipla. In campo ambientale è intuitivo che l'approccio di classificazione è quello più applicato, in quanto consente di esaminare similitudini tra i dati.

Lo studio delle metodologie statistiche rivolte all'analisi congiunta di più variabili può essere fatto risalire ai primi anni del secolo con i contributi di Spearman e Pearson, anche se fu agli inizi degli anni '30 che Hotelling formalizzò i principi metodologici che diventarono le basi dell'analisi dei dati. A quel tempo la mancanza di adeguati strumenti di calcolo rappresentava un ostacolo alla piena diffusione di tali metodi, che si svilupparono in un contesto quasi esclusivamente teorico dando vita a quel filone di studi noto come Analisi Multivariata.

La scuola anglosassone utilizza il termine "analisi multivariata" per enfatizzare il ruolo delle molteplici variabili e delle rispettive distribuzioni osservate e teoriche, dando maggiore importanza all'inferenza statistica.

L'analisi multivariata si distingue dalla definizione francese di analisi dei dati, detta Analisi Multidimensionale, che si propone di evidenziare la struttura latente sottostante al sistema in esame tramite una riduzione della dimensionalità dello spazio di rappresentazione delle variabili o di quello delle unità statistiche, in modo che l'informazione strutturale estratta possa ritenersi ottimale in relazione ad un criterio prefissato.

Per diversi anni la contrapposizione tra la scuola anglosassone di analisi multivariata e la scuola francese di analisi multidimensionale fu molto aspra; il passar degli anni ha contribuito ad ammorbidire le posizioni e ad attenuare le distinzioni.

L'analisi multidimensionale a partire degli anni'80 si è mossa nella direzione di un bilanciamento tra i due approcci.

L'analisi multidimensionale dei dati consente:

- Il trattamento simultaneo di numerose variabili ed osservazioni,
- La visualizzazione di associazioni complesse,
- La riduzione del numero di variabili e di modalità osservate,
- La ricostruzione di tipologie di osservazioni,
- L'analisi di fenomeni evolutivi complessi,
- La validazione dei dati,
- L'identificazione di modelli.

Non si tratta, quindi, solo di *presentare* dei dati, ma di *analizzare*, *scoprire*, e a volte *verificare* o *rifiutare* determinate ipotesi. Le tecniche di analisi multidimensionale non hanno più la semplicità dei metodi di statistica descrittiva elementare, perché non si richiede più solo la semplificazione di una realtà complessa, ma anche l'esplorazione di una realtà nascosta. Infatti, la matematica dell'analisi multivariata è basata sull'algebra delle matrici, certamente più complessa della matematica incontrata nell'analisi univariata.

Particolare rilevanza assumono quindi la fase di preparazione e codifica dei dati, e la definizione di regole d'interpretazione e di validazione delle rappresentazioni fornite dalle tecniche utilizzate. Il punto di partenza è il sistema osservato, mentre particolare importanza assumono le conoscenze a priori che si fanno sul fenomeno indagato, in quanto possono condizionare la fase della raccolta dei dati. I dati vengono raccolti in tabelle (o matrici) e con l'ausilio di metodi di analisi multidimensionale si giunge all'analisi simultanea delle interrelazioni tra molte variabili correlate, con l'obiettivo di visualizzazione e interpretazione della struttura di vasti insiemi di dati. Il tipo di matrice in cui vengono raccolti i dati e il metodo di analisi è condizionato dalla natura degli stessi e dagli obiettivi che si intendono raggiungere. Se non si hanno particolari conoscenze sul fenomeno, la matrice sarà di tipo non strutturato e l'analisi sarà di tipo esplorativo, rivolta cioè allo studio delle relazioni tra l'insieme delle variabili e/o all'insieme delle unità. Se si hanno conoscenze preliminari la matrice sarà di tipo strutturato e l'approccio sarà si tipo esplicativo o confermativo. Tra i metodi esplorativi di base per l'analisi di variabili quantitative ricordiamo l'Analisi delle Componenti Principali.

