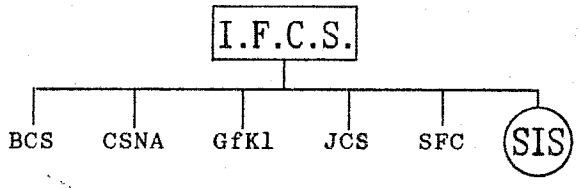
SOCIETA' ITALIANA DI STATISTICA
GRUPPO ITALIANO ADERENTE ALL'I.F.C.S.
(International Federation of Classification Societies)

ATTI DELLE GIORNATE DI STUDIO DEL GRUPPO ITALIANO ADERENTE ALL'I.F.C.S.

ERICE - TRAPANI
24 - 25 Ottobre 1988



CLUSTER ANALYSIS PER MATRICI DI DATI A DUE O PIU' INDICI E CON LIVELLO DI MISURAZIONE MISTO

AGOSTINO DI CIACCIO
Dipartimento di Statistica, Probabilita' e Statistiche Applicate
Universita' di Roma "La Sapienza"

JAN DE LEEUW
Dpt. of Mathematics, UCLA - Los Angeles

1. Introduzione

In questo lavoro proponiamo alcuni metodi di classificazione che hanno particolare rilevanza quando si intenda analizzare uno o piu' insiemi di caratteri qualitativi o misti. Ci riferiremo in particolare all'analisi di una o piu' matrici di dati di tipo *unita'* x *caratteri*, ossia matrici a due indici e due modi, o di tipo *unita'* x *caratteri* x *occasioni*, ossia a tre indici e tre modi (cfr. Coppi e Bolasco eds., 1989). Considereremo inoltre solo problemi di classificazione di tipo non gerarchico delle unita'.

Un approccio molto generale in questo contesto puo' essere basato sulla definizione di opportuni Criteri di Massima Associazione (cfr Saporta 1988) e sull'Optimal Scaling dei caratteri (cfr. Gifi 1981).

Consideriamo innanzitutto matrici di dati a due indici. In sostanza il problema della ricerca di una classificazione di I unita' statistiche rispetto a J caratteri X_j osservati puo' essere visto equivalente alla individuazione delle determinazioni di una variabile nominale Z latente che sia massimamente "*associata*" ai caratteri osservati. Cio' puo' essere scritto

$$\max_{z^*} \sum_{j=1}^{J} \phi(\, z^* ; x_j^*) \qquad (1)$$

in cui ϕ e' una opportuna misura di associazione, z^* e' l'insieme delle determinazioni incognite del carattere Z ed infine x_j^* e' l'insieme delle determinazioni del carattere osservato X_j.

Si potrebbe pensare a questo punto che la scelta di ϕ e' necessariamente condizionata al livello di misurazione dei caratteri in esame, in realta' l'introduzione dell'Optimal Scaling permette l'utilizzo di misure specifiche per il caso quantitativo, anche in presenza di caratteri qualitativi o misti. In particolare vedremo che la scelta del quadrato del coefficiente di correlazione risulta particolarmente significativa ed interpretabile sul piano teorico.

Un altro approccio, basato anch'esso sui criteri di massima associazione ma non sull'Optimal Scaling e limitato all'analisi dei caratteri nominali, e' presente in

Marcotorchino (1986), dove vengono proposti dei criteri in cui la misura ϕ e' applicabile al caso nominale.

Un'ulteriore possibilita' consiste nel trasformare la matrice dei dati *unita'* x *caratteri* in una matrice di prossimita' tra unita' utilizzando indici che si adattino a tale situazione, in modo da ottenere una matrice di dati omogenei a cui applicare uno dei metodi di classificazione gia' noti. L'analisi '*indiretta*' che ne deriva introduce pero' notevoli elementi di arbitrarieta', sia perche' il risultato dipende fortemente dall'indice scelto sia perche' la bonta' del risultato ottenuto e' difficilmente verificabile in termini della matrice dei dati originari.

Nel prosieguo ci occuperemo in dettaglio dei criteri basati sul quadrato del coefficiente di correlazione e mostreremo come essi permettono la formulazione di tecniche particolarmente efficaci e potenti. Vedremo tra l'altro come l'approccio seguito permetta di ritrovare come caso particolare un metodo di classificazione proposto da Diday e coll. (1979) che costituisce la prima applicazione dell'Optimal Scaling alla Cluster Analysis, il metodo GROUPALS di Van Buuren e Heiser (1989), la Forced Classification di Nishisato (1984), la Principal Cluster Analysis di Ibrahim e Schecktman (1986), il metodo SYNCLUS di De Sarbo e altri (1984).

