I A C

ISTITUTO PER LE APPLICAZIONI DEL CALCOLO

"MAURO PICONE"

CONSIGLIO NAZIONALE DELLE RICERCHE

QUADERNI

Serie III - N. 132

R. ABBONDANZA - P.G. GHERARDINI - N. LATTANZI

Metodi grafici per la cluster analysis

ROMA

1981

Finito di stampare nel mese di dicembre 1981

Tipo-litografia Marves
Via Mecenate, 35 - Roma - Tel. 730.061

INDICE

1. Introduzione	Pag.	5
2. Definizioni	"	6
3. Il metodo delle curve di Andrews	"	7
4. Il metodo delle facce di Chernoff	"	9
5. Descrizione di un package per l'analisi grafica di dati multivariati	"	12
5.1. Struttura del sistema	"	12
5.2. Utilizzazione interattiva	"	14
6. Applicazioni	"	18
7. Considerazioni conclusive	"	30
Bibliografia	"	32
Liste dei programmi	"	33

INDICE

1. Introduzione Pag. 3

2. Definizioni " 6

3. Il metodo della curva di Andrews " 7

4. Il metodo delle facce di Chernoff " 9

5. Descrizione di un package per l'analisi grafica di dati multivariati " 12

5.1. Struttura del sistema " 12

5.2. Utilizzazione interattiva " 14

6. Applicazioni " 18

7. Considerazioni conclusive " 30

Bibliografia " 32

Lista dei programmi " 33

METODI GRAFICI PER LA CLUSTER ANALYSIS(*)

Rita ABBONDANZA - Pier Giorgio GHERARDINI - Natalino LATTANZI

1. Introduzione

Nell'analisi di dati multivariati, soprattutto in una fase iniziale volta ad "esplorare" la loro struttura, puo' essere di qualche aiuto la possibilita' di una rappresentazione grafica dei dati.

Come e' noto, numerose tecniche di analisi multivariata tendono, attraverso trasformazioni e semplificazioni opportune, a ridurre la dimensione originale dei dati, e quindi rappresentare la configurazione iniziale di "punti" dello spazio p-dimensionale con configurazioni analizzabili in spazi di dimensione inferiore.

Con alcuni fra questi metodi, per esempio quello delle componenti principali e del multidimensional scaling, la riduzione della dimensione e' ottenuta trascurando particolari componenti della variabilita' totale: una conseguenza, certo non l'unica ne' la piu' importante, di questa riduzione della dimensione puo' essere spesso quella di consentire una rappresentazione grafica della configurazione trasformata.

Altre tecniche invece consentono di ottenere una rappresentazione grafica come risultato specifico, associando a ciascun punto dello spazio p-dimensionale un disegno ottenuto come risultato di una o piu' trasformazioni, di tipo funzionale, delle coordinate del punto.

In questo lavoro sono state prese in considerazione solamente due tra le tecniche grafiche piu' note, il metodo delle *curve di Andrews* e quello delle *facce di Chernoff*. Entrambe vengono presentate in modo molto sintetico (nei par. 3 e 4 risp.), al solo scopo di fornire gli elementi essenziali e la terminologia cui si fara' riferimento nei paragrafi successivi.

Non e' stato nelle nostre intenzioni, infatti, di fornire un'esposizione dettagliata dei metodi considerati, ma piuttosto di studiare la loro efficacia in problemi di classificazione di dati multivariati, e fornire allo sperimentatore uno strumento di lavoro particolarmente semplice nell'uso.

Il software sviluppato allo scopo é descritto nel para-

(*) Pervenuto in redazione nel mese di novembre 1981.

grafo 5. Esso e' costituito da un sistema che oltre a permettere, interattivamente, la scelta del metodo e dell'unita' di uscita grafica (video o plotter), offre possibilita' accessorie, quali la facolta' di includere solo un sottoinsieme di variabili o di trasformare opportunamente i dati iniziali.

Alcune esperienze di utilizzazione del sistema sono illustrate nel paragrafo 6, ove si mostrano anche dei confronti tra i risultati conseguiti con queste tecniche grafiche e quelli ottenuti applicando metodi analitici di cluster analysis.

2. Definizioni

Il contesto cui ci riferiamo in questo lavoro e' quello costituito da una situazione sperimentale in cui ciascuna unita' sotto osservazione e' caratterizzata da un ben definito insieme di variabili, e in cui un problema che si pone e' quello di stabilire se sussistono, in modo piu' o meno evidente, raggruppamenti di unita'. Il significato da attribuire al termine raggruppamento e' quello intuitivo di sottoinsieme di unita' in qualche modo omogeneo, e comunque tale che due unita' appartenenti a un dato sottoinsieme siano tra loro piu' simili di quanto siano due unita' di sottoinsiemi diversi.

La situazione sperimentale é descritta convenientemente da una configurazione iniziale di dati costituita da una matrice $X=(x_{ij})$ di dimensioni $n \times p$, ove l'indice di riga i rappresenta il n. d'ordine, o etichetta convenzionale, dell'unita' sperimentale, e l'indice di colonna j denota la generica variabile associata a ciascuna unita'.

Per ogni $i=1,\ldots,n$ il vettore $(x_{i1}, x_{i2}, \ldots, x_{ip})$ rappresenta quindi l'insieme delle osservazioni, su p variabili, associate all'unita' i-ma.

Il termine "individuo" sara' usato come sinonimo di unita'; le variabili prese in considerazione sono le stesse, ovviamente, per tutti gli individui.

E'ben noto che per applicare alcuni metodi "classici" della cluster analysis (i metodi gerarchici) occorre costruire, a partire dalla matrice X, una matrice simmetrica $n \times n$ contenente i valori di somiglianza per ciascuna coppia di individui: la definizione dell'indice di somiglianza e' un problema strettamente legato al tipo di variabili (continue, discrete, qualitative,...) prese in considerazione. Nelle tecniche grafiche descritte nel seguito, tuttavia, questo aspetto non ha una rilevanza specifica, e cio' potrebbe, a nostro avviso, invalidare

in alcune circostanze i risultati conseguibili: riprenderemo questo problema nel paragrafo 7.

3. Il metodo delle curve di Andrews

La rappresentazione grafica prodotta con questo metodo, introdotto da Andrews (1972), fornisce per ogni individuo una curva del piano, e quindi consente di raggruppare gli individui a seconda dell'andamento di queste curve.

L'individuo *i-mo*, caratterizzato come si e' detto dal vettore $\underline{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, e' rappresentato con la funzione

$$(3.1)\quad f_{\underline{x}_i}(t) = x_{i1}/\sqrt{2} + x_{i2}\,sen\,t + x_{i3}\,cos\,t + x_{i4}\,sen\,2t + x_{i5}\,cos\,2t + \ldots$$

per $-\pi < t < \pi$.

Le proprieta' principali di questa rappresentazione, che ne giustificano la scelta (in realta' potrebbero essere considerate anche altre funzioni del tipo della (3.1), cioe' combinazioni lineari di funzioni ortonormali, v. l'articolo citato di Andrews), sono:

1. Se si definisce come distanza d_{ij} tra due di tali funzioni la grandezza

$$d_{ij} = \int_{-\pi}^{\pi} [f_{\underline{x}_i}(t) - f_{\underline{x}_j}(t)]^2 dt$$

allora risulta

$$d_{ij} = \pi \left\| \underline{x}_i - \underline{x}_j \right\|^2$$

e quindi a punti "vicini" nel senso della metrica euclidea, nello spazio p-dimensionale, corrispondono curve "vicine" nel piano.

2. Detto $\underline{x}_*$ il vettore delle medie: $\underline{x}_* = \frac{1}{n}\sum_1^n \underline{x}_i$, risulta

$$f_{\underline{x}_*}(t) = \frac{1}{n}\sum_1^n f_{\underline{x}_i}(t)$$

La rappresentazione conserva quindi il valor medio.

3. Per ogni fissato valore t_0, posto

$$\underline{a} = (1/\sqrt{2},\ sen\ t_0,\ cos\ t_0,\ sen\ 2t_0,\ cos\ 2t_0, \ldots)$$

risulta

$$f_{\underline{x}_i}(t_0) = \frac{\underline{x}_i' \underline{a}}{\| a \|} \cdot \| a \|$$

In ogni punto dunque il valore della funzione è proporzionale alla proiezione del vettore $\underline{x}_i$ su una fissata direzione (i grafici rappresentano l'andamento di tali proiezioni) e la configurazione data dai diversi individui può essere studiata anche in sottospazi particolari.