3.3 Analisi delle Componenti Principali

L'analisi delle componenti principali (PCA, ovvero Principal Components Analysis) è uno dei primi metodi nati per il trattamento di dati multidimensionali quantitativi. Nasce con l'obiettivo di analizzare i dati tenendo conto della multidimensionalità. L'algoritmo della PCA fu sviluppato nei primi anni del XIX secolo, ma la tecnica ha cominciato ad essere intensivamente applicata soltanto dopo gli anni 70 a seguito della diffusione dei personal computer, in quanto la matematica connessa alla PCA richiede calcoli piuttosto complicati anche per matrici di dati aventi dimensioni ridotte. La caratteristica portante dell'Analisi delle Componenti Principali è la capacità di determinare similitudini tra campioni, indicando

simultaneamente anche le variabili che determinano similitudini o dissimilitudini. 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 in cui 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. La giustificazione teorica che sta alla base di questa analisi risiede nella possibilità di rendere "visibile" una nuvola di punti a *p* dimensioni, al fine di cogliere un qualche legame tra le variabili oggetto d'indagine. In un grafico bidimensionale o, al limite, tridimensionale è immediato cogliere l'esistenza o meno di un possibile legame tra le variabili. Quando, però, ci si sposta in ambiti dimensionali superiori a tre, i punti che definiscono le osservazioni non sono più rappresentabili in un piano e quindi occorre effettuare una riduzione dimensionale: ciò è possibile se tali punti a *p* dimensioni vengono proiettati in uno spazio bidimensionale. Dal momento che l'operazione di proiezione riduce le distanze originali, allora si massimizzano le proiezioni e si vincolano gli assi fattoriali, su cui si proiettano i punti, ad essere ortogonali tra loro. La proiezione tende a schiacciare i punti, rendendo le distanze proiettate più piccole di quelle nello spazio originario. L'obiettivo della PCA è rendere minima tale distorsione per tutte le coppie di punti e quindi definire un particolare piano in cui le distanze siano conservate il più possibile nella loro forma originaria e sia rappresentata al meglio la variabilità della nube dei punti osservati. L'Analisi delle Componenti Principali cerca di individuare nella nube dei punti (*p*-dimensionale) un sistema di assi ortogonali di riferimento, detti assi principali.
In una matrice (*n, p*), dove n sono le unità e p le variabili quantitative rilevate, la PCA si propone di determinare le *variabili di sintesi* che costituiscono la struttura di base delle relazioni osservate.
Si ipotizza che tali variabili, dette **Componenti Principali (CP),** siano legate linearmente alle variabili originarie (combinazione lineare delle variabili iniziali a mezzo degli assi principali o autovettori) e siano in numero minore di queste ultime, consentendo una riduzione dimensionale. Geometricamente le componenti principali sono assi che attraversano un set di dati multivariati, minimizzando la varianza delle vecchie variabili, cioè la distanza che si ottiene proiettando gli oggetti sugli assi delle nuove variabili. Il primo asse (o componente principale) giustificherà la porzione più grande di varianza e gli assi successivi saranno quelli che giustificheranno porzioni sempre più piccole della varianza. Un importantissimo aspetto è che ogni asse è comunque indipendente (ortogonale) dagli altri, quindi in termini strettamente matematici, ogni asse apporta un contributo "personale e indipendente " alla varianza e quindi alla descrizione del sistema.