Nel caso in cui si abbia a disposizione non una matrice unita' x caratteri ma una matrice di dati a tre indici e tre modi, ossia del tipo unita' x caratteri x occasioni, le tecniche usuali risultano inadeguate ed e' necessario considerare tecniche specifiche di analisi. Infatti in tale situazione la classificazione delle unita' puo' rispondere ad obiettivi diversi, anche in dipendenza delle caratteristiche della matrice che si sta analizzando. Nonostante la notevole rilevanza applicativa, non vi e', a nostra conoscenza, in letteratura un'esame dei diversi obiettivi ne' tantomeno proposte di metodi in grado di soddisfare le diverse necessita' di analisi che si possono presentare. La ricerca si e' rivolta infatti quasi esclusivamente all'analisi di piu' matrici di prossimita' (cfr. Carroll e Arabie 1983, Bellacicco 1989).

2. Parametrizzazione dei dati e Optimal Scaling

L'informazione fornita dai dati a nostra disposizione puo' essere vantaggiosamente espressa in forma parametrica seguendo l'approccio dell'Optimal Scaling. Accenniamo qui brevemente solo ad alcuni aspetti essenziali, rimandando per un'esposizione piu' dettagliata all'ampia bibliografia disponibile (cfr. ad es. Gifi 1981, Young 1981, Di Ciaccio 1989).

E' noto che per analizzare un carattere nominale si possono utilizzare le *variabili indicatrici*, ossia variabili *dummy* che indicano la presenza/assenza di una determinata modalita' in corrispondenza di ciascuna unita' statistica. In tal modo il carattere nominale viene analizzato considerando le possibili combinazioni lineari delle sue variabili indicatrici. Tale approccio e' implicito in numerose tecniche di analisi di caratteri nominali, ad es. ANOVA e Analisi delle Corrispondenze. D'altra parte si vede

facilmente che ricercare una particolare combinazione lineare di variabili indicatrici coincide con la ricerca di una opportuna quantificazione del carattere, dove i coefficienti della combinazione lineare sono le quantificazioni delle modalita' corrispondenti.

Se il carattere non e' nominale ma ordinale, allora e' consigliabile imporre dei vincoli alle quantificazioni in modo che queste siano *coerenti* con tale informazione a priori (cfr. Herzel 1974). L'imposizione dei vincoli di ordinamento nel caso di caratteri ordinali, rispecchia ovviamente il piu' ricco livello informativo a nostra disposizione rispetto al livello nominale.

Se il carattere e' quantitativo, evidentemente non e' necessario *quantificarlo*, tuttavia anche in questo caso il dato puo' essere utilmente trasformato (riquantificato). Ad esempio in alcune situazioni puo' essere utile considerare trasformazioni non-lineari per applicare un modello di tipo lineare in presenza di relazioni nonlineari tra i caratteri.

Tali considerazioni indicano la presenza di un grado di *indeterminatezza* nella matrice dei dati in esame, poiche' per ogni carattere possiamo avere piu' quantificazioni o trasformazioni "*equivalenti*". Cio' puo' essere reso in forma parametrica attraverso un'opportuna formulazione. Se ad esempio abbiamo un carattere nominale con M"modalita' osservato su I unita' statistiche, possiamo scrivere il generico vettore dei dati quantificati come

$$\mathbf{x} = \mathbf{F}_{(n)}\,\alpha$$

in cui $\mathbf{F}_{(n)}$ e' la matrice che ha per colonne le variabili indicatrici, α e' un vettore a M elementi contenente le quantificazioni delle modalita'. L'insieme dei vettori x quantificati (ossia delle possibili combinazioni lineari delle variabili indicatrici) individua allora un sottospazio lineare $\Re^M \subset \Re^I$. Tale sottospazio contiene quindi i vettori che esprimono un'informazione "*equivalente*".

Nel caso di un carattere ordinale richiediamo che le quantificazioni rispettino l'ordinamento presente tra le M modalita'. L'imposizione di questo vincolo porta quindi ad individuare un sottoinsieme del sottospazio $\Re^M$, e precisamente un cono poliedrico convesso i cui generatori sono noti. Infatti in tal caso potremo scrivere il generico vettore come

$$\mathbf{x} = \mathbf{F}_{(o)}\,\alpha \qquad \text{con} \quad \alpha \geq 0$$

in cui $\mathbf{F}_{(o)}$ e' la "*matrice indicatrice di ordine*" (Bouroche e altri 1977).

Se la variabile e' quantitativa e' possibile considerare delle trasformazioni nonlineari di tipo spline. In questo caso, utilizzando il fuzzy coding e le Basic-splines, e' possibile costruire una matrice $\mathbf{F}_{(q)}$ tramite cui esprimere il vettore dei dati trasformati come

$$\mathbf{x} = \mathbf{F}_{(q)}\alpha \qquad \text{con} \quad \alpha \geq 0$$

in cui il vincolo di non-negativita' sul vettore α e' necessario solo ove si richieda la monotonicita' della trasformazione (cfr. Di Ciaccio 1989, Van Rijckevorsel e De Leeuw 1988). Occorre anche osservare che non essendo interessati a quantificazioni costanti o

banali, cio' porta ovviamente alla considerazione, per tutti i casi descritti, delle sole quantificazioni che si trovano sul sottospazio ortogonale al vettore unitario.