4. Nell'ipotesi che le variabili siano non correlate e con varianza comune, si può mostrare che la varianza di $f_{\underline{x}_i}(t)$ è costante se p è dispari, quasi costante se p è pari. Questa circostanza, peraltro poco realistica e difficilmente verificabile in pratica, permette, se si suppone ulteriormente la normalità dei dati, di costruire assai semplicemente intervalli di confidenza e test di ipotesi (per dettagli v. Andrews, art. cit.).

Buona parte del successo di questo metodo nell'individuare raggruppamenti risiede nel fatto che effettivamente i dati si presentino in una configurazione con un numero limitato di cluster, abbastanza ben definiti. Se i punti sono molto sparsi le curve che li rappresentano possono intersecarsi più volte, rendendo il grafico assolutamente ingarbugliato.

Un'altra limitazione consiste nel fatto che, per facilità di interpretazione, solo un numero limitato di curve (diciamo al più 10) possono essere rappresentate nello stesso sistema di riferimento: se si hanno molti individui è necessario rappresentarli in gruppi separati, e quindi confrontare i diversi grafici ottenuti.

Quale che sia la struttura iniziale dei dati e la numerosità del campione, può essere utile eseguire la rappresentazione sia partendo dai dati originari, sia trasformando preliminarmente le osservazioni con il metodo delle *componenti principali*, e calcolando le curve a partire da queste, associando la prima componente al primo coefficiente x che compare nell'espressione di $f_{\underline{x}_i}(t)$, la seconda componente al secondo coefficiente, e così via. In questo modo infatti le componenti che

spiegano la maggiore percentuale di variabilita' sono associate, nella rappresentazione attraverso $f_{\underline{x}_i}(t)$, ai coefficienti relativi alle frequenze piu' basse, che sono evidenziate nel grafico in modo piu' netto: cio' dovrebbe risultare in una separazione maggiore tra curve rappresentanti individui dissimili. La trasformazione delle componenti principali (a partire dalla matrice di correlazione) e' inoltre essenziale se le unita' di misura delle variabili sono disomogenee e si e' in presenza di variabili di tipo diverso (qualitative e quantitative).

Per un approfondimento delle possibilita' offerte dal metodo delle *curve di Andrews* nell'analisi di dati multivariati si rimanda a Gnanadesikan (1977) sez. 6.2.

Si noti che questo metodo, contrariamente a quello descritto nel paragrafo che segue, permette la rappresentazione di dati di dimensione p qualsiasi.

4. Il metodo delle facce di Chernoff

Questo metodo, introdotto inizialmente in Chernoff (1971), si basa su un'idea molto semplice: dopo aver definito un disegno stilizzato di "faccia" attraverso funzioni parametriche rappresentanti i diversi elementi della faccia stessa (contorno, occhi, naso, bocca,...), ciascun individuo viene rappresentato con una di queste facce, associando ai parametri da cui dipendono gli elementi somatici i valori osservati delle variabili, cioe' gli elementi del vettore $\underline{x}_i$.

Le facce cosi' ottenute vengono quindi raggruppate a seconda delle "somiglianze" (nel senso piu' comune del termine) riscontrate.

Osservando ad esempio la fig.2 del par.6, si riconosce che in questi disegni il contorno e' dato da due porzioni di ellisse congiungentisi a circa meta' dell'altezza totale, gli occhi sono rappresentati da ellissi, le pupille e le orecchie da cerchi, il naso da un triangolo, la bocca da un arco di circonferenza, le sopracciglia da segmenti.

L'elenco dei parametri da cui questi elementi dipendono e' dato nella tavola seguente, in cui si suppone di riferire il disegno della faccia a un sistema di assi ortogonali con origine nel centro della faccia stessa.

Tav. I - Parametri delle facce di Chernoff

1. Larghezza	e' la distanza tra il centro della faccia e il punto di incontro delle porzioni di ellisse del contorno
2. Posizione dell'orecchio	e' l'angolo tra il segmento di cui sopra e l'asse X
3. Altezza parte superiore	e' la distanza tra il centro e il punto piu' alto della faccia
4. Eccentricita' del contorno sup.	e' l'eccentricita' dell'ellisse superiore
5. Eccentricita' del contorno inf.	e' l'eccentricita' dell'ellisse inferiore
6. Lunghezza del naso	
7. Altezza del centro della bocca	per centro si intende il punto mediano dell'arco di circonferenza rappresentante la bocca
8. Curvatura della bocca	
9. Lunghezza della bocca	
10. Altezza centro degli occhi	e' l'ordinata del centro delle ellissi rappresentanti gli occhi
11. Separazione occhi	e' l'ascissa (in valore assoluto) del centro degli occhi
12. Inclinazione occhi	e' l'angolo, rispetto all'asse X, dell'asse maggiore dell'ellisse degli occhi
12. Eccentricita' occhi	
14. Lunghezza occhi	semiasse maggiore dell'ellisse
15. Posizione delle pupille	ascissa (valore assoluto) del centro degli occhi
16. Altezza delle sopracciglia	distanza tra il centro degli occhi e il centro del segmento rappresentante le sopracciglia
17. Angolo delle sopracciglia	
18. Lunghezza delle sopracciglia	
19. Raggio orecchie	
20. Larghezza naso	

Si puo' notare che non tutte le grandezze occorrenti sono rappresentate in questa lista (per esempio non compare il raggio delle pupille); difatti alcune grandezze vengono calcolate in funzione di quelle specificate. Con questa rappresentazione di dati multivariati e' possibile quindi trattare individui caratterizzati da al piu' 20 variabili: se le variabili sono in numero inferiore, ai parametri in sovrappiu' vengono assegnati valori di *default*, rendendo cosi' comunque possibile il disegno.

L'associazione fra variabili e parametri delle facce (e quindi anche la decisione di assegnare i valori di *default* ad alcuni parametri) e' arbitraria: a volte puo' risultare conveniente associare alcune variabili ritenute piu' importanti (con una valutazione a priori sufficientemente motivata) ai parametri che, nell'opinione dello sperimentatore, mettono in risalto maggiormente le differenze somatiche, per esempio la curvatura della bocca o l'altezza della faccia.

Uno dei maggiori inconvenienti di questa rappresentazione consiste tuttavia non tanto nel limite di 20 per la dimensione dei dati (o nella componente di soggettivita' presente, come si puo' immaginare, nella fase di valutazione delle somiglianze grafiche), quanto nel fatto che alcune caratteristiche somatiche sono interdipendenti. Ad esempio, nel calcolo della bocca vengono usati, oltre ai parametri rilevanti (7, 8 e 9), i parametri 1 (larghezza della faccia), 2 (posizione dell'orecchio), 3 (altezza della parte superiore), 5 (eccentricita' del contorno inferiore), 6 (lunghezza del naso). Cio' e' dovuto all'esigenza di proporzionare reciprocamente alcuni elementi, in modo da ottenere sembianze non troppo distorte.

Nella tavola 2 e' riportato l'elenco completo delle dipendenze.

Tav. 2 - Dipendenze tra gli elementi della faccia

	dipende da
Bocca	1,2,3,5,6 (oltre che dai param. specifici 7, 8 e 9)
Altezza degli occhi	3,6 (oltre che dal param. specifico 10)
Separazione degli occhi	3,4,6 (oltre che dal param. specifico 11)
Altezza delle orecchie	3 (oltre che dal param. specifico 2)

Le pupille e le sopracciglia dipendono dal calcolo degli occhi solo per assicurare un loro corretto posizionamento (dentro e sopra l'occhio risp.).

Per ovviare, almeno in parte, al pericolo di rappresentare dipendenze tra le variabili estranee alla natura dei dati e' opportuno ripetere il disegno delle facce facendo variare in modo pseudo-casuale l'assegnazione dei valori $(x_{i1}, x_{i2}, \ldots, x_{ip})$ ai parametri, e analizzare i cambiamenti nelle diverse classificazioni risultanti.

Se la dimensione p dei dati e' piccola, una soluzione possibile a questo problema e' data dal mantenere costanti (usare quindi i valori di *default*) alcuni dei parametri maggiormente coinvolti nelle dipendenze. Cio' puó essere fatto in due modi: (i) assegnare i valori di *default* ai parametri "dipendenti", cioe' 7,8,9,10,11; (ii) mantenere fissi i parametri da cui questo gruppo dipende, cioe' quelli dall'1 al 6 (si noti che il parametro 2 appartiene ad entrambi i gruppi).