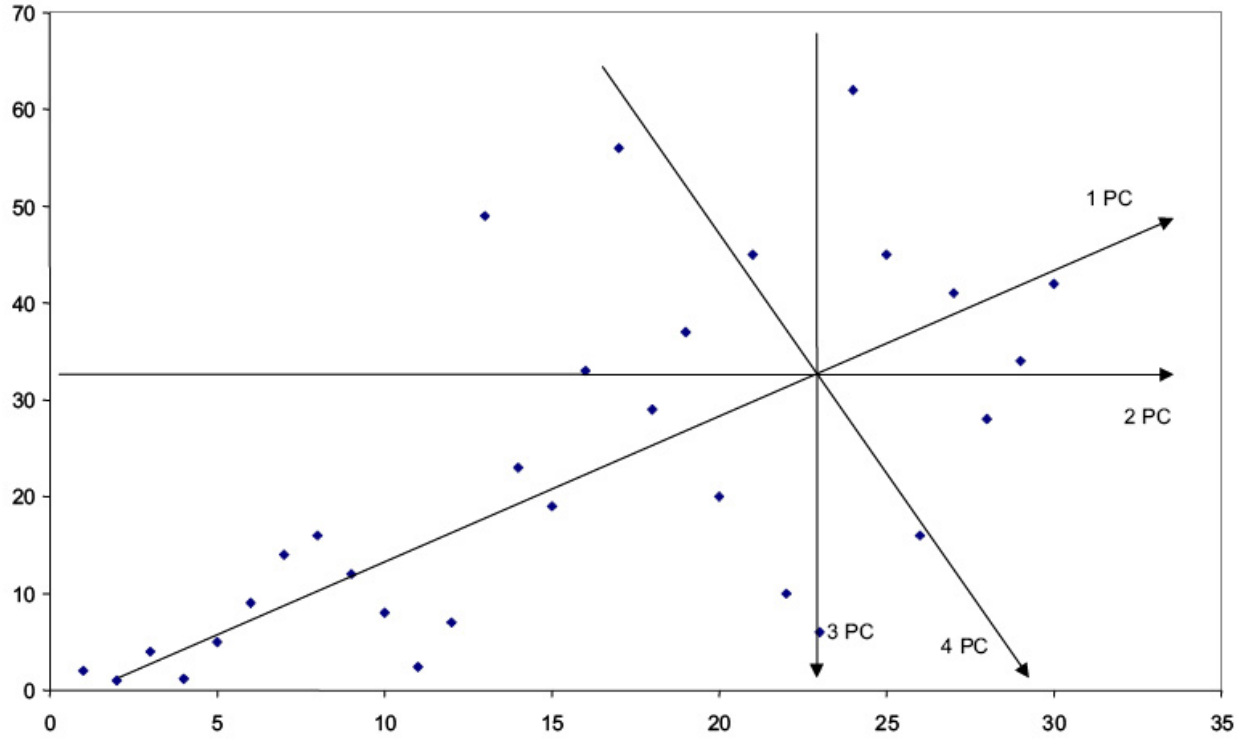

Figura **3.3:** Spiegazione geometrica della PCA.
Ogni nuovo asse (componente principale) spiega una porzione della varianza totale del sistema.

3.3.1 Definizione e determinazione delle componenti principali

Come detto, l'Analisi delle Componenti Principali provvede a trasformare in modo lineare le variabili originarie in nuove variabili.
Data una matrice dei dati riferiti ad *n* individui e *p* variabili quantitative**,** si sintetizzano i dati, nel senso di pervenire ad una riduzione delle colonne della matrice dei dati X, definendo un numero *q (q < p)* di variabili artificiali. La riduzione del numero delle variabili consente alle volte più agevoli sintesi interpretative. Dal punto di vista geometrico, la matrice dei dati Xn,p è rappresentabile come n punti nello spazio Rp. Si tratta di proiettare gli n punti in un sottospazio Rq, individuato in modo tale che la nuvola degli n punti in Rp sia deformata il meno possibile. Pertanto, con riferimento a *p* variabili, X_1, X_2,.., X_p (vettore casuale multivariato), la PCA consente di individuare altrettante *p* variabili (diverse dalle prime), Y_1, Y_2, …, Y_p (vettore multivariato), ognuna combinazione lineare delle *p* variabili di partenza. L'obiettivo della PCA consiste nel definire opportune trasformazioni lineari Y_i delle variabili osservate, facilmente interpretabili e capaci di evidenziare e sintetizzare l'informazione insita nella matrice iniziale 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. Per far ciò, come prima cosa occorre organizzare i dati da analizzare in una matrice X, avente per colonne i descrittori

e per righe gli oggetti. Gli oggetti possono essere campioni, stazioni di misura, o punti di campionamento, mentre i descrittori riportati lungo le colonne possono essere variabili, attributi o caratteristiche. Per una corretta applicazione della PCA è necessario che i descrittori siano di tipo quantitativo e che la loro distribuzione sia di tipo normale. Si assume, inoltre, che essi siano legati da relazioni lineari e che la matrice di dati non contenga un numero eccessivo di zeri.

Non sempre, però si ha la fortuna di trovarsi in presenza di una distribuzione di probabilità normale, perciò spesso occorre standardizzare le variabili e condurre l'analisi delle componenti principali non più sulle variabili originali bensì su quelle standardizzate.

