Università degli Studi di Napoli
Federico II

Riduzione delle Modalità Nell'Analisi delle Corrispondenze Multiple

Attraverso una ricodifica sequenziale automatica

Pietro Mascia

Tesi di Dottorato in
Statistica

XIX Ciclo

Riduzione delle Modalità
Nell'Analisi delle Corrispondenze Multiple

Napoli. 30 novembre 2006

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

In questa tesi, viene presentata la Ricodifica Sequenziale Automatica delle Modalità, (SAR) (Mascia, Mola 2006). La SAR può essere vista come una procedura generale, applicabile ogni qualvolta si abbiano variabili con un elevato numero di modalità e si renda necessario ridurne il numero. La portata generale della metodologia, la rende potenzialmente applicabile a qualunque metodologia statistica, ad ogni modo, nel presente lavoro lo scopo è quello di presentarne un'implementazione in grado di rendere più agevole l'interpretazione del piano fattoriale nell'Analisi delle Corrispondenze Multiple (ACM) e di fornire una procedura di ricodifica oggettiva delle modalità. La procedura proposta si presenta particolarmente utile nel caso di variabili con un numero eccessivo di modalità. Nel primo capitolo viene presentata una trattazione formalizzata dell'ACM allo scopo di rendere chiara e coerente la simbologia utilizzata nei capitoli seguenti. Particolare attenzione è posta sugli aspetti critici dell'ACM, come ad esempio l'influenza del numero delle modalità sui tassi d'inerzia, l'influenza delle modalità a bassa frequenza sulla stabilità dei risultati e la necessità di un bilanciamento nel numero di modalità in ciascuna variabile. Nel secondo capitolo sono descritti i problemi che la complessità computazionale e alcuni tipi di ricodifica possono creare alle principali metodologie statistiche. La complessità computazionale viene scomposta nei tre aspetti principali:

- complessità nel numero di osservazioni;
- complessità nel numero di variabili;
- complessità nel numero di modalità.

Dalla constatazione che una riduzione del numero di modalità riduce in modo sostanziale alcuni problemi legati alla complessità e dalla necessità di una ricodifica oggettiva, s'introduce la Ricodifica Sequenziale Automatica delle Modalità

(SAR). Punto centrale della SAR è la quasi totale indipendenza della ricodifica dalle opinioni dell'analista e la totale dipendenza dai risultati della metodologia alla quale viene applicata. L'obiettivo del terzo capitolo è l'applicazione della Ricodifica Sequenziale Automatica delle Modalità per la risoluzione di alcuni problemi ricorrenti nella ACM. In particolare viene proposto l'utilizzo della SAR per:

- la riduzione del numero di modalità allo scopo di rendere più agevole l'interpretazione del piano fattoriale. La riduzione avviene attraverso un accorpamento delle modalità. Vengono proposte tre varianti, una per variabili sconnesse, una per variabili ordinali ed una per variabili numeriche;
- l'eliminazione del problema delle modalità con frequenze eccessivamente basse attraverso la definizione di modalità ''*Semi-Attive*''. Questo tipo di modalità riduce l'arbitrarietà dell'assegnazione casuale.
- la ricodifica automatica in classi delle variabili continue;

Il quarto capitolo propone alcune possibili interpretazioni della metodologia proposta attraverso l'applicazione a casi reali. La prima applicazione illustra come la SAR possa ridurre notevolmente il numero delle modalità rendendo più leggibile il piano fattoriale e rendendone più immediata l'interpretazione. Nella seconda sezione del quarto capitolo, la SAR viene impiegata come strumento di Data mining. Si mostra come in presenza di un grandissimo numero di variabili anche una ricodifica possa risultare inefficace, ma che imponendo una gerarchia alle variabili e proiettandole sul piano una alla volta e contestualmente ricodificandole, si possano comunque individuare le relazioni più importanti presenti in una matrice di dati. Nelle conclusioni, si mostra come la procedura proposta possa essere interpretata sia come una variante metodologica dell'Analisi delle Corrispondenze Multiple o più in generale una variante dei metodi di riduzione delle dimensioni, sia come uno strumento di supporto alle decisioni o all'interpretazione del piano fattoriale. Si prospettano infine le linee di ricerca future che riguardano, il miglioramento degli algoritmi e lo sviluppo di software grafici che permettano una migliore visualizzazione del piano fattoriale e la manipolazione in tempo reale delle variabili da parte del ricercatore. La Ricodifica Sequenziale Automatica si è mostrata uno strumento utile e flessibile che può essere esteso anche ad altre metodologie che soffrano la presenza di variabili con un numero eccessivo di modalità.

Capitolo 1

Il modello dell'analisi fattoriale esplorativa

1.1 Introduzione

Negli ultimi anni la grande disponibilità di dati e la possibilità d'uso di potenti calcolatori ha evidenziato alcuni limiti della statistica classica. Concepita in un periodo caratterizzato dalla carenza sia di dati che di strumenti per la loro elaborazione, la statistica classica era stata concepita prevalentemente per dare risposta proprio a queste carenze. I suoi fondamenti teorici si basavano prevalentemente sull'uso del calcolo probabilistico attraverso il quale si cercava di indurre dal caso particolare, conosciuto, al caso generale e sconosciuto. Oggigiorno, la disponibilità di dati e mezzi, sia per la loro acquisizione che per l'elaborazione, ha completamente rovesciato la situazione mostrando tutti i limiti della statistica classica e dando impulso allo sviluppo di un nuovo settore della statistica: L'analisi multidimensionale dei dati. L'obbiettivo dell'analisi multidimensionali dei dati (AMD) è quello di studiare simultaneamente una grande quantità di informazioni con lo scopo di fornire una descrizione complessiva del fenomeno e trovare l'andamento di fondo dello stesso. Gli elementi caratterizzanti dell'AMD sono stati schematizzati da Benzecrì e si possono cosi sintetizzare:

1. la statistica non è calcolo delle probabilità;

2. il modello deve adattarsi ai dati e non viceversa;

3. una visione esaustiva della struttura sottostante il fenomeno è possibile solo attraverso il trattamento simultaneo delle informazioni inerenti il fenomeno stesso;

4. elemento importantissimo è la rappresentazione grafica del risultato, ottenuto attraverso le proprietà geometriche delle tecniche di analisi multidimensionale;

Storicamente l'AMD si è sviluppata su due grandi filoni: i metodi fattoriali, che hanno come scopo la rappresentazione di una nuvola di punti nello spazio multidimensionale in un sottospazio di dimensioni ridotte e i metodi di classificazione, che hanno come scopo quello di classificare gli individui analizzati in un certo numero di gruppi massimamente omogenei al loro interno e massimamente eterogenei all'esterno. I metodi fattoriali storicamente più importanti sono l'Analisi in Componenti Principali (ACP), l'Analisi delle Corrispondenze semplici (AC) e l'Analisi delle Corrispondenze Multiple (ACM). L'ACP consente di ridurre il numero delle variabili che descrivono le unità e riprodurre le caratteristiche di queste attraverso nuove variabili (componenti principali) che sono combinazioni lineari delle variabili di partenza, conservando il più possibile delle relazioni di partenza. Le nuove variabili ottenute sono per loro costruzione incorrelate. L'ACP si realizza attraverso i seguenti passi:

1. definizione di una misura della distanza tra le unità

2. ricerca dei nuovi assi in modo da ottenere la migliore proiezione di tali distanze su un sottospazio di riferimento, ottenuto massimizzando l'inerzia dei punti rispetto al baricentro

3. individuazione su tali assi delle coordinate degli individui e delle variabili

4. analisi del cerchio delle correlazioni, e analisi della posizione degli individui.

Nella prima parte del capitolo verrà descritta l'analisi fattoriale generale; successivamente verranno illustrati i fondamenti dell'Analisi delle Corrispondenze, per concludere con la descrizione formalizzata dell'Analisi delle Corrispondenze Multiple.

1.2 Il modello generale

Si consideri una matrice di dati $\mathbf{X}$ di dimensione $(n \times p)$ in cui ciascuna riga corrisponde ad un'unità statistica e ciascuna colonna rappresenta una variabile. Nel caso si stiano trattando caratteri quantitativi, le colonne della matrice $\mathbf{X}$ rappresenteranno p misurazioni su ciascuna delle n unità statistiche, mentre nel caso di caratteri qualitativi, $\mathbf{X}$ potrà essere o una tabella di contingenza o una matrice disgiuntiva completa.

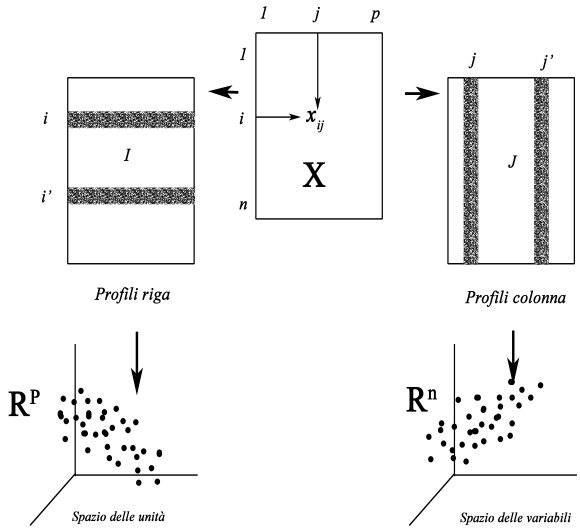


Figura 1.1: Nuvole dei punti nei diversi spazi ambiente

Come mostra la figura 1.1, se si definisce I come l'insieme delle osservazioni, è sempre possibile rappresentare questo insieme in uno spazio ambiente di tipo vettoriale, in cui l'insieme I è rappresentabile attraverso una nuvola di punti *N(I)* nello spazio $\mathbf{R}^p$ detto spazio delle unità. Analogamente, si può definire l'insieme delle variabili J e rappresenatare la corrispondente nuvola *N(J)* in uno spazio $\mathbf{R}^n$ chiamato spazio delle variabili. Lo studio delle unita nello spazio delle unità è totalmente definito da una matrice dei dati $\mathbf{X}$, da un vettore dei pesi delle unità $\mathbf{D}$ e da un criterio di riponderazione delle variabili $\mathbf{M}$. Generalmente $\mathbf{M}$ ha la forma di una matrice diagonale contenente una misura della variabilità

delle variabili considerate nell'analisi. **M** è detta *metrica* dello spazio di rappresentazione delle unità, in quanto da essa dipendono le distanze tra i punti nello spazio considerato. Lo studio multidimensionale è pertanto totalmente definito da tre matrici, **(X,M,D)**. Naturalmente la forma e la composizione di queste matrici sarà diversa a seconda della scala di misura utilizzata per la definizione delle variabili oggetto di studio. E' altresì ovvia l'impossibilità di visualizzare direttamente le nuvole dei punti *N(I)* e *N(J)* quando n o p sono maggiori di tre. Dato che questa è la situazione ordinaria, ben difficilmente si affronterà uno studio multidimensionale con tre variabili e ben difficilmente si affronterà un qualsiasi studio statistico con tre unità, nasce l'esigenza di fornire una rappresentazione semplificata, ma allo stesso tempo utile ed efficace, di tali insiemi di informazioni in sottospazi ottimali di dimensione ridotta generati dai cosiddetti assi fattoriali. L'obiettivo principale delle tecniche fattoriali di tipo esplorativo, consiste quindi nel descrivere la matrice originaria dei dati attraverso la visualizzazione della struttura esistente sugli elementi delle righe e sulle colonne, o, in altre parole, la rappresentazione della forma delle nuvole dei punti da essi generate. Le tecniche fattoriali hanno quindi come finalità:

- ridurre la dimensionalità della matrice attraverso la definizione di nuove variabili (fattori) tra loro incorrelate;

- costruire delle dimensioni sintetiche e originariamente inosservabili (assi fattoriali) che rappresentino dei modelli teorici in grado sia di interpretare il fenomeno sia di offrire un punto di vista originale dello stesso.

