ANALISI DEI CLUSTER

In questo documento presentiamo alcune opzioni analitiche della procedura di analisi de cluster di SPSS che non sono state incluse nel testo pubblicato. Si tratta di opzioni che, pur non essendo utilizzate correntemente per effettuare l'analisi fattoriale, consentono di completare lo spettro delle possibilità di analisi offerte da SPSS. In particolare, il documento completa la descrizione della procedura di analisi dei cluster con in seguenti paragrafi:

1. Analisi dei cluster con il linguaggio SINTASSI
2. Criteri di fusione
3. Indici di distanza e di similarità/dissimilarità

1. Analisi dei cluster con il linguaggio SINTASSI

1.1. Cluster analysis gerarchica

Di seguito presentiamo i programmi SINTASSI per le analisi descritte nel testo.

```
**** Cluster analysis gerarchica.

CLUSTER  n e o a c
  /MEASURE= SEUCLID
  /METHOD WARD
  /PRINT SCHEDULE CLUSTER(2,5) DISTANCE
  /PLOT DENDROGRAM
  /SAVE CLUSTER(2,5) .
```

Illustriamo brevemente i comandi specificati nella procedura CLUSTER.

Dopo il comando CLUSTER che richiama la procedura di cluster analysis gerarchica vanno inserite le variabili che vengono utilizzate per definire la soluzione di analisi dei cluster.

Il comando `/MEASURE= SEUCLID` serve per specificare l'indice che viene utilizzato per misurare la prossimità dei casi. L'opzione riportata è la distanza euclidea al quadrato, che rappresenta l'opzione di default: nel paragrafo 3 di questo volume presentiamo molte delle altre opzioni utilizzabili in SPSS a seconda del livello di misurazione delle variabili.

Il comando `/METHOD WARD` serve per specificare il metodo di classificazione che viene utilizzato, ovvero il "criterio di fusione" che il programma utilizza per calcolare la distanza tra i cluster ad ogni livello della gerarchia di partizioni che viene ad essere definita. L'opzione riportata è relativa al metodo di Ward, che in questo caso non costituisce il valore di default: nel paragrafo 2 di questo volume presentiamo le altre opzioni utilizzabili in SPSS per specificare il criterio di fusione a seconda dell'interessa del ricercatore, e del livello di misurazione delle variabili.

Il comando `/PRINT SCHEDULE CLUSTER(2,5) DISTANCE` serve per ottenere in output indici e informazioni che consentono di interpretare al meglio la soluzione. In particolare "SCHEDULE" (che rappresenta l'opzione di default) permette di ottenere in output il programma di agglomerazione (vedi il testo, pagg. 217-219). Le altre due opzioni invece consentono di ottenere in output l'appartenenza ai gruppi di tutti i casi nelle soluzioni a 2, 3 , 4 e 5 gruppi, e la matrice delle distanze tra i casi. Specificando "NONE" invece nessuna di queste informazioni verrà inserita nell'output.

Il comando `/PLOT DENDROGRAM` serve per ottenere in output la rappresentazione grafica del dendrogramma. Specificando invece VICICLE(min,max,inc) e HICICLE(min,max,inc) verranno prodotti i grafici a stalattite rispettivamente verticale e orizzontale, per tutti i cluster se non vengono riportati i valori tra parentesi, oppure per un intervallo di cluster che va da *min* a *max* con un incremento pari al valore *inc*. Anche in questo caso, specificando "NONE" invece nessuna di queste informazioni verrà inserita nell'output.

Il comando `/SAVE CLUSTER(2,5)` serve per salvare nel file attivo l'appartenenza del soggetto ai gruppi specificati tra parentesi (nel nostro caso verranno salvate le appartenenze per le soluzioni a 2, 3, 4 e 5 gruppi).

Completano il quadro della SINTASSI i comandi:

- /MISSING che gestisce il trattamento dei casi con valori mancanti (le opzioni sono EXCLUDE, che è il valore di default ed esclude tutti i casi con almeno un valore mancante, e INCLUDE che invece include i casi con valori mancanti);

- /MATRIX che consente di leggere un file in formato matriciale (IN) o di salvare un file in formato di matrice (OUT).

1.2. Cluster analysis non-gerarchica

Di seguito presentiamo i programmi SINTASSI per le analisi descritte nel testo.

```
**** Cluster analysis non-gerarchica.

QUICK CLUSTER
  n e o a c
  /MISSING=LISTWISE
  /CRITERIA= CLUSTER(4) MXITER(10) CONVERGE(0)
  /METHOD=KMEANS(UPDATE )
  /SAVE CLUSTER DISTANCE
  /PRINT INITIAL ANOVA CLUSTER DISTAN.
```

Illustriamo brevemente i comandi specificati nella procedura CLUSTER.

Dopo il comando QUICK CLUSTER che richiama la omonima procedura di cluster analysis non-gerarchica vanno inserite le variabili che vengono utilizzate per definire la soluzione di analisi dei cluster.

Il comando `/MISSING=LISTWISE` gestisce il trattamento dei casi con valori mancanti. Le opzioni sono le note alternative LISTWISE e PAIRWISE descritte a pag. 230 del testo. Completa il quadro l'opzione INCLUDE che permette di includere in analisi tutti i casi con valori mancanti.

Il comando `/CRITERIA= CLUSTER(4) MXITER(10) CONVERGE(0)` prevede una serie di opzioni che consentono di controllare il processo di analisi. Nel nostro esempio `CLUSTER(4)` specifica che verrà creata una partizione di 4 gruppi, `MXITER(10)` stabilisce il numero massimo di iterazioni previste per raggiungere la convergenza, `CONVERGE(0)` serve per determinare il cambiamento minimo nei centroidi dei cluster affinché il processo di convergenza possa considerarsi concluso. Le opzioni sono le note alternative LISTWISE e PAIRWISE descritte a pag.

230 del testo. Completa il quadro l'opzione INCLUDE che permette di includere in analisi tutti i casi con valori mancanti. Un'ulteriore importante opzione, che non è riportata in queste linee di sintassi, è "NONINITIAL": se tale opzione viene specificata, il programma prenderà come centroidi iniziali per generare la partizione i primi n casi senza valori mancanti nel data file (dove n è il numero di gruppi in cui verranno suddivisi i soggetti. L'opzione di default invece prevede che il programma esamini i dati una prima volta per scegliere come centroidi iniziali dei gruppi gli n soggetti che sono più distanti.

Il comando `/METHOD=KMEANS(UPDATE )` consente di governare le operazione relative al ri-calcolo dei centroidi dei gruppi dopo l'assegnazione dei soggetti ai gruppi. Se si lascia l'opzione di default "NOUPDATE" i centroidi sono ri-calcolati dopo che tutti i casi sono stati assegnati, alla fine dell'iterazione. Se invece viene scelta l'opzione "UPDATE" (come nel nostro caso) i centroidi sono ri-calcolati dopo che ogni caso viene assegnato ad un gruppo: si tratta dunque del metodo delle cosiddette "medie mobili", che prevede un aggiornamento dinamico dei centri dei gruppi. Con l'opzione "CLASSIFY" infine i casi vengono assegnati ai gruppi più vicini, non si effettua nessuna iterazione e i centroidi dei gruppi vengono ri-calcolati quando tutti i soggetti sono stati classificati.

Il comando `/SAVE CLUSTER DISTANCE` prevede le due opzioni che consentono di salvare nel file attivo rispettivamente il numero del cluster in cui il soggetto è stato classificato e la distanza dal centroide del gruppo di appartenenza.

Il comando `/PRINT INITIAL ANOVA CLUSTER DISTAN` prevede alcune opzioni che consentono di ottenere in output diverse informazioni utili per interpretare la soluzione. In particolare, l'opzione `INITIAL` consente di avere i centroidi iniziali dei cluster, l'opzione `ANOVA` consente di esaminare la significatività statistica della differenza tra le medie delle variabili attraverso i gruppi, l'opzione `CLUSTER` consente di avere in output una tabella con specificati per ogni casi il gruppo cui appartiene e la distanza dal centroide, l'opzione `DISTAN` consente di avere in output una tabella con le distanze tra i centroidi dei cluster. Completa questo comando l'opzione "ID(nome della variabile)" che consente di utilizzare il valore della variabile specificata

(solitamente si tratta di una variabile alfanumerica con un etichetta che identifica il caso) come identificatore supplementare oltre al numero del caso assegnato di default nel file dati.

Come per comandi analoghi nella procedura gerarchica, anche in questa procedura, specificando "NONE" nessuna di queste informazioni verrà inserita nell'output.

Ci sono infine tre ulteriori comandi che possono essere specificati nella procedura QUICK CLUSTER. Il comando "/INITIAL()" serve a specificare i valori dei centroidi iniziali; in particolare vanno inserire nella parentesi le medie di ciascuna variabile dal primo gruppo all'ultimo gruppo (nel nostro esempio dovremmo fornire 5*4= 20 differenti valori, dove 5 è il numero delle variabili, e 4 il numero dei gruppi). I valori delle medie possono essere letti anche da un file esterno utilizzando il comando "/FILE=nomefile": ovviamente il file deve essere in formato SPSS. Nell'esempio discusso nel testo alle pagine 238-239 abbiamo utilizzato un file esterno per i centroidi iniziali: questo file è riportato nella figura 5.14 ed è scaricabile da questo sito (il nome del file è CLUSTER_3_CENTRI.SAV). Infine il comando "/OUTFILE" consente di salvare i valori finali dei centroidi in un file esterno in formato SPSS, che appare come quello citato poc'anzi.

2. Criteri di fusione nella procedura di cluster analysis di gerarchica SPSS

Come è noto nella cluster analysis gerarchica, ad ogni passo si associano in un nuovo cluster gli oggetti (o cluster) più vicini. Se questo vale per tutti i metodi, cambia però il criterio rispetto al quale vengono calcolate le distanze tra i gruppi. Ogni metodo diverso infatti prevede un criterio diverso. Presentiamo di seguito i principali criteri di fusione utilizzati nella procedura della cluster analysis gerarchica di SPSS.

Metodo del legame singolo: la distanza tra due cluster è uguale alla distanza dei due individui nei due differenti cluster che risultano più vicini. Questo metodo viene definito tramite il comando "`/METHOD SINGLE`".

Metodo del legame completo: la distanza tra due cluster è uguale alla distanza dei due individui nei due differenti cluster che risultano più lontani. Questo metodo viene definito tramite il comando "`/METHOD COMPLETE`".

Metodo del legame medio: la distanza fra due cluster diversi corrisponde alla media aritmetica delle distanze definite su tutte le coppie di oggetti nei due cluster. In SPSS esistono due varianti di questo metodo. Nel metodo del legame medio *fra i gruppi* la distanza tra due gruppi è uguale alla media delle distanze tra ogni coppia di elementi appartenenti a gruppi differenti. Questo metodo viene definito tramite il comando "`/METHOD BAVERAGE`", ed è il metodo di default di SPSS. Nel metodo del legame medio *entro i gruppi* la distanza tra due gruppi è uguale alla media delle distanze tra *ogni* coppia di elementi, incluse le coppie di elementi che appartengono allo stesso gruppo. Questo metodo viene definito tramite il comando "`/METHOD WAVERAGE`".

Metodo del Centroide: la distanza fra due cluster è definita dalla distanza fra i rispettivi centroidi. Questo metodo viene definito tramite il comando "`/METHOD CENTROID`".

Metodo della Mediana: la distanza fra due cluster è definita dalla distanza fra le rispettive mediane. Questo metodo viene definito tramite il comando "`/METHOD MEDIAN`".

Metodo di Ward: in questo metodo il procedimento di associazione fra due cluster diversi è basato sulla minimizzazione della devianza entro i gruppi (ovvero, la massimizzazione delle distanze tra i centroidi dei gruppi). La devianza entro i gruppi è uguale a 0 quando tutti i casi sono separati ed è

massima quando essi appartengono tutti a un gruppo unico. La coppia di cluster da aggregare in un certo passo è quella che determina un incremento minimo della varianza interna ai cluster. La distanza euclidea tra due oggetti viene calcolata con una funzione che considera sia la numerosità dei gruppi sia la distanza euclidea al quadrato tra i centroidi dei due gruppi (vedi Barbaranelli, 2003). Questo metodo viene definito tramite il comando "`/METHOD WARD`".

3. Indici di distanza e di similarità/dissimilarità nella procedura di cluster analysis gerarchica di SPSS

Anche nel caso degli indici di distanza la procedura per la cluster analysis gerarchica di SPSS prevede una serie di opzioni differenti, a seconda del livello di misurazione delle variabili prese in esame. Presentiamo di seguito i principali (si veda il manuale del programma e le funzioni di aiuto del programma per un quadro completo degli indici).

Variabili a intervalli equivalenti:

Distanza Euclidea: viene richiesta con il comando "`/MEASURE= EUCLID`"

Distanza Euclidea al quadrato: viene richiesta con il comando "`/MEASURE= SEUCLID`" (è l'opzione di default)

Distanza di Minkowski: viene richiesta con il comando "`/MEASURE= MINKOWSKI(p)`" dove p rappresenta l'esponente utilizzato per elevare alla p-esima potenza la differenza tra i punteggi dei soggetti, sulla la cui somma viene calcolata la radice p-esima.

Variabili che rappresentano Frequenze:

Chi-quadrato: viene richiesta con il comando "`/MEASURE= CHISQ`"

Phi-quadrato: viene richiesta con il comando "`/MEASURE= PH2`"

Variabili dicotomiche:

Indice di somiglianza è l'indice di Russell e Rao: "`/MEASURE= RR`"

Coefficiente di concordanza semplice di Sokal e Michener: "`/MEASURE= SM`"

Indice di distanza euclidea "`/MEASURE= BEUCLID`"

Distanza euclidea al quadrato "`/MEASURE= BSEUCLID`"

Riferimenti bibliografici

Barbaranelli, C. (2003). *Analisi dei dati*. Milano: LED