5. Descrizione di un package per l'analisi grafica di dati multivariati

5.1. Struttura del sistema

Forniamo solo l'elenco dei sottoprogrammi costituenti il sistema e un breve cenno alla funzione svolta da ciascuno di essi[1]. Nella parte iniziale della lista del programma principale (MGCA) sono inserite le istruzioni per modificare la prima dimensione (n. di righe) della matrice dei dati (e quella del vettore Y) attualmente fissata a 50; il numero di colonne di tale matrice e' fissato a 20 e la sua modifica, non consigliata, comporterebbe sostanziali modifiche alla subroutine CHERN (e a quelle da essa richiamate), poiche' i parametri che definiscono una faccia di Chernoff (v. par.4) sono appunto 20.

MGCA: rappresenta il "main" del sistema. Visualizza su terminale video la presentazione del sistema, apre la seduta interattiva con la richiesta all'utente del metodo e del terminale grafico desiderato, richiama tutti i sottoprogrammi necessari per il calcolo e la rappresentazione grafica dei dati, consente infine di ripetere la rappresentazione con scelte diverse relative ai metodi e/o ai terminali.

---

(1) I sottoprogrammi CORR, RJCEI, ORD, TRAS, PERC, CORCOV, LOAD sono tratti da Brambilla (1978). Il sottoprogramma CHERN e' una rielaborazione di un programma pubblicato in Bruckner e Mills (1979). La libreria grafica di base per il video Tektronix 4014 e' documentata in Abbondanza, Moltedo e Ocello (1978), mentre quella per il plotter e' descritta nella documentazione originaria fornita dalla Calcomp.

CHERN: richiede interattivamente l'introduzione di alcuni parametri (v. punti (4)-(7) del par.5) e quindi calcola le coordinate dei punti di tutti gli elementi di una faccia di Chernoff.

RANGE: calcola il valore minimo e massimo di un vettore reale.

CIRCLE: calcola le coordinate dei punti di una circonferenza, dato il centro e il raggio.

VDCHER: attraverso l'utilizzo delle funzioni della libreria grafica TEK14 (v.nota (1)), pone il terminale video nelle condizioni necessarie al tracciamento di un grafico, creando la corrispondenza richiesta tra il range dei dati introdotti e i punti schermo su cui rappresentare i dati stessi; richiama quindi il sottoprogramma TRACON per effettuare il disegno dei singoli elementi di una faccia.

TRACON: effettua il tracciamento su video dei vettori che rappresentano gli elementi di una faccia.

PLCHER: utilizzando le funzioni standard della libreria grafica del plotter Calcomp, pone il plotter nelle condizioni necessarie al tracciamento di un grafico ed effettua quindi, attraverso richiami successivi, il disegno della singola faccia.

CODIFI: trasforma un valore numerico in carattere alfanumerico.

ANDR: richiede interattivamente l'introduzione di alcuni parametri (v. punti (10)-(12) del par.5) e quindi calcola per ogni individuo la funzione (3.1) in un insieme di punti equidistanti dell'intervallo $(-\pi,\pi)$.

CORR: calcola la matrice di correlazione (o di covarianza) delle variabili date.

RJCEI: calcola gli autovalori e gli autovettori della matrice di cui sopra.

ORD: ordina in senso discendente gli autovalori.

TRAS: calcola le componenti principali.

PERC: calcola la percentuale di varianza spiegata da ciascun autovalore.

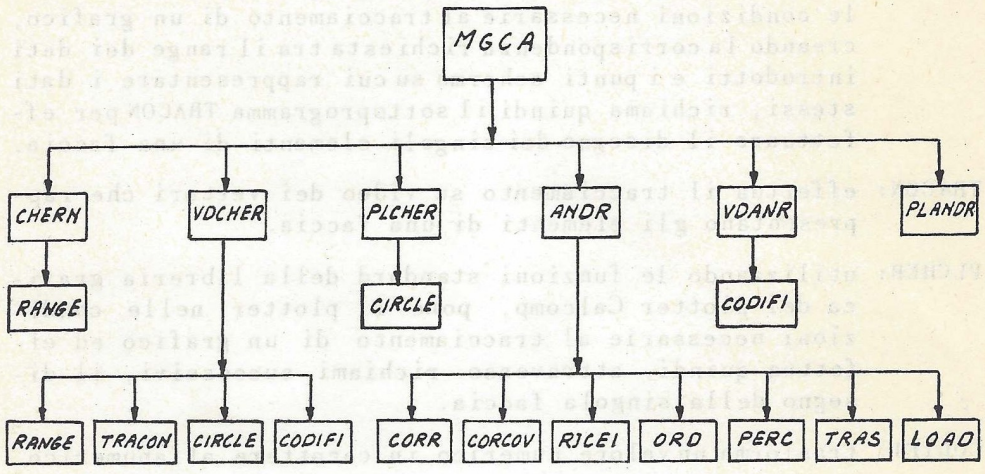
CORCOV: stampa la matrice di correlazione (o di covarianza).

LOAD: calcola le correlazioni fra le variabili originarie e le componenti estratte.

VDANDR: dopo aver predisposto il terminale video-grafico, analogamente al sottoprogramma VDCHER, effettua il tracciamento delle *curve di Andrews*.

PLANDR: dopo aver calcolato tutti gli elementi richiesti dal plotter analogamente al sottoprogramma PLCHER, effettua il disegno delle *curve di Andrews*.

La struttura del sistema e i collegamenti tra le diverse componenti sono descritti dal grafico seguente:



5.2. Utilizzazione interattiva

Riportiamo lo schema dell'utilizzazione interattiva, elencando i messaggi forniti dal sistema, il loro significato e il tipo di informazioni da introdurre di volta in volta.

	MESSAGGIO	*DATI DA INTRODURRE*
1.	SISTEMA PER LA RAPPRESENTAZIONE GRAFICA DI DATI MULTIVARIATI SECONDO I METODI DELLE FACCE DI CHERNOFF E DELLE CURVE DI ANDREWS	
2.	OPZIONI: 1 FACCE DI CHERNOFF 2 CURVE DI ANDREWS	1 o 2, a seconda del metodo desiderato
3.	OPZIONI: 1 VIDEO 2 PLOTTER	1 o 2, a seconda del terminale grafico scelto

MESSAGGIO	*DATI DA INTRODURRE*
A) METODO DELLE FACCE DI CHERNOFF	
4. NUMERO DI VARIABILI DA INCLUDERE	Numero di variabili da includere nell'analisi fra quelle introdotte inizialmente
5. ATTENZIONE: IL N. INTRODOTTO E' MAGGIORE DEL N. DI VARIABILI. INTRODURRE IL VALORE CORRETTO.	Segnalazione di errore nel caso che il numero introdotto in (4) sia >NVREAD, parametro attualmente posto uguale a 20. L'esecuzione continua con la ripetizione di (4)
6. ASSOCIAZIONE TRA I PARAMETRI DELLA FACCIA E LE VARIABILI (INTRODURRE Ø SE SI VUOLE ASSUMERE IL VALORE DI DEFAULT) 1 LARGHEZZA FACCIA 2 POSIZIONE ORECCHIO 3 ALTEZZA META' FACCIA 4 ECCENTRICITA' CONTORNO SUPER. 5 ECCENTRICITA' CONTORNO INFER. 6 LUNGHEZZA NASO 7 ALTEZZA CENTRO BOCCA 8 CURVATURA BOCCA 9 LUNGHEZZA BOCCA 10 ALTEZZA CENTRO OCCHI 11 SEPARAZIONE OCCHI 12 INCLINAZIONE OCCHI 13 ECCENTRICITA' OCCHI 14 LUNGHEZZA OCCHI 15 POSIZIONE PUPILLE 16 ALTEZZA SOPRACCIGLIA 17 ANGOLO SOPRACCIGLIA 18 LUNGHEZZA SOPRACCIGLIA 19 RAGGIO ORECCHIO 20 LARGHEZZA NASO	Il numero d'ordine della variabile che si vuole associare a ciascun parametro; con l'introduzione di Ø viene associato al parametro il valore di default specificato nella procedura (v. assegnazione di valori al vettore DEFAULT in MGCA). Il numero complessivo di associazioni da specificare (escluse quelle di default) è pari al numero introdotto in (4)
7A. ATTENZIONE: L'ULTIMO VALORE INTRODOTTO SUPERA IL N. DI VARIABILI. INTRODURRE IL VALORE CORRETTO	Segnalazione di errore nel caso che il numero appena introdotto in (6) sia > NVREAD. L'esecuzione continua presentando di nuovo la stringa da correggere
7B. ATTENZIONE: SONO STATE USATE SOLO ... VARIABILI, ANZICHE' ... COME PRECEDENTEMENTE INDICATO	Segnalazione di errore nel caso che siano state associate ai parametri della faccia meno variabili di quelle indicate in (4)