Ogni nuova dimensione fattoriale costituisce un riassunto dell'informazione originaria, pertanto i metodi fattoriali possono essere interpretati, e nei fatti lo sono, come modelli di riduzione dei dati e riduzione del rumore in esso presente. Le nuvole dei punti *N(I)* e *N(J)*, come mostra la figura 1.2 possono assumere nello spazio diverse forme che caratterizzano la natura e l'intensità delle relazioni esistenti tra i punti della matrice dei dati. Per rendere visibili queste forme, l'analisi mediante metodi fattoriali, consiste nel proiettare queste forme su rette o piani minimizzando quanto possibile la deformazione derivante da questa proiezione.

L'obiettivo è quindi la ricerca del sottospazio Δ, che massimizza la somma dei quadrati delle distanze tra le proiezioni su Δ di tutte le coppie di punti (i, i').

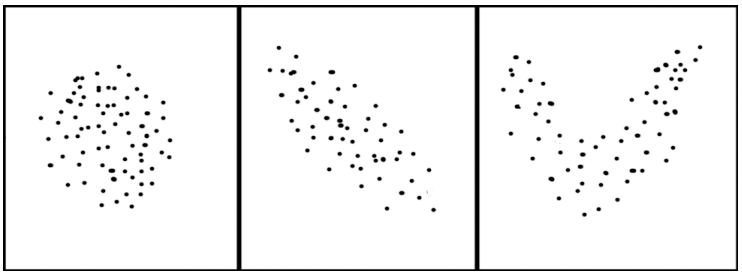


Figura 1.2: Diverse forme assunte dalle nuvole dei punti nello spazio

$$max_{(\Delta)} \left\{ \sum_i \sum_{i'} d^2(i, i') \right\} \tag{1.1}$$

Se ciascun punto è pesato per una quantità p_i, otteniamo:

$$max_{(\Delta)} \left\{ \sum_i \sum_{i'} p_i p_{i'} d^2(i, i') \right\} \tag{1.2}$$

Considerare le distanze tra ciascuna coppia di unità, equivale a considerare l'insieme delle distanze dei punti dal baricentro G, ossia:

$$max_{(\Delta)} \left\{ \sum_i p_i d^2(i, G) \right\} \tag{1.3}$$

1.2.1 Ricerca del sottospazio ottimale per le unità

Considerando la nuvola dei punti *N(I)*, l'obiettivo consiste dunque nel cercare il sottospazio di R^P ad una dimensione, $(\Delta \boldsymbol{u})$, che riproduca nel miglior modo possibile la nuvola *N(I)*. Intuitivamente il procedimento consiste nel far passare una retta nel mezzo della nuvola dei punti muovendola fintanto che non risulti massimizzata la proiezione della distanza tra punti sulla retta stessa. In maniera più formale: sia $\boldsymbol{u}$ un vettore di norma unitaria dello spazio R^P che individua questa retta (Δu). La proiezione ortogonale OH_i dell'i-esimo individuo OM_i sulla retta di vettore unitario $\boldsymbol{u}$ è uguale al prodotto scalare tra $OM_i = x_i$ e il vettore di norma unitaria $\boldsymbol{u}$:

$$OH_i = x_i^{'}u = \sum_{j=1}^{p} x_{ij}u_j \tag{1.4}$$

Graficamente si avrà:

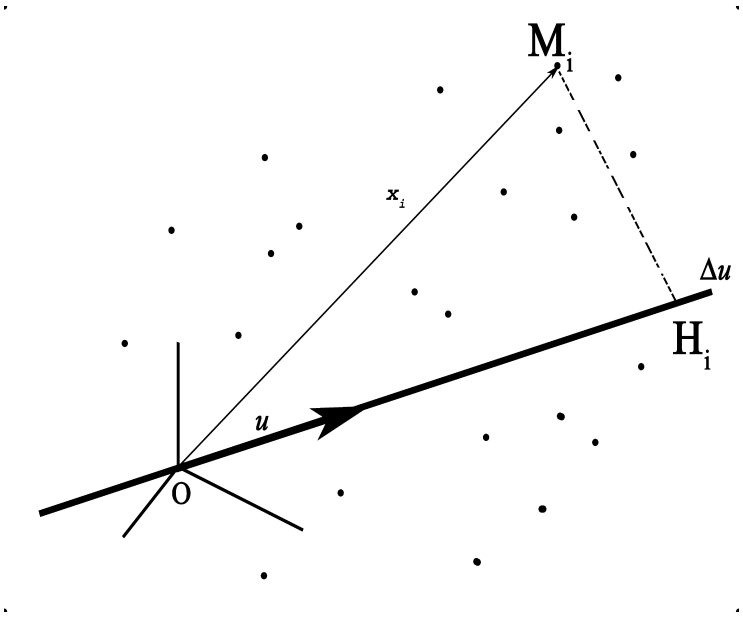


Figura 1.3: Proiezione dei punti unità nel sottospazio ottimale

Mentre esprimendo la proiezione dell'intera nuvola dei punti *N(I)* sulla retta $(\Delta \boldsymbol{u})$ in forma matriciale si ottiene:

$$\mathbf{X}u = \begin{vmatrix} x_{11} & ... & x_{1p} \\ ... & ... & ... \\ ... & x_{ij} & ... \\ ... & ... & ... \\ x_{n1} & ... & x_{np} \end{vmatrix} \begin{vmatrix} u_1 \\ ... \\ u_j \\ ... \\ u_p \end{vmatrix} = \begin{vmatrix} ... \\ \sum_{j=1}^{p} x_{ij}u_j \\ ... \\ ... \end{vmatrix}$$

Per la ricerca del migliore adattamento del sottospazio cercato alla nuvola dei punti si ricorre al metodo dei minimi quadrati, consistente nel cercare la retta dalla quale risulti minima la somma dei quadrati delle distanze dei punti, indi-

cata con

$$\sum_{i=1}^{n}(M_iH_i)^2 \tag{1.5}$$

Se si applica il teorema di Pitagora a ciascuno degli n triangoli identificati dai vertici M_iH_iO figura 1.3 si ottiene:

$$\sum_{i=1}^{n}(OM_i)^2 = \sum_{i=1}^{n}(M_iH_i)^2 + \sum_{i=1}^{n}(OH_i)^2 \tag{1.6}$$

Esplicitando rispetto alla quantità d'interesse si ottiene:

$$\sum_{i=1}^{n}(M_iH_i)^2 = \sum_{i=1}^{n}(OM_i)^2 - \sum_{i=1}^{n}(OH_i)^2 \tag{1.7}$$

Essendo $\sum_{i=1}^{n}(OM_i)^2$ una quantità data ed indipendente dal vettore cercato $\boldsymbol{u}$, minimizzare $\sum_{i=1}^{n}(M_iH_i)^2$ equivale a massimizzare $\sum_{i=1}^{n}(OH_i)^2$.
Contestualizzando i principi geometrici appena esposti nell'ambito della matrice di dati precedentemente esposta, sezione 1.2, ed esprimendo il tutto in funzione di $\mathbf{X}$ ed $\boldsymbol{u}$ si ottiene:

$$\sum_{i=1}^{n}(OH_i)^2 = (\mathbf{X}\boldsymbol{u})'(\mathbf{X}\boldsymbol{u}) = \boldsymbol{u}'\mathbf{X}'\mathbf{X}\boldsymbol{u} \tag{1.8}$$

Per trovare il vettore $\boldsymbol{u}$ occorre dunque cercare il massimo di $\boldsymbol{u}'\mathbf{X}'\mathbf{X}\boldsymbol{u}$ vincolato al fatto che $\boldsymbol{u}$ abbia norma unitaria. Ossia:

$$\begin{cases} max_{(u)}\left\{\boldsymbol{u}'\mathbf{X}'\mathbf{X}\boldsymbol{u}\right\} \\ \boldsymbol{u}'\boldsymbol{u} = 1 \end{cases} \tag{1.9}$$

La ricerca dell'asse migliore $\Delta\boldsymbol{u}$ di versore $\boldsymbol{u}$, ossia della retta per la quale risulta massimizzata la somma delle proiezione dei punti, si effettua attraverso la risoluzione di un'equazione agli autovalori del tipo

$$\mathbf{X}'\mathbf{X}\boldsymbol{u} = \lambda\boldsymbol{u} \tag{1.10}$$

Sia ora $\boldsymbol{u}_1$ il primo vettore cercato. Il vettore $\boldsymbol{u}_1$ è l'autovettore della matrice **X'X** di ordine *(p,p)* corrispondente al più grande autovalore λ_1. Una volta trovato il vettore $\boldsymbol{u}_1$, si cerca il vettore di norma unitaria $\boldsymbol{u}_2$ ortogonale a $\boldsymbol{u}_1$ e associato al secondo autovalore λ_2 tale che renda massima l'espressione:

$$\boldsymbol{u}_2'\mathbf{X}'\mathbf{X}\boldsymbol{u}_2 \tag{1.11}$$

Si procede di questo passo cercando il terzo autovettore $\boldsymbol{u}_3$ ortogonale ai primi due fino ad ottenere il numero di dimensioni desiderato l, con l comunque minore di p. Le l componenti trovate sono di importanza decrescente e forniscono un nuovo sistema di riferimento nello spazio delle unità che *passa* il più vicino possibile alla nuvola originaria *N(I)*. Una volta definito il nuovo sottospazio ottimale, ossia l'insieme degli assi $\Delta\boldsymbol{u}$ che individuano la base

$$\{\boldsymbol{u}_1 ... \boldsymbol{u}_\alpha ... \boldsymbol{u}_l\} \tag{1.12}$$

l'individuo i-esimo avrà per l'α-esimo asse una coordinata pari all'estremità della proiezione ortogonale OH_i, ossia:

$$c_\alpha(i) = \boldsymbol{x}_i'\boldsymbol{u}_\alpha \tag{1.13}$$

1.2.2 Ricerca del sottospazio ottimale per le variabili

Per la ricerca del sottospazio ottimale nello spazio delle variabili per la nuvola *N(J)*, si segue un procedimento analogo a quello seguito per la nuvola *N(I)*. Nello spazio R^n, si ricerca quel vettore $\boldsymbol{v}$ che consente la migliore proiezione, sempre seguendo il criterio dei minimi quadrati, della nuvola *N(J)*, dei p punti variabile, nel sottospazio ad una dimensione $\Delta\boldsymbol{v}$ in R^n. Questo procedimento, ancora una volta, equivale a rendere massima la somma dei quadrati delle p proiezioni su $\boldsymbol{v}$, corrispondenti alle p componenti del vettore $c^* = \mathbf{X}'\boldsymbol{v}$, ossia:

$$(\mathbf{X}'\boldsymbol{v})'(\mathbf{X}'\boldsymbol{v}) = \boldsymbol{v}'\mathbf{X}\mathbf{X}'\boldsymbol{v} \tag{1.14}$$

In modo analogo a quanto visto nello spazio delle unità, si devono trovare gli l autovettori, corrispondenti ai primi l autovalori della matrice $\mathbf{XX}$' di dimensione $(n \times n)$. Indicando con$\boldsymbol{v}_\alpha$ il generico autovettore di $\mathbf{XX}$' corrispondente all'autovalore μ_α, l'equazione agli autovalori si esprime con:

$$\mathbf{XX}'\boldsymbol{v}_\alpha = \mu_\alpha \boldsymbol{v}_\alpha \tag{1.15}$$