La standardizzazione è un procedimento che riconduce una variabile aleatoria distribuita secondo una media μ e varianza σ^2, ad una variabile aleatoria con distribuzione "standard", ossia di media zero e varianza pari a 1. La nuova matrice dei dati standardizzata conterrà i valori di z_i calcolati con la seguente formula:

$$z_i = \frac{(\chi_i - \mu)}{\sigma}$$

Le conseguenze di realizzare una PCA standardizzata sono:

- L'analisi inizia dalla matrice di correlazione delle variabili, piuttosto che dalla matrice delle varianze/covarianze. Il calcolo della prima implica la standardizzazione necessaria, per cui i dati devono essere prima divisi per la deviazione standard,
- L'ovvia conseguenza della standardizzazione è che tutte le variabili avranno varianza unitaria, quindi la varianza totale è *p*, cioè il numero delle variabili, proprio come la somma degli autovalori,
- Le componenti con varianza (autovalore) inferiore a 1 possono essere tranquillamente trascurate nell'interpretazione dei risultati della PCA. In questo caso una componente non apporta informazioni utili, poiché la sua varianza è inferiore a quella di qualunque variabile standardizzata.

Una volta standardizzati i valori delle variabili osservate, che avranno dunque varianza uguale a 1 e media uguale a 0, la i-esima componente principale (*Yi*) può essere scritta come $Y_i = w_{i1}X_1 + w_{i2}X_2 + \ldots + w_{ip}X_p$, dove X_1, X_2, …, X_p sono le *p* variabili originarie standardizzate e $w_{i1}, w_{i2}, .. w_{ip}$ sono i valori dei pesi associati a ciascuna di esse.

Per la ricerca delle componenti principali è necessario applicare trasformazioni lineari in grado di diagonalizzare la matrice di correlazione. La diagonalizzazione si ottiene con una rotazione delle coordinate nella base degli autovettori (componenti principali).

Sulla diagonale della matrice di correlazione si incontrano una serie di 1, che presuppongono l'esistenza di una perfetta correlazione di ogni variabile con se stessa.

La matrice che diagonalizza la matrice di correlazione è la *matrice degli autovettori*, mentre la matrice diagonalizzata si chiama *matrice degli autovalori*. Ad ogni autovettore è associato un autovalore a cui corrisponde la varianza della componente principale associata. Se le variabili originarie erano parzialmente correlate tra loro alcuni autovalori avranno un valore trascurabile. Gli autovettori corrispondenti possono

essere trascurati e la rappresentazione può essere limitata solo agli autovettori con gli autovalori più grandi. In pratica, risolvere il problema degli autovalori e degli autovettori significa trovare p assi principali della generica distribuzione multinormale a p-dimensioni. Ad ogni autovettore corrisponde un autovalore ben determinato, che rappresenta la varianza spiegata dalla nuova variabile.

L'autovalore l_i corrispondente all'i-esima componente principale si ottiene risolvendo il seguente calcolo matriciale: $(R - l_i I)\, a_i = 0$, dove R è la matrice di correlazione tra le variabili originarie, I è la matrice identità e a_i è l'autovettore formato dai coefficienti che determinando la i-esima componente principale come combinazione lineare delle variabili originarie. Gli autovettori e gli autovalori derivano da una soluzione iterativa.

Gli autovalori l_i forniscono l'informazione su come si ripartisce la varianza sugli assi principali, dove la percentuale di varianza spiegata dall'i-esimo asse principale è data da

$$P_i = \frac{l_i}{\sum_i l_i}$$

Come risultato dell'Analisi delle Componenti Principali si ottengono due matrici: la prima in cui sono riportati gli *score*, ovvero i coefficienti delle CP, e la seconda che contiene i *loadings*, ossia i pesi assegnati alle variabili originarie nella definizione delle componenti principali.

Per scegliere il numero di Componenti Principali sufficiente a riprodurre con buona approssimazione i dati di partenza si considerano solo le componenti principali corrispondenti agli auto valori, che sono in grado di spiegare circa l'80% della varianza dei dati. Pertanto, da p variabili mutuamente correlate si passa a q nuove variabili indipendenti (solitamente 3 o 4), che da sole riescono a spiegare la gran parte della variabilità totale.