MESSAGGIO	*DATI DA INTRODURRE*
A1) RAPPRESENTAZIONE SU VIDEO	
8. SEGNALE ACUSTICO	Premendo RETURN lo schermo viene cancellato ed inizia il disegno delle facce. La rappresentazione grafica prevede 30 facce su ogni pagina video (5 righe e 6 colonne); riempita la prima pagina, con il tasto RETURN si cancella lo schermo e si prosegue su una nuova pagina. Terminata la rappresentazione di tutti gli individui, l'esecuzione continua con (18)
A2) RAPPRESENTAZIONE SU PLOTTER	
9.	La rappresentazione grafica prevede la suddivisione del numero totale di individui (facce) in gruppi di 20, ciascuno su 4 righe e 5 colonne. Terminato il disegno, l'esecuzione continua con (18)
B) METODO DELLE CURVE DI ANDREWS	
10. VARIABILI DA INCLUDERE: SCRIVERE Ø SE SI VUOLE UTILIZZARE L'INTERA MATRICE, ALTRIMENTI SPECIFICARE LE VARIABILI RICHIESTE, INTRODUCENDO I NUMERI CORRISPONDENTI	Introducendo Ø viene utilizzata l'intera matrice dei dati; se invece si vuole selezionare un numero inferiore di variabili (colonne della matrice iniziale), occorre introdurre i numeri d'ordine corrispondenti, separati da virgola
11 OPZIONI: Ø NESSUNA TRASFORMAZIONE 1 COMPONENTI PRINCIPALI UTILIZZANDO LA MATRICE DI COVARIANZA 2 COMPONENTI PRINCIPALI UTILIZZANDO LA MATRICE DI CORRELAZIONE	Ø, 1 o 2, a seconda dell'opzione scelta. Solo nel caso in cui si richiede il calcolo delle componenti principali l'esecuzione continua con (12), altrimenti passa direttamente a (13)
12. OPZIONI: Ø NESSUNA STAMPA	Ø o 1, a seconda dell'opzione scelta. Se si introduce 1 vengono stampati, per ogni variabile, la media

MESSAGGIO	*DATI DA INTRODURRE*
1 STAMPA DI INFORMAZIONI RIASSUNTIVE SULLE VARIABILI E SULLE COMPONENTI PRINCIPALI	e la deviazione standard. Inoltre vengono dati: la matrice di covarianza (o di correlazione, a seconda della scelta precedente), la varianza di ciascuna componente ricavata e la percentuale della varianza totale, la correlazione tra le variabili originarie e le componenti estratte
B1) RAPPRESENTAZIONE SU VIDEO	
13. N. DI CURVE SU OGNI GRAFICO	Numero massimo di curve da disegnare su ogni grafico (pagina video). Introdotto il numero, lo schermo viene cancellato, vengono quindi tracciati gli assi coordinati. Sull'asse x i valori variano da $-\pi$ a π; i valori sull'asse y sono scalati automaticamente in modo tale da essere compresi nell'intervallo $[-10,10]$ (in realta' il fattore di scala é ricavato dai valori della prima curva e mantenuto costante per tutte quelle successive: puo' capitare quindi che i valori associati a qualche individuo fuoriescano dall'intervallo anzidetto). Vengono tracciate poi le singole curve, ciascuna delle quali e' contrassegnata automaticamente dal numero d'ordine dell'individuo corrispondente. Completata la prima pagina video con il tasto **RETURN** si cancella lo schermo e si prosegue con pagine successive fino al completamento della rappresentazione. L'esecuzione passa poi a (18)
B2) RAPPRESENTAZIONE SU PLOTTER	
14. N. DI CURVE SU OGNI GRAFICO	Numero massimo di curve da disegnare su ogni grafico

MESSAGGIO	*DATI DA INTRODURRE*
15. LUNGHEZZA ASSE X LUNGHEZZA ASSE Y	Lunghezza richiesta, in centimetri, per gli assi x e y
16. FATTORE DI RIDUZIONE	Con il fattore di riduzione f e' possibile ridurre o ingrandire il disegno, poiche' per f vengono moltiplicate le coordinate di ogni punto (da tenere presente che l'unita' di misura sul plotter e' il centimetro)
17. ATTIVAZIONE DEL PLOTTER	La rappresentazione, che inizia con il disegno degli assi coordinati (per quanto riguarda la scala, v. (13)), prevede la suddivisione del numero totale di individui (curve) in gruppi, la cui numerosita' e' quella specificata in (14). Terminata la rappresentazione, il controllo torna al terminale video, lo schermo viene cancellato e l'esecuzione continua con (18)
18. INTRODURRE: 1 DISEGNO SU PLOTTER 2 MODIFICA DELLE ASSOCIAZIONI TRA PARAMETRI E VARIABILI (FACCE DI CHERNOFF) 3 USO DI VARIABILI DIVERSE (CURVE DI ANDREWS) 4 SCELTA DEL METODO 5 STOP	1,2,3,4,5, a seconda della opzione scelta. OPZ 1: l'esecuzione riprende da (4) o (10), a seconda del metodo che si e' scelto in precedenza OPZ 2: l'esecuzione riprende da (3) e continua con (4) e succes. OPZ 3: l'esecuzione riprende da (3) e continua con (10) e succes. OPZ 4: l'esecuzione riprende da (2) OPZ 5: termine dell'esecuzione.

6. Applicazioni

Con gli esempi che proponiamo in questo paragrafo si e' inteso sia illustrare l'utilizzazione del package descritto, sia presentare i risultati di un confronto tra metodi grafici e metodi analitici "classici". Per applicare questi ultimi si e' usato il package GENSTAT, descritto in Brambilla, C. e Gherardini, P.G. (1981): vengono forniti solo i risultati delle ana-

lisi eseguite, non la specifica dei programmi GENSTAT utilizzati.

ESEMPIO 1 - I dati utilizzati in questo esempio sono quelli analizzati da Andrews (1972); essi si riferiscono ai valori medi di 8 misurazioni su un dente premolare, prese su individui appartenenti a 9 gruppi, di cui i primi 3 esseri umani, i rimanenti dati da diversi tipi di scimmie antropomorfe. Precisamente tali gruppi sono:

1. Africano occidentale
2. Britannico
3. Aborigeno australiano
4. Gorilla maschio
5. " femmina
6. Orang-utan maschio
7. " femmina
8. Scimpanze' maschio
9. " femmina

I metodi applicati sono: (1) *Curve di Andrews*; (2) *Facce di Chernoff*; (3) Metodo dell'*unione semplice*; (4) Metodo di *ottimizzazione*.

(1) *Curve di Andrews*

Il grafico ottenuto e' rappresentato in fig. 1.

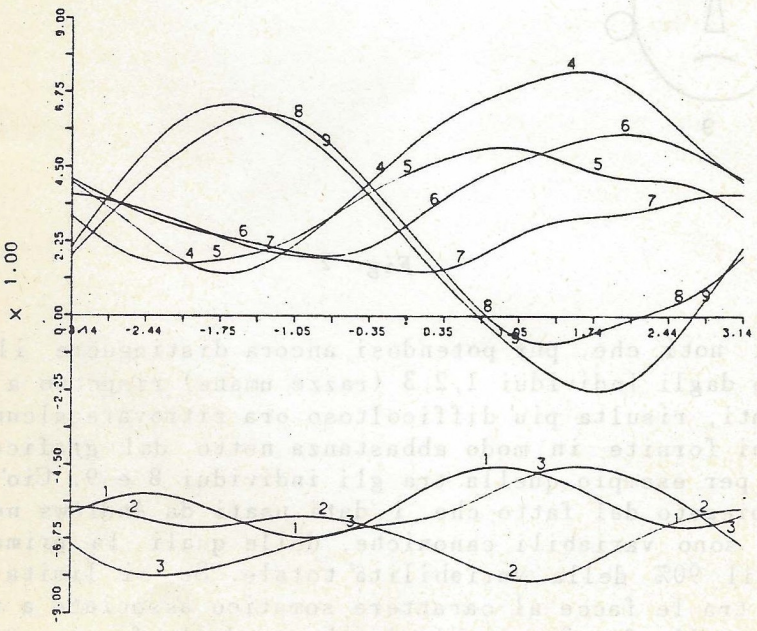


Fig. 1

(2) *Facce di Chernoff*

Associando le 8 variabili ai parametri da 1 a 8 della Tav.1 (nell'ordine) si ottiene il disegno di fig.2.

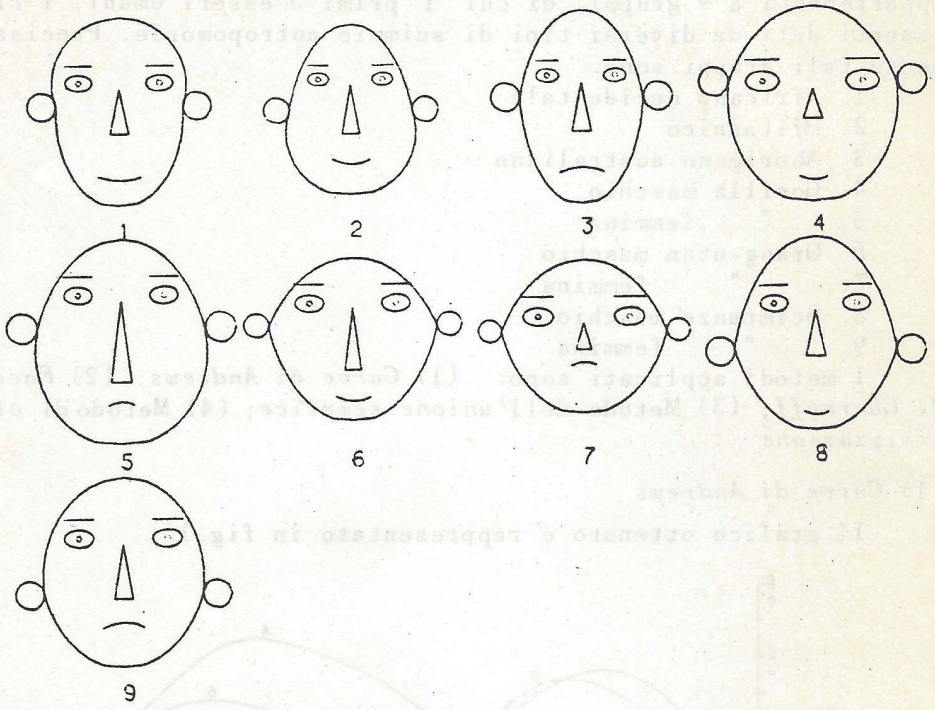


Fig. 2

Si noti che, pur potendosi ancora distinguere il gruppo formato dagli individui 1,2,3 (razze umane) rispetto a tutti i rimanenti, risulta piu' difficoltoso ora ritrovare alcune similitudini fornite in modo abbastanza netto dal grafico precedente, per esempio quella tra gli individui 8 e 9. Cio' puo' essere spiegato dal fatto che i dati usati da Andrews nell'art. citato sono variabili canoniche, delle quali la prima spiega oltre il 90% della variabilita' totale. Se si limita il confronto tra le facce al carattere somatico associato a tale variabile nella fig.3, cioe' la larghezza della faccia, si ritrovano i gruppi ben distinti dati dalla fig.2.

(3) Metodo dell'*unione semplice*

Il dendrogramma relativo e' rappresentato nella fig.3.

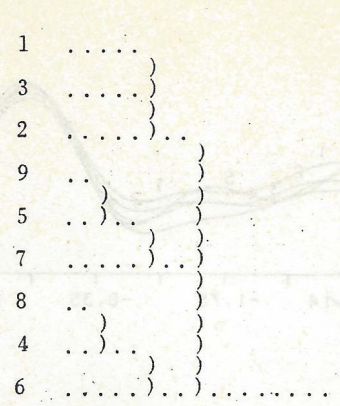


Fig.3

Ancora una volta risalta il gruppo dei primi tre individui, mentre si notano i raggruppamenti (5,9) e (4,8) che non appaiono nelle figure precedenti.

Anche i metodi della *mediana* e dell'*unione media* (di cui non si riportano i dendrogrammi) danno lo stesso risultato, cioe' i tre gruppi (1,2,3), (4,8,6) e (5,9,7).

(4) Metodo di *ottimizzazione*

Il criterio scelto e' quello di minimizzare la somma delle distanze euclidee tra gli individui di una stessa classe. Con tre classi si ottengono i raggruppamenti (1,2,3), (4,5,6,7), (8,9); ovviamente questo metodo fornisce risultati del tutto equivalenti a quelli dati da (1), in virtu' della proprieta' 1 ricordata nel par.3.

ESEMPIO 2 - In questa applicazione gli "individui" sono 20 campioni di roccia estratti da un corpo igneo complesso e apparentemente disomogeneo (dati tratti da Davis (1973), p.528). Per ciascuno di essi sono state misurate le percentuali presenti di 8 importanti costituenti chimici (ossidi), che in pratica esauriscono la composizione di ciascun frammento. Si ha dunque una matrice iniziale 20×8.

L'analisi e' stata condotta con i metodi seguenti: (1) *Curve di Andrews*; (2) *Facce di Chernoff*; (3) Metodo dell'*unione semplice*; (4) Metodo dell'*unione completa*; (5) Metodo dei *centroidi*.

(1) *Curve di Andrews*

I grafici di fig.4 mostrano le curve ottenute partendo dai dati originari. Si noti come sia assai piu' difficile in questo caso individuare raggruppamenti distinti.

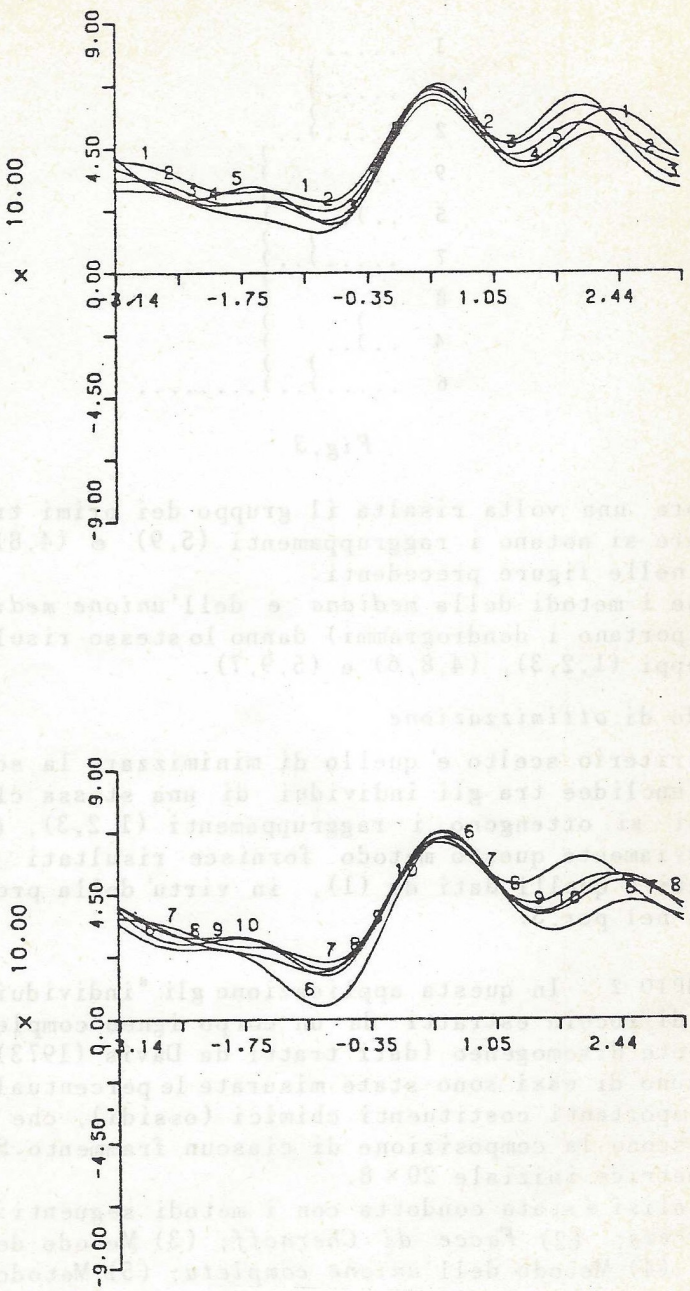


Fig. 4a

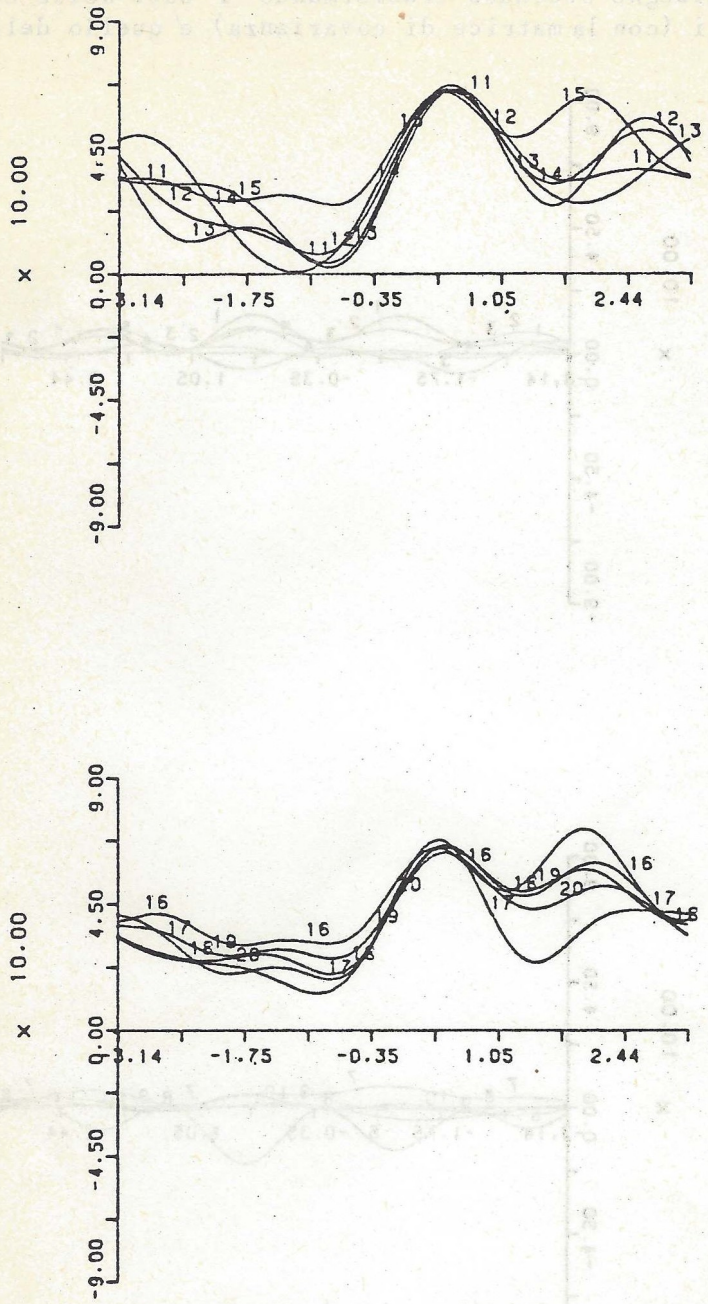


Fig. 4b

Il disegno ottenuto trasformando i dati nelle componenti principali (con la matrice di covarianza) e' quello della fig. 5.

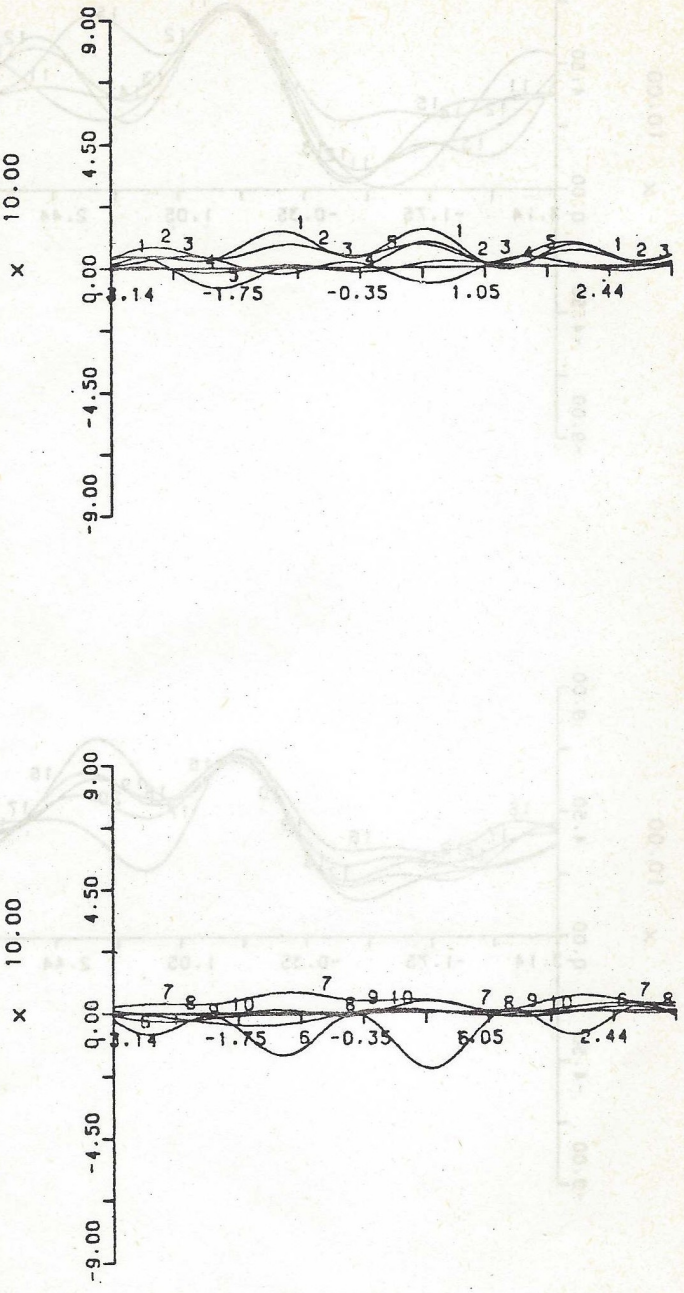


Fig. 5a

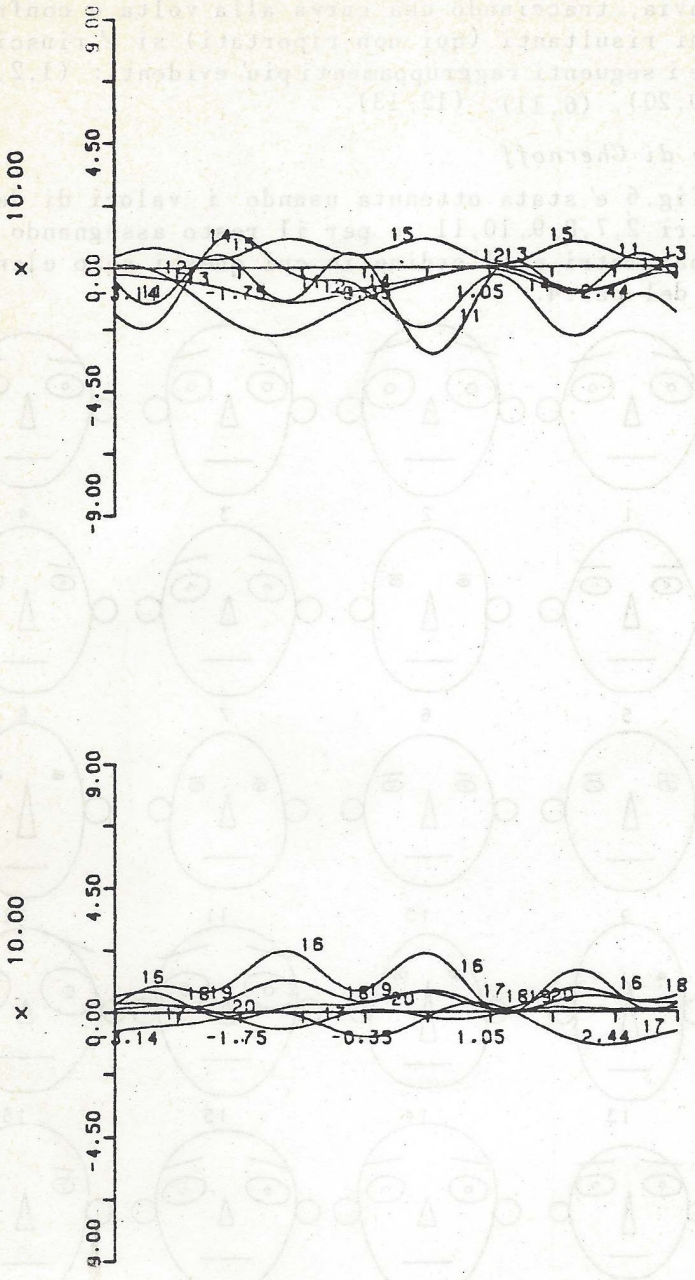


Fig. 5b

Anche questi grafici non sono di immediata interpretazione; tuttavia, tracciando una curva alla volta e confrontando i 20 disegni risultanti (qui non riportati) si e' riusciti ad individuare i seguenti raggruppamenti piu' evidenti: (1,2,7,15,19), (4,8,9,10,20), (6,11), (12,13).

(2) *Facce di Chernoff*

La fig.6 e' stata ottenuta usando i valori di default per i parametri 2,7,8,9,10,11, e per il resto assegnando le variabili ai parametri nell'ordine in cui questi sono elencati nella Tav.1 del par.4.

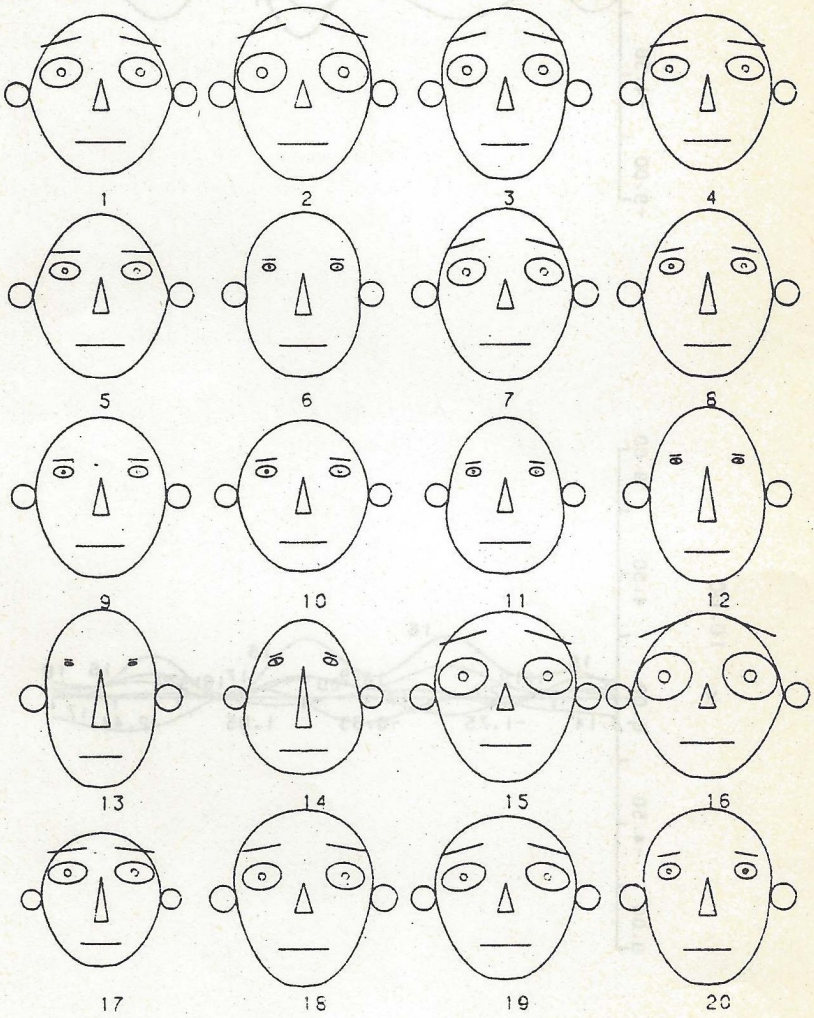


Fig. 6

Rispetto al metodo precedente sembra piu' agevole sia indicare dei raggruppamenti (per esempio (1,7,18,19), (2,15), (6,12,13), (9,10,20)), sia notare individui chiaramente diversi (per esempio (14,16 e 17)).

(3) Metodo dell'*unione semplice*

Anche con questo metodo si ottiene una scarsa discriminazione, come mostrato dal dendrogramma di fig.7.

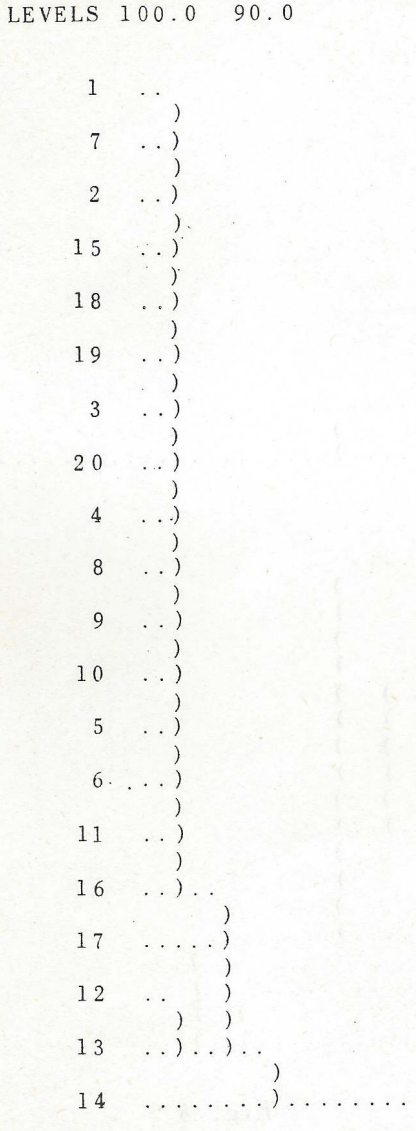


Fig. 7

(4) Metodo dell'*unione completa*

Si e' ottenuto il dendrogramma di fig.8, in cui si notano i gruppi (1,7,4,19,2,15,3,18), (5,8), (6,11,9,10,20), (12,13), e gli individui 16, 14 e 17 non raggruppati.

Questi risultati concordano sostanzialmente con quelli forniti dalle *facce di Chernoff*.

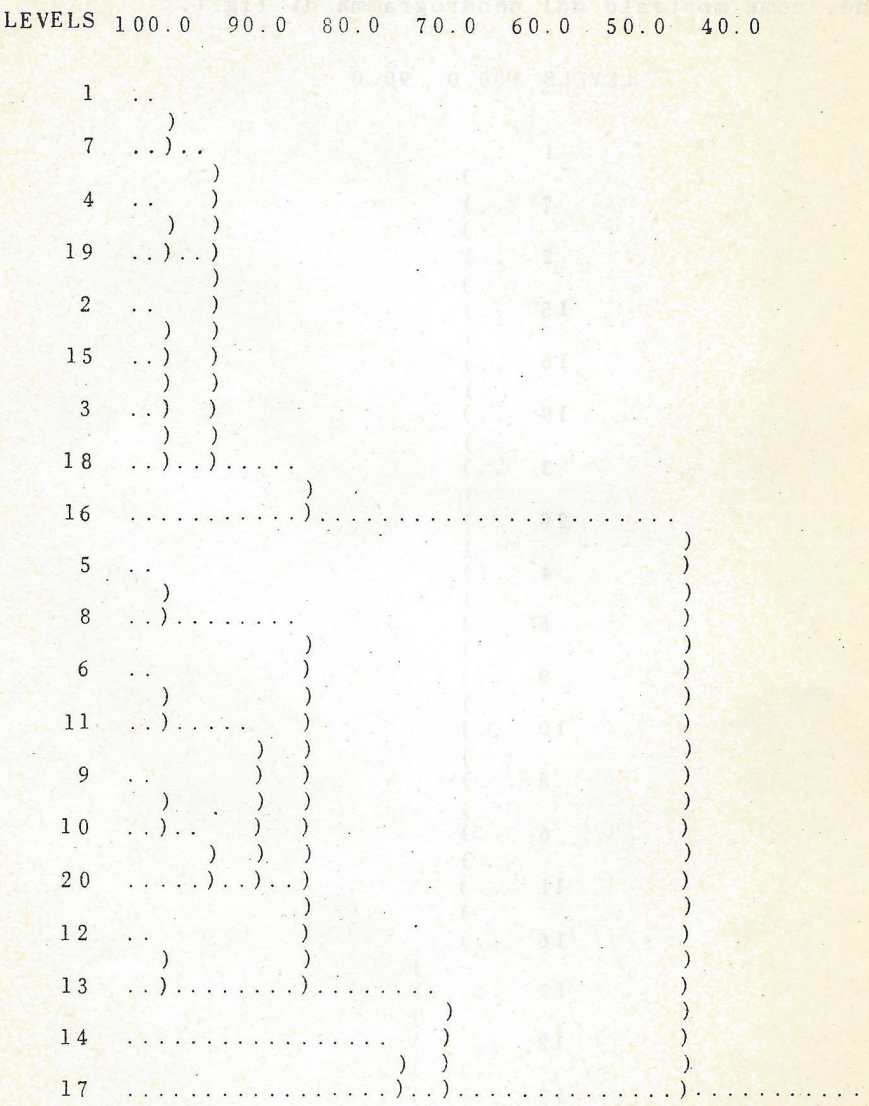


Fig. 8

(5) Metodo dei *centroidi*

I gruppi forniti da questo metodo sono abbastanza simili a quelli dati dal precedente; ispezionando il dendrogramma di fig. 9 si riconoscono infatti i raggruppamenti (1,7 2,15,18,3, 4,19), (6,11), (8,9,10), (12,13), mentre rimangono non classificati gli individui 16, 5, 20, 17, 14.

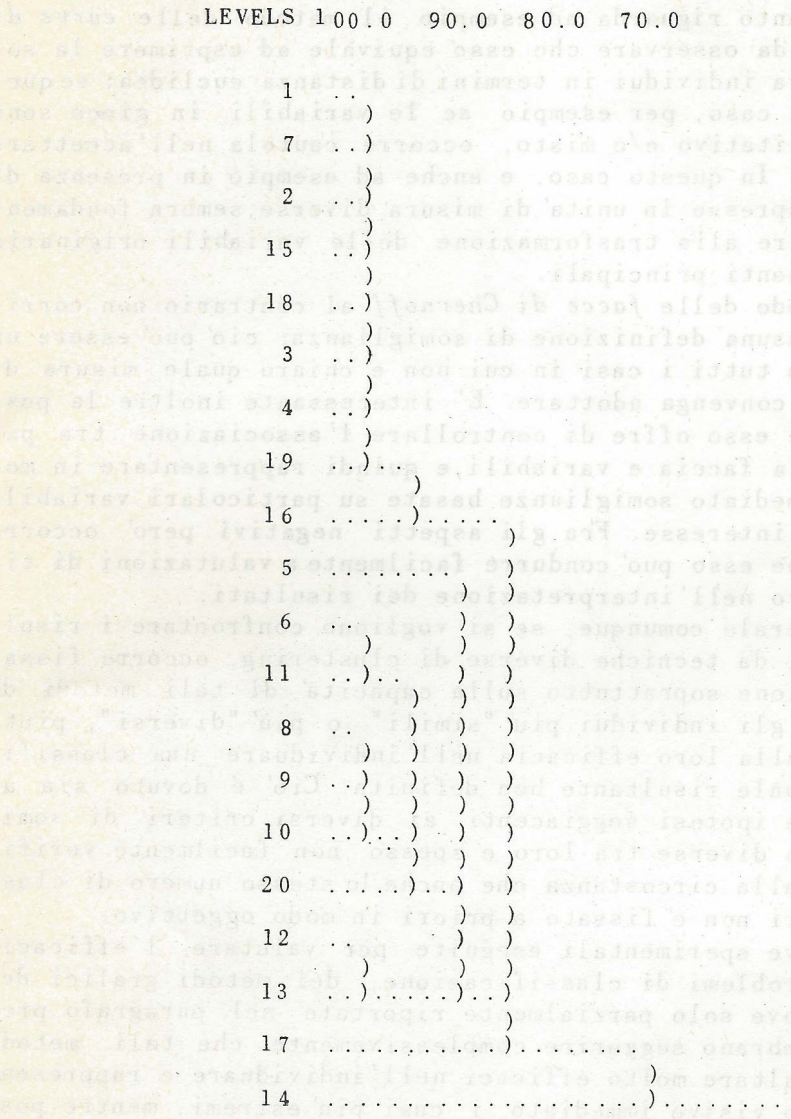


Fig. 9

7. Considerazioni conclusive

Come si è accennato nel par.2, i metodi grafici qui presentati non richiedono la definizione iniziale di una matrice di somiglianza. Cio' puo' essere un vantaggio con certi insiemi di dati, mentre puo' creare problemi di interpretazione con altri.

Per quanto riguarda ad esempio il metodo delle *curve di Andrews* c'e' da osservare che esso equivale ad esprimere la somiglianza tra individui in termini di distanza euclidea: se questo non e' il caso, per esempio se le variabili in gioco sono di tipo qualitativo e/o misto, occorre cautela nell'accettare i risultati. In questo caso, e anche ad esempio in presenza di variabili espresse in unita' di misura diverse, sembra fondamentale ricorrere alla trasformazione delle variabili originarie nelle componenti principali.

Il metodo delle *facce di Chernoff* al contrario non corrisponde a nessuna definizione di somiglianza; cio' puo' essere un vantaggio in tutti i casi in cui non e' chiaro quale misura di somiglianza convenga adottare. E' interessante inoltre la possibilita' che esso offre di controllare l'associazione tra parametri della faccia e variabili, e quindi rappresentare in modo molto immediato somiglianze basate su particolari variabili di maggiore interesse. Fra gli aspetti negativi pero' occorre ricordare che esso puo' condurre facilmente a valutazioni di tipo soggettivo nell'interpretazione dei risultati.

In generale comunque, se si vogliono confrontare i risultati forniti da tecniche diverse di clustering, occorre fissare l'attenzione soprattutto sulla capacita' di tali metodi di evidenziare gli individui piu' "simili" o piu' "diversi", piuttosto che sulla loro efficacia nell'individuare una classificazione globale risultante ben definita. Cio' è dovuto sia al fatto che le ipotesi soggiacenti ai diversi criteri di somiglianza sono diverse tra loro e spesso non facilmente verificabili, sia alla circostanza che anche lo stesso numero di classi risultanti non e' fissato a priori in modo oggettivo.

Le prove sperimentali eseguite per valutare l'efficacia reale, in problemi di classificazione, dei metodi grafici descritti (prove solo parzialmente riportate nel paragrafo precedente) sembrano suggerire complessivamente che tali metodi possono risultare molto efficaci nell'individuare e rappresentare in modo visivo immediato i casi piu' estremi, mentre possono dare risultati globali deludenti, soprattutto quando gli individui sono in gran numero.

Concludendo occorre ribadire che la rappresentazione gra-

fica di dati multivariati allo scopo di individuare raggruppamenti puo' essere utile soprattutto in una fase di analisi informale, in cui si vuole avere un'idea delle caratteristiche piu' evidenti dell'insieme di individui analizzato. Indipendentemente dalle difficolta' di ordine pratico specifiche di ciascuno dei metodi presentati, é importante che questa analisi informale possa trovare conferma attraverso metodi di tipo analitico.

Bibliografia

ABBONDANZA, R. - MOLTEDO, L. - OCELLO, N.: Estensione di un insieme di sottoprogrammi base per l'utilizzo specifico del terminale grafico TEKTRONIX 4014, *Quaderno IAC, s.III*, n.72 (1978).

ANDREWS, D.F.: Plots of High Dimensional Data, *Biometrics, 28*, 125-136 (1972).

BRAMBILLA, C.: Un codice di calcolo per l'ordinamento di dati multivariati, *Quaderno IAC, s.III*, n.74 (1978).

BRAMBILLA, C. - GHERARDINI, P.G.: Il sistema GENSTAT, *Quaderno IAC, s.III*, n.125 (1981).

BRUCKNER, L.A. - MILLS, C.F.: The Interactive Use of Computer Drawn Faces to Study Multidimensional Data, *Los Alamos Scientific Laboratory Informal Report* n. LA-7752-MS, Los Alamos (1979).

CHERNOFF, H.: The Use of Faces to Represent Points in n-dimensional Space Graphically, *Dept. of Statistics Technical Report n.71, Stanford University*, Stanford (1971).

DAVIS, J.C.: *Statistics and data analysis in geology*, Wiley, New York (1973).

GNANADESIKAN, R.: *Methods for statistical data analysis of multivariate observations*, Wiley, New York (1977).