Una volta definito il sottospazio ottimale, la coordinata della generica variabile j, pari alla proiezione corrispondente su $\Delta\boldsymbol{v}_\alpha$ sarà data da:

$$c_\alpha^*(j) = \boldsymbol{x}_j'\boldsymbol{v}_\alpha \tag{1.16}$$

1.2.3 Relazione tra lo spazio delle unità $\mathbf{R}^p$ e lo spazio delle variabili $\mathbf{R}^n$

Nonostante sia possibile identificare spazi diversi a seconda che si consideri la matrice dei dati secondo le righe o secondo le colonne, la matrice è pur sempre la stessa, per cui è evidente la dualità presente nelle due analisi precedentemente esposte. Le due equazioni agli autovalori, possono essere così riscritte:

$$\mathbf{X}'\mathbf{X}\boldsymbol{u}_\alpha = \lambda_\alpha \boldsymbol{u}_\alpha \; in \; \mathbf{R}^p \tag{1.17}$$

$$\mathbf{XX}'\boldsymbol{v}_\alpha = \mu_\alpha \boldsymbol{v}_\alpha \; in \; \mathbf{R}^n \tag{1.18}$$

premoltiplicando la 1.17 per $\mathbf{X}$ si ottiene:

$$\left(\mathbf{XX}'\right)\mathbf{X}\boldsymbol{u}_\alpha = \lambda_\alpha \left(\mathbf{X}\boldsymbol{u}_\alpha\right) \tag{1.19}$$

Questa relazione mostra che ad ogni autovettore μ_α di **X'X** relativo all'autovalore non nullo λ_α corrisponde un autovettore $\mathbf{X}\boldsymbol{u}_\alpha$ di **X'X** relativo allo stesso autovalore λ_α (Bolasco 1999). Poichè dalla relazione 1.18 si è indicato con μ_1 il più grande autovalore di **X'X**, si deve necessariamente avere

$$\lambda_1 \leq \mu_1 \tag{1.20}$$

Si premoltiplichi ora l'equazione 1.18 per **X**', si ottiene cosi

$$\left(\mathbf{X}'\boldsymbol{X}\right)\mathbf{X}'\boldsymbol{v}_\alpha = \mu_\alpha\left(\mathbf{X}'\boldsymbol{v}_\alpha\right) \tag{1.21}$$

Si può cosi notare che $\mathbf{X}'\boldsymbol{v}_\alpha$ è un autovettore di **X'X** relativamente all'autovalore μ_1, così come lo è μ_α nella relazione 1.17, per cui deve valere anche $\mu_1 \leq \lambda_1$. Ma dovendo essere vere entrambe, non può che essere:

$$\lambda_1 = \mu_1 \tag{1.22}$$

Questa relazione vale inoltre per tutti gli autovalori, ossia

$$\lambda_\alpha = \mu_\alpha \tag{1.23}$$

Si può inoltre notare che il vettore $\mathbf{X}\boldsymbol{u}_\alpha$ ha norma quadratica uguale a λ_α, si ha infatti che

$$\boldsymbol{u}'_\alpha\mathbf{X}'\mathbf{X}\boldsymbol{u}_\alpha = \lambda_\alpha \tag{1.24}$$

ossia

$$var(c_\alpha) = \|c_\alpha\|^2 = \lambda_\alpha \tag{1.25}$$

Si può quindi dedurre che l'autovettore $\boldsymbol{v}_\alpha$ della relazione 1.18 coincide con l'autovettore $\mathbf{X}\boldsymbol{u}_\alpha$ della relazione 1.19, in quanto entrambi corrispondenti allo stesso autovalore λ_α. Quindi $\boldsymbol{v}_\alpha$ è facilmente calcolabile in funzione di $\boldsymbol{u}_\alpha$. Essendo però il vettore $\boldsymbol{v}_\alpha$ unitario, si deve rendere unitario $\mathbf{X}\boldsymbol{u}_\alpha$ che si ottiene dividendolo per la sua norma, per cui:

$$\boldsymbol{v}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\mathbf{X}\boldsymbol{u}_\alpha \tag{1.26}$$

La relazione tra lo spazio delle unità $\mathbf{R}^p$ e lo spazio delle variabili $\mathbf{R}^n$ è definita dalle seguenti **formule di transizione**:

$$\begin{cases} \boldsymbol{v}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{X} \boldsymbol{u}_\alpha \\ \\ \boldsymbol{u}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{X}' \boldsymbol{v}_\alpha \end{cases} \tag{1.27}$$

Considerando lo spazio delle unità e definendo c_α come l'insieme delle coordinate delle unità, si ha che $c_\alpha = \mathbf{X}\boldsymbol{u}_\alpha$ mentre nello spazio delle variabili, definendo c^*_α come le coordinate delle variabili, si ha

$$c^*_\alpha = \mathbf{X}'\boldsymbol{v}_\alpha \tag{1.28}$$

Sostituendo nelle formule di transizione si ricava che:

$$c_\alpha = \sqrt{\lambda_\alpha}\boldsymbol{v}_\alpha \tag{1.29}$$

$$c^*_\alpha = \sqrt{\lambda_\alpha}\boldsymbol{u}_\alpha \tag{1.30}$$

Per cui le coordinate delle variabili sono calcolabili direttamente a partire dagli autovettori ottenuti nello spazio delle unità. Inoltre nel sottospazio di $\mathbf{R}^p$ generato da $\boldsymbol{u}_\alpha$, le coordinate dei punti della nuvola *N(I)* delle unità sono le componenti di $\mathbf{X}\boldsymbol{u}_\alpha$ e sono anche le componenti di

$$\sqrt{\lambda_\alpha}\boldsymbol{v}_\alpha \tag{1.31}$$

Per cui le coordinate $c_\alpha(i)$ dei punti unità su un generico asse fattoriale in $\mathbf{R}^p$ sono proporzionali alle componenti $\boldsymbol{v}_\alpha(i)$ dell'asse fattoriale $\boldsymbol{v}_\alpha$ in $\mathbf{R}^n$, corrispondenti all'autovalore λ_α. Un analogo discorso può essere fatto per le coordinate della nuvola delle variabili *N(J)*, (Bolasco 1999).

1.2.4 Rappresentazione nello spazio vettoriale di elementi non-attivi o supplementari

La matrice originaria dei dati $\mathbf{X}$ può essere divisa idealmente in due sottomatrici, o meglio, nell'anzidetta matrice possono essere individuati due tipologie di elementi. I primi chiamati elementi **attivi** sono quelli che concorrono alla ricerca del sottospazio ottimale, nel senso che entrano quali elementi costitutivi

nella costruzione del modello di rappresentazione dei dati. I secondi, chiamati **supplementari** non entrano invece quali elementi costitutivi del modello specificato. Nonostante questo è pur sempre possibile posizionare questi elementi nel sottospazio trovato. Gli elementi supplementari concorrono alla interpretazione degli assi trovati è, nonostante siano elementi supplementari, mostrano un'indubbia importanza nei modelli fattoriali. Bisogna ad ogni modo rimarcare che essi si trovano in posizione subordinata rispetto agli elementi attivi non concorrendo comunque a determinare la soluzione ottimale ma solo a meglio interpretarla a posteriori, per questo motivo sono denominati supplementari, **illustrativi** o impropriamente **fuori analisi**. Gli elementi supplementari possono appartenere indifferentemente sia all'insieme I che all'insieme J. Se si indica con X_+ la matrice dell'insieme degli individui supplementari, si veda la figura 1.4, le coordinate degli individui supplementari saranno date da

$$c_\alpha^s = X_+ \boldsymbol{u}_\alpha \tag{1.32}$$

mentre le coordinate delle variabili supplementari saranno

$$c_\alpha^{*s} = X^{+'} \boldsymbol{v}_\alpha \tag{1.33}$$

Per concludere si ricorda che gli elementi supplementari possono essere anche interpretati come degli elementi attivi ma senza massa. La loro inerzia è conseguentemente nulla, a rimarcarne la minore importanza rispetto agli elementi attivi.

1.3 L'Analisi delle Corrispondenze

L'Analisi delle Corrispondenze conosciuta anche come Analisi delle Corrispondenze Semplici o Binarie, è forse la più nota tra le metodologie per l'analisi dei dati di tipo qualitativo. Le origini di questa metodologia sono difficilmente databili a causa delle continue ridefinizioni del metodo stesso. Nonostante si possano considerare alcuni lavori di Fisher (Fisher 1940), come le origini teoriche di riferimento, e poi proposta sotto diversi punti di vista da Guttman (Guttman 1941), da Hayashi (Hayashi 1950), (Hayashi 1956) e intuita gia dal 1935 da Hirschfeld, (Hirschfeld 1935), è agli inizi degli anni '60 grazie a Benzécri

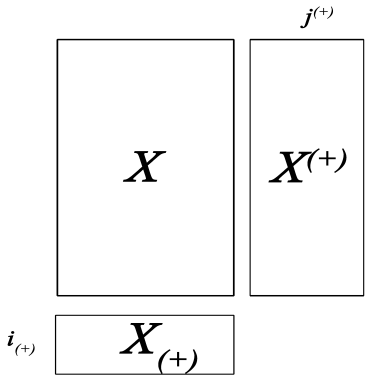


Figura 1.4: Rappresentazione matriciale degli elementi attivi e non attivi

(Benzécri 1973) e alla scuola francese che il metodo ha assunto la sua connotazione più moderna. La scuola francese, in contrapposizione all'impostazione inferenziale classica, propone un'impostazione che utilizza le proprietà algebriche e geometriche utilizzandole come strumento prevalentemente descrittivo. L'Analisi delle Corrispondenze, fu, in seguito alle posizioni epistemologiche espresse da Benzécri, al centro di accese discussioni tra le due principali scuole sopra citate. Il punto più controverso è individuabile nel principio espresso da Benzécri secondo il quale: I**l modello deve seguire i dati e non viceversa**. Questo punto di vista rispecchia un'esigenza all'epoca particolarmente sentita: di fronte alla complessità dei fenomeni reali, la possibilità di un'analisi globale del fenomeno grazie alla definizione di variabili non direttamente osservabili, consente di **mettere momentaneamente da parte le proprie conoscenze a priori**, e di osservare in maniera non preconcetta le informazioni che i dati possono fornire. Sottostante a questa proposizione si trova un altro principio enunciato Benzecrì: C**onviene trattare simultaneamente informazioni concernenti il maggior numero possibile di dimensioni**. Solo successivamente, si vedrà se ciò che è emerso può generare nuovi modelli d'interpretazione del reale o se esso può essere spiegato con modelli già conosciuti. Se quest'impostazione nasce dall'accettazione della complessità dei fenomeni analizzati, oggi quest'esigenza è ancora più sentita. La grandissima disponibilità di dati permette, spesso, di analizzare i fenomeni in quasi tutta la loro complessità. Oltre alla disponibilità

di un gran numero di variabili, oggi si dispone anche di una grandissima quantità di osservazioni che rendono spesso superflue le usuali procedure inferenziali. Si pensi a tutte le volte che il p-value assume valori fuori scala nei moderni software per l'analisi dei dati. Questo a causa del fatto che oramai ci si trova a lavorare più che con campioni con dei piccoli universi. Tutte queste circostanze non fanno che rinvigorire i principi enunciati da Benzécri e rendere ancora più moderna ed attuale tutta l'impostazione della scuola francese dei dati e ribadire la modernità dell'Analisi delle Corrispondenze.

1.3.1 Considerazioni generali, matrice dei dati e spazio di riferimento

Si consideri una generica tabella di contingenza $\boldsymbol{T}(r,c)$, dove r indica il numero delle righe e c il numero di colonne. Sia inoltre $\boldsymbol{A}$ il carattere posizionato nella colonna madre e $\boldsymbol{B}$ il carattere disposto in testata. Si supponga che i due caratteri $\boldsymbol{A}$, e $\boldsymbol{B}$ non siano tra loro indipendenti in senso statistico. La misurazione del grado di dipendenza può trovare risposta attraverso l'indice X^2, mentre se si vuole conoscere a cosa sia dovuta la ragione della dipendenza, si può utilizzare l'Analisi delle Corrispondenze. L'obiettivo consiste nello studiare la struttura della relazione d'insieme, ovvero l'interdipendenza tra i due caratteri e illustrarne gli aspetti principali scoprendo quali sono gli assi principali di inerzia. In altre parole l'applicazione dell'Analisi delle corrispondenze permette di individuare in quali celle della tabella si hanno scostamenti tra le frequenze attese in caso di indipendenza e frequenze effettive (contingenze positive).

