Università di Bologna
Sede di Rimini

Anno Accademico 2009-2010

Analisi statistica multivariata

(Alessandro Lubisco)

Materiale relativo al primo periodo:

- Analisi delle componenti principali

INDICE

ANALISI DELLE COMPONENTI PRINCIPALI

PREMESSA

Il campo dell'analisi multivariata consiste in quelle tecniche statistiche che considerano due o più variabili casuali correlate come una singola entità e cercano di produrre un risultato complessivo che tenga conto delle relazioni esistenti tra le variabili.

Molte delle tecniche multivariate inferenziali sono generalizzazioni delle classiche procedure univariate. Si pensi, per esempio, alla regressione multivariata.

C'è, tuttavia, un'altra classe di tecniche che è unica nel panorama delle tecniche multivariate. Il coefficiente di correlazione è un esempio calzante. Sebbene queste tecniche possano essere anche utilizzate nella statistica inferenziale, la maggior parte delle loro applicazioni è come strumento di analisi dei dati. In particolare, sono utilizzate come tecniche che cercano di descrivere la struttura multivariata dei dati. La PCA è una tecnica di questo tipo che, sebbene usata principalmente come tecnica descrittiva, può essere utilizzata altrettanto in procedure inferenziali.

L'analisi delle componenti principali (ACP) appartiene alle tecniche statistiche di analisi multivariata, ognuna delle quali ha caratteristiche proprie, presentando tuttavia connessioni concettuali e possibilità complementari di utilizzo, ciascuna rispetto alle altre.

I metodi dell'ACP vengono adottati nell'ambito di varie discipline, particolarmente la sociologia e la psicologia, ma anche la biologia, la medicina e l'economia. Tali metodi si applicano a un insieme di osservazioni iniziali (variabili osservate), eseguite al fine di disporre di un vasto ambito di informazioni, conseguenti a indagini condotte su campioni più o meno vasti. Così, per esempio, se per eseguire un'indagine, poniamo sui consumi, vengono scelte venti variabili cosiddette rilevanti (età, scolarità, numero dei familiari conviventi, reddito, numero di elettrodomestici posseduti, ecc.) e supponendo di intervistare mille persone, si ottiene una massa di ventimila dati, i quali, considerati in blocco, non sono immediatamente idonei a fornire informazioni sintetiche e riassuntive, anche se abbondano le notizie di dettaglio. L'obiettivo dell'ACP consiste allora nell'individu-

are alcuni (pochi) fattori di fondo che spieghino e diano ragione dei dati stessi. Tali fattori di fondo, o componenti, rappresentano delle dimensioni "ideali" dotate di significato. Così, nell'ambito del comportamento di consumo, potrebbero emergere componenti quali: "desiderio di essere accettati dal gruppo cui si aspira appartenere", "tendenza a incrementare il proprio senso di sicurezza", "desiderio di essere percepiti come appartenenti a un determinato status sociale", "motivazione al successo", "ottimismo e pessimismo verso il proprio futuro", e via dicendo.

In ambito economico-aziendale, l'ACP viene impiegata per:
- sintesi di indici di bilancio aziendale - valutazione della performance aziendale sulla base di *p* indici di bilancio
- sintesi di valutazioni espresse da consumatori/utenti con riferimento a un certo bene/servizio/azienda/punto vendita.
- "riduzione" della "dimensione" dell'informazione preliminare all'analisi dei gruppi (impiegata per la segmentazione del mercato)
- valutazione sintetica di caratteristiche territoriali: valutazione della qualità della vita con riferimento a un certo territorio, sulla base di *p* indicatori di reddito, consumo, dotazione di servizi, indicatori ambientali, ecc.

Intervistiamo i cittadini di una certa area su vari aspetti relativi ai servizi forniti, ad esempio, dal Servizio Sanitario Nazionale (competenza del personale, stato della struttura, orari di apertura, ecc.). Sulla base delle osservazioni raccolte possiamo trarre degli indici sintetici (meglio: un unico indice) di "gradimento" del servizio.

Il ricercatore tende dunque a scoprire delle dimensioni sottese (emergerà anche il concetto di "latenti") atte a dar ragione di un fenomeno collettivo altrimenti difficilmente decifrabile.

Ma, partiamo dall'inizio. Già verso la fine dell'800 ci si ponevano domande sulla possibilità di esprimere in forma sintetica insiemi di variabili tra loro correlate, magari proprio riducendo il numero delle variabili stesse. Le prime soluzioni sono state proposte da Galton nel 1889 e da Edgeworth nel 1891 lavorando su misure antropometriche. Tali studiosi cercarono di ricombinare tali misure individuando strutture lineari indipendenti che potevano essere interpretate come indici antropometrici non correlati.

All'inizio del 900 anche Karl Pearson determinò una soluzione simi-

le, anche se stava lavorando su un obiettivo differente. Il suo approccio era di tipo geometrico e intendeva determinare rette e piani che potessero approssimare un insieme di punti in uno spazio p-dimensionale.

In realtà, la formulazione più nota è attualmente quella proposta da Hotelling nel 1933 e si basa sull'ipotesi che i valori di un insieme di p variabili originarie siano determinati da un più ristretto insieme di variabili tra loro indipendenti. Tali variabili vengono determinate come combinazione lineare delle variabili originarie in modo tale da massimizzare il loro successivo contributo alla varianza totale dell'insieme.

Date P variabili *X* osservare, tali combinazioni lineari sono del tipo

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1i}X_i + \ldots + a_{1p}X_p$$

$$\ldots$$

$$Y_k = a_{k1}X_1 + a_{k2}X_2 + \ldots + a_{ki}X_i + \ldots + a_{kp}X_p$$

$$\ldots$$

$$Y_p = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pi}X_i + \ldots + a_{pp}X_p$$

e vengono dette "componenti principali".

Devono soddisfare i seguenti requisiti:

- non devono essere correlate, cioè

  $\mathrm{Cov}(Y_{k'},Y_k) = 0 \qquad \forall\ k' \neq k$

- devono essere ordinate in relazione alla quantità di variabilità complessiva che ciascuna di esse può sintetizzare, cioè in relazione a quanto è l'apporto di ciascuna componente alla variabilità complessiva. Tale requisito può essere indicato con

  $V(Y_1) \geq V(Y_2) \geq \ldots \geq V(Y_i) \geq \ldots \geq V(Y_p)$

  sempre tenendo conto del fatto che la variabilità complessiva dei due diversi sistemi di riferimento deve coincidere, cioè:

$$\sum_{k=1}^{p} V(Y_k) = \sum_{i=1}^{p} V(X_i)$$

Le componenti principali così determinate, come già accennato inizialmente, non corrispondono a caratteristiche direttamente osservabili, e vanno di volta in volta interpretate.

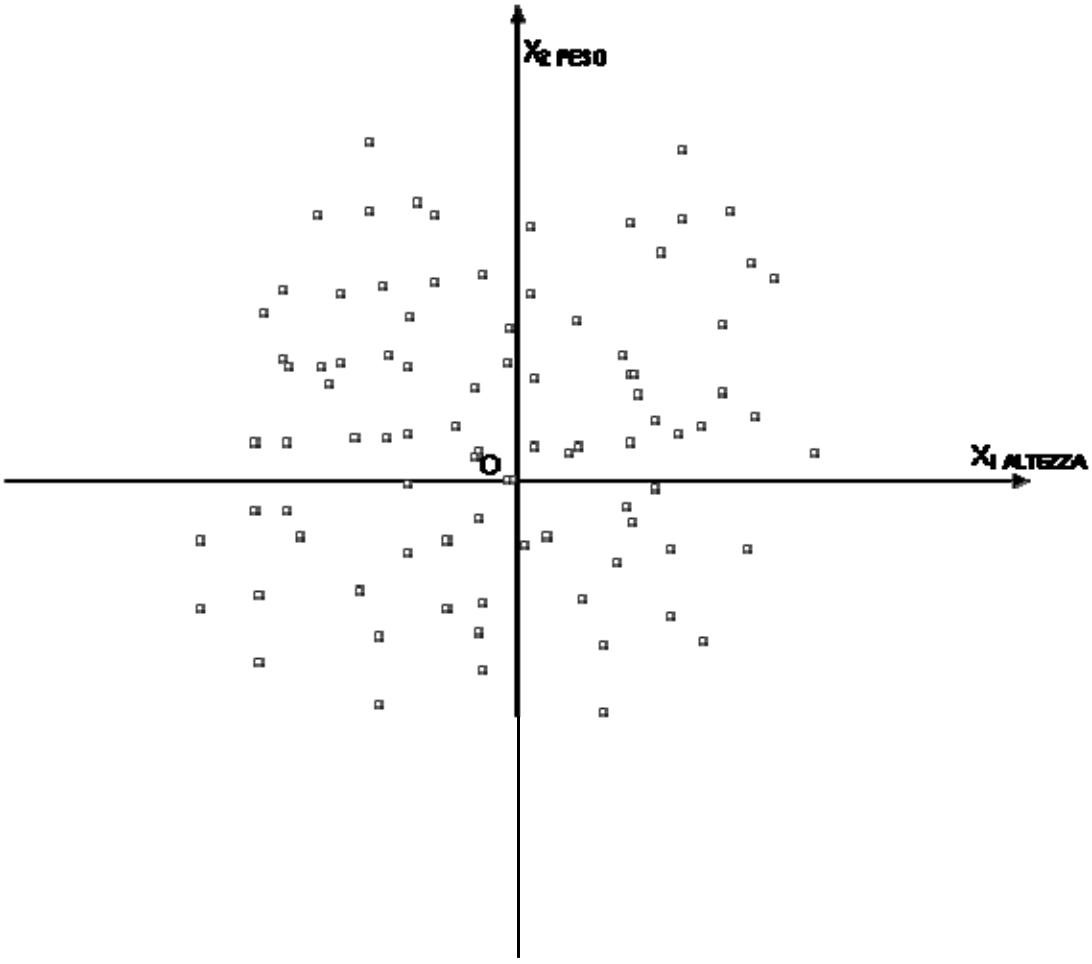
Inoltre, dal momento che le variabili X_i sono tra loro correlate, il sistema sarà in qualche misura ridondante. Per questo motivo, scegliendole opportunamente, non sarà necessario utilizzare tutte le componenti principali per riassumere in modo adeguato i dati, ma sarà sufficiente considerarne le prime m, con m inferiore a p, qualora queste forniscano un adeguato contenuto informativo.

UN'INTERPRETAZIONE GEOMETRICA

Facciamo un piccolo esempio. Si supponga di disporre di una n-pla di individui e di aver rilevato su di essi due variabili: peso e altezza.

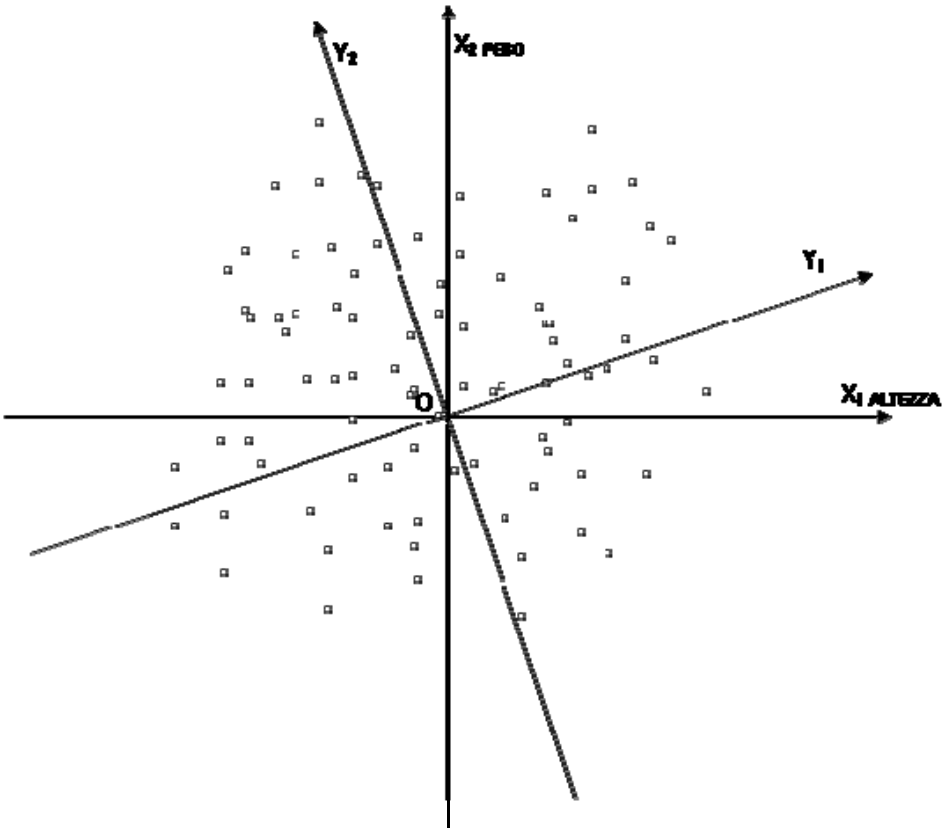
X_1 Altezza	X_2 Peso
165	59
168	77
172	65
192	85
170	58
159	55
163	60
180	90
175	78
…	…

Qual è la prima cosa può venire in mente di fare disponendo di dati simili? A parte gli indicatori sintetici univariati, si potrebbe partire con un grafico che consenta di visualizzare entrambe le variabili…



Quindi, come procedere? Si fa una regressione? Poi però bisogna stabilire quale variabile è da considerare dipendente e quale indipendente. Ci troveremmo perciò di fronte alla determinazione di due modelli differenti, nel caso si decidesse di prendere come predittiva una variabile piuttosto che l'altra.

E se, invece, semplicemente si facesse una rotazione degli assi passando dal sistema X_1, X_2 a un nuovo sistema che chiamiamo Y_1, Y_2? In questa maniera, la forma della nube di punti resta immutata, mentre cambiano le coordinate dei punti. Ciò che può accadere è che, se letti rispetto al nuovo sistema di riferimento, i punti consentano di trarre maggiori informazioni (o semplicemente informazioni più chiare) rispetto al vecchio sistema.



Nel caso di peso e altezza, se facciamo riferimento alle nuove coordinate troveremo che gli individui che si trovano sull'estremità destra dell'asse OY_1 saranno soggetti con peso e statura elevati, mentre quelli che si trovano dalla parte opposta sono individui di basso peso e bassa statura.

Come si potrebbe interpretare quindi tale nuova dimensione? La posizione di un individuo lungo l'asse OY_1 dà indicazioni sulla sua **taglia**.

Facciamo ora riferimento all'altro nuovo asse: i punti in alto a sinistra lungo questo asse rappresentano individui di peso elevato, ma bassa statura. Analogamente, individui lungo la parte inferiore dell'asse sono alti ma di peso basso. La posizione lungo l'asse OY_2 che indicazioni ci dà? Ciò che si ottiene è un'indicazione di **forma**.

Se x_1 e x_2 rappresentano altezza e peso di un individuo del collettivo preso in esame, chiameremo y_1 e y_2 le coordinate di quell'individuo nel nuovo sistema basato sugli assi OY_1 e OY_2. Dal momento che il nuovo sistema di assi è stato ottenuto con una semplice rotazione rispetto all'origine, la relazione tra i due sistemi di riferimento può essere espressa nella forma:

$$y_1 = x_1 \cos \alpha + x_2 \text{ sen } \alpha$$

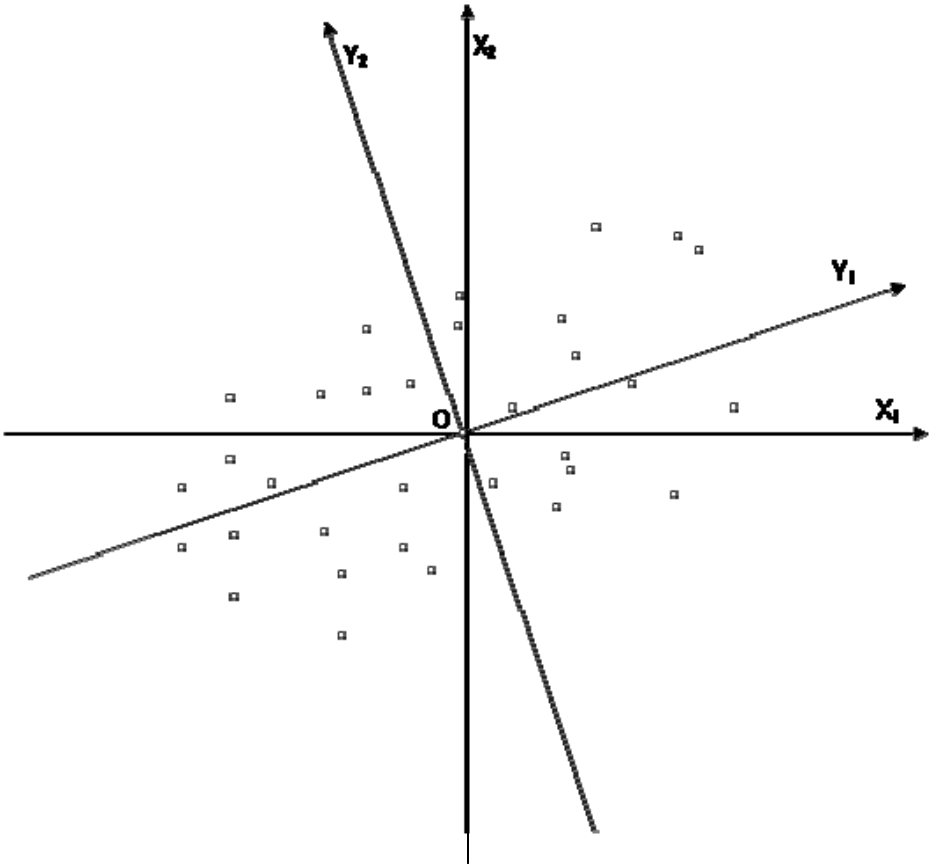
$$y_2 = -x_1 \text{ sen } \alpha + x_2 \cos \alpha$$

dove α è l'angolo tra OX_1 e OY_1 (ossia tra OX_2 e OY_2).

Ne consegue che la "taglia" è una combinazione lineare di peso e altezza (cioè $y_1 = a_1x_1 + a_2x_2$ dove a_1 e a_2 hanno entrambi lo stesso segno), mentre la "forma" è un contrasto lineare tra statura e peso (cioè $y_2 = b_1x_1 + b_2x_2$ dove b_1 e b_2 sono di segno opposto).

In questo modo abbiamo solamente compiuto una "rilettura" del collettivo rispetto a due nuovi indicatori che non abbiamo osservato direttamente, ma che abbiamo in qualche modo interpretato dalla relazione esistente con le variabili osservate. Finora non c'è stata alcuna sintesi e il sistema continua a essere a due dimensioni.

Proviamo ora a prendere in considerazione un secondo collettivo di n soggetti, sul quale sono sempre stati rilevati altezza e peso. E facciamo sempre riferimento al nuovo sistema di assi individuato precedentemente dove OY_1 indica la taglia e OY_2 la forma.



Che tipo di considerazioni possiamo fare osservando una nuvola di punti del genere? La dispersione dei punti lungo l'asse OY_1 è molto elevata, ma quella lungo l'asse OY_2 è invece piuttosto ridotta. Cosa significa ciò? Significa che gli individui di questo collettivo hanno "taglie" molto differenti, ma "forme" molto simili.

Le variabili X_1 e X_2 (altezza e peso) danno luogo a una distribuzione bivariata. Anche le variabili Y_1 e Y_2 (taglia e forma) danno luogo a una distribuzione bivariata.

Se consideriamo le due distribuzioni poste su piani diversi, è facile osservare che esiste una corrispondenza biunivoca tra le due distribuzioni, sicché ciascun punto del piano OX_1OX_2 si connette a un corrispondente punto del piano OY_1OY_2.

X_1 e X_2 sono però variabili correlate e così, visto che nel nostro caso le due variabili sono correlate positivamente, a valori elevati di X_1 tendono a corrispondere valori elevati di X_2 e, viceversa, a valori bassi di X_1 tendono a corrispondere valori bassi di X_2.

Dato che invece Y_1 e Y_2 sono variabili incorrelate, scelto un valore elevato di Y_1 (taglia) non ci si attende che a esso si associ, in media, un valore elevato di Y_2 (forma), e viceversa.

La stessa popolazione può essere analizzata sulla base di due distinte distribuzioni bivariate: la prima è basata sulla classificazione per aspetti distinti e correlati (altezza e peso), mentre la seconda è basata su aspetti astratti e non correlati. Y_1 e Y_2 assumono significati da precisare di volta in volta; nel caso in esame, assumono il significato di taglia e forma.

Y_1 e Y_2 tendono quindi a discriminare le unità statistiche in modo differente. Mentre però Y_1 tende a discriminare cospicuamente i vari soggetti in ragione della sua elevata varianza, ciò non avviene con Y_2.

Queste considerazioni portano a ritenere che, se noi impiegassimo solo la dimensione Y_1 per rappresentare i dati, cioè utilizzassimo la rappresentazione unidimensionale ottenuta proiettando i punti sull'asse OY_1, tale rappresentazione sarebbe una buona approssimazione della nuvola di punti bidimensionale.

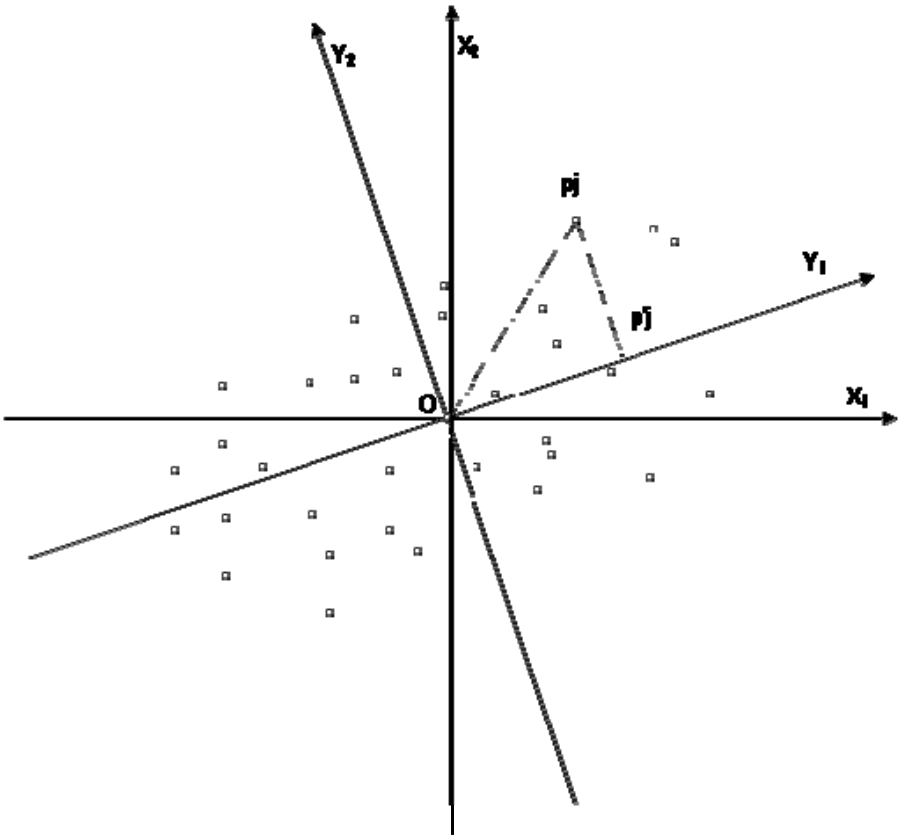
In altre parole, le differenze tra gli n soggetti di questo secondo campione, vengono espresse in maniera sufficientemente accurata anche se, invece di indicare per ciascuno di essi sia la statura x_1 che il peso x_2, si considera l'indicatore **taglia** $y_1 = x_1\cos\alpha + x_2\text{sen}\alpha$. Sostituendo alle variabili originarie X_1 e X_2 la nuova variabile Y_1, si ottiene una riduzione di dimensioni da 2 a 1, dal momento che i dati ora possono essere rappresentati graficamente solo in funzione dei valori di Y_1.

A seconda dei valori di α che si sceglieranno, cioè in base alla rotazione di assi che si sceglierà, si otterranno differenti variabili Y_1 e conseguentemente si otterranno differenti rappresentazioni grafiche unidimensionali.

Fra tutte queste rappresentazioni ne esisterà una che potrà essere considerata la migliore approssimazione (perché, avendo ridotto il sistema da due a una dimensione, parliamo appunto di approssimazione) della relazione che esiste tra gli n punti nello spazio bidimensionale.

Teniamo presente che siamo in una situazione nella quale stiamo cercando di valutare quale trasformazione, da uno spazio bidimensionale a uno unidimensionale, sia migliore. Passando da uno spazio bidimensionale a uno unidimensionale, sostanzialmente facciamo una proiezione dei punti sul nuovo asse che abbiamo individuato.

La trasformazione migliore (nel nostro caso, rotazione) deriverà da quel valore di α che rende minimo lo "spostamento" dei punti dalla posizione originale nel processo di proiezione. Dal momento che le coordinate dei punti rispetto a Y_1 sono le loro proiezioni ortogonali sull'asse OY_1, la soluzione sarà data dalla retta la cui distanza dai punti è minima.



Indichiamo con P_j un punto generico e con P'_j la sua proiezione ortogonale sull'asse OY_1. La retta migliore sarà quella che rende minima la somma:

$$\sum_{j=1}^{n} \left(P_j P'_j\right)^2$$

(si noti la differenza tra questa condizione e i minimi quadrati)

Applicando il teorema di Pitagora al triangolo $OP_jP'_j$ si ottiene:

$$(OP_j)^2 = (OP'_j)^2 + (P_jP'_j)^2$$

da cui, sommando per tutti gli n punti, consegue

$$\sum_{j=1}^{n} \left(OP_j\right)^2 = \sum_{j=1}^{n} \left(OP'_j\right)^2 + \sum_{j=1}^{n} \left(P_jP'_j\right)^2$$

Dividendo entrambi i membri per n-1 si ottiene infine:

$$\frac{1}{n-1}\sum_{j=1}^{n}\left(OP_j\right)^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(OP'_j\right)^2 + \frac{1}{n-1}\sum_{j=1}^{n}\left(P_jP'_j\right)^2$$

Il primo membro della precedente uguaglianza è costante per ogni dato campione e indipendente dal sistema di riferimento (infatti, non compaiono le coordinate P'_j che sono quelle del nuovo sistema di riferimento). Scegliere l'asse OY_1 in modo da minimizzare

$$\frac{1}{n-1}\sum_{j=1}^{n}\left(P_jP'_j\right)^2$$

sarà quindi equivalente a scegliere OY_1 così da massimizzare la quantità

$$\frac{1}{n-1}\sum_{j=1}^{n}\left(OP'_j\right)^2 .$$

Poiché O è il baricentro della nube di punti (perché abbiamo considerato variabili scarto dalla media) la quantità

$$\frac{1}{n-1}\sum_{j=1}^{n}\left(OP'_j\right)^2$$

è semplicemente la varianza delle proiezioni delle unità campionarie sull'asse OY_1.

Trovare l'asse OY_1 che renda minima la somma dei quadrati delle distanze (perpendicolari) da essa è quindi equivalente a trovare la retta OY_1 tale che le proiezioni su di essa abbiano massima varianza. E' su questa constatazione che si basa la determinazione delle *componenti principali* nella formulazione fornita da Hotelling.

SVILUPPI FORMALI

Proviamo ora a estendere quanto sopra descritto in un contesto bi-dimensionale a una situazione multivariata.

Si prenda in considerazione una n-pla di soggetti sui quali siano state osservate p variabili. La matrice dei dati avrà dimensioni (nxp) e gli n punti potranno essere rappresentati in uno spazio a *p* dimensioni.

Così come nell'esempio precedente i due assi OX_1 e OX_2 rappresentavano "altezza" e "peso", nello spazio p-dimensionale che stiamo ora considerando, ciascun asse identifica una delle p variabili osservate.

Analogamente a quanto fatto precedentemente, si definisca una retta OY_1 di questo spazio tale che la dispersione delle proiezioni dei punti su di essa sia massima. Questo equivale alla costruzione di una variabile del tipo

$$Y_1 = a_1X_1 + a_2X_2 + \ldots + a_pX_p$$

con coefficienti a_i che soddisfano il vincolo

$$\sum_{i=1}^{p} a_i^2 = 1$$

(perché basta un vettore a norma unitaria per identificare una retta in uno spazio) e determinata, inoltre, dalla condizione che la sua varianza sia massima.

Una volta individuata OY_1, si considera il sottospazio di dimensioni (p-1) ortogonale a OY_1 e, come abbiamo fatto per OY_1, si cerca la retta OY_2 di questo sottospazio con massima dispersione delle proiezioni dei punti lungo di essa. Tale retta dovrà soddisfare anch'essa la condizione

$$\sum_{i=1}^{p} a_i^2 = 1.$$

Fino a quando può proseguire questo processo? Il processo può continuare fino a quando non si siano ottenute *p* rette OY_k ($k = 1, \ldots, p$) fra loro ortogonali.

Ciascuna di queste rette definisce una variabile

$Y_k = a_{k1}X_1 + a_{k2}X_2 + \ldots + a_{kp}X_p$

in cui le costanti a_{ki} sono determinate dalla condizione che la varianza di Y_k sia massima sotto il vincolo di ortogonalità con ciascuna altra variabile $Y_{k'}$ dove ($k' < k$), e sotto il vincolo

$$\sum_{i=1}^{p} a_{ki}^2 = 1 \text{ per ogni } k.$$

Prima di procedere a una ulteriore formalizzazione del processo che porta a determinare le componenti principali, facciamo un richiamo ad alcuni elementi di geometria delle matrici.

Matrici ortonormali e ortogonali

Una matrice ortonormale è una matrice quadrata con le seguenti proprietà:

1) $|\mathbf{A}| = \pm 1$, dove $|\mathbf{A}|$ è il determinante di **A**

2) $\sum_{i=1}^{p} a_{ij}^2 = \sum_{j=1}^{p} a_{ij}^2 = 1 \quad (\forall\, i, j)$: la somma dei quadrati di qualunque riga o colonna è pari a 1

3) $\sum_{i=1}^{p} a_{ij}a_{ik} = 0 \quad (\forall\, j \neq k)$: i prodotti vettoriali di qualunque coppia di due colonne è pari a 0; ciò implica che le coordinate degli assi, che queste colonne rappresentano, si intersecano con un angolo di 90°.

Questo significa che se la matrice **A** è ortonormale, valgono le seguenti uguaglianze:

$\mathbf{AA}' = \mathbf{I}$

$\mathbf{A}^{-1} = \mathbf{A}'$ dove $\mathbf{A}^{-1}$ è l'inversa di **A**.

Una matrice che soddisfi la condizione 3, ma non le condizioni 1 e 2 si dice **ortogonale**.

Determinante di una matrice

Il determinante $|\mathbf{A}|$ di una matrice quadrata mxm **A**, è un singolo

valore associato alla matrice quadrata stessa e in molti casi è in relazione a una misura di volume. Per matrici 2x2 il detrerminante può essere calcolato agilmente con il metodo dell'incrocio.

Disponendo della seguente matrice 2x2

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

Allora il determinante di **A** sarà dato da

$$|\mathbf{A}| = a_{11}a_{22} - a_{21}a_{12}$$

Esiste un metodo analogo anche per una matrice 3x3.

Se una qualunque coppia di righe o colonne della matrice quadrata sono dipendenti, cioè sono una combinazione lineare dell'altra, allora il determinante della matrice è pari a zero.

Con il determinante si determina il **rango** della matrice. Se il rango è diverso da zero, allora la matrice si dice **non-singolare**.

(da Edward Jackson, pag. 7 e segg.)

Il metodo delle componenti principali è basato sull'algebra delle matrici: una matrice *p*x*p* simmetrica e non-singolare, come potrebbe essere la matrice di covarianza, può essere ridotta a una matrice diagonale **L** premoltiplicando e postmoltiplicando tale matrice con una particolare matrice ortonormale **A** tale che

$$\mathbf{A'SA} = \mathbf{L}$$

Gli elementi della diagonale di **L**, l_1, l_2, …, l_p sono chiamati **radici caratteristiche**, **radici latenti** o **autovalori** di **S**. Le colonne di **A** sono i **vettori caratteristici** o **autovettori**. Gli autovalori possono essere ottenuti dalla soluzione dalla seguente equazione, chiamata **equazione caratteristica**:

$$|\mathbf{S} - l\mathbf{I}| = 0$$

dove **I** è la matrice identità. L'equazione fornisce un polinomio di grado *p* dal quale vengono determinati i valori l_1, l_2, …, l_p.

DETERMINAZIONE DELLE COMPONENTI PRINCIPALI

Premesso ciò, si supponga quindi di avere un vettore di variabili $\mathbf{X}' = (X_1, X_2, \ldots, X_p)$ di dimensioni (1x*p*) osservato su *n* individui. Ciò genera, come è noto, una matrice di dati di dimensione (*n*x*p*) il cui generico elemento è x_{ij}.

I valori relativi al *j*-esimo individuo sono contenuti nel vettore $\mathbf{x}_j$ composto da *p* elementi.

Il vettore $\bar{x}' = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)$ contiene i valori medi delle *p* variabili.

E' noto altresì che la varianza della *i*-esima variabile è data da:

$$s_{ii} = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_i\right)^2$$ (siamo in un contesto campionario)

mentre la covarianza fra la *i*-esima e la *i'*-esima variabile è data da

$$s_{ii'} = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)\left(x_{ji'} - \bar{x}_{i'}\right)$$

Le varianze e le covarianze possono essere riassunte nella matrice **S** (detta appunto matrice di varianze e covarianze) i cui elementi vengono indicati con $s_{ii'}$; è possibile, altresì, esprimere tale matrice nel seguente modo:

$$\mathbf{S} = \frac{1}{n-1}\sum_{j=1}^{n}\left(\mathbf{x}_j - \bar{\mathbf{x}}\right)\left(\mathbf{x}_j - \bar{\mathbf{x}}\right)'$$

Omettendo l'indicatore k, la generica variabile scalare

$Y = a_1X_1 + a_2X_2 + \ldots + a_pX_p$ (*X* e *Y* maiuscole)

ottenuta come combinazione lineare delle variabili originarie X_i, può essere scritta nella forma più compatta come

$Y = \mathbf{a}'\mathbf{X}$

dove $\mathbf{a}' = (a_1, a_2, \ldots, a_p)$.

Il valore di *Y* per il *j*-esimo individuo osservato sarà:

$y_j = a_1x_{j1} + a_2x_{j2} + \ldots + a_px_{jp} = \mathbf{a}'\mathbf{x}_j$ (*X* e *Y* minuscole)

mentre la media della variabile *Y* nel collettivo esaminato sarà

$$\bar{y} = a_1 \bar{x}_1 + a_2 \bar{x}_2 + \ldots + a_p \bar{x}_p = \mathbf{a}'\bar{\mathbf{x}}$$

La varianza di *Y* è

$$s_y^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(y_j - \bar{y}\right)^2$$

Ma quanto vale $y_j - \bar{y}$? Per quanto sopra visto:

$$y_j - \bar{y} = \mathbf{a}'\mathbf{x}_j - \mathbf{a}'\bar{\mathbf{x}} = \mathbf{a}'(\mathbf{x}_j - \bar{\mathbf{x}})$$

o, essendo il trasposto di uno scalare uguale allo scalare stesso,

$$(y_j - \bar{y}) = (\mathbf{x}_j - \bar{\mathbf{x}})\mathbf{a}.$$

Moltiplicando le due espressioni si ottiene che

$$(y_j - \bar{y})^2 = \mathbf{a}'(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})\mathbf{a}$$

e quindi, sommando rispetto all'indice *j* e dividendo per (*n*-1) si ha:

$$s_y^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(y_j - \bar{y}\right)^2 = \mathbf{a}'\left\{\frac{1}{n-1}\sum_{j=1}^{n}\left(\mathbf{x}_j - \bar{\mathbf{x}}\right)\left(\mathbf{x}_j - \bar{\mathbf{x}}\right)'\right\}\mathbf{a} = \boxed{\mathbf{a}'\mathbf{S}\mathbf{a}}$$

Teniamo presente questa relazione perché sarà un punto di arrivo nei prossimi passaggi.

Se ora formalizziamo quanto prima era stato detto in termini geometrici, si può definire la prima componente principale come la combinazione lineare $Y_1 = \mathbf{a}'_1\mathbf{X}$ delle variabili originarie che dà luogo al massimo valore di $\mathbf{a}'_1\mathbf{S}\mathbf{a}_1$ sotto il vincolo $\mathbf{a}'_1\mathbf{a}_1 = 1$.

N.B. La quantità $\mathbf{a}'_1\mathbf{S}\mathbf{a}_1$ è una forma quadratica definita positiva (significa che tutti gli autovalori sono positivi) e perciò non ammette un massimo finito. Per massimizzare tale quantità è necessario imporre un vincolo sul vettore dei coefficienti.

Tale vincolo (appunto $\mathbf{a}'_1\mathbf{a}_1 = 1$) consente di ottenere una soluzione unica costituita da un vettore normalizzato.

Si cercherà perciò di determinare il vettore $\mathbf{a}_1$ che soddisfi tale condizione

Trovare il vettore che massimizza $\mathbf{a}'_1\mathbf{S}\mathbf{a}_1$, posto $\mathbf{a}'_1\mathbf{a}_1 = 1$, equivale a trovare il vettore che rende massima la quantità

$$\mathbf{a}'_1\mathbf{S}\mathbf{a}_1 - l_1(\mathbf{a}'_1\mathbf{a}_1 - 1)$$

dove l_1 è noto come il **moltiplicatore di Lagrange**.

Il problema viene risolto ponendo a zero le derivate parziali rispetto agli elementi di $\mathbf{a}_1$, ottenendo dopo alcuni passaggi:

$$\mathbf{S}\mathbf{a}_1 = l_1\mathbf{a}_1$$

da cui si ricava il seguente sistema omogeneo di equazioni, la cui soluzione è il vettore $\mathbf{a}'_1$:

$(\mathbf{S} - l_1\mathbf{I})\mathbf{a}_1 = 0$ (dove $\mathbf{I}$ è la matrice identità)

Affinché il sistema ammetta soluzione non banale è però necessario che il determinante della matrice dei coefficienti sia uguale a zero:

$$|\mathbf{S} - l_1\mathbf{I}| = 0$$

Quindi l_1 è un autovalore di $\mathbf{S}$ e la soluzione $\mathbf{a}_1$ il corrispondente autovettore normalizzato.

Quanti autovalori e corrispondenti autovettori normalizzati ha la matrice $\mathbf{S}$? Ne ha p. Bisogna perciò determinare quale sia quello che stiamo cercando.

Premoltiplicando l'uguaglianza $\mathbf{S}\mathbf{a}_1 = l_1\mathbf{a}_1$ per $\mathbf{a}'_1$ si ottiene

$$\mathbf{a}'_1\mathbf{S}\mathbf{a}_1 = l_1\mathbf{a}'_1\mathbf{a}_1$$

e poiché $\mathbf{a}'_1\mathbf{a}_1 = 1$ sarà

$$l_1 = \mathbf{a}'_1\mathbf{S}\mathbf{a}_1$$

Quindi l_1 è la varianza di Y_1 che si voleva massimizzare; questo problema ha allora soluzione se si considera il più elevato autovalore di $\mathbf{S}$. Ne segue che i coefficienti $\mathbf{a}_1$ nella prima componente principale $Y_1 = \mathbf{a}'_1\mathbf{X}$ sono gli elementi dell'autovettore di $\mathbf{S}$ che corrisponde all'autovalore maggiore.

Per la seconda componente e per le successive si fa un ragionamento analogo; si tratta di massimizzare la varianza di ciascuna componente sotto il vincolo di norma unitaria dei vettori dei coeffi-

cienti $\mathbf{a}'_k\mathbf{a}_k = 1$, ponendo l'ulteriore condizione di ortogonalità con le precedenti componenti

$$\mathbf{a}'_{k'}\mathbf{a}_k = 0 \qquad (k' < k)$$

Per la k-esima componente principale i coefficienti $\mathbf{a}'_k$ sono gli elementi dell'autovettore che corrisponde al k-esimo autovalore l_k nella graduatoria decrescente degli autovalori di **S**.

Ogni autovalore può allora essere letto come varianza della corrispondente componente principale:

$$V(Y_k) = l_k$$

Inoltre

$$\Sigma_k l_k = \Sigma_k V(Y_k) = \Sigma_k V(X_i)$$

cioè, la somma delle varianze delle componenti principali è uguale alla somma delle varianze delle variabili osservate.

L'importanza di ogni componente in termini di variabilità "spiegata" all'interno del sistema si misura con il rapporto

$$\frac{V(Y_k)}{\sum V(Y_k)} = \frac{l_k}{\sum l_k}$$

Il risultato, moltiplicato per 100, indica infatti la percentuale di varianza complessiva espressa dalla *k*-esima componente.

CAMBIAMENTI DI SCALA

Qualora le variabili differiscano per unità di misura o, pur essendo espresse nella medesima unità di misura, presentino variabilità notevolmente diversa, è opportuno determinare le componenti principali non già sulle variabili originarie, bensì sulle corrispondenti variabili standardizzate.

In presenza di forti differenze fra le varianze, infatti, le variabili con varianza maggiore tendono a dominare le prime componenti principali. Cosa vuol dire questo? Vuol dire che nelle prime componenti principali, i coefficienti relativi alle variabili *X* osservate con maggiore varianza saranno molto più elevati di quelli relativi alle altre variabili osservate.

In tal caso si opererà sulla matrice di correlazione **R** piuttosto che sulla matrice di varianze covarianze **S**. I risultati che si ottengono determinando autovalori e autovettori dell'una o dell'altra matrice sono sensibilmente diversi e non esiste alcuna relazione che leghi gli uni agli altri.

La scelta di operare sui dati standardizzati anziché sui dati grezzi va perciò attentamente ponderata, come pure va fatta con attenzione la scelta delle unità di misura rispetto a cui vengono espressi i caratteri osservati.

Infatti, **le componenti principali non sono insensibili a cambiamenti di scala**.

A tal proposito, valga il seguente esempio. Si supponga che siano state rilevate due sole variabili X_1 e X_2 e che X_1 possa essere espressa sia in *cl* (centilitri) che in *ml* (millilitri).

Nei due casi le matrici di varianza e covarianza potrebbero essere, ad esempio:

$$\mathbf{S}_1 = \begin{pmatrix} 90 & 50 \\ 50 & 90 \end{pmatrix} \quad \text{e} \quad \mathbf{S}_2 = \begin{pmatrix} 9000 & 500 \\ 500 & 90 \end{pmatrix}$$

La prima componente principale di $\mathbf{S}_1$ è quindi

$$Y_1 = 0{,}707X_1 + 0{,}707X_2$$

e l'autovalore a essa associata, cioè la sua varianza è 140. Quanta parte della varianza complessiva spiega tale componente? Tale componente spiega il 77,78% della variabilità complessiva.

Il calcolo avviene facendo (140/(90+90))*100.

La prima componente principale di $\mathbf{S}_2$ è invece

$$Y_1 = 0{,}998X_1 + 0{,}055X_2$$

e la sua varianza è 9027,97, cioè il 99,32% della variabilità complessiva del sistema.

Il calcolo avviene facendo (9027,97/(9000+90))*100.

Un semplice cambiamento di unità di misura ha quindi l'effetto di trasformare una componente principale che dà ugual peso a X_1 e X_2 in una componente dominata completamente da in una componente dominata completamente da X_1.

Generalizzando quanto appena mostrato al caso di $p>2$ si avrà che, in presenza di forte eteroschedasticità, le componenti principali sono pressoché equivalenti alle variabili originarie disposte in ordine di varianza decrescente.

Una ulteriore giustificazione al calcolo delle componenti principali sulla matrice **R** piuttosto che su **S** sta nella maggiore facilità di confronti con altre situazioni.

In conclusione, è però opportuno ricordare che, quando si voglia fare inferenza sulle componenti principali della popolazione a partire da quelle calcolate su un campione, i risultati valgono solo se le componenti principali sono calcolate a partire dalla matrice di varianze e covarianze.

È infine opportuno ricordare che il numero delle componenti lineari che risultano dalla soluzione dell'equazione caratteristica dipende dal rango della matrice di varianza-covarianza. Se la matrice è a pieno rango, si possono determinare fino a p componenti principali ordinate secondo l'ammontare del loro contributo alla variabilità complessiva. Nella pratica, si è soliti considerare solo le m componenti che "spiegano" una frazione sufficiente di variabilità, ritenendo le restanti $p - m$ componenti praticamente irrilevanti per la descrizione del fenomeno.

LA SCELTA DEL NUMERO DI COMPONENTI

La percentuale di varianza spiegata

Il criterio più immediato per la scelta del numero *m* di componenti rispetto a cui rappresentare il fenomeno presuppone che venga prefissatala proporzione della variabilità complessiva che si desidera tali componenti spieghino, ad esempio il 70/80%; *m* sarà perciò il numero più piccolo di componenti per le quali tale percentuale è superata.

I software che calcolano le componenti principali forniscono tale indicazione e quindi, se lo si decide, si può tranquillamente utilizzare questa strategia.

Regola di Kaiser

Un altro criterio, noto anche con il nome di "regola di Kaiser", utile quando si estraggono le componenti dalla matrice di correlazione, suggerisce di trattenere solo le componenti corrispondenti ad autovalori maggiori o uguali all'unità. L'idea sottostante questo criterio è che se le variabili X_i, i = 1, …, *p*, fossero indipendenti e standardizzate, le componenti coinciderebbero con tali variabili e avrebbero varianza unitaria; quindi ogni componente con varianza inferiore a 1 contiene meno informazioni di una qualsiasi delle variabili originarie.

Alcuni autori ritengono che questo criterio sottostimi il numero di componenti necessario per una adeguata rappresentazione della variabilità del fenomeno osservato e propongono di abbassare la soglia a 0,7 così da tener conto della variabilità campionaria.

Qualora l'analisi sia svolta sulla matrice di varianza-covarianza la media degli autovalori

$$\bar{l} = \sum_{k=1}^{p} \frac{l_k}{p}$$

può essere utilizzata come valore critico; solo le componenti con varianza maggiore di $\bar{l}$ verranno trattenute.

Scree diagram

Ancora, nella scelta di *m* (il numero di componenti sufficienti a ri-

produrre con buona approssimazione i dati di partenza) ci si può servire del grafico degli autovalori rispetto al loro ordine di estrazione (noto come *scree diagram* o *scree plot*) e sceglierequel numero *m* di componenti in corrispondenza del quale il grafico presenta un “gomito”.