Senza perdita di generalita', nel prosieguo indicheremo la generica quantificazione/trasformazione dell'insieme delle determinazioni del carattere X semplicemente come $x = F\alpha$, in cui la particolare struttura della matrice F e l'eventuale vincolo di non-negativita' per il vettore α, dipendenti dal livello di misurazione della variabile, non saranno specificati. Indicheremo invece con il simbolo G una matrice che ha per colonne L variabili indicatrici.

Una volta ottenuta un'opportuna esplicitazione parametrica del dato, occorre applicare un procedimento che permetta l'individuazione dei parametri introdotti, e quindi delle quantificazioni/trasformazioni dei caratteri. Cio' puo' essere ottenuto considerando un procedimento di ottimizzazione che determini complessivamente sia i parametri relativi ai dati sia quelli relativi al metodo che si vuole applicare. In tal modo e' possibile estendere all'analisi dei caratteri misti le usuali tecniche di analisi multivariata per caratteri quantitativi. Tale approccio risulta estremamente efficace e versatile se si formula in termini di *criteri di massima associazione* (Di Ciaccio 1989).

3. Criteri di massima correlazione e Cluster Analysis

In questo paragrafo ci riferiremo esplicitamente a criteri basati sul coefficiente di correlazione e sul rapporto di correlazione, poiche' essi permettono di evidenziare il legame esistente tra modellistica multilineare e classificazione non gerarchica con le proprieta' che ne conseguono.

Considereremo quindi il criterio implicito nei metodi di classificazione basati sulla massimizzazione della traccia della matrice di covarianza o codevianza tra le classi, facendo particolare riferimento al ben noto metodo delle K-medie (MacQueen 1967).

Tali metodi analizzano una matrice unita' x variabili quantitative al fine di individuare una suddivisione in L classi, dove il valore di L deve essere fissato a priori. Consideriamo in particolare J variabili quantitative osservate $X_1, X_2, \ldots, X_j, \ldots, X_J$, con vettori di determinazioni $x_1, x_2, \ldots, x_j, \ldots, x_J$.
La funzione obiettivo puo' essere scritta come

$$\max \; Tr(B)$$

dove B e' la matrice di codevianza tra le classi. Il termine generico di B puo' essere scritto come

$$B = ((\sum_l I_l \, (\bar{x}_{lj}-\bar{x}_j) \, (\bar{x}_{lj'}-\bar{x}_{j'})))_{j,j'}$$

in cui $\qquad \bar{x}_j = \frac{1}{I}\sum_i x_{ij} \qquad\qquad \bar{x}_{lj} = \frac{1}{I_l} \sum_{i \varepsilon I_l} x_{ij}$

avendo indicato con I_l l'insieme delle unita' che appartengono alla classe l-esima ed anche, con lo stesso simbolo, la cardinalita' di tale insieme.

Si ricava quindi

$$\max \; Tr(\,B\,) = \max \; \sum_j \sum_l I_l \, (\bar{x}_{lj} - \bar{x}_j)^2 \qquad (2)$$

in cui il max e' preso rispetto all'insieme delle possibili classificazioni delle I unita' in L classi.

Sia Z una variabile nominale con L modalita' con insieme di determinazioni z^*. E' noto che

$$\sum_j \eta^2(\, z^* \,;\, x_j \,) = \sum_j \frac{\sum_l I_l \, (\bar{x}_{lj} - \bar{x}_j)^2}{\sum_i (x_{ij} - \bar{x}_j)^2} \qquad (3)$$

quindi se le variabili X_j sono standardizzate e considerando che una classificazione in L classi di un insieme di unita' puo' essere vista come l'insieme delle determinazioni di una variabile nominale con L modalita', potremmo scrivere considerando z^* incognita

$$\max_{z^*} \; \sum_j \eta^2(\, z^* \,;\, x_j) \;=\; \max_{z^*} \, Tr(\,B\,) \qquad (4)$$

in cui z^* appartiene all'insieme delle possibili classificazioni di I unita' in L classi.

D'altra parte per variabili X_j standardizzate si dimostra che vale anche la seguente relazione (Di Ciaccio 1989)

$$\max_{z^*} \sum_j \eta^2(\, z^* \,;\, x_j \,) \;=\; \max_{G,\gamma(J)} \sum_j r^2(\, G\gamma_j \,;\, x_j \,) = \max_{z^*} Tr(B) \qquad (5)$$

in cui G puo' variare nell'insieme delle matrici indicatrici di ordine IxL, γ_j e' un vettore di L parametri incogniti, $\gamma'(J) = \{\, \gamma'_1, \gamma'_2, \ldots\ldots\ldots, \gamma'_J \}$.
Il criterio

$$\max_{G,\gamma(J)} \sum_j r^2(G\gamma_j \,;\, x_j) \qquad (5b)$$

in base a quanto detto, coincide quindi con quello dei metodi di Cluster Analysis basati sulla massimizzazione della traccia della matrice di codevianza o covarianza tra le classi purche' le variabili siano standardizzate.

D'altra parte se vogliamo che le variabili abbiano lo stesso ruolo nell'analisi (e nel caso di disomogeneita' delle variabili e ancor piu' nel caso misto cio' ha particolare importanza) e' opportuno utilizzare variabili standardizzate.

Un procedimento euristico per la soluzione della (5b) e' dato proprio dall'algoritmo delle K-medie applicato alla matrice X dei dati (che stiamo supponendo quantitativi). In tal caso la partizione individuata definisce G mentre i centroidi dei clusters sono dati proprio dai parametri $\gamma_l = (\gamma_{1l}, \gamma_{2l}, \ldots\ldots\ldots, \gamma_{Jl})$.

Le relazioni precedenti sono interessanti in quanto suggeriscono una possibile formulazione generale che include i criteri precedenti come casi particolari e permette il collegamento con la modellistica multilineare e nonlineare.
Si consideri il criterio

$$\max \; \sum_j r^2(G\gamma_j : x_j) \qquad (6)$$

dove i parametri incogniti possono essere definiti a seconda degli scopi dell'analisi come

a) i vettori γ_j
b) i vettori γ_j e α_j in cui $x_j = F_j \alpha_j$ (***Forced Classification***)

c) i vettori γ_j e la matrice G (*K-Medie*)
d) i vettori γ_j e α_j e la matrice G, in cui $x_j = F_j\alpha_j$ (*CLUSTALS*)
Se a questo punto consideriamo che le matrici G e F_j possono essere definite a seconda del livello di misura del carattere, si comprende la generalita' del criterio.

4. Il metodo CLUSTALS e la PCA con variabili strumentali

Si consideri il criterio (6) con la definizione d), ossia

$$\max_{G,\gamma(J),\alpha(J)} \sum_j r^2(G\gamma_j ; F_j\alpha_j) \qquad (7)$$

in cui abbiamo posto
$\alpha'(J) = \{ \alpha'_1, \alpha'_2, \ldots\ldots, \alpha'_J \}$, $\gamma'(J) = \{ \gamma'_1, \gamma'_2, \ldots\ldots, \gamma'_J \}$
ed abbiamo introdotto la quantificazione/trasformazione dei caratteri X_j attraverso
$x_j = \mathcal{T}_j(x_j^*) = F_j\alpha_j$ con i vincoli $||x_j||=1$ e $1'x_j=0$ (7a)
Il criterio (7) estende il metodo delle K-medie all'analisi di caratteri qualitativi e misti. Si noti che i vincoli (7a) rendono equivalente il criterio basato sulla misura r^2 a quello basato sulla *covarianza*.

Chiameremo CLUSTALS il metodo di classificazione per caratteri qualitativi e misti definito dalla (7). Elemento caratterizzante della tecnica e' che CLUSTALS individua le quantificazioni/trasformazioni ottimali dei caratteri per i metodi basati sulla funzione obiettivo *max* Tr(B).

E' possibile formulare criteri che individuano differenti quantificazioni/trasformazioni dei caratteri. Si consideri ad esempio

$$\max_{G,\gamma(S),\alpha(J,S)} \sum_{s=1}^{S} \sum_{j=1}^{J} r^2(G\gamma_s ; F_j\alpha_{js}) \qquad (8)$$

in cui $z_s = G\gamma_s$ con i vincoli $1'z_s=0$, $|| z_s||=1$, $z_s' z_s = 0$.
Si noti che nella (8) abbiamo ora S quantificazioni ortogonali della variabile latente Z ed S trasformazioni per ogni variabile X_j.

La (8) e' una riformulazione della proposta GROUPALS (Van Buuren & Heiser 1989) ossia un metodo di classificazione per caratteri misti. La (8) e' anche un criterio che generalizza la Redundancy Analysis o ACP con variabili strumentali (cfr. ad es. Escoufier 1987), nel senso che le variabili strumentali sono considerate non osservate e sono ammesse trasformazioni (multiple) non-lineari dei caratteri osservati. Si noti anche che se le variabili X_j sono quantitative e se non vengono ammesse trasformazioni, otteniamo una procedura equivalente alla Principal Cluster Analysis di Ibrahim e Schecktman (1986). Se invece di $z_s = G\gamma_s$ nella (8) avessimo un qualsiasi vettore y_s non osservato e appartenente ad $\mathbb{R}^I$ allora otterremmo la funzione obiettivo della ACP non lineare (PRINCALS, cfr. Gifi 1981).

Il criterio (8) individua la stessa classificazione di CLUSTALS (ma trasformazioni differenti) purche' si ponga S=L-1 (ossia il numero di componenti deve

essere uguale al numero di cluster meno 1) e non si considerino quantificazioni multiple dei caratteri. Se la classificazione e' data (e quindi G e' nota), otteniamo la Forced Classification (Nishisato 1984). Con il metodo CLUSTALS, sotto alcune condizioni (uguale peso degli individui) e a meno di una differente normalizzazione dei parametri, possiamo ottenere il metodo proposto da Diday e coll. (1979) basato anch'esso sull'Optimal Scaling (chiamato *codage optimal adaptatif*).