L'analisi della tabella di contingenza $\boldsymbol{T}$, è condotta sulle frequenze relative, consentendo cosi di rendere confrontabili le diverse modalità di una stessa variabile. Le prossimità sul risultante piano fattoriale indicheranno la similarità tra le modalità, ossia la similitudine fra le distribuzioni parziali loro associate. L'importanza delle modalità all'interno della tabella viene determinata dalla loro frequenza relativa. Dato che la somma delle frequenze relative è sempre uguale ad uno, la nuvola dei punti giace in uno spazio a *c-1* dimensioni mentre, per la nuvola dei profili colonna sarà individuata in uno spazio a *r-1* dimensioni. La scelta di operare sulle tabelle dei profili porta ad utilizzare, nel calcolo delle distanze tra due punti, una metrica diversa da quella euclidea. Infatti, se si vuole che il computo della distanza tra due profili tenga conto, da un lato della similitudine tra le distribuzioni, e dall'altro dell'importanza di ciascuna delle

modalità, la distanza euclidea

$$d^2(i,i') = \sum_{j=1}^{c} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}} \right)^2 \tag{1.34}$$

rende si conto della similitudine tra due profili, ma non dell'importanza delle singole modalità. Per ovviare a questo inconveniente, si pondera ogni componente della sommatoria con un peso inverso alla massa della modalità corrispondente. La distanza tra due profili riga, così ottenuta è detta distanza del χ^2, ed è espressa da:

$$d^2(i,i') = \sum_{j=1}^{c} \frac{n}{n_{.j}} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}} \right)^2 \tag{1.35}$$

In questo modo si da un peso maggiore alle componenti a più bassa frequenza ridimensionando così quelle con le frequenze più elevate. Si deve ad ogni modo evitare di avere nella tabella modalità con frequenze eccessivamente basse in quanto avrebbero un peso eccessivo nel calcolo delle distanze risultando nel piano fattoriale con un'importanza eccessiva a dispetto della loro effettiva importanza nella spiegazione del fenomeno. Anche nell'Analisi delle Corrispondenze valgono le formule di transizione introdotte per il modello generale nella sezione 1.2.3, che in questo caso generano delle importanti relazioni dette **relazioni quasi baricentriche** che mostrano come la coordinata di una modalità su un generico asse, a meno di un fattore di scala pari all'inverso della radice quadrata dell'autovalore, sia una media, ponderata per le frequenze relative, delle coordinate delle modalità dell'altro carattere sullo stesso asse. Il termine corrispondenze sta ad indicare, come precedentente accennato, il fatto che l'analisi tende a mettere in corrispondenza tra loro quelle modalità che forniscono il maggior contributo alla relazione tra le due variabili. In conclusione se due modalità di una stessa variabile sono in posizione ravvicinata sul piano, significa che i due corrispondenti profili hanno una struttura simile, mentre una forte lontananza, in termini di opposizione rispetto all'origine indica una struttura nettamente diversa. Se due modalità delle due diverse variabili sono vicine sul piano, significa invece che esse si caratterizzano a vicenda. L'Analisi delle Corrispondenze può essere vista come un'Analisi in Componenti Principali che consideri le righe come unità e le colonne come variabili. Una volta definito il

miglior asse fattoriale, le coordinate degli individui si ricavano dal prodotto

$$\boldsymbol{c} = \mathbf{M}\boldsymbol{u} \tag{1.36}$$

mentre la coordinata di dell'i-esimo individuo si può analiticamente esprimere come

$$c_\alpha(i) = \sum_{j=1}^{c} \frac{n_{ij}}{n_{i.}} \frac{n}{n_{.j}} u_{\alpha j} \tag{1.37}$$

ossia come somma delle sue coordinate ciascuna ponderata con l'inverso della singola componente originaria. E' dunque un baricentro, la cui posizione sull'asse fattoriale $\boldsymbol{u}_\alpha$ è influenzata dall'importanza che nello specifico profilo hanno le modalità dell'altro carattere.

Una trattazione formalizzata

Precedentemente si sono elencati gli aspetti fondamentali sottostanti l'Analisi delle Corrispondenze. Si presenterà ora una trattazione formalizzata dei passi necessari per giungere ai risultati necessari per la comprensione del fenomeno studiato. Si indichino i profili riga con $\mathbf{P}_r$, i profili colonna con $\mathbf{P}_c$. Sia inoltre $\mathbf{D}_r = \text{diag}(...n_{i.}...)$ la matrice dei totali di riga e $\mathbf{D}_c = \text{diag}(...n_{.j}...)$ la matrice contenente i totali di colonna. La matrice dei profili riga assume allora la forma:

$$\mathbf{P}_r = \mathbf{D}_r^{-1}\mathbf{T} \tag{1.38}$$

mentre matrice dei profili colonna assume la forma:

$$\mathbf{P}_c = \mathbf{D}_c^{-1}\mathbf{T}' \tag{1.39}$$

Svolgere un'analisi delle corrispondenze sulla tabella $\mathbf{T}$ secondo l'ottica delle unità, equivale a svolgere un'Analisi in Componenti Principali, per ciascun carattere sull'insieme delle matrici ($\mathbf{X}$,$\mathbf{M}$,$\mathbf{D}$) opportunamente trasformate. Queste matrici assumono forme diverse per i due carartteri della tabella considerata. Per il carartteré A definito in $R^c = R^p$, si ha:

$$\begin{cases} \mathbf{X} = \mathbf{P}_r(r,c) = \mathbf{D}_r^{-1}\mathbf{T}' \\ \mathbf{M} = \mathbf{M}_r(c,c) = n\mathbf{D}_c^{-1} \\ \mathbf{D} = n^{-1}\mathbf{D}_r(r,r) \end{cases} \tag{1.40}$$

mentre per il cararttere B definito in $R^r = R^n$, si ha:

$$\begin{cases} \mathbf{X} = \mathbf{P}_c(c,r) = \mathbf{D}_c^{-1}\mathbf{T}' \\ \mathbf{M} = \mathbf{M}_c(r,r) = n\mathbf{D}_r^{-1} \\ \mathbf{D} = n^{-1}\mathbf{D}_c(c,c) \end{cases} \tag{1.41}$$

1.3.2 La distanza scelta

Come precedentemente specificato la distanza scelta è la distanza del χ^2, che tiene in considerazione sia la similitudine tra i profili, sia il peso di ciascuna modalità all'interno della tabella di contingenza. Pertanto la distanza tra due profili riga sarà data da

$$d^2(i,i') = \sum_{j=1}^{c} \frac{n}{n_{.j}} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}} \right)^2 \tag{1.42}$$

Mentre la distanza tra due profili colonna sarà:

$$d^2(j,j') = \sum_{i=1}^{r} \frac{n}{n_{i.}} \left(\frac{n_{ij}}{n_{.j}} - \frac{n_{i'j}}{n_{.j'}} \right)^2 \tag{1.43}$$

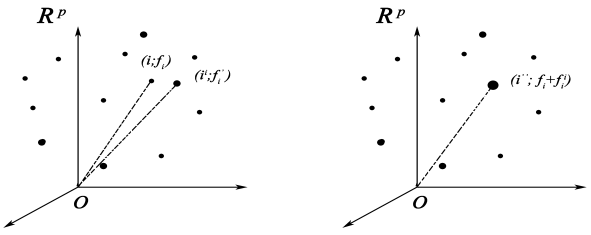


Figura 1.5: Rappresentazione grafica del principio dell'equivalenza distributiva

Si vuole mettere inoltre in evidenza un'importantissima, soprattutto ai fini del presente lavoro, proprietà di tale distanza. La proprietà dell'**equivalenza distributiva**. Dal punto di vista puramente applicativo, tale proprietà **permette di sommare profili simili**, sapendo che ciò non inficia significativamente la struttura delle distanze. Al contrario, la fusione di profili tra loro diversi, provocherebbe una netta perdita di informazione. Ciò **incoraggia la fusione delle modalità di scarso peso in altre modalità**, più importanti, purché aventi dei profili simili, si veda la figura 1.5, (Bolasco 1999).

1.3.3 La funzione obiettivo

Data la trasformazione effettuata sui dati da valori assoluti a profili, l'informazione è ora espressa mediante caratteri quantitativi. In tale contesto è pertanto possibile utilizzare il modello delle componenti principali. Per quanto concerne i profili riga, l'operatore di dispersione utilizzato per descrivere la somma ponderata delle distanze tra i punti e l'origine è espresso da:

$$\mathbf{X}'\mathbf{D}\mathbf{X} \tag{1.44}$$

Nel caso dell'Analisi delle Corrispondenze, il suddetto operatore viene espresso come $\mathbf{P}_r'\mathbf{D}\mathbf{X}$. Sviluppandosi l'analisi rispetto all'origine, si vuole rendere massima la proiezione di questa quantità in un opportuno sottospazio $\boldsymbol{u}$, con il vincolo $\boldsymbol{u}'\mathbf{M}\boldsymbol{u} = 1$, cioe:

$$\begin{cases} max_{(u)} \left\{ \sum_i p_i d_u^2(i, O) \right\} \\ \\ \boldsymbol{u}'\boldsymbol{M}\boldsymbol{u} = 1 \end{cases} \tag{1.45}$$

Applicando quanto detto nel caso dell'analisi generale, al caso dell'Analisi delle Corrispondenze la quantità da massimizzare riconduce alla ricerca degli autovalori di una matrice A del tipo $\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{M}$. Questa diventa:

$$\mathbf{P}_r'\mathbf{D}\mathbf{P}_r\mathbf{M}_r \tag{1.46}$$

Sostituendo le quantità precedentemente definite nelle formule 1.38 e 1.39, si ottiene

$$\mathbf{A} = (\mathbf{D}_r^{-1}\mathbf{T})' n^{-1} \mathbf{D}_r \mathbf{D}_r^{-1} \mathbf{T} n \mathbf{D}_c^{-1} \tag{1.47}$$

ed infine

$$\mathbf{A} = \mathbf{T}'\mathbf{D}_r^{-1}\mathbf{T}\mathbf{D}_c^{-1} \tag{1.48}$$

1.3.4 Rappresentazione dei punti nel sottospazio fattoriale

Nell'Analisi delle Corrispondenze, le coordinate dei punti unità su un generico asse fattoriale, sempre seguendo il modello generale, sono date da $c = \mathbf{XMu}$ per i profili riga esse assumono dunque la forma

$$c_\alpha(\mathbf{P}_r) = \mathbf{P}_r\mathbf{M}_r\boldsymbol{u}_\alpha = \mathbf{D}_r^{-1}\mathbf{T}n\mathbf{D}_c^{-1}\boldsymbol{u}_\alpha \tag{1.49}$$

Ricordando che $\mathbf{M}_r$ è una matrice diagonale con tutti gli elementi diversi da zero, si può esprimere analiticamente la coordinata dell'i-esimo elemento (carattere $\boldsymbol{A}$) come:

$$c_\alpha(i) = \sum_{j=1}^{c} \frac{n_{ij}}{n_{i.}} \frac{n}{n_{.j}} u_{\alpha j} \tag{1.50}$$

mentre le modalità del carattere $\boldsymbol{B}$, possono cosi essere espresse:

$$c_\alpha^*(j) = \sum_{i=1}^{r} \frac{n_{ij}}{n_{.j}} \frac{n}{n_{i.}} v_{\alpha j} \tag{1.51}$$

