Capitolo 9. Statistica multivariata. Relazioni tra osservazioni. Relazioni tra variabili. Sur-ray-plot. Draftsman plot. Analisi delle componenti principali. Analisi discriminante. Analisi dei gruppi.

Statistica multivariata (SM)

Come sappiamo, è impossibile costruire una mappa della Terra che rispetti tutte le distanze tra i diversi punti. Tutte le carte geografiche si basano su un certo criterio di proiezione per rappresentare al meglio, a seconda degli scopi, la superficie terrestre 3D su un foglio di carta 2D. Ogni proiezione implica un certo compromesso tra **in**formazione e **de**formazione. Ricordate ad es., certi planisferi di forma ovale, altri rettangolari, altri formati da due dischi accostati. Anche quando osserviamo un oggetto 3D al computer, ci è spesso consentito ruotare l'oggetto per trovare la migliore rappresentazione 2D sullo schermo. Anche in statistica succede di dover analizzare fenomeni descritti da 3 o più variabili. Anche in statistica si pone quindi lo stesso problema della proiezione e rappresentazione di fenomeni 3D o nD (descritti da 3 o più variabili) su scale o grafici ad 1 o 2 dimensioni (sempre accettando una certa deformazione).

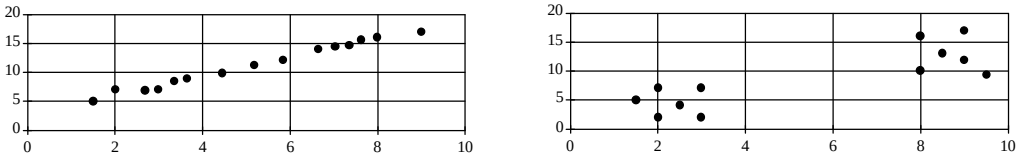
La SM considera i fenomeni (biologici, clinici, fisici, economici, sociali, ecc.) nella loro interezza, tenendo quindi conto delle diverse caratteristiche (v. variabili) che meglio servono a descrivere i fenomeni stessi: da 3-4 variabili sino, in casi particolari, a diverse centinaia di variabili. Per tale vocazione la SM 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. In queste poche pagine tenteremo di tenere l'orizzonte il più ampio possibile.

La possibilità di analizzare più variabili simultaneamente non significa tuttavia che sia sempre meglio mettere quante più variabili si può. Se alle variabili di interesse (supposto) vengono aggiunte altre di scarso significato si causa una diluizione dell'informazione che rende più difficile l'analisi dei dati. Esistono, è vero, anche metodi che aiutano a valutare l'importanza relativa delle diverse variabili, ma non sostituiscono mai l'esperienza ed il buon senso. Trascuriamo per ora il problema della selezione delle variabili ed assumiamo di includere variabili di interesse ai fini dell'analisi.

Per molti aspetti, la SM può essere considerata un'estensione della statistica mono- e bivariata (per alcuni sono invece queste ultime da considerarsi una 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.

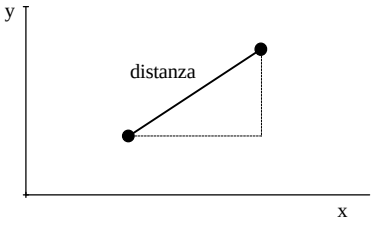
Sino ad ora, la nostra esperienza nel campo bivariato è stata quella di esaminare la relazione tra due variabili mediante l'analisi della regressione lineare, espressa graficamente dalla retta di regressione. Esiste tuttavia un aspetto molto importante che l'analisi della regressione non rileva: la relazione tra i soggetti, rappresentati graficamente dai punti del grafico. Le **relazioni tra variabili** e **relazioni tra soggetti** rappresentano due prospettive da cui si osserva lo stesso fenomeno (ogni

soggetto esprime diverse variabili; ogni variabile è espressa in diversi soggetti). Ora, strano ma vero, si può facilmente osservare che quanto più è forte la relazione tra le variabili, tanto meno forte è la relazione tra i soggetti e viceversa:



Il grafico a sinistra dimostra una forte correlazione tra le due variabili. I soggetti (punti del grafico) si *piegano* alla forza della relazione tra le variabili e non si intravedono raggruppamenti particolari. Viceversa il grafico a destra dimostra una evidenti relazioni (gruppi, parentele) tra i soggetti a scapito della relazione tra le variabili. Infatti nessuno potrebbe pensare di rappresentare il grafico a destra mediante una retta di regressione. Possiamo dire quindi che variabili fortemente correlate sono poco utili a svelare relazioni tra i soggetti. Viceversa variabili poco o zero correlate (indipendenti) mettono spesso in luce interessanti le relazioni tra soggetti.

La relazione tra due soggetti può valutarsi nella loro distanza:



Disponendo di un sistema di assi ortogonali, la distanza euclidea corrisponde all'ipotenusa di un triangolo rettangolo, i cui cateti sono quindi le differenze tra i valori delle due variabili:

cateto orizzontale: $x_2 - x_1$

cateto verticale: $y_2 - y_1$

distanza = ipotenusa: $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

Quanto detto vale se e solo se:

1. Gli assi (**variabili**) sono **ortogonali** (altrimenti sarebbe come applicare il teorema di Pitagora ad un triangolo non rettangolo). In termini di variabili, la condizione di ortogonalità (90°) corrisponde alla indipendenza delle variabili. Quanto più due variabili sono correlate, tanto più il loro spazio vettoriale si chiude con un angolo sempre più stretto. Paradossalmente, se si applica la stessa variabile a entrambe gli assi, lo spazio vettoriale si riduce ad una retta. Con

metodi matematici è possibile rendere le variabili ortogonali, in modo da calcolare valide distanze. Sono state proposte diverse distanze multivariate che superano i limiti della distanza euclidea. Anche se non possiamo soffermarci su questo punto, citiamo una delle più utilizzate: la **distanza di Mahalanobis**.

2. Gli assi hanno la **stessa metrica** (altrimenti sarebbe come applicare il teorema di Pitagora avendo un cateto in centimetri e l'altro in millimetri). E' difficile dire quando le variabili siano metricamente omogenee. E inoltre non basta che siano espresse con la stessa unità di misura. Consideriamo ad es. il peso in grammi del fegato e del pancreas. Pur essendo entrambi espressi in grammi, la variabilità del peso del fegato è molto maggiore di quella del pancreas. Occorre che la dispersione relativa sia uguale. Ciò si ottiene solo standardizzando i dati.

Quando le variabili sono più di due, dobbiamo immaginare uno spazio rappresentato da tante dimensioni quante sono le variabili. Si tratta di una rappresentazione matematica ed astratta. La formula che calcola la distanza tra due punti in uno spazio a n dimensioni (n variabili) è una estensione di quella che calcola la distanza tra due punti sul piano:

$$d_{a,b} = \sqrt{(v_{1a} - v_{1b})^2 + (v_{2a} - v_{2b})^2 + (v_{3a} - v_{3b})^2 + \ldots + (v_{na} - v_{nb})^2}$$

Per ciascuna variabile, si calcola la differenza tra i due soggetti, si eleva al quadrato, si fa la somma ed infine si estrae la radice quadrata.

Generalmente i dati sono organizzati in matrice con i soggetti nelle righe e le variabili nelle colonne:

	Età	Altezza	Peso	Piede
Paolo	19	178	83	43
Fred	34	177	82	41
Pippo	20	179	82	41
Minni	19	165	57	39
Betty	19	164	52	37
Wilma	18	162	51	38
Poldo	19	170	69	41
Clara	19	163	40	33
Dino	18	169	67	45
Alice	19	170	50	36

Si calcolano i principali parametri:

	Età	Altezza	Peso	Piede
n	10	10	10	10
media	20.4	169.7	63.3	39.4
s	4.812	6.395	15.535	3.534

Dopo la necessaria standardizzazione delle variabili (una per una) la matrice diventa:

	Età norm.	Altezza norm.	Peso norm.	Piede norm.
Paolo	-0.291	+1.298	+1.268	+1.019
Fred	+2.826	+1.141	+1.204	+0.453
Pippo	-0.083	+1.454	+1.204	+0.453
Minni	-0.291	-0.735	-0.406	-0.113
Betty	-0.291	-0.891	-0.727	-0.679
Wilma	-0.499	-1.204	-0.792	-0.396
Poldo	-0.291	+0.047	+0.367	+0.453
Clara	-0.291	-1.048	-1.500	-1.811
Dino	-0.499	-0.109	+0.238	+1.585
Alice	-0.291	+0.047	-0.856	-0.962

Non occorrerebbe dirlo: i parametri delle variabili standardizzate sono:

	Età	Altezza	Peso	Piede
media	0	0	0	0
s	1	1	1	1

Dai dati standardizzati possiamo ottenere

1. La mappa (matrice simmetrica) delle correlazioni tra tutte le variabili prese a due a due (notare la diagonale costituita tutta da valori 1 in quanto si tratta delle correlazioni tra le stesse variabili):

	Età	Altezza	Peso	Piede
Età	**1.00**	.46	.46	.14
Altezza	.46	**1.00**	.91	.60
Peso	.46	.91	**1.00**	.81
Piede	.14	.60	.81	**1.00**

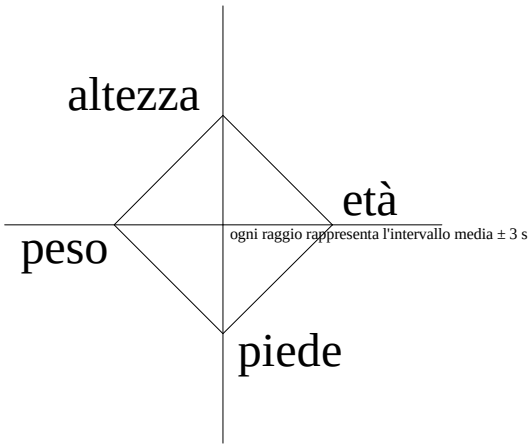
2. La mappa (matrice simmetrica) delle distanze euclidee tra tutti i soggetti a due a due (notare la diagonale costituita tutta da valori 0 in quanto si tratta delle distanze nulle tra gli stessi soggetti):

	Paolo	Fred	Pippo	Minni	Betty	Wilma	Poldo	Clara	Dino	Alice
Paolo	**0**	10.07	0.39	8.21	11.66	12.55	2.70	21.17	3.40	10.00
Fred	10.07	**0**	8.56	16.15	18.86	21.26	11.62	26.94	14.83	17.16
Pippo	0.39	8.56	**0**	7.75	10.55	11.94	2.72	18.74	4.83	8.27
Minni	8.21	16.15	7.75	**0**	0.45	0.49	1.53	4.18	3.73	1.53
Betty	11.66	18.86	10.55	0.45	**0**	0.23	3.36	1.90	6.71	0.98
Wilma	12.55	21.26	11.94	0.49	0.23	**0**	3.67	2.57	6.18	1.93
Poldo	2.70	11.62	2.72	1.53	3.36	3.67	**0**	9.81	1.37	3.50
Clara	21.17	26.94	18.74	4.18	1.90	2.57	9.81	**0**	15.47	2.33
Dino	3.40	14.83	4.83	3.73	6.71	6.18	1.37	15.47	**0**	7.75
Alice	10.00	17.16	8.27	1.53	0.98	1.93	3.50	2.33	7.75	**0**

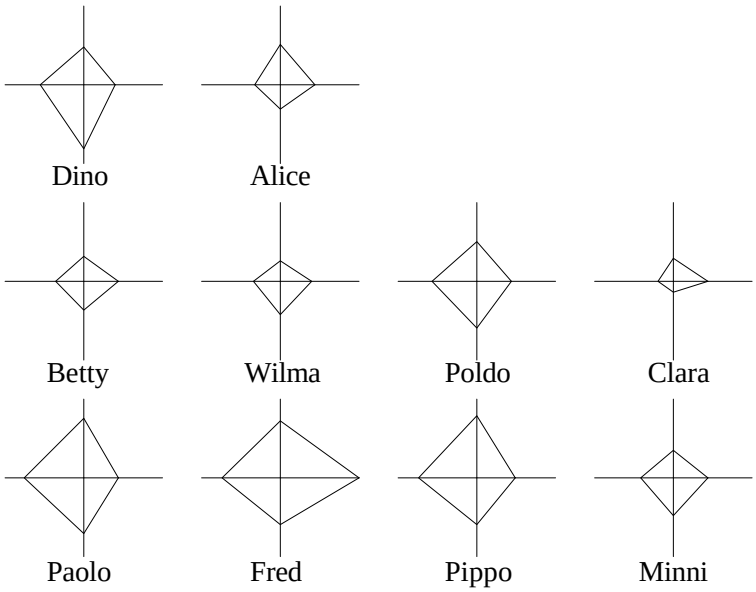
Poter disporre di tali informazioni è già un notevole passo avanti, rispetto al semplice tabulato dei dati originali.

Sun-ray-plot

Per ogni variabile si ha un raggio che rappresenta l'intervallo media ± 3 deviazioni standard, con la media al centro del raggio, i valori bassi verso il centro ed i valori alti verso l'esterno.



Ogni soggetto produce quindi un poligono a seconda dei valori posseduti.
Dall'esame delle varie forme e dimensioni dei poligoni è semplice ricavare valutazioni e confronti nell'ambito multivariato.



Sono evidenti:

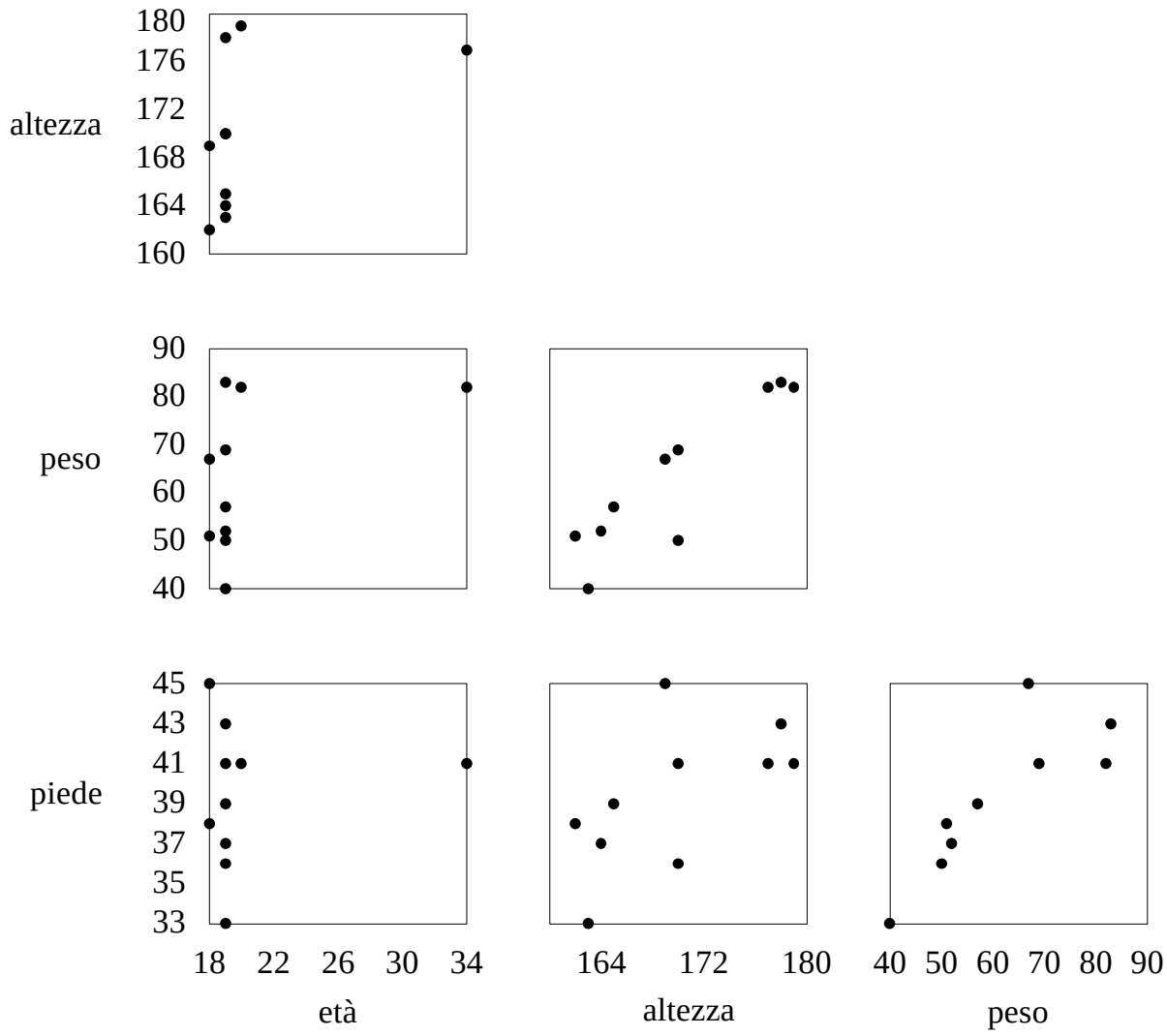
- l'età attempata di Fred
- la taglia minuta di Clara
- il gran piede di Dino
- la somiglianza tra Pippo e Poldo
- la somiglianza tra Minni e Wilma
- ...e varie altre cose da scoprire.

Per ovvie ragioni, il Sun-ray-plot va bene sino a che le variabili non sono più di 10.

Draftsmanplot

E' concettualmente banale, ma efficace: si tratta di tutti i possibili grafici a punti tra tutte le coppie di variabili. Accompagna quindi bene la matrice delle correlazioni (vedi prima).

Per ovvie ragioni, anche questa tecnica è utile quando il numero di variabili è limitato.

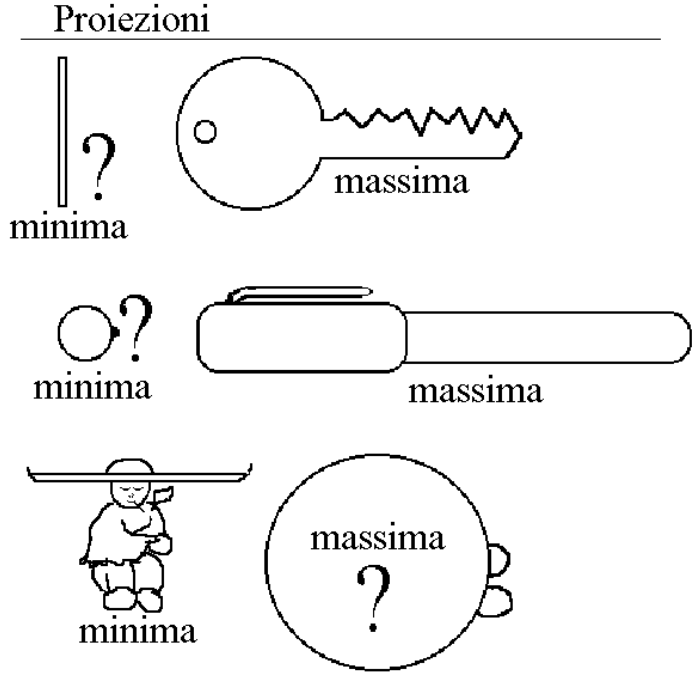


Per ragioni di spazio non abbiamo aggiunto i nomi vicino ai punti. Sarebbe stato interessante osservare la posizione di ogni soggetto nei vari grafici.

Analisi delle Componenti Principali (CP)

E' una classica e fondamentale tecnica multivariata che ruota i punti descritti nello spazio a n variabili in modo da trovare altrettante proiezioni con dispersione (=varianza) massima e tra loro ortogonali. Immaginiamo di dover riconoscere un oggetto tridimensionale dalla sua ombra. L'ombra è una proiezione di tutti i punti dell'oggetto (3D) su un piano (2D). Noi possiamo posizionare cioè ruotare diversamente l'oggetto ottenendo diverse ombre di diversa estensione. Quasi sempre, l'ombra di estensione massima è spesso quella più informativa sulla natura dell'oggetto. L'ombra con la massima estensione non è altro che la proiezione con la massima dispersione (=massima varianza) dei dati sul piano. Nei disegni vediamo ad es. come sia semplice riconoscere un oggetto 3D dal suo profilo 2D, quando l'area di proiezione è massima (vedi esempio della chiave e delle penna). Ci sono ovviamente delle eccezioni. Una famosa è quella alla base di un filone di giochi di interpretazione di immagini. Ad esempio, interpretare l'immagine di un cerchio con un piccolo particolare a fianco. La soluzione in genere è che si tratta di un messicano col sombrero visto dall'alto mentre sta seduto (nel disegno), o va in bicicletta, o fuma una sigaretta, ecc. In questo caso, statisticamente sfortunato, per interpretare l'oggetto occorre non già la proiezione massima (il sombrero copre l'oggetto) ma una proiezione minore (ad es., il messicano visto di fronte). Proprio per queste situazioni i metodi multivariati sono anche in grado di ruotare i dati secondo diversi livelli di varianza (v. analisi fattoriale).

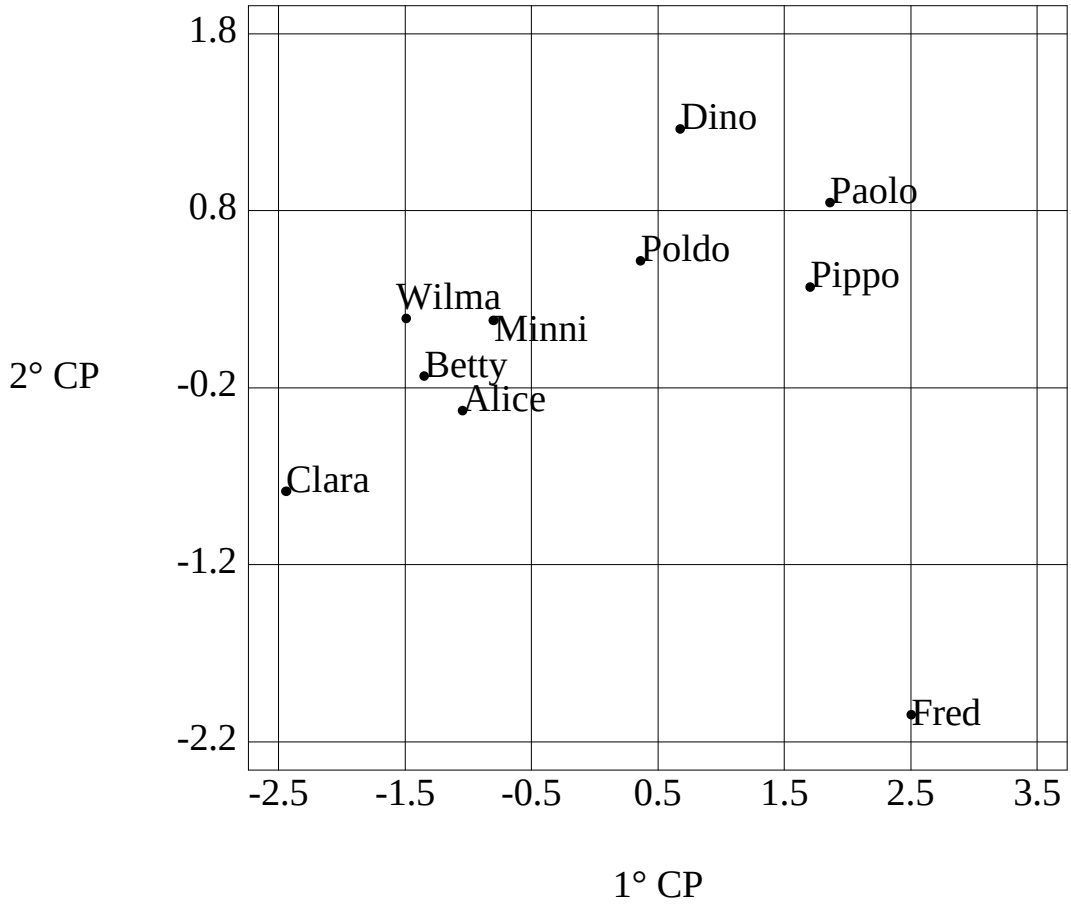
L'analisi delle CP è applicata in altri campi con nomi diversi: analisi di Hotelling, analisi di Karhunen-Loève, analisi degli autovettori, ecc.



Nell'ambito n-dimensionale, le proiezioni che realizzano la maggior varianza quasi sempre riassumono la maggior parte dell'informazione contenuta in numerose variabili. Tali assi di proiezione non corrispondono più ad alcuna variabile originaria e non hanno un significato immediato (anche se la loro interpretazione è interessante). Sono detti appunto componenti principali. Una volta trovata la prima CP (l'asse di proiezione che realizza la massima varianza dei dati) si trova la seconda CP (ortogonale alla prima e realizzante la massima quota di varianza residua), e così via. Alla fine del procedimento si ottengono tante CP quante erano le variabili originarie, con la maggior parte dell'informazione concentrata nelle prime CP come appare nei risultati dell'analisi applicata al nostro esempio.

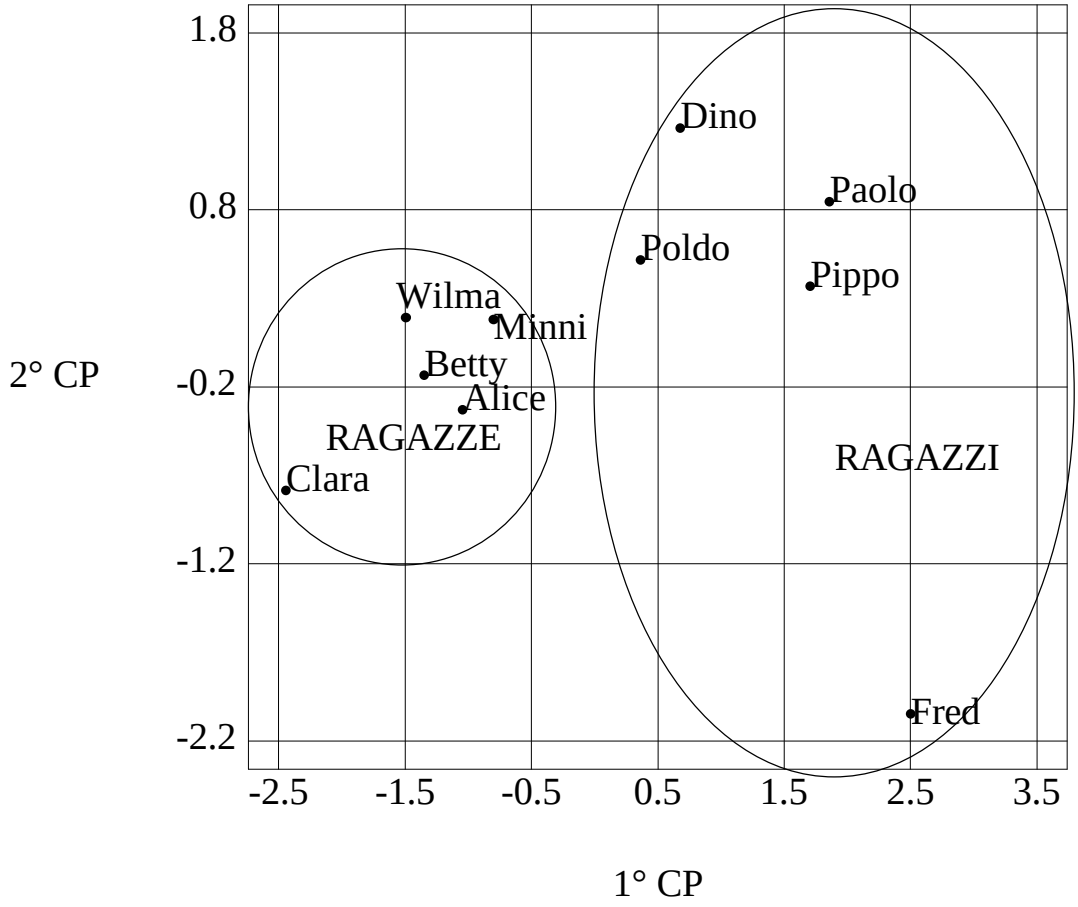
Componente Numero	Percentuale di Varianza	Percentuale Cumulativa
1	69.33	69.33
2	22.07	91.40
3	7.87	99.27
4	0.73	100.00

Il grafico delle prime due CP, che assieme totalizzano il 91.4% della varianza totale, è il seguente:



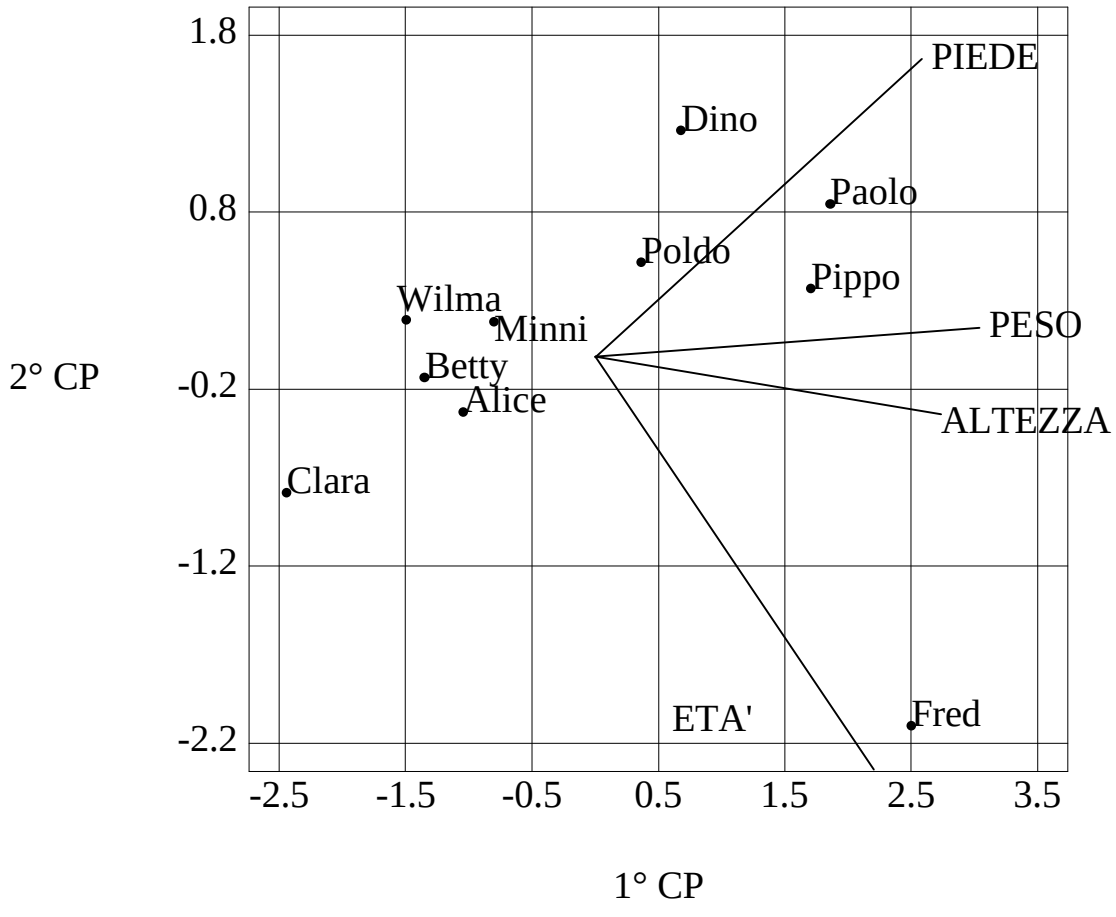
Se poi si è interessati a vedere come le stesse variabili concorrono a determinare le CP, è possibile proiettarle nel grafico assieme ai soggetti. Si vedrà ad es. che nel

nostro caso è il piede a separare meglio i ragazzi dalle ragazze, anche se la prima CP è più 'stirata' dal peso e dall'altezza. L'età non ha granché relazione con le altre tre variabili, anche se serve a mettere in evidenza Fred, un po' fuori norma.



Se dovessimo dare un'interpretazione a queste prime due CP, potremo dire che la 1° CP può rappresentare la dimensione dell'individuo, mentre la 2° CP, più difficile da interpretare, può rappresentare l'armonicità del soggetto.

Osserviamo infine la proiezione delle nostre quattro variabili sulle due componenti principali.
Le quattro linee rappresentano gli assi positivi delle variabili. Il punto d'incontro o vertice o origine degli assi rappresenta la media delle quattro variabili. Questo è collocato più o meno al centro, intermedio tra i due gruppi di ragazzi e ragazze. Gli assi positivi vanno verso i ragazzi. Ciò vuol dire che ragazzi avevano in genere valori superiori alla media. Gli assi negativi (non tracciati per ragioni di chiarezza) si ottengono prolungando gli assi positivi. Gli assi negativi vanno verso le ragazze. In particolare, notiamo che le variabile peso e piede sono quelle meglio allineate sulla linea ideale che unisce ragazzi e ragazze. Invece la variabile età è piuttosto ortogonale alla stessa direttrice. Diremmo allora che peso e piede sono più associate al fattore sesso, mentre l'altezza e soprattutto l'età si discostano da tale associazione.



Analisi discriminante (AD)

L'AD, a differenza dell'analisi delle CP, parte dalla conoscenza pregressa di gruppi di individui e mira a stabilire a quale gruppo appartenga un nuovo individuo. Il problema è sempre quello di ruotare i dati, questa volta per massimizzare non già la varianza totale (come era nell'analisi delle CP) ma il rapporto varianza tra/varianza entro gruppi. Ovviamente, si parla di varianza multivariata. Una volta trovata la rotazione giusta che realizza tale condizione sui soggetti appartenenti a gruppi conosciuti, si applica la stessa rotazione ai nuovi soggetti da classificare.

Seguendo il nostro caso, possiamo prendere come esempio i due gruppi formati da 5 ragazzi e da 5 ragazze. Trattandosi di 2 soli gruppi, basterà un solo asse a separare i ragazzi dalle ragazze. Vediamo i valori dei coefficienti della prima ed unica (in questo caso) funzione discriminante:

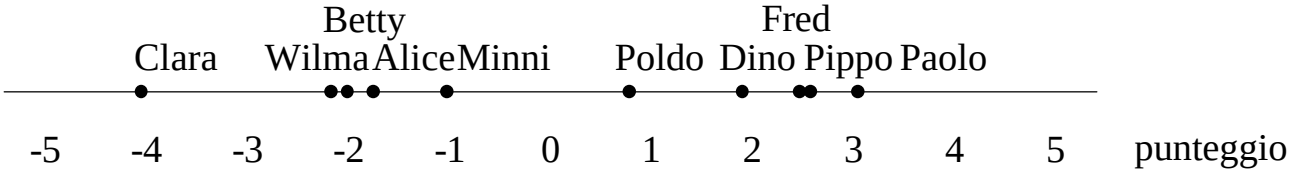
costante	-36.7426
età	0.02029
altezza	0.11916
peso	0.04506
piede	0.33643

Il valore ottenuto dalla funzione:

$-36.7426 + (\text{età} \times 0.02029) + (\text{altezza} \times 0.11916) + (\text{peso} \times 0.04506) + (\text{piede} \times 0.33643)$

darà un punteggio o score che assegnerà ogni nuovo soggetto al gruppo dei ragazzi o a quello delle ragazze, a seconda della minor distanza dal centroide delle ragazze o a quello dei ragazzi. I centroidi dei due gruppi sono:

ragazze: -2.15761
ragazzi: 2.15761



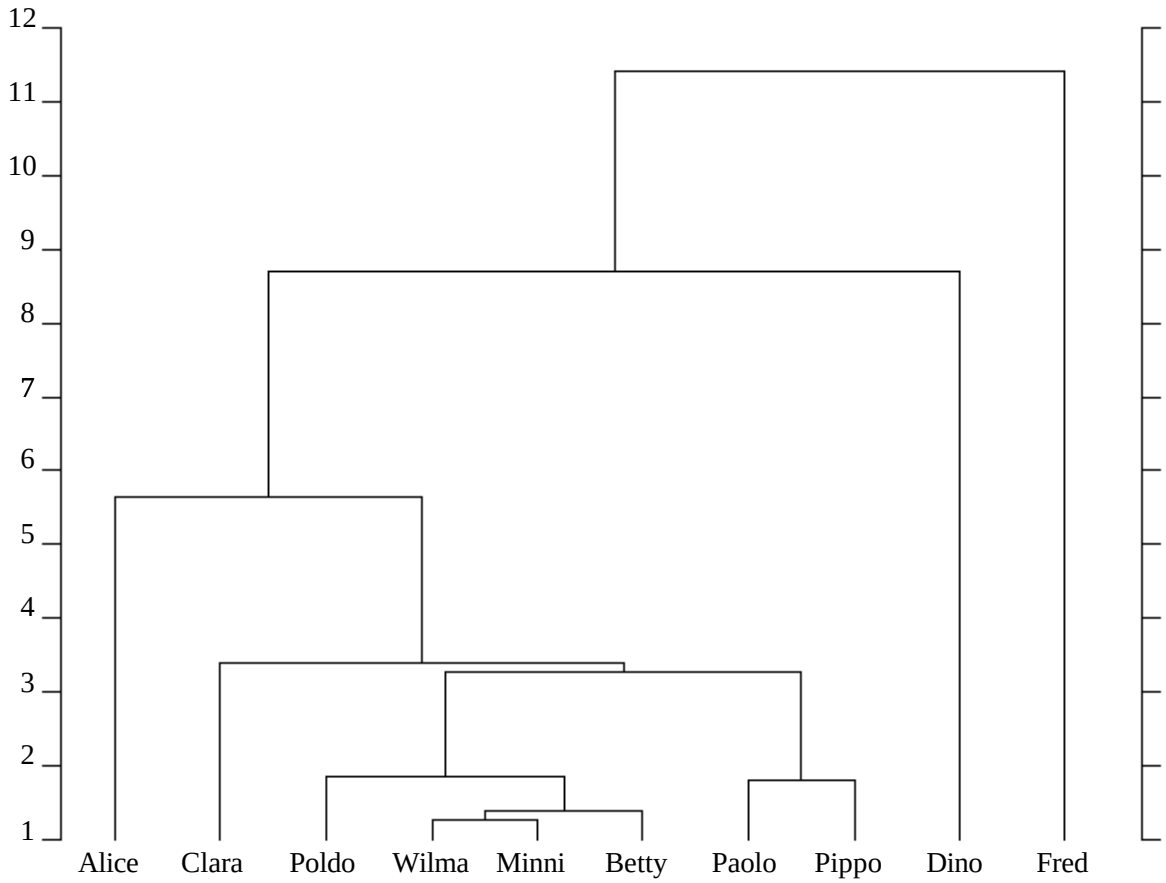
Lo zero è la soglia discriminante. I punteggi positivi classificano i soggetti come maschi, quelli negativi come femmine. Notare che la funzione è stata trovata in base a soggetti di sesso noto (set di apprendimento). L'indagine vera e propria inizia solo ora, con l'applicazione della funzione a soggetti di sesso non noto (set di applicazione).

Analisi dei gruppi (ADG)

Sinonimi: analisi dei grappoli, analisi dei clusters, cluster analysis.

L'ADG, qui intesa come analisi gerarchica dei gruppi, prescinde anch'essa da qualsiasi conoscenza a priori sul tipo di struttura dei dati. La sua è una classificazione spontanea, automatica. Il suo scopo è quello di suddividere l'intero set di dati in gruppi e sottogruppi, sino ai singoli soggetti, ottenendo una classificazione che in qualche modo rappresenti un'immagine del differenziamento del fenomeno sotto esame. Tipico risultato dell'ADG è il dendrogramma, o diagramma ad albero, spesso utilizzato in modo descrittivo anche da altre discipline (v. mappe filogenetiche).

Per dare un'idea di come l'ADG funziona, riesaminiamo la mappa delle distanze tra i soggetti a due a due. Il procedimento vuole che ogni nuovo gruppo, a partire dal primo, sia costituito dai due soggetti più vicini tra loro. Una volta costituito un gruppo, questo si considera a tutti gli effetti un nuovo soggetto che rappresenta i suoi componenti attraverso un certo criterio (v. legame singolo, medio, completo, ecc.). Il processo è iterativo nel senso che la ricerca dei due soggetti più vicini e la loro fusione in gruppo si ripete sino ad includere tutti i soggetti in un unico gruppo. Senza entrare in ulteriori dettagli, osserviamo il dendrogramma ottenuto per il set di 10 soggetti (criterio del legame minimo).



Si noterà che la classificazione indicata dal dendrogramma non separa bene ragazzi da ragazze. Questo è il punto fondamentale: l'ADG non considera nessuna catalogazione precostituita e tratta tutti i soggetti allo stesso modo. Pensiamo comunque che tra i nostri 10 soggetti vi sono tanti altri fattori di raggruppamento oltre a quello dell'essere maschio o femmina. L'ADG classifica i soggetti in base all'equilibrio esistente tra tutte le caratteristiche descritte dalle variabili. E' quindi una tecnica esplorativa e, paradossalmente, può segnalare raggruppamenti che non corrispondono alle tipologie più note. Anche questo è un risultato interessante, in quanto stimola la ricerca di quelle componenti latenti che hanno prodotto una determinata classificazione.

Nota: data la notevole correlazione tra alcune variabili, si è utilizzata la distanza di Mahalanobis al posto di quella euclidea (v. introduzione). Per questo motivo i gruppi sono leggermente diversi da quelli previsti in base ai dati della matrice delle distanza euclidee vista all'inizio del capitolo. Ad esempio, qui, in base alle distanze di Mahalanobis, Wilma e Minni costituiscono il gruppo con maggiore affinità, a cui poi si aggiunge Betty. Se invece consideriamo le distanze euclidee, Wilma e Betti costituiscono il gruppo con maggiore affinità, a cui poi si aggiunge Minni.