APPENDICE TECNICO-STATISTICA AL CAPITOLO 17

di Francesco Giovanni Truglia

La scelta di presentare un capitolo dedicato alle tecniche di analisi multidimensionale riducendo al minimo la formalizzazione statistico-matematica è giustificata dalla sua impostazione manualistica e dal suo carattere introduttivo alla pratica di elaborazione statistica dei dati. Tuttavia, per non rischiare di promuovere un uso automatizzato e poco consapevole di tale procedura da parte del neofita, si ritiene opportuno corredare il capitolo di un'appendice tecnico-statistica che ne approfondisca alcuni passaggi procedurali, ricorrendo, questa volta, al linguaggio più formalizzato dell'algebra delle matrici.

Analisi in componenti principali (ACP): alcuni aspetti statistico-matematici

Per introdurre gli aspetti teorici dell'ACP è opportuno prendere le mosse dal concetto di combinazione lineare. Le componenti principali, infatti, non sono altro che combinazioni lineari di variabili, ciascuna delle quali è moltiplicata per un coefficiente. Se si opera nello spazio delle variabili, da una matrice $\mathbf{X}_{np}$ possono essere estratte p componenti principali (in pratica tante componenti quante sono le variabili).

Dato un set di variabili x_j la combinazione lineare y_k, che rappresenta la generica componente principale, può essere formalizzata nel seguente modo:

$$y_k = a_{k1}x_1 + \cdots + a_{kj}x_j + \cdots + a_{kp}x_p$$

Dove i paramenti a_{kj} sono coefficienti che registrano il "peso" di ogni x_j.

La varianza di y_k è:

$$var(y_k) = a_{k1}^2 var(x_1) + \cdots + a_{kp}^2 var(x_p) + 2cov(x_1x_2)a_{k1}a_{k2} + \cdots + 2cov(x_{p-1}x_p)a_{kp-1}a_k$$

Questa espressione mette in evidenza due aspetti. Il primo riguarda i coefficienti a_{kj} che è un fattore di scala, il cui valore influenza la varianza di y_k. Il secondo è che essa non è la somma delle sole varianze, ma anche delle covarianze tra variabili, in cui il valore è 0 se le variabili sono indipendenti.

Le statistiche che compaiono nella parte destra dell'equazione, ad eccezione dei coefficienti a_{kj}, sono gli elementi della matrice di varianza-covarianza **S** di dimensione pp, per cui, in forma di algebra matriciale, la precedente espressione può essere tradotta nel seguente modo:

$$\mathrm{var}(\mathrm{y_k}) = \mathrm{var}(\mathrm{X}a_\mathrm{k}) = a'_\mathrm{k}\mathrm{S}a_\mathrm{k}$$

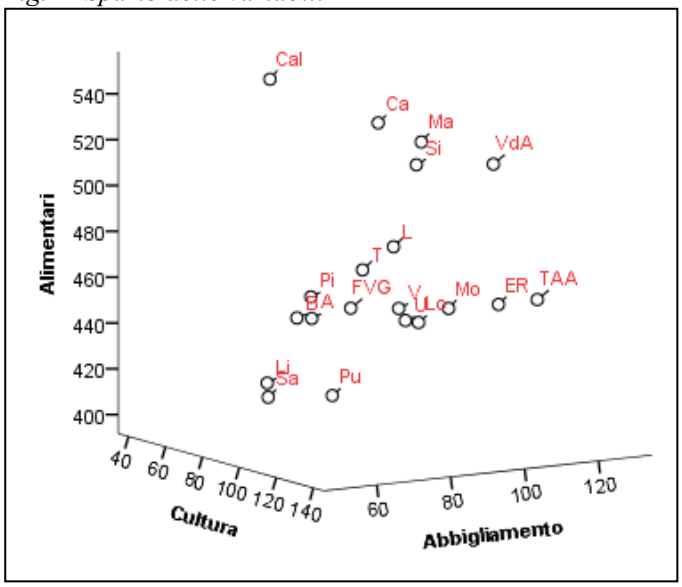
Nel linguaggio dell'ACP i coefficienti a_{kj} sono le *saturazioni* o *pesi fattoriali* (*factor loading*) e sono calcolati in modo tale da massimizzare la varianza riprodotta da ciascuna componente. Sotto l'aspetto geometrico, questi coefficienti forniscono indicazioni relative alla direzione e al posizionamento degli assi fattoriali. Essi sono calcolati imponendo tre condizioni:

1. la varianza riprodotta da ciascuna componente non è influenzata dal loro valore per cui la somma dei quadrati delle saturazioni è pari a 1 (in termini matricali: $\boldsymbol{a_k'a_k}=1$)[1];
2. le componenti devono essere tra di loro indipendenti (ortogonali) → $cov(y_k, y_{k+1}) = 0$;
3. l'ordine delle componenti replica la loro capacità informativa → $var(y_1) > var(y_2) > \ldots var(y_p)$;

[1] Con $\boldsymbol{a}_k$ si indica il vettore di lunghezza k che contiene i coefficienti a_{kj} , il simbolo ' indica la "trasposizione" che può riguardare sia le matrici che un vettore.

4. Per soddisfare questi tre vincoli è necessario utilizzare specifici metodi matematici, per i quali si rimanda ai testi più specialistici (Zani, Cerioli,2007). Tuttavia, ai fini di questo scritto è importante sottolineare come l'ACP si risolva nel calcolo di questi coefficienti. Da essi infatti dipende sia la bontà e sia l'interpretazione sostantiva della soluzione fattoriale. Per gli scopi di questo scritto e per il pubblico a cui è indirizzato, si è evitato di ricorrere a un'eccessiva formalizzazione matematica e si è scelto di *tradurre* sul piano operativo gli aspetti tecnico-statistici. A tale scopo, a titolo di esempio, sono utilizzati tre indicatori della spesa media mensile delle famiglie a livello regionale destinata ai consumi: alimentari e alle bevande (Alimentari); all'abbigliamento e alle scarpe (Abbigliamento); alle attività ricreative e culturali (Cultura) (fonte Istat, 2019). I dati sono organizzati in una matrice casi per variabili $X_{3,20}$ dove i pedici si riferiscono rispettivamente al numero di variabili e al numero di regioni. Sotto l'aspetto geometrico le regioni – in questo caso identificate con le sigle - possono essere proiettati in uno spazio a tre dimensioni (R^3) (Fig.1).

Fig.1 - Spazio delle variabili



I parametri a_{ki}, ottenuti utilizzando la libreria *FactoMineR* del software *R*[2], per le tre componenti principali sono riportate di seguito in forma tabellare (Tab. 1).

Tab. 1 - Saturazioni o *pesi fattoriali*

	y1	y2	y3
Alimentari	0,716	0,811	0,422
Abbigliamento	0,476	-0,031	0,869
Cultura	-0,509	0,584	0,258

I coefficienti a_{kj}, come i coefficienti di correlazione, variano tra -1 e 1 e registrano l'intensità del legame tra ciascuna variabile e la componente principale.

La variabile che maggiormente incide sulla formazione della prima e della seconda componente principale è quella delle spese alimentari, mentre l'abbigliamento incide in modo più marcato sulla terza componente principale.

Il quadrato delle *saturazioni* esprime la proporzione di varianza che le variabili e le componenti hanno in comune. La somma dei quadrati dei coefficienti a_{kj} riferiti a ciascuna variabile corrisponde alla quota di informazione riprodotta da tutte le componenti principali ed è nota come *comunalità* (h_j):

$$\sum_{i}^{p} a_{ij}^2 = h_j \; per \; j = Alimentari \rightarrow 0{,}716^2 + 0{,}811^2 + 0{,}422^2 = 1{,}348$$

Al contrario, se questi coefficienti vengono sommati per colonna si ottengono gli autovalori (λ_k) di ciascuna componente principale che registra la quota di varianza (informazione) totale riprodotta da ciascuna componente e quindi l'importanza di ciascuna di esse nel processo di riduzione della complessità.

[2] Tutte le elaborazioni presentate in questo scritto sono eseguite con questo pacchetto di R:

$$\sum_{k}^{p} a_{kj}^2 = \lambda_k \; per \; k = 1 \rightarrow 0{,}716^2 + 0{,}476^2 + (-0{,}509)^2 = 1{,}000$$

Gli altri due autovalori associati alla seconda e terza componente principale sono rispettivamente 0,893 e 0,667. Come già detto, l'ordine delle componenti rispecchia l'importanza di ciascuna componente in funzione della varianza riprodotta. Per cui la prima componente principale è la più importante in quanto riproduce circa il 46% della varianza totale (1,348/1,348+0,893+0,667), mentre le altre due riproducono il 31% e il 23%.

Le combinazioni lineari che esprimo le tre componenti principali possono essere riportate nel seguente modo:

y_1 =0,716Alimentari+0,476Abbigliamento -0,509Cultura
y_2 = 0,811Alimentari-0,031Abbigliamento+0,584Cultura
y_3=0,422Alimentari+0,869Abbigliamento+0,258Cultura

Tab. 2- Correlazione tra componenti principali e variabili

	y_1	y_2	y_3
Alimentari	0,717	0,477	-0,509
Abbigliamento	0,811	0,031	0,585
Cultura	- 0,422	0,869	0,258

Fino ad ora le analisi hanno riguardato le variabili, ma è chiaro che ad essere modificata è anche la configurazione della nuvola dei punti, i quali possono essere proiettati in un nuovo spazio, i cui assi sono appunto le componenti principali. La posizione di ogni punto nello spazio fattoriale è determinato da un nuovo *set* di coordinate o *punteggi fattoriali* (*scores*) che per l'*i*-esimo caso e la *k*-esima componente principale è:

$$y_{ik} = \sum_{k}^{p} a_{ik} z_{ik}$$

Dove il primo pedice indica l'*i*-esimo caso e il secondo la *k*-esima componente principale. Questi punteggi hanno tutti media pari a 0, ma la loro varianza è diversa a seconda della componente principale alla quale fanno riferimento e che riproduce, come più volte detto, una quota diversa di varianza.

Per questo motivo, quindi, gli *scores* delle diverse componenti principali non possono essere direttamente confrontati. Una soluzione a questo problema consiste nel normalizzare gli *scores* dividendoli per la radice quadrata dei rispettivi autovalori:

$$y_{ik} = \frac{\sum_{k}^{p} a_{ik} z_{ki}}{\sqrt{\lambda_k}}$$

Nella figura 2 le regioni sono proiettate sul piano fattoriale formato dal primo e secondo asse che rappresentano rispettivamente la spesa per consumi "primari bassi *vs* alti (Y_1)" e quella dei consumi "secondari alti *vs* bassi (Y_2)" delle famiglie.

Come per le variabili, anche per gli individui è possibile valutare la *qualità* della loro rappresentazione sullo spazio fattoriale.

A tale scopo le statistiche a cui guardare sono i *contributi relativi* (quadrati dei coseni tra asse e individuo) che hanno valori comprese tra 0 (l'asse non rappresenta l'individuo) e 1 (rappresentazione ottima).

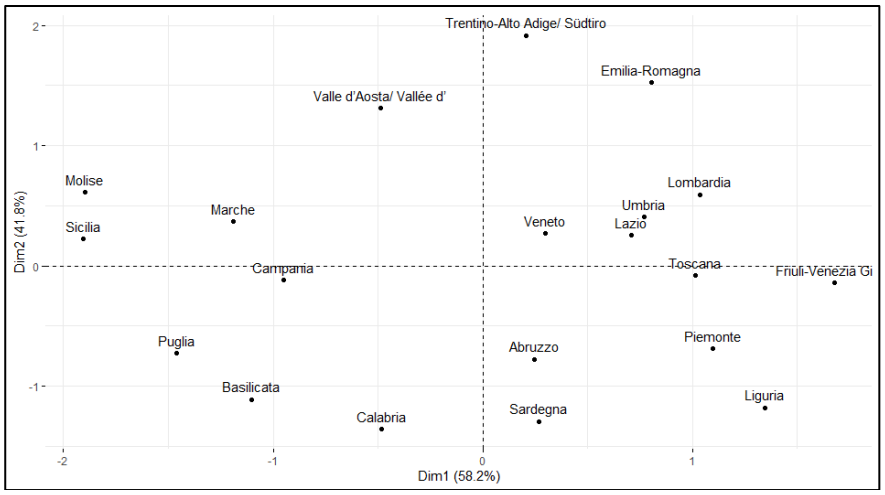


Fig. 2 – Nuvola dei punti-regione sul piano fattoriale

Nell'esempio riportato sopra, da una matrice **X** composta da tre variabili sono state estratte tre componenti che possono essere organizzate in una matrice **Y**. In questo caso quindi non c'è alcuna riduzione fattoriale, per cui l'informazione globale è uguale nelle due matrici, ma distribuita in modo diverso tra i vettori colonna della **Y** (come si è visto sopra le tre componenti riproducono il 100% dell'informazione originaria), per cui:

$$\mathbf{S_y = AS_xA'}$$

Dove $\mathbf{S_y}$ e $\mathbf{S_x}$ sono le matrici di varianza-covarianza di **Y** e **X** entrambe di dimensione *np* ed **A** è la matrice degli autovettori di dimensione *pp*.

La matrice $\mathbf{S_y}$ ha quindi sulla diagonale le varianze riprodotte da ciascuna componente principale - che non sono altro che i rispettivi autovalori - e fuori dalla diagonale le covarianze tra le componenti che, proprio perché ortogonali, sono tutte pari a 0.

Analisi delle corrispondenze multiple (ACM): Matrice logico-disgiuntiva, matrice di Burt e spazio fattoriale delle unità

La ricodifica delle variabili-originali in *variabili-indicatrici* è la condizione per poter trattare con tecniche fattoriali variabili categoriali. Si tratta, quindi, di trasformare la matrice $\mathbf{X}_{np}$ (casi per variabili) in una matrice *logico-disgiuntiva-completa* $\mathbf{Z}_{nq}$ (casi per modalità) dove ogni elemento z_{ij} è pari a 1 o 0 a seconda se l'*i*-esimo individuo possiede o meno la *j*-esima modalità di una certa variabile.

Questo tipo di ricodifica è detta:

- *disgiuntiva* perché le diverse modalità della stessa variabile si escludono a vicenda;
- *completa* in quanto ad ogni individuo è sicuramente attribuito un valore 1 o 0 per ciascuna modalità di ciascuna variabile.

Anche se alla ricodifica delle variabili in forma *disgiuntiva completa* provvedono in automatico i software dedicati a questo tipo di analisi, sembra opportuno soffermarsi sul passaggio dalla matrice $\mathbf{X}_{np}$ a quella $\mathbf{Z}_{nq}$ (Fig.1).

A tale scopo sono utilizzate alcune delle informazioni sugli incidenti stradali, contenute nei verbali dalla Polizia di Roma. I dati riguardano 20 incidenti stradali e descrivono:

1) la località dove l'incidente è avvenuto (3 modalità);
2) la fascia oraria nella quale è accaduto l'evento (4 modalità);
3) le conseguenze dell'incidente (2 modalità).

Per meglio chiarire alcuni *meccanismi* di calcolo a fianco e sotto la matrice $\mathbf{Z}_{nq}$ (da ora in avanti le matrici sono scritte senza pedici) sono aggiunti i totali dei profili-riga $z_{i.}$ e dei profili-colonna $z_{.j}$. Come si può constatare la somma di ciascun profilo-riga è costante ed è pari al numero delle variabili (*p*). Mentre i totali dei profili-colonna possono variare tra 1 e *n*. Infine il totale degli "1" è pari a *np*.

Fig. 3 - Costruzione delle matrici logico-disgiuntiva-completa

X20,3

TipoLuogo	FasciaOraria	Feriti
Altro	09,00-12,59	no
Altro	09,00-12,59	no
Altro	17,00-20,59	si
Incrocio	01,00-04,59	si
Incrocio	01,00-04,59	si
Incrocio	13,00-16,59	no
Incrocio	13,00-16,59	si
Incrocio	17,00-20,59	no
Rettilineo	01,00-04,59	no
Rettilineo	09,00-12,59	si
Rettilineo	09,00-12,59	no
Rettilineo	09,00-12,59	no
Rettilineo	09,00-12,59	no
Rettilineo	13,00-16,59	si
Rettilineo	13,00-16,59	no
Rettilineo	13,00-16,59	no
Rettilineo	17,00-20,59	si
Rettilineo	17,00-20,59	no
Rettilineo	17,00-20,59	no
Rettilineo	17,00-20,59	no

Z20,9

Rettilineo	Incrocio	Altro	04,59	12,59	16,59	20,59	Si	No	zi.
0	0	1	0	1	0	0	0	1	3
0	0	1	0	1	0	0	0	1	3
0	0	1	0	0	0	1	1	0	3
0	1	0	1	0	0	0	1	0	3
0	1	0	1	0	0	0	1	0	3
0	1	0	0	0	1	0	0	1	3
0	1	0	0	0	1	0	1	0	3
0	1	0	0	0	0	1	0	1	3
1	0	0	1	0	0	0	1	0	3
1	0	0	0	1	0	0	0	1	3
1	0	0	0	1	0	0	1	0	3
1	0	0	0	1	0	0	0	1	3
1	0	0	0	0	1	0	0	1	3
1	0	0	0	0	1	0	0	1	3
1	0	0	0	0	1	0	1	0	3
1	0	0	0	0	0	1	0	1	3
1	0	0	0	0	0	1	1	0	3
1	0	0	0	0	0	1	0	1	3
1	0	0	0	0	0	1	0	1	3
1	0	0	0	0	0	1	0	1	3
z.j	12	5	3	3	5	5	7	8	12	60

Un secondo passaggio riguarda la trasformazione dei dati in punteggi relativi. Nelle consuete tabelle di contingenza questa operazione è il risultato del rapporto tra ciascuna frequenza vincolata o di cella (n_{ij}) per il rispettivo valore di riga ($n_{i.}$) o colonna ($n_{.j}$). Nella matrice **Z**, che è una tabella di variabili-indicatrici, questa operazione non è consentita. In questo caso si ricorre ad una procedura di "assegnazione" che sostanzialmente consiste nel dividere ciascun elemento z_{ij} per i rispettivi marginali di riga e colonna ($z_{i.}$ e $z_{.j}$) e ottenere in tal modo gli elementi r_{ij} e colonna c_{ij} dei profili riga e colonna:

$$r_{ij} = \frac{z_{ij}}{z_{i.}} = \frac{z_{ij}}{p} \; ; c_{ij} = \frac{z_{ij}}{z_{.j}}$$

mentre i profili marginali di riga e colonna, cioè le *masse*, sono dati dai seguenti rapporti:

$$m_{i.} = \frac{z_{i.}}{np} = \frac{p}{np} = \frac{1}{n} \; ; m_{.j} = \frac{z_{.j}}{np}$$

La *massa*, quindi, non è altro che il *peso* utilizzato per la ponderazione ed è costante per i profili-riga (tutti gli individui hanno lo stesso peso). Mentre per dei profili-colonna, la *massa* varia tra $1/q$ e 1 in quanto le modalità forniscono un contributo diverso. Infine, la somma delle *masse* sia di riga che colonna è pari a 1.

Un secondo modo per analizzare i profili è operare sulla matrice di Burt **B (**di dimensione qq**)** che nell'ACM ha lo stesso ruolo della matrice di varianza-covarianza nell'ACP.

La matrice **B** si ottiene da quella logico-disgiuntiva (**B=Z'Z**) (Fig.2) ed è un *ipercubo* nel quale le modalità di una variabile sono incrociate con quelle di tutte le altre variabili e può essere intesa come un *contenitore* di tabelle di contingenza (il totale delle tabelle contenute in **B** è pari a p^2) che sono di due tipi.

La prima è formata dalle tabelle nelle quali si incrociano le modalità della stessa variabile e sono quindi quadrate con valori non nulli solo sulla diagonale; valori che non sono altro che frequenze marginali di riga e colonna. All'interno della Matrice **B** queste tabelle si localizzano sulla diagonale. Il secondo tipo di tabelle è dato dall'incrocio tra ciascuna variabile e tutte le altre. La traccia della matrice **B** è pari a np.

X20,3

TipoLuogo	Fascia Oraria	Feriti
Altro	09,00-12	no
Altro	09,00-12	no
Altro	17,00-20	si
Incrocio	01,00-04	si
Incrocio	01,00-04	si
Incrocio	13,00-16	no
Incrocio	13,00-16	si
Incrocio	17,00-20	no
Rettilineo	01,00-04	no
Rettilineo	09,00-12	si
Rettilineo	09,00-12	no
Rettilineo	09,00-12	no
Rettilineo	09,00-12	no
Rettilineo	13,00-16	si
Rettilineo	13,00-16	no
Rettilineo	13,00-16	no
Rettilineo	17,00-20	si
Rettilineo	17,00-20	no
Rettilineo	17,00-20	no
Rettilineo	17,00-20	no

Z20,9

Rettilineo	Incrocio	Altro	01,00-04,59	09,00-12,59	13,00-16,59	17,00-20,59	Si	No
0	0	1	0	1	0	0	0	1
0	0	1	0	1	0	0	0	1
0	0	1	0	0	0	1	1	0
0	1	0	1	0	0	0	1	0
0	1	0	1	0	0	0	1	0
0	1	0	0	0	1	0	0	1
0	1	0	0	0	1	0	1	0
0	1	0	0	0	0	1	0	1
1	0	0	1	0	0	0	1	0
1	0	0	0	1	0	0	0	1
1	0	0	0	1	0	0	1	0
1	0	0	0	1	0	0	0	1
1	0	0	0	0	1	0	0	1
1	0	0	0	0	1	0	0	1
1	0	0	0	0	1	0	1	0
1	0	0	0	0	0	1	0	1
1	0	0	0	0	0	1	1	0
1	0	0	0	0	0	1	0	1
1	0	0	0	0	0	1	0	1
1	0	0	0	0	0	1	0	1

B9,9

	Altro	Incro	Retti	04,59	12,59	16,59	20,59	No	Si
Altro	3				2		1	2	1
Incrocio		5		2		2	1	2	3
Rettilineo			12	1	4	3	4	9	3
01,00-04,59		2	1	3				1	2
09,00-12,59	2		4		6			5	1
13,00-16,59		2	3			5		3	2
17,00-20,59	1	1	4				6	4	2
No	2	2	9	1	5	3	4	13	
Si	1	3	3	2	1	2	2		7

Fig. 4 – Costruzione della matrice di Burt

La Matrice **X**, **Z** e **B** *contengono*, quindi, la stessa informazione**,** ma organizzata in modo diverso. Per cui l'analisi può essere effettuata a partire dalla matrice **Z** o dalla matrice **B**. Nel primo caso due modalità (colonne) sono simili se la sequenza di 0 e 1 è uguale in tutti gli individui. Nel secondo caso la somiglianza/dissomiglianza tra profili si basa sul confronto tra ogni modalità e tutte le altre. In altri termini, come per le tabelle di contingenza, due profili sono simili se, *tendenzialmente*, le modalità di una variabile si associano sempre con le stesse modalità delle altre variabili.

Sia che si opera sulla matrice **Z** o sulla **B** il confronto tra i profili- riga o profili-colonna per valutare il grado di somiglianza/diversità si utilizzata la metrica del χ^2.

Infine, è bene evidenziare due aspetti. Il primo è che nella matrice **B** non compaiono i singoli individui per cui, a differenza della matrice **Z**, non è possibile identificare i singoli profili e quindi la corrispondenza tra individui e modalità. Il secondo riguarda il numero di fattori che si possono estrarre che è lo stesso si se si utilizza la matrice **B** che **Z**, ma nel primo caso gli autovalori sono il quadrato di quelli ottenuti con la matrice **Z** (Mussino,2001, p.116).

Più volte, nell'ambito di questo scritto si è fatto riferimento all'*inerzia* - che nell'ACM è il termine con cui si indica la varianza – per cui registra il livello d'informazione contenuta nella matrice dei dati.

Nell'ACM, come già segnalato, l'inerzia può essere calcolata sia rispetto all'origine dello spazio di riferimento e sia rispetto al baricentro della nuvola dei punti. In entrambi i casi l'obiettivo è proiettare i punti in uno spazio di dimensioni ridotte. Come nell'ACP, anche in questo caso gli assi fattoriali sono il risultato di una operazione matematica che consiste, come nel caso dell'ACP, nel risolvere un sistema di equazioni rispettando il vincolo di:

1. massimizzare l'inerzia riprodotta da ogni fattore;
2. ortogonalità tra gli assi.

I risultati dell'analisi non cambiano se si opera nello spazio delle unità (R^n)o in quello delle modalità (R^q), per cui è possibile *transitare* dalla soluzione fattoriale dello spazio delle unità a quello della modalità per *via* matematica; ciò rende superfluo fare due analisi separate (Bolasco, 1999, p.136).

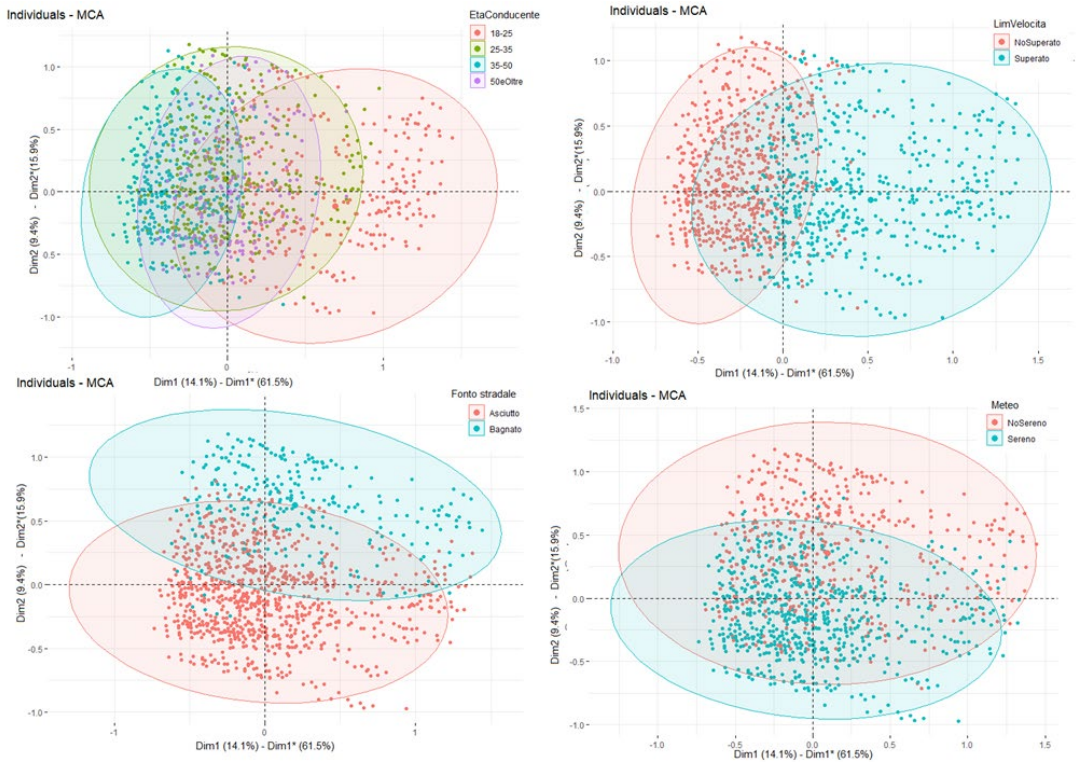


Fig. 5 – Configurazione degli incidenti stradali nel comune di Roma sul primo piano fattoriale

La forma della nuvola dei punti-modalità sul piano fattoriale, così come per altri versi quella dei punti-individuo, possono essere di aiuto per l'interpretazione dell'output. In tal senso, le configurazioni più tipiche sono quelle a forma di:

- ellisse, che indica una connessione tra le modalità e il primo asse;
- ferro di cavallo, che segnala una sostanziale unidimensionalità nella struttura dei dati. In questo caso il secondo fattore è interpretabile in funzione del primo;
- triangolo, che segnala una variazione inversa tra le modalità del primo fattore rispetto a quelle del primo;
- nuvole separate, che indicano la presenza di sottogruppi tra loro poco connessi per cui è preferibile procedere con analisi separate (Bolasco,1999, p.123-124).

Cluster Analysis metriche per il calcolo delle distanze tra le unità

A corredo dell'illustrazione delle procedure di cluster analysis, di seguito sono riportate le formule per il calcolo delle distanze tra unità, nel caso in cui le variabili utilizzate per l'analisi dei gruppi siano cardinali o quasi-cardinali:

1. Blocchi o di Manhattan

$$d_{1(a,b)} = \sum_j^p |x_{aj-}x_{bj}|;$$

2. Euclidea

$$d_{2(a,b)} = \sqrt{\sum_j^p (x_{aj-}x_{bj})^2} \;;$$

queste due metriche possono essere ricondotte ad una funzione più generale, messa a punto da Minkowsky di ordine k:

$$d_{k(a,b)} = \sqrt[k]{\sum_{j}^{p} (x_{aj-}x_{bj})^{k}}$$

dove per *k*=1 si ha la distanza a blocchi, per k=2 la distanza euclidea, per *k*=3 la distanza cubica, ecc. Nella metrica a blocchi c'è una sorta di "effetto compensazione" tra le grandi e le piccole distanze che si attenua se si utilizza distanza euclidea nella quale, proprio perché elevati al quadrato, acquistano più peso le grandi distanze e meno le piccole distanze. È chiaro che le grandi distanze diventano ancora più incisive, e questo effetto scopare, man mano che aumenta l'esponente *k* della formula di Minkowsky.

Una metrica che non rientra nella funzione di Minkowsky è quella di Mahalanobis che prende in considerazione anche la correlazione tra le variabili (Zani e Cerioli, *op.cit.*, pp.335-343).

Di seguito, a completamento dell'esposizione dell'esempio riportato nel capitolo, sono riportati i grafici delle distanze di Manhattan ed euclidee calcolate sia sui 16 indicatori socioeconomici-culturali delle regioni italiane sia sulle prime tre componenti principali ottenute su questo set di indicatori.

La configurazione delle distanze descritte in forma grafica nelle figure 6 e 7 sembra indicare una sostanziale invarianza sia rispetto alla metrica utilizzata e sia rispetto al set di indicatori utilizzati.

Fig. 6 – Distanze di Manhattan e euclidee calcolate sulle variabili standardizzate

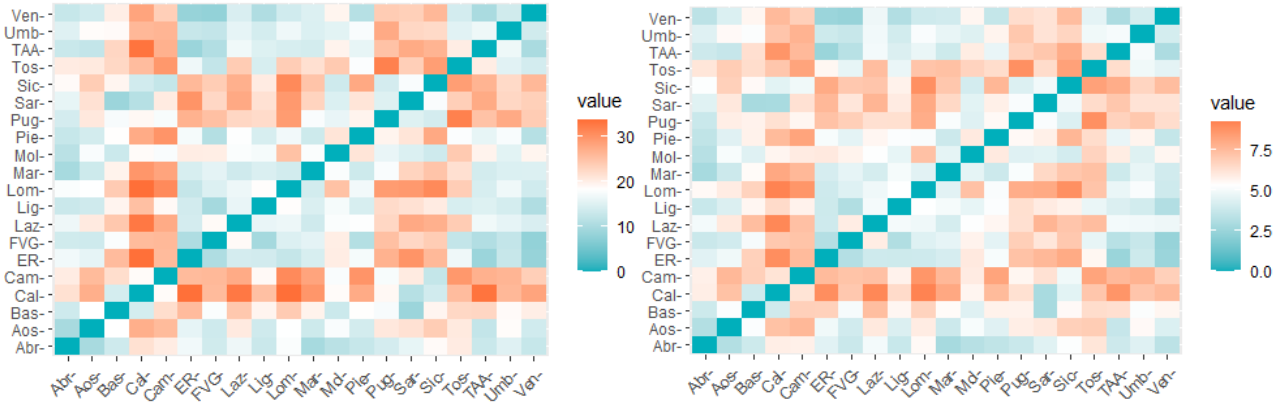


Fig.7 – Distanze di Manhattan ed euclidee calcolate sulle prime tre componenti principali