Le formule di transizione sono:

$$c_\alpha(i) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{c} \frac{n_{ij}}{n_{i.}} c_{\alpha j}^* \tag{1.52}$$

e analogamente

$$c_\alpha^*(j) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{r} \frac{n_{ij}}{n_{.j}} c_{\alpha i} \tag{1.53}$$

1.3.5 Valutazione dei risultati

Interpretazione statistica dell'inerzia

Nell'Analisi delle Corrispondenze la traccia della matrice da diagonalizzare $\mathbf{A}$, assume un particolare ed importante significato. L'inerzia totale dei punti rispetto all'origine, può essere cosi definita:

$$\Psi = \sum_{i=1}^{r} p_i d^2(i, O) = \sum_{j=1}^{c} p_j d^2(j, O) \tag{1.54}$$

Si ricorda inoltre che l'inerzia dei punti è uguale alla traccia della matrice **A**. Inoltre nel caso dei profili di riga si osserva che

$$tr(P_c P_r) = \sum_{\alpha=1}^{c} \lambda_\alpha = \sum_j \sum_i \frac{n_{ij}}{n_{.j}} \frac{n_{ij}}{n_{i.}} = \sum_i \sum_j \frac{n_{ji}^2}{n_{i.} n_{.j}} \tag{1.55}$$

Come si può facilmente osservare, questa quantità richiama decisamente l'indice χ^2, poiché

$$\sum \sum \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1 = \phi^2 = \frac{\chi^2}{N} \tag{1.56}$$

quindi

$$\sum \sum \frac{n_{ij}^2}{n_{i.} n_{.j}} = \phi^2 = \frac{\chi^2}{N} + 1 \tag{1.57}$$

Data la relazione esistente tra le componenti dei punti unità, cioè

$$\sum_j \frac{n_{ij}}{n_{i.}} = 1 \tag{1.58}$$

La nuvola dei punti appartiene in realtà allo spazio $\mathbf{R}^{c-1}$. Si tratta di un iperpiano ortogonale alla direzione che unisce l'origine al baricentro e che contiene sia l'intera nuvola centrata dei punti sia l'insieme degli assi fattoriali. L'inerzia della nuvola centrata è per costruzione nulla, mentre vale sempre uno l'entità della distanza del baricentro dall'origine. Tale inerzia unitaria è in pratica totalmente spiegata da qualsiasi sottospazio vettoriale. Pertanto nella traccia esiste sempre un autovalore uguale ad uno, detto autovalore banale, che viene trascurato poiché non apporta nessuna informazione utile al tipo di associazione esprimendo solo la distanza tra l'origine e il sottospazio in cui si trova la nuvola dei punti centrata al baricentro. Allora la somma degli autovalori si può cosi esprimere:

$$\sum_{\alpha=1}^{c} \lambda_\alpha = \lambda_1 + \sum_{\alpha=2}^{c} \lambda_\alpha = 1 + \sum_{\alpha=2}^{c} \lambda_\alpha \tag{1.59}$$

Si deduce pertanto che la traccia significativa della matrice $\mathbf{P}_c \mathbf{P}_r$ è uguale alla

misura dell'interdipendenza tra due variabili misurata dall'indice ϕ^2

$$\Psi_G = \sum_{i=1}^{c-1} \lambda_\alpha = tr(P_c P_r) - 1 = \frac{\chi^2}{N} \tag{1.60}$$

Quindi attraverso il valore della traccia significativa della matrice $\mathbf{P}_c\mathbf{P}_r$, è possibile risalire all'intensita della relazione tra i due caratteri.

$$n\Psi_G = \chi^2_T \tag{1.61}$$

Rapportando l'inerzia spiegata dai primi k assi dell'Analisi delle Corrispondenze all'inerzia totale, si ottiene una misura della capacità di questi assi nella spiegazione del fenomeno studiato

$$\frac{\sum_{i=1}^{k} \lambda_\alpha}{\sum_{i=1} \lambda_\alpha} \tag{1.62}$$

Qualità della rappresentazione delle modalità

La proiezione dei punti nel sottospazio ottimale crea pur sempre una distorsione. E' pertanto utile fornire una misura di quanto un punto è ben rappresentato sul piano fattoriale, ossia fornire una misura della qualità della sua rappresentazione. Questa misura è fornita dai cosiddetti contributi relativi

$$QLT_{F_1,F_2}(i) = \frac{\sum_{\alpha=1}^{2} c^2_\alpha(i)}{\sum_{\alpha=1}^{c-1} c^2_\alpha(i)} \tag{1.63}$$

Essendo i punti rappresentati sul piano attraverso una proiezione ortogonale, la norma riprodotta è funzione dell'angolo che il vettore originario forma con il sottospazio di riferimento che è funzione del coseno dell'angolo ed è pari al rapporto tra la norma riprodotta e quella originaria. La somma dei coseni quadrati di una modalità su tutti gli assi è uguale ad uno. Una modalità è ben rappresentata su un asse, quando il suo contributo relativo è alto. Bisogna peraltro rimarcare che questo valore dipende anche dal numero di assi e pertanto dal grado di compressione applicato all'informazione.

Contributo delle modalità alla costruzione di un fattore

Oltre ad essere ben rappresentata sul piano fattoriale, una modalità può contribuire in maniera più o meno marcata alla costruzione di un asse fattoriale. A tal

fine è di fondamentale importanza l'introduzione di una misura che permetta di valutare quanta parte ha avuto una data modalità nel determinare la direzione dell'asse fattoriale. Questa misura chiamata contributo assoluto, è data da:

$$CTA_\alpha(i) = \frac{p_i C_\alpha^2(i)}{\lambda_a} \tag{1.64}$$

Le modalità che presentano i contributi assoluti più alti sono quelle che maggiormente hanno contribuito ad orientare l'asse fattoriale. La somma dei contributi di tutte le modalità di un carattere sullo stesso asse vale necessariamente uno. Il contributo di una modalità può essere elevato sia a causa della massa elevata dell'elemento sia a causa della sua distanza dall'origine. La coordinata di una modalità a sua volta dipende dalla sua norma originaria ed è pertanto correlata con la sua qualità di rappresentazione. In conclusione i contributi relativi forniscono una misura di quanto una modalità è ben spiegata da un asse fattoriale mentre i contributi assoluti quanto una modalità contribuisce a spiegare un asse.

1.4 L'Analisi delle Corrispondenze Multiple

L'Analisi delle Corrispondenze Multiple (ACM) è l'estensione dell'Analisi delle Corrispondenze semplici (AC) allo studio simultaneo di più di due caratteri. L'interesse per questo tipo di analisi è dovuto alla possibilità di studiare simultaneamente sia caratteri quantitativi sia caratteri qualitativi. Il campo d'applicazione per eccellenza dell'ACM, indicato in letteratura è l'analisi di dati provenienti da questionari, anche se, in tempi più recenti si applica sempre più spesso anche a dati provenienti da archivi amministrativi. Il passaggio da una matrice eterogenea ad una matrice adatta all'applicazione dell'ACM, presuppone un processo di trasformazione dei caratteri quantitativi in variabili qualitative, suddividendolo in classi non vuote e la ricodifica delle modalità a bassissima frequenza in classi più ampie, compatibilmente con la proprietà dell'equivalenza distributiva (Lebart, Morineau, & Piron 1997). Terminata la fase di ricodifica delle variabili, si ottiene una matrice unità per variabili in forma di codifica ridotta, successivamente si passa alla matrice in forma disgiuntiva completa, ovvero una tabella booleana composta da s blocchi, tanti quante sono le variabili considerate. Questa matrice può essere vista come una particolare tabella di frequenza. Pertanto una volta trasformata la matrice disgiuntiva in matrici definenti i profili riga e i profili colonna si può applicare l'AC in modo da definire un sottospazio ottimale, secondo l'usuale criterio delle proiezioni orto-

gonali massimizzando il tasso di inerzia delle suddette proiezioni. La soluzione ottimale del problema è fornita dalla diagonalizzazione della matrice di Burt, costituita da s^2 tabelle doppie, che esprimono le tabelle di frequenza tra tutte le possibili coppie di variabili presenti nella matrice originaria. La matrice di Burt rappresenta l'insieme delle facce dell'ipercubo di contingenza. Rappresenta però solo le distribuzioni marginali doppie e semplici dell'ipercubo limitandosi insomma solo alle tabelle derivanti le interazioni di ordine zero e tralasciando le interazioni di ordine superiore. La denominazione di ACM sta infatti generalmente ad indicare lo studio simultaneo delle corrispondenze binarie tra modalità di variabili diverse. Le coordinate sugli assi fattoriali sono di notevolissima importanza dal momento che, in base alle proprietà quasi baricentriche illustrate sia nel modello generale che nel caso dell'AC, ogni unita si posiziona, a meno di un fattore di espansione, nel baricentro delle modalità che possiede mentre ogni modalità rappresenta un baricentro per gli individui che la possiedono.

1.4.1 Introduzione metodologica e passi dell'analisi

Descrizione formalizzata dell'analisi, definizione della notazione e introduzione delle matrici fondamentali

Si definiscano prelirminarmente la matrice in forma codificata ridotta $\mathbf{R}$ e la matrice in forma disgiuntiva completa $\mathbf{Z}$. La prima è una matrice in cui ciascuna riga rappresenta una osservazione e in cui ciascuna colonna rappresenta una variabile. Nella matrice in forma disgiuntiva completa ciascuna riga continua a rappresentare una osservazione mentre ciascuna colonna rappresenta una modalità. All'interno di ciascuna colonna di quest'ultima matrice sono contemplati due soli valori: uno nel caso l'unità considerata possieda la corrispondente modalità; zero altrimenti. Si consideri ora un insieme I composto da n osservazioni su s variabili. In questo caso la matrice in forma di codifica ridotta ha dimensioni $n \times s$. Sia q la generica variabile V^q e m_q il numero delle sue modalità, si ha allora che:

$$\sum_{q=1}^{s} m_q = p \tag{1.65}$$

per cui la matrice disgiuntiva completa ha dimensioni $(n \times p)$. La matrice disgiuntiva completa $\mathbf{Z}$ può allora essere vista come costituita da s blocchi di

variabili indicatrici la presenza o l'assenza per ogni osservazione della modalità ad essa associata

$$Z = \{\mathbf{Z}_1, \mathbf{Z}_2, ...\mathbf{Z}_q, ...\mathbf{Z}_s\} \tag{1.66}$$

Siano inoltre

$$z_{i.} = \sum_{j=1}^{p} z_{ij} = s \tag{1.67}$$

i totali di riga e

$$z_{.j} = \sum_{i=1}^{n} z_{ij} = n_{.j} \tag{1.68}$$

le osservazioni con modalità j corrispondenti all'elemento marginale semplice di colonna in ogni sottotabella $\mathbf{Z}_q$ e

$$z = \sum_{i=1}^{n} \sum_{i=1}^{n} z_{ij} = ns \tag{1.69}$$

il totale generale di $\mathbf{Z}$.
La matrice risultante dal sottostante prodotto, con generico elemento rappresentato tra parentesi

$$\mathbf{B} = \mathbf{Z}'\mathbf{Z} \text{ di elemento generico } \{...b_{ij}...\} = \left\{ ... \sum_{i=1}^{n} z_{ij} z_{ij'} ... \right\} \tag{1.70}$$