Infine e' interessante notare che anche CLUSTALS puo' essere formulato come una generalizzazione della Redundancy Analysis al trattamento di caratteri misti e alla considerazione di variabili strumentali incognite. Infatti la (7) puo' essere scritta come

$$\max_{G,\alpha(J)} \sum_j R^2(G ; F_j\alpha_j) \qquad (8a)$$

Se G e $\alpha(J)$ sono note, allora $\frac{1}{J}\sum_j R^2(G ; x_j)$ e' l'indice di ridondanza introdotto da Stewart e Love (1968) come misura della capacita' predittiva di un insieme di variabili rispetto ad un altro. Tale espressione conduce in pratica ad effettuare J regressioni multiple tra ognuno dei caratteri dipendenti X_j (quantificati) e le variabili esplicative date dalle colonne della matrice indicatrice G.

Quando G e $\alpha(J)$ non sono note, come nella (8a), si ricerca il massimo di tale indice rispetto alle possibili classificazioni delle unita' statistiche e le possibili quantificazioni dei caratteri.

Un vantaggio importante dell'impostazione introdotta con CLUSTALS e' la sua estendibilita' alla considerazione di matrici di dati a tre indici.

5. Estensione di CLUSTALS al trattamento di matrici di dati a tre indici

Consideriamo una matrice di dati di tipo unita' x caratteri x occasioni. In tale situazione e' possibile effettuare una classificazione delle unita' o dei caratteri o delle occasioni (cfr. ad es. Rizzi 1989). Noi considereremo in particolare il primo obiettivo, che si presenta con maggiore frequenza nelle applicazioni.

E' opportuno distinguere innanzitutto tra tre diverse strutture della matrice dei dati in esame. In particolare considereremo il caso in cui siano state osservate:

*S*1) le stesse unita' e gli stessi caratteri
*S*2) differenti unita' ma gli stessi caratteri
*S*3) le stesse unita' ma differenti caratteri

A seconda della struttura della matrice in esame possiamo avere obiettivi diversi nell'analisi. Potremmo infatti richiedere un'unica classificazione per le diverse occasioni, o K distinte classificazioni. In particolare:

a) Se abbiamo osservato gli stessi caratteri in ciascuna occasione (casi S1 e S2), allora potremmo individuare K classificazioni delle unita' (una per ogni occasione) richiedendo che il significato ed il numero dei clusters rimanga costante da una classificazione all'altra, ma lasciando che i clusters corrispondenti possano contenere anche unita' diverse.

b) Se abbiamo osservato le stesse unita' in ciascuna occasione (casi S1 e S3) allora potremmo individuare un'unica classificazione per le K occasioni, richiedendo che ciascuna occasione abbia lo stesso peso nell'individuazione della classificazione, mentre i caratteri all'interno delle occasioni possono assumere un peso diverso.

c) Se abbiamo osservato sia gli stessi caratteri che le stesse unita' statistiche (caso S1) allora potremmo individuare un'unica classificazione delle unita' che colga gli elementi di costanza nei dati osservati al variare delle occasioni.

Consideriamo un esempio di combinazione tra situazione S2 e obiettivo a):
si e' osservato per i comuni delle 5 provincie del Lazio lo stesso insieme di caratteri relativo ad un censimento generale, l'obiettivo a) consiste nel ricercare una classificazione per ognuna delle provincie in modo tale che i clusters nelle diverse classificazioni abbiano lo stesso significato. In tal modo siamo infatti in grado di confrontare comuni o insiemi di comuni appartenenti a provincie diverse grazie allo schema comune di interpretazione delle classi.

Per semplificare la notazione successiva, ma senza perdita di generalita', supporremo nel prosieguo che ogni insieme contenga lo stesso numero J di caratteri (ma non necessariamente gli *stessi* caratteri).

L'obiettivo c) e' il piu' vincolante. Puo' essere scritto generalizzando la (7) come

CR1: $$\max_{G,\gamma(J),\alpha(J,K)} \sum_k \sum_j r^2(G\gamma_j ; F_{jk}\alpha_{jk}) \qquad (9)$$

con i vincoli di normalizzazione analoghi ai (7a).
Si noti che per ogni valore di j abbiamo un problema di ACP (non lineare), ossia ogni y_j deve essere massimamente correlata con le trasformazioni dei caratteri $x_{j1}, x_{j2}, \ldots, x_{jk}$. Questo implica che deve valere l'uguaglianza

$$y_j = \frac{1}{K}\sum_k x_{jk}$$

Si noti che la (9) implica la ricerca dei clusters che presentino la minima variabilita' interna sia all'interno di ogni occasione, sia al variare delle occasioni. In tal senso CR1 e' piu' vincolante di CLUSTALS applicato alla matrice dei dati giustapposti di dimensione Ix(JxK).

L'obiettivo a) puo' essere formulato tramite il seguente criterio

CR2: $$\max_{G(K),\gamma(J),\alpha(J,K)} \sum_k \sum_j r^2(G_k\gamma_j ; F_{jk}\alpha_{jk}) \qquad (10)$$

con gli usuali vincoli di normalizzazione (7a).
In tal caso come abbiamo gia' detto, vogliamo ottenere K differenti classificazioni in cui il significato dei clusters, rappresentato dai centroidi

$$\mu_l = (\gamma_{l1}, \gamma_{l2}, \ldots, \gamma_{lj}, \ldots, \gamma_{lJ})$$

rimanga costante.

Si noti che, considerando le proiezioni y_{jk} dei vettori $x_{jk} = F_{jk}\alpha_{jk}$ si ha

$$y_{jk} = P_{G_k} x_{jk} = G_k (G'_k G_k) G'_k x_{jk} = G_k\beta_{jk}$$

quindi la (10) si puo' anche scrivere

$$\min \sum_k \sum_j \| G_k\gamma_j - G_k\beta_{jk} \|^2$$

in cui il minimo e' preso rispetto agli stessi parametri incogniti della (10). Da tale espressione si ricava quindi che si deve avere $\gamma_j = \frac{1}{K}\sum_k \beta_{jk}$.

Infine l'obiettivo b), questo puo' essere formulato come

CR3: $$\max_{G,\gamma(K),\alpha(J,K),\beta(K)} \sum_k r^2(G\gamma_k ; \sum_j \beta_{jk}F_{jk}\alpha_{jk}) \qquad (11)$$

con i vincoli di normalizzazione analoghi ai (7a).

L'espressione (11) se G e' nota (ossia la variabile Z e' osservata) si riduce in pratica a K Analisi Discriminanti indipendenti. Cioe' alla determinazione di K combinazioni lineari, una per ognuno dei K insiemi, tali che sia massimo il quadrato della correlazione tra $y_k = G\gamma_k$ e $\nu_k = \sum_j \beta_{jk}F_{jk}\alpha_{jk}$. Viceversa se G e' incognita come stiamo supponendo, non abbiamo piu' K analisi indipendenti, ma un unico problema di classificazione. La partizione ottima in questo caso sara' quella per cui esiste per ognuno dei K insiemi, una combinazione lineare dei caratteri che permetta una buona discriminazione dei gruppi. Tale procedura porta quindi ad individuare dei coeffienti β_{jk} che indicano l'importanza di ogni carattere considerato nella determinazione della classificazione. Si ritrova quindi la logica del metodo SYNCLUS (De Sarbo e altri 1984). Tale metodo ha numerosi punti di contattto con il criterio CR3, infatti i due criteri possono essere considerati sostanzialmente equivalenti, a meno di una diversa normalizzazione dei parametri, purche' si abbiano solo caratteri quantitativi e non si consideri alcuna trasformazione.

I criteri CR1, CR2, CR3 possono essere tradotti numericamente tramite algoritmi che combinano un metodo di calcolo di tipo *minimi quadrati alternati*, con un metodo di classificazione di tipo *K-medie* (Di Ciaccio 1989).

6. Cluster Analysis di una matrice di dati a tre indici: applicazione a dati di tipo qualitativo ordinale

In questa applicazione considereremo dati riguardanti la valutazione di 15 tipi di vino nei sei anni 1964-1969. L'esempio e' tratto da Hartigan (1975, pag. 144). Per ognuno degli anni considerati e' stato assegnato un giudizio sulla qualita' del vino ottenuta. Si avevano a disposizione 5 modalita' di giudizio: Disastrosa (D), Cattiva (P), Media (A), Buona (G), Eccellente (E), anche se non tutte le modalita' compaiono in ogni anno di osservazione. Abbiamo quindi a disposizione dati di tipo qualitativo ordinale. A differenza dell'Autore, per ottenere una matrice di dati a tre indici, abbiamo considerato una suddivisione dei caratteri in due insiemi. In particolare abbiamo considerato 2 periodi, 1964-1966 e 1967-1969. Il nostro scopo e' stato quindi quello di individuare due distinte classificazioni (una per ogni periodo) tali che i clusters conservino da un periodo all'altro lo stesso *profilo* e la stessa interpretazione.

Non esiste a nostra conoscenza un'altra tecnica in grado di risolvere tale problema, mentre non e' difficile mostrare l'interesse applicativo che essa riveste.

Tav. 1
Applicazione dell'algoritmo CLUSTAL3:
dati riordinati in base al cluster di appartenenza

Primo Insieme				Sec. Insieme		
Cl.	Vini	'64	'65	'66	Cl. Vini	'67	'68	'69
1	V1	G	D	G	1 V4	G	P	G
1	V2	G	P	G	1 V12	G	P	G
1	V4	G	D	G	1 V13	G	P	G
1	V5	G	P	G	1 V14	G	P	G
1	V7	G	P	G	1 V3	G	P	G
1	V8	G	D	G				
1	V12	E	P	G				
1	V13	G	P	G				
1	V14	E	P	G				
2	V10	G	A	A	2 V15	G	G	A
2	V15	P	G	A				
3	V3	D	D	A	3 V5	A	D	G
					3 V8	A	P	G
4	V11	G	P	A	4 V1	G	P	A
					4 V2	G	P	A
5	V6	G	A	G	5 V6	G	A	G
5	V9	G	A	G	5 V7	G	A	G
					5 V9	G	A	G
					5 V10	G	A	G
					5 V11	G	G	G

Tav. 2
Applicazione dell'algoritmo CLUSTAL3: centroidi dei clusters

Centroidi medi

cluster	var.1	var.2	var.3
1..	0.0851	-0.1837	0.1424
2..	0.0851	0.3722	-0.4723
3..	-0.8122	-0.1854	-0.1495
4..	0.0851	-0.1884	-0.4723
5..	0.0851	0.3725	0.1424

Centroidi 1 insieme

cluster	var.1	var.2	var.3
1..	0.0690	-0.1560	0.1567
2..	0.0690	0.4282	-0.4282
3..	-0.9661	-0.1562	-0.4282
4..	0.0690	-0.1562	-0.4282
5..	0.0690	0.4282	0.1557

Centroidi 2 insieme

cluster	var.1	var.2	var.3
1..	0.1013	-0.2003	0.1291
2..	0.1013	0.3173	-0.5164
3..	-0.6583	-0.2116	0.1291
4..	0.1013	-0.2116	-0.5164
5..	0.1013	0.3161	0.1291

Tav. CLUSTAL3 - Quantificazioni delle modalita'

	primo insieme			secondo insieme		
modalita' /	'64	'65	'66	'67	'68	'69
D	-.96	-.15	*	*	-.21	*
P	.07	-.15	*	*	-.21	*
A	*	.42	-.42	-.66	.32	-.52
G	.07	.42	.15	.10	.32	.13
E	.07	*	*	*	*	*

Il risultato esposto nella Tav. 1 si riferisce ad una classificazione in 5 classi. I vini per ragioni di spazio sono stati indicati con le etichette V1 - V15. Si puo' notare che nel primo periodo ('64-'69) abbiamo un primo cluster ampio comprendente 9 vini caratterizzati da un buon giudizio al primo anno. un cattivo giudizio al secondo ed un buon giudizio al terzo. Il corrispondente cluster per il secondo periodo ha un *profilo* del tutto analogo, mentre vi e' una notevole diversita' in quanto a composizione e numero di vini. Dalla tav. 2 si puo' notare che in effetti i centroidi del primo cluster nei due insiemi sono quasi identici (che e' quanto si voleva ottenere). La situazione e' molto buona anche per gli altri clusters, come si puo' verificare considerando le differenze tra i centroidi nei due periodi considerati.

Si potrebbe osservare che il secondo cluster del primo periodo presenta due vini con profilo piuttosto diverso, d'altra parte se consideriamo le quantificazioni delle modalita' riportate nella tavola 3 si nota che la procedura ha accorpato (uguale quantificazione) per l'anno '64 le modalita' E, G, P (la modalita' A era assente) rendendo quindi tale variabile un elemento di discriminazione unicamente tra il cluster 3 e gli altri clusters. Analogamente anche la variabile '65 viene resa binaria accorpando le modalita' G con A e P con D, in tal modo essa discrimina tra i clusters 1,3,4 ed i clusters 2,5.

A qesto punto possiamo analizzare il risultato interpretando i clusters ottenuti in base al loro *profilo*. Il cluster che presenta il profilo migliore e' ovviamente il cluster 5, si osservi che solo i vini V6 e V9 si trovano in tale cluster in ambedue i periodi considerati. Un profilo buono ma inferiore rispetto al precedente e' dato dal cluster 2 che nel terzo anno presenta il giudizio piu' basso osservato. Una situazione equivalente viene presentata dal cluster 1, in cui il secondo anno presenta un giudizio negativo. Infine un profilo inferiore e' presentato dal cluster 4, con due anni negativi, mentre la situazione peggiore e' presente nel cluster 3. Quest'ultimo presenta i giudizi peggiori nel primo periodo mentre nel secondo periodo si puo' osservare un giudizio positivo nel terzo anno. Tale discordanza emerge facilmente anche confrontando i centroidi del cluster nei due insiemi (tav. 2), e costituisce l'unico elemento che impedisce una classificazione perfetta (loss=0) rispetto al criterio utilizzato.

BIBLIOGRAFIA

Bellacicco,A. (1989) Linear representation of clusters in multiway matrices, *in Multiway Data Analysis* (*Coppi e Bolasco eds.*), North Holland, Amsterdam.

Bouroche,J.M.,Saporta,G.,Tenenhaus,M. (1977) Some methods of qualitative data analysis, *In Barra,J.R. et al.* (*Eds.*), *Recent Developments in Statistics. Amsterdam*, the Netherlands:North-Holland, 749-755.

Carroll,J.D.,Arabie,P. (1983) INDCLUS: An individual differences generalization of the adclus model and the MAPCLUS algorithm, *Psychometrika*, 48, *n.*2, 157-169.

Coppi,R.,Bolasco,S.(eds.) (1989) Multiway data analysis, North Holland, Amsterdam.

De Sarbo,W.S.,Carroll,J.D.,Clark,L.A.,Green,P.E. (1984) Synthesized Clustering: a method for amalgamating alternative clustering bases with differential weighting of variables, *Psychometrika, v.* 49 *n.* 1, 57-78.

Di Ciaccio,A. (1989) Criteri di Massima Associazione e Optimal Scaling - un approccio unificato al trattamento dei caratteri qualitativi e misti, *tesi di dottorato, Dpt. di Statistica, Probabilita' e Stat. Applicate*, Universita' di Roma "La Sapienza".

Diday,E. e coll. (Eds.). (1979) Optimisation en Classification Automatique, *Tome II.. France: I.N.R.I.A.*

Escoufier,Y. (1987) Principal Component Analysis with respect to instrumental variables, *in Methods for Multidimensional Data Analysis*, ECAS, Napoli.

Gifi,A. (1981) Nonlinear Multivariate Analysis, Dept. Data Theory ,Leiden.

Hartigan,J.A. (1975) Clustering Algorithms, *John Wiley & Sons*, New York.

Herzel,A. (1974) Un criterio di quantificazione. Aspetti statistici, *Metron vol. XXXII, n.*1-4.

Ibrahim,A.,Shektman,Y. (1986) Principal Cluster Analysis, *in Classification As A Tool Of Research (Gaul W. al. eds.)*, North Holland, 217-223.

MacQueen,J. (1967) Some methods for classification and analysis of multivariate observations, *Proceedings of Fifth Berkeley Symposium on Math. Statistics and Probability, Vol.*1, 231-297.

Marcotorchino,f. (1986) Maximal association as a tool for classification, *In Gaul, W. and Schader,M. (Eds.), Classification as a Tool of Research (pp.* 275-288). *Amsterdam*, North Holland.

Nishisato,S. (1984) Forced classification: A simple application of a quantification method, *Psychometrika*,49, 25-36.

Rizzi,A. (1989) Clustering per le matrici a tre vie, *Dip. Statist., Prob. e Stat. Applicate, serie A - ricerche*, Univ. di Roma "La Sapienza".

Saporta,G. (1988) About Maximal Association Criteria In Linear Analysis And In Cluster Analysis, *in Classification and related methods of data analysis (Bock ed.)*.

Stewart,D.,Love,W. (1968) A general canonical correlation index, *Psychological Bulletin*, 70, 160-163.

Van Buuren,S., Heiser,W.J. (1989) Clustering N objects into K groups under optimal scaling of variables, *Psychometrika*, in press.

Van Rijckevorsel,J., De Leeuw,J. (1988) Component and Correspondence Analysis, Wiley, Chichester.

CLUSTER ANALYSIS FOR TWO-WAY AND MULTI-WAY DATA WITH MIXED MEASUREMENT LEVEL

A. Di Ciaccio
Dip. di Statistica, Prob. e Stat. Appl.
Univ. di Roma "La Sapienza"

J. De Leeuw
Dip. of Mathematics
UCLA, Los Angeles

SUMMARY

In this paper we consider the problem of clustering a set of objects when one or more data matrices (objects x variables) have been observed. In addition it is considered the case, very frequent in the reality, in which the variables are both qualitative and quantitative.

For clustering two-way multivariate data, we propose the algorithm CLUSTALS, which is an extension of the K-means method in order to consider both qualitative and quantitative variables. This extension is obtained using Optimal Scaling with an Alternate Least Squares approach. Other proposals for clustering with Optimal Scaling of the variables (Diday and coll. 1979, van Bureen and Heiser 1989) are considered in the paper.

The particular approach that we introduced, based on "maximal association criteria", allows to extend quite easily the method to the analysis of three-way data matrices (objects x variables x occasions). We distinguish the following cases:
1) we observed the same objects and the same variables on each occasion
2) we observed the same objects but different variables on each occasion
3) we observed the same variables but different objects on each occasion
In the first case there is only one set of objects and only one set of variables, in the second case there is again one set of objects but several sets of variables, in the third case we have one set of variables but several sets of objects.

Three clustering methods (CLUSTAL2, CLUSTAL3, CLUSTAL4), are proposed which allow an interesting analysis of the different cases. The relation of the new methods with SYNCLUS (De Sarbo et al. 1984) is also considered. Finally we show an application of the method CLUSTAL3 to the analysis a set of real data already considered in the literature.

Keywords : Cluster Analysis, Multiway data matrices, Optimal Scaling, maximal association criteria