è detta matrice di Burt. La matrice di Burt è composta da tutti i possibili incroci ottenibili dalle variabili utilizzate nella matrice originaria mentre, nella diagonale principale si trovano altrettante matrici diagonali esprimenti le frequenze per ogni modalità. La matrice di Burt $\mathbf{B}$ ha dimensioni (p,p) e i suoi elementi si esprimono in funzione di quelli di $\mathbf{Z}$ nel modo seguente:

$$b_{jj'} = \sum_{i=1}^{n} z_{ij} z_{ij'} \tag{1.71}$$

$$b_j = \sum_{j'=1}^{p} b_{jj'} = s z_{.j} \tag{1.72}$$

$$b = s^2 n \tag{1.73}$$

L'Analisi delle Corrispondenze Multiple (ACM), può essere interpretata come un'Analisi delle Corrispondenze Semplice applicata alla matrice $\mathbf{Z}$, concepita come una particolare tabella di frequenza. Le matrici dei profili riga e colonna, assumeranno rispettivamente la seguente forma

$$P_r = \frac{1}{s}\mathbf{Z} \quad \text{e} \quad P_c = \mathbf{D}^{-1}\mathbf{Z}' \tag{1.74}$$

Le matrici di pesi, p_i o p_j sono i corrispondenti marginali relativi della matrice $\mathbf{Z}$

$$\mathbf{D}_r = p_i = \frac{s}{ns} = \frac{1}{n} = \frac{1}{n}\mathbf{I} \quad \text{e} \quad \mathbf{D}_c = p_j = \frac{z_{.j}}{ns} = \frac{\mathbf{D}}{ns} \tag{1.75}$$

Le distanze del χ^2 tra coppie di elementi assumono rispettivamente le seguenti forme

$$d^2(i,i') = \sum_{j=1}^{p} \frac{ns}{z_{.j}} \left(\frac{z_{ij}}{s} - \frac{z_{i'j}}{s} \right)^2 = \frac{1}{s} \sum_{j=1}^{p} \frac{n}{z_{.j}} \left(z_{ij} - z_{i'j} \right)^2 \tag{1.76}$$

$$d^2(j,j') = \sum_{i=1}^{n} \frac{ns}{z} \left(\frac{z_{ij}}{z_{.j}} - \frac{z_{ij'}}{z_{.j'}} \right)^2 \tag{1.77}$$

e generano come matrici che definiscono i relativi prodotti scalari, le seguenti metriche :

$$\mathbf{M}_r = ns \begin{vmatrix} \frac{1}{z_1} & & 0 \\ & \frac{1}{z_{.j}} & \\ 0 & & \frac{1}{z_{.p}} \end{vmatrix} = ns\mathbf{D}^{-1} \quad \mathbf{M}_c = ns \begin{vmatrix} \frac{1}{s} & & 0 \\ & \frac{1}{s} & \\ 0 & & \frac{1}{s} \end{vmatrix} = \left(\frac{1}{n}\mathbf{I} \right)^{-1} \tag{1.78}$$

1.4.2 Definizione della funzione obiettivo e rappresentazione nel sottospazio vettoriale

Dal modello generale dell'analisi fattoriale, si cerca di massimizzare la forma quadratica

$$\boldsymbol{u}'\mathbf{M}\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{M}\boldsymbol{u} \tag{1.79}$$

con il vincolo di normalizzazione del vettore $\boldsymbol{u}$. Come precedentemente illustrato

nel caso generale, la soluzione si ottiene attraverso un'equazione agli autovalori del tipo

$$\mathbf{A}\boldsymbol{u} = \lambda\boldsymbol{u} \tag{1.80}$$

La matrice $\mathbf{A}$ da diagonalizzare $\mathbf{X}'\mathbf{DXM}$ del modello generale modificata per rispondere alle esigenze dell'Analisi delle Corrispondenze semplici forniva la seguente espressione

$$\mathbf{A} = \mathbf{P}_r'\mathbf{D}_r\mathbf{P}_r\mathbf{M}_r \tag{1.81}$$

Mentre tradotta nelle notazioni appropriate all'Analisi delle Corrispondenze Multiple assume la forma seguente:

$$\frac{1}{s}\mathbf{Z}'\frac{1}{n}\mathbf{I}\frac{1}{s}\mathbf{Z}ns\mathbf{D}^{-1} = \frac{1}{s}\mathbf{Z}'\mathbf{Z}\mathbf{D}^{-1} = \frac{1}{s}\mathbf{B}\mathbf{D}^{-1} \tag{1.82}$$

infine:

$$\mathbf{A} = \frac{1}{s}\mathbf{B}\mathbf{D}^{-1} \tag{1.83}$$

dove il termine generico assume la forma:

$$a_{jj'} = \frac{1}{sz_{.j'}}\sum_{i=1}^{n} z_{ij}z_{ij'} \tag{1.84}$$

per cui, in $\mathbf{R}^p$, l'equazione agli autovalori in termini di assi fattoriali si esprime come:

$$\frac{1}{s}\mathbf{Z}'\mathbf{Z}\mathbf{D}^{-1}\boldsymbol{u}_\alpha = \lambda_\alpha\boldsymbol{u}_\alpha \tag{1.85}$$

Il vettore delle coordinate dei punti sul generico asse fattoriale risulta essere:

$$c_\alpha = \mathbf{X}\mathbf{M}\boldsymbol{u}_\alpha = \mathbf{P}_r\mathbf{M}_r\boldsymbol{u}_\alpha = \frac{1}{s}\mathbf{Z}ns\mathbf{D}^{-1}\boldsymbol{u}_\alpha = n\mathbf{Z}\mathbf{D}^{-1}\boldsymbol{u}_\alpha \tag{1.86}$$

Attraverso le usuali formule di transizione si possono ricavare le formule quasi baricentriche, che per le unità forniscono:

$$c_\alpha(i) = \frac{1}{\sqrt{\lambda_\alpha}}\sum_{j=1}^{p}\frac{z_{ij}}{z_{i.}}c^*_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}}\sum_{j=1}^{p}\frac{z_{ij}}{s}c^*_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}}\sum_{j\in m(i)} z_{ij}c^*_{\alpha j} \tag{1.87}$$

dove con *m(i)* si indicano le modalità possedute dall'individuo i-esimo. Quindi a meno di un coefficiente $\frac{1}{\sqrt{\lambda_\alpha}}$ l'individuo i-esimo si trova nel punto medio della nuvola delle modalità che da esso sono state scelte, in altre parole si trova nel baricentro dei suoi attributi. Per le modalità si ottiene invece

$$c_\alpha(j)^* = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{n} \frac{z_{ij}}{z_{.j}} c_{\alpha i} = \frac{1}{z_{.j}\sqrt{\lambda_\alpha}} \sum_{i \in I(j)} z_{ij} c_{\alpha i} \tag{1.88}$$

Dove $I(j)$ indica l'insieme di individui che corrispondono alla modalità j-esima. Ossia la modalità si trova nel baricentro degli individui che la possiedono.

1.4.3 Inerzia totale, tassi di inerzia e valutazione del risultato

La traccia della matrice A da diagonalizzare eguaglia l'inerzia della nuvola dei punti. Il numero massimo di autovalori estraibili nel caso dell'ACM è pari a *(p-s+1)* nel caso si esegua l'analisi della nuvola rispetto all'origine mentre sono *(p-s)* se si considera l'analisi rispetto al baricentro eliminando dunque l'autovalore banale. La nuvola *N(J)* è composta da s sottoinsiemi relativi ai diversi blocchi di $\mathbf{Z}$. Le componenti del baricentro della nuvola del q-esimo blocco di modalità valgono

$$G_{qi} = \sum_{j=1}^{m_q} \frac{z_{.j}}{n} \frac{z_{ij}}{z_{.j}} = \frac{1}{n} = G_i \tag{1.89}$$

per cui la distanza

$$d^2(j, G) = n \sum_{i=1}^{n} \left(\frac{z_{ij}}{z_{.j}} - \frac{1}{n} \right)^2 = \frac{n}{z_{.J}} - 1 \tag{1.90}$$

di una modalità dal baricentro è tanto maggiore quanto minore è la sua frequenza. L'inerzia di una modalità, o variabile indicatrice, è data da

$$I(j) = p_j d^2(j, G) = \frac{z_{.j}}{ns} \left(\frac{n}{z_{.j}} - 1 \right) = \frac{1}{s} - \frac{z_{.j}}{ns} = \frac{1}{s} \left(1 - \frac{z_{.j}}{n} \right) \tag{1.91}$$

ovvero l'inerzia della j-esima variabile indicatrice aumenta al diminuire di $z_{.j}$ e il suo massimo risulta pertanto uguale a *1/s*. L'inerzia del q-esimo carattere vale

$$I_q = \sum_{j=1}^{m_q} I(j) = \sum_{j=1}^{m_q} \frac{1}{s}\left(1 - \frac{z_{.j}}{ns}\right) = \frac{1}{s}(m_q - 1) \tag{1.92}$$

Quindi aumenta all'aumentare del numero di modalità. Infine l'inerzia complessiva della nuvola *N(J)* è funzione delle variabili e delle modalità, ovvero del numero medio di modalità, e non ha significato statistico:

$$I = \sum_q I_q = \sum_{j=1}^{p} \frac{z_{.j}}{ns} d^2(j, G) = \sum_{j=1}^{p} \frac{z_{.j}}{ns}\left(\frac{n}{z_{.j}} - 1\right) = \sum_{j=1}^{p}\left(\frac{1}{s} - \frac{z_{.j}}{ns}\right) = \left(\frac{p}{s} - 1\right) \tag{1.93}$$

In particolare, vale uno quando tutte le variabili hanno solo due modalità. Si pone allora il problema di come valutare il potere esplicativo degli assi fattoriali. In letteratura vengono indicate alcune strade da seguire. La prima parte dal presupposto della quantità di inerzia che può spiegare un asse, soppesando la soluzione in termini di distacco dal valore massimo che può assumere un autovalore. Ogni autovalore non può superare il valore massimo

$$\lambda_{max} = \frac{1}{traccia} \tag{1.94}$$

Nel caso di variabili con molte modalità il valore massimo che l'autovalore può assumere è sempre piuttosto basso per cui la valutazione della capacità esplicativa dell'analisi è sempre piuttosto pessimistica. La seconda strada suggerisce di considerare solo gli autovalori superiori all'autovalore medio, che risulta

$$\lambda_{medio} = \frac{\text{Inerzia totale}}{\text{n° di autovalori non banali}} = \frac{(p/s - 1)}{(p - s)} = \frac{1}{s} \tag{1.95}$$

Il metodo usualmente utilizzato consiste però nel rapportare semplicemente la variabilità spiegata dai primi k assi di interesse a quella totale. La percentuale di variabilità spiegata da un fattore è rappresentata dalla quantità

$$\frac{\lambda_\alpha}{\sum_{i=1}^{s-p} \lambda_\alpha} 100 \tag{1.96}$$

Questa quantità è come già detto, è una misura eccessivamente pessimistica dell'effettivo potere esplicativo dei fattori. Questo perché la codifica disgiuntiva impone una sfericità artificiale nella nube dei punti. Benzécri propone allora di rivalutare il tasso di inerzia mediante la quantità

$$\rho(\lambda) = \left(\frac{s}{s-1}\right)^2 \left(\lambda - \frac{1}{s}\right)^2 \quad \text{per cui} \quad \tau(\lambda) = \frac{\rho(\lambda)}{\sum_{\lambda > \bar{\lambda}} \rho(\lambda)} \tag{1.97}$$

Capitolo 2

Introduzione alla Ricodifica Sequenziale delle Modalità

2.1 Il contesto di riferimento

Come evidenziato nell'introduzione, il grande successo dell'Analisi Multidimensionali dei Dati (AMD) è dovuto da un lato alla grande disponibilità di dati oggigiorno fruibili, che tendono spesso a mettere in luce alcune lacune dell'analisi statistica classica di impianto probabilistico, e dall'altra la disponibilità di moderni calcolatori che rendono oggi applicabili metodologie che seppur da tempo conosciute restavano inapplicabili a causa dell'enorme sforzo computazionale necessario. Oggi sempre più, quasi tutte le metodologie vengono adattate per poter essere utilizzabili con enormi moli dati. Queste enormi moli di dati, non sempre provengono da questionari, ma sempre più spesso vengono estratte da database aziendali o amministrativi. Il tipo di dati usualmente utilizzato proviene dai più svariati campi come per esempio: transazioni d'affari; dati scientifici memorizzati negli archivi di laboratori specializzati; dati provenienti da registrazioni continue di fenomeni fisici; dati relativi agli accessi ai siti web. In tutti questi esempi l'eccessiva mole di dati rende quasi impossibile la loro analisi. È evidente la necessità di progettare strumenti appositamente studiati per superare queste difficoltà. Proprio in risposta a questa esigenze, si è sviluppato un nuovo filone chiamato Data Mining. Il termine Data Mining proviene dal paragonare il processo di estrazione di informazioni da enormi moli di dati al lavoro di estrazione di materiali svolto dai minatori nelle miniere. In questa accezione,

come da più parti sottolineato, sarebbe più opportuno parlare di estrazione della conoscenza piuttosto che estrazione di dati. L'obiettivo principale del Data Mining riguarda l'estrazione di conoscenza utile ed interessante (regole, pattern regolari, vincoli) da grandi quantità di dati presenti nei database, data warehouse o in altre strutture di memorizzazione. Interessanti in quanto le informazioni recuperate devono essere non banali, (ovvero non presenti nell'input) e devono portare all'ottenimento di nuova conoscenza, potenzialmente utile. Dal punto di vista della performance, gli obiettivi primari sono: efficienza computazionale; efficacia dei parametri che regolano il grado di interesse delle informazioni estratte; efficacia nel modo in cui i dati ricavati vengono presentati, in quanto i dati riguardanti la nuova conoscenza devono essere ben visibili ai possibili osservatori (quasi mai coincidenti con coloro che svolgono l'analisi). In questo contesto le tecniche di elaborazione dei dati, come per esempio il Data Mining, fanno parte di un processo più ampio di elaborazione dei dati detto Knowledge Discovery Process, il cui schema è riportato in Figura 2.1.

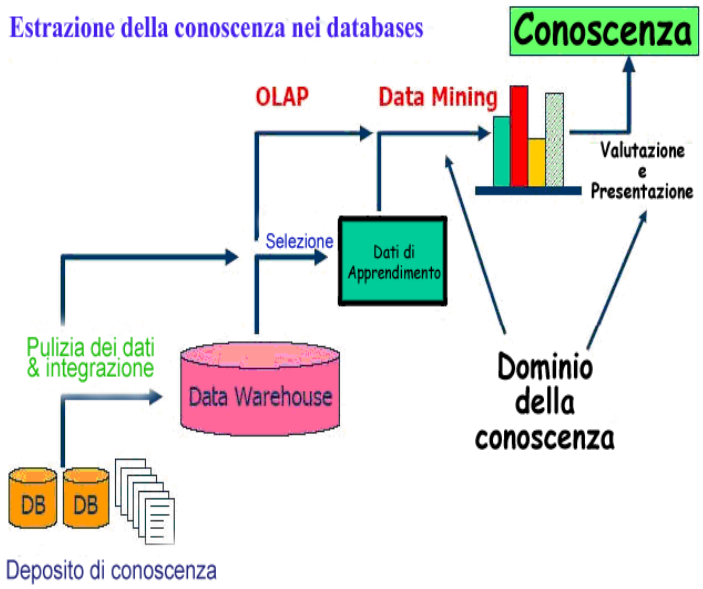


Figura 2.1: Knowledge Discovery Process nei database

Dall'analisi della figura 2.1, si possono evidenziare le fasi più importanti del processo di estrazione della conoscenza:

1. Selezione dei dati: ossia la selezione dei dati che sono maggiormente rilevanti per l'analisi

2. Pre-trattamento dei dati: rimozione del rumore e rimozione di informazioni inconsistenti, errate, inutili ed eventuale integrazione

3. Ricodifica dei dati: ossia la trasformazione più appropriata per i tipi di metodologie statistiche che s'intendono applicare

4. Data Mining: estrazione dei patterns dai dati applicando le opportune tecniche statistiche

5. Presentazione della conoscenza estratta: ossia presentazione all'utente finale delle più importanti relazioni trovate, in modo facilmente comprensibile, utilizzando opportune tecniche di visualizzazione dei risultati

Finora si è discusso dei problemi inerenti la moderna analisi dei dati imputandoli principalmente alla grande disponibilità di dati senza specificare in cosa consista esattamente e come influenzi le metodologie statistiche. Ovviamente un'analisi completa di tutte le possibili complessità che si possono verificare nel trattamento statistico dei dati e di tutte le distorsioni che vengono generate in tutte le metodologie statistiche sarebbe un compito piuttosto arduo. Ogni metodologia ha sue specifiche caratteristiche che possono risentire in maniera diversa a seconda di quanto i dati reali differiscano da quelli ideali prospettati da chi ha messo a punto, e migliorato nel tempo, una data metodologia. Nonostante queste considerazioni si possono identificare tre diverse determinanti che generano il cosiddetto problema computazionale:

1. problemi dovuti all'elevato numero di unità; problemi in n

2. problemi dovuti all'eccessivo numero di variabili; problemi in s

3. problemi dovuti all'eccessivo numero di modalità; problemi in p

Ancora una volta ognuno di questi tre aspetti potrebbe essere scomposto in altri sottoaspetti generanti altrettanti sottoproblemi. Si indicheranno pertanto quelli più ricorrenti prestando particolare attenzione a quelli che hanno maggiormente stimolato il seguente lavoro.

2.1.1 Problemi in n

Ovviamente il problema principale derivante dal dover elaborare matrici di dati con un elevato numero di osservazioni è un problema di puro calcolo, che trova limite solo nella capacità di memoria dell'elaboratore destinato all'elaborazione. Considerando anche solo il calcolo della correlazione tra due variabili, il dover analizzare un terabyte di dati può rappresentare, oggi, per la maggior parte degli elaboratori comunemente utilizzati, un problema insormontabile. Essendo unicamente un problema determinato dalla capacità di calcolo dell'elaboratore questo genere di problemi è solo marginalmente un problema statistico. Paradossalmente però, non prendendo ora in considerazione quanto appena detto, anche una situazione che dovrebbe rappresentare un chiaro miglioramento del contesto generale, maggiore disponibilità di osservazioni, ha mandato in crisi prassi ormai consolidate. Si prenda in considerazione l'ambito della statistica inferenziale, che si può concretizzare in due procedimenti: la stima dei parametri e la verifica delle ipotesi. La verifica delle ipotesi è un procedimento che consiste nel fare una congettura o un'ipotesi su un parametro θ e nel decidere, sulla base di un campione se è condivisibile o meno. Per decidere se essa sia accettabile o meno si utilizza una regola chiamata statistica test. Si consideri una semplice verifica di ipotesi sulla media con varianza nota. Siano

$$\bar{x} = 2 \quad \sigma = 0.27 \quad n = 20 \quad \mu^* = 1.8 \tag{2.1}$$

dove μ rappresenta il parametro dell'universo e μ^* il valore del parametro ipotizzato. Si vuole verificare se sia verosimile sull'evidenza dei dati campionari una media della popolazione uguale a μ data una media campionaria uguale a $\bar{x}$. La struttura della verifica di ipostesi assume pertanto la forma:

$$H_0 : \mu = \mu^* \quad H_1 : \mu > \mu^* \tag{2.2}$$

mentre il test utilizzato sarà:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{2.3}$$

Si supponga ora di lasciare invariati tutti i parametri del problema, di far variare unicamente n e verificare i risultati sia sul valore assunto dalla statistica test sia dal p-value. In tabella 2.1 sono riportati i risultati. Quando n è uguale a 10, il test oscilla sulla soglia della significatività, a seconda che si scelga come livello di significatività il 5 o il 10 percento. Quando n passa a 40 il test è comunque

significativo qualunque valore di significatività si scelga.

Tabella 2.1: Valori assunti dalla statistica test Z e dal p-value al crescere di n

n	z	p-value
10	1.75	0.040059
20	2.48	0.006569
40	3.51	0.000224
80	4.96	0.000003
160	7.02	0.000000
320	9.93	0.000000
640	14.05	0.000000
1280	19.87	0.000000
2560	28.10	0.000000
5120	39.75	0.000000
32400	100.00	0.000000

Ad un valore di n uguale ad 80, sia il valore del test che del p-value sono ancora commentabili: nel senso che si può ancora trovare qualche tavola cartacea in cui sia contemplato un valore pari a 4.96 e qualche software che mostri un numero di cifre decimali tali da non mostrare solo zeri. Da 160 fino 5120 il p-value ha perso ogni significato cosi come il valore del test, seppur ancora leggibile. Ad un valore di n uguale a 32400 il valore test diventa di tre cifre e non verrà più visualizzato da nessun software. Ovviamente già per un valore di n superiore a 160 diventa inutile il commento del valore test e quindi superflua la sua visualizzazione. La maggior parte delle metodologie statistiche si basa su almeno un test statistico, regressione lineare, modelli logit ed in generale tutta l'analisi multivariata. Matrici di dati di 5000 unità oggi non sono una rarità bensì quasi la norma. In simili situazioni un numero eccessivo di osservazioni crea forti problemi rendendo spesso inutilizzabili a fini decisionali le usuali procedure inferenziali.

2.1.2 Problemi in *s*

Da quanto detto nella sezione 2.1 e riportato in figura 2.1, si evince che qualunque indagine statistica deve essere alla fine divulgata. Da questo punto di vista, un numero elevato di osservazioni non crea particolari problemi. Un numero elevato di variabili può invece portare a vari inconvenienti. L'investigazione di un fenomeno dovrebbe partire, qualunque esso sia, con la descrizione del data-

set utilizzato. La descrizione più elementare che si possa fare, e generalmente si fa, consiste nella rappresentazione tabellare e grafica di ciascuna variabile. Successivamente si incrociano le variabili di maggiore interesse. Se si hanno a disposizione 50 variabili, non tantissime, il tempo necessario per la lettura, comprensione e valutazione di 50 tabelle univariate, e anche solo qualche decina di tabelle a doppia entrata, può richiedere svariate ore. Oltre alla complessità interpretativa, inoltre, si deve rimarcare che alcune metodologie soffrono particolarmente la presenza di un numero eccessivo di variabili. Nella regressione statistica l'aggiunta indiscriminata di variabili crea un R^2 artificialmente elevato che non corrisponde ad una effettiva capacità esplicativa del modello. Negli alberi di regressione può generare un eccessivo adattamento ai dati perdendo cosi il modello, la capacità di generalizzare le conclusioni oltre i dati analizzati. Nei metodi che prevedono la visualizzazione dell'output, si creano inevitabilmente problemi di visualizzazione dei risultati. La visualizzazione avviene di solito o su base cartacea o sul monitor dell'elaboratore. Quasi tutte le produzioni scientifiche hanno formati standard che difficilmente superano il formato 20×30 cosi come i monitor degli elaboratori raramente superano i 20 pollici. Se il numero di informazioni da visualizzare diventa eccessivo, rimanendo fisse le dimensioni dei supporti diventa problematica l'interpretazione dei risultati dell'elaborazione.

2.1.3 Problemi in p

Per quanto riguarda le modalità si può fare un discorso analogo a quanto fatto per le variabili. Un numero eccessivo di modalità crea sicuramente problemi al momento della divulgazione dei risultati. La statistica è pur sempre un momento di sintesi, e la sua ragion d'essere è l'incapacità della mente umana di sintetizzare immediatamente le informazioni complesse. Variabili con decine di modalità raramente riescono a fornire una visione immediata del fenomeno. Questo, in verità, è forse il minore dei mali e, se finissero qui, non ci sarebbero grossi problemi. Inconvenienti ben più gravi derivano dall'influenza di un eccessivo numero di modalità sulla stabilità delle metodologie statistiche. Sono ben noti i problemi, ad esempio, che variabili con troppe modalità creano agli alberi di classificazione, rendendo difficoltosa la loro interpretazione o quelli derivanti dal trattamento di variabili qualitative, attraverso le dummy, sui modelli di regressione o ancora sull'Analisi delle Corrispondenze Multiple. Ulteriori problemi derivano dal fatto che troppe modalità per ogni variabile tendono a frammentare eccessivamente il campione analizzato creando spesso sottocategorie vuote. In

ultimo, modalità con frequenze troppo basse, situazione quasi inevitabile quando si oltrepassa un certo grado di dettaglio nella rilevazione, rendono alcune metodologie statistiche poco robuste e favoriscono la presenza di outliers.

2.2 ACM, Knowledge Discovery e Data Mining

Per quanto detto nella sezione 1.4 del capitolo precedente, ossia la capacita di trattare simultaneamente caratteri di tipo qualitativo e quantitativo, l'assenza d'ipotesi distribuzionali, la facilità d'interpretazione dei risultati, l'Analisi delle Corrispondenze Multiple rimane oggi uno dei più importanti e utilizzati strumenti per l'analisi e la descrizione grafica di tabelle di contingenza multiple (Bolasco 1999). L'ACM permette la visualizzazione grafica sia delle unità sia delle variabili nei sotto-spazi ottimali, identificati attraverso le procedure indicate nel capitolo 1. Pur essendo possibile la rappresentazione sia dello spazio delle unità, sia dello spazio delle variabili, generalmente l'attenzione è concentrata sulla visualizzazione dei profili colonna ossia dello spazio delle variabili. Il problema della complessità computazionale illustrato nella sezione 2.1 affligge anche l'Analisi delle Corrispondenze Multiple. Se è vero che un grande numero di osservazioni può creare problemi unicamente per quanto riguarda la memoria dell'elaboratore, è anche vero che un numero eccessivo di modalità o di variabili può rendere piuttosto ardua l'interpretazione dei risultati o perfino falsarne il risultato finale. Per introdurre il problema, si prendano in considerazione alcune frasi enunciate dal prof. Michael Greenacre e tratte dal libro *Theory and application of Correspondence Analysis* (Greenacre 1984). *L'esempio più comune di matrice multidimensionale emerge dal risultato di un'indagine campionaria, dove I individui rispondono a Q domande di un questionario. Ci sono molti modi di condurre un'indagine, per esempio, una domanda potrebbe essere posta con un numero di risposte alternative dalle quali il rispondente deve selezionarne esattamente una. In qualche caso risulta difficile specificare preliminarmente tutte le possibili risposte, cosicché la domanda è lasciata aperta e una categorizzazione deve essere fatta dopo che il questionario è stato completato e studiato. Questa ultima strategia, è più problematica ed implica una grande mole di lavoro perfino prima che l'analisi statistica vera e propria cominci.* Inoltre come Han e Kamber hanno recentemente affermato (Han, Kamber 2001):*Oggigiorno, la nostra capacità di generare e raccogliere dati è incrementata rapidamente gra-*

zie all'informatizzazione di molte transazioni d'affari, scientifiche, governative e l'oramai comune utilizzo del World Wide Web. Tutti questi sistemi informativi, ci hanno inondati di un incredibile ammontare di dati. Questa esplosione di informazioni ha generato un urgente richiesta di nuove tecniche e strumenti automatizzati che possano, intelligentemente, assisterci nel trasformare questo vasto ammontare di dati in utili informazioni e conoscenza. Considerando queste premesse inerenti, le due fondamentali tipologie di dati che normalmente si incontrano ogniqualvolta si conduce un'analisi mediante l'ACM, che confermano quanto illustrato nella sezione 2.1, e integrando queste considerazioni con i principi che regolano l'ACM, si possono identificare gli inconvenienti che questi tipi di dati creano nell'Analisi delle Corrispondenze Multiple. Presupposto fondamentale dell'Analisi Multidimensionale dei Dati è che la matrice analizzata debba contenere un coerente ed *esteso* numero di variabili che sono essenziali per la comprensione del fenomeno investigato. Questo principio, oltre ad essere suffragato dall'intuizione è una chiara indicazione dei principi benzecriani che ispirano l'Analisi dei Dati 1.3. D'altra parte il tentativo di inserire più variabili possibili nell'ACM, o se si vuole di dimensioni, si scontra sia con l'evidenza pratica che fortemente sconsiglia l'inserimento di troppe variabili o di troppe modalità al momento di compiere l'analisi, sia con l'evidenza analitica.

Problemi in p

Secondo Lebart (Lebart et al. 1997) il giusto numero di modalità per ogni variabile dovrebbe essere compreso tra 3 ed 8, sia in modo da far si che sia possibile la comparazione dei diversi contributi delle diverse variabili, sia per evitare modalità con frequenze eccessivamente basse. Infatti, sia $I(q)$ l'inerzia della variabile q, essa è data da:

$$I_q = \sum_{j=1}^{m_q} I(j) = \sum_{j=1}^{m_q} \frac{1}{s}\left(1 - \frac{z_{.j}}{ns}\right) = \frac{1}{s}(m_q - 1) \tag{2.4}$$

L'inerzia di una variabile è quindi direttamente proporzionale al numero di modalità della variabile stessa, sezione 1.4.3. Questo è un risultato importante che deve essere tenuto in considerazione in quanto impone l'obbligo di evitare che ci sia un forte squilibrio di modalità tra le diverse variabili. Inoltre la distanza di una modalità dal baricentro è tanto maggiore quanto minore è la sua frequenza, si veda ancora la sezione 1.4.3. Da cui si ricava che l'inerzia di una modalità è tanto maggiore quanto più bassa è la sua frequenza:

$$I(j) = p_j d^2(j, G) = \frac{z_{.j}}{ns}\left(\frac{n}{z_{.j}} - 1\right) = \frac{1}{s} - \frac{z_{.j}}{ns} = \frac{1}{s}\left(1 - \frac{z_{.j}}{n}\right) \tag{2.5}$$

Si dovrà pertanto evitare la presenza di modalità con frequenze molto basse che potrebbero condizionare la direzione degli assi.

Problemi in s

La prima, e più ovvia considerazione, nasce dallo stretto legame esistente tra numero di variabili e numero di modalità. A meno di situazioni estreme all'aumentare del numero delle variabili, aumenta anche il numero delle modalità, per cui c'è una stretta relazione tra i due tipi di complessità: complessità in p e complessità in s. L'altra considerazione, quasi altrettanto ovvia, nasce dai problemi di visualizzazione, già accennati nella sezione 2.1, in cui si incorre se il numero di variabili simultaneamente visualizzate è eccessivo.

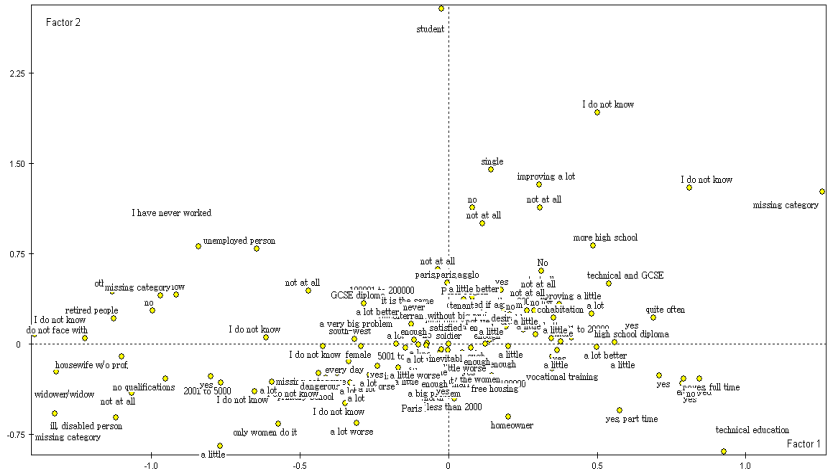


Figura 2.2: Rappresentazione fattoriale di un dataset composto da 29 variabili e 138 modalità.

Si consideri il seguente dataset composto da 29 variabili di 138 modalità complessive per una media di circa 5 modalità per variabile. Il dataset è sicuramente di modeste dimensioni, ma, come si evince dal grafico 2.2 è del tutto impossibile l'interpretazione di alcune parti del piano fattoriale a causa della sovrapposi-

zione delle etichette delle modalità. Non è difficile immaginare cosa succeda con datasets di maggiori dimensioni. Solitamente per ovviare a questo inconveniente, si preferisce analizzare i questionari per parti omogenee, in modo da evitare la sovrapposizione di troppi punti sul piano fattoriale e renderne più facile l'interpretazione. Questo procedimento pur facilitando non di poco l'analisi, ha però delle forti controindicazioni. Per convincersene è sufficiente ricordare i principi ispiratori dell'Analisi dei Dati introdotti nel capitolo 1.1 e qui riportati:

- **una visione esaustiva della struttura sottostante il fenomeno è possibile solo attraverso il trattamento simultaneo** delle informazioni inerenti il fenomeno stesso
- **elemento importantissimo è la rappresentazione grafica del risultato**, ottenuto attraverso le proprietà geometriche delle tecniche di analisi multidimensionale.

E' chiaro che una simile procedura, seppure spesso indispensabile, contraddice alla base due principi basilari della moderna analisi dei dati, la simultaneità del trattamento dell'informazione e la rappresentazione grafica del risultato. L'impossibilità di una chiara lettura del piano rende problematica la sua interpretazione, che cosi viene meno. Inoltre, viene violata anche una delle caratteristiche del Knowledge Discovery Process, ossia la possibilità di fornire all'utente finale una chiara ed immediatamente comprensibile interpretazione dei dati analizzati. Si può riassumere ora tutto il discorso mettendo in evidenza le similarità dei problemi e le possibili soluzioni. La matrice dei dati dovrebbe contenere un esteso e coerente numero di variabili necessario per la comprensione del fenomeno studiato. Riferendosi ai dati provenienti da databases, è chiaro che questi tipi di dati non sono solitamente concepiti a fini statistici quindi se è molto probabile che siano estesi, è molto meno probabile che siano coerenti con tutte le limitazioni sopra citate. Inoltre, è molto improbabile che la variabili provenienti da databases abbiano il giusto numero di modalità per condurre un'analisi attraverso l'ACM. A questo punto, **la domanda centrale è la seguente**: cosa succede se 8 modalità per una variabile di un questionario non sono sufficienti per investigare l'intero significato di un fenomeno o il numero di modalità di una variabile proveniente da un archivio è eccessivo? Nella prima situazione il ricercatore riduce il numero di possibili risposte della variabile del questionario (aggregando nella sua mente alcune modalità sulla base di qualche

criterio), nell'ultimo egli deve accorpare alcune modalità (sempre sulla base di un criterio) in modo da ridurre il loro numero. Quando si lavora con questionari, l'accorpamento ha luogo prima che il lavoro inizi, spesso è contestuale alla stesura del questionario stesso ed avviene, in molti casi, quasi inconsciamente. Quando, invece, si trattano dati provenienti da archivi, questa procedura avviene successivamente alla preparazione della base di dati da elaborare e comporta il sacrificio conscio di alcune modalità. Sostanzialmente la maggior parte di problemi identificati si può risolvere attraverso una riduzione delle modalità. Riducendo in modo ponderato le modalità:

- si evita la presenza di variabili con un numero di modalità troppo diverso che falserebbe l'analisi;
- si evita l'eccessiva frammentazione di una variabile evitando cosi di avere modalità con frequenze troppo basse;
- si evita di saturare il piano fattoriale di punti ottenendo una lettura ed interpretazione più immediata;
- si rende più semplice l'interpretazione anche ai non esperti favorendo la divulgazione e comprensione dei risultati;

E' chiaro quindi che una riduzione delle modalità attraverso il loro accorpamento avviene sempre e comunque. Potrà essere nelle mani dell'analista, potrà essere nella sua mente, ma una riduzione delle modalità avviene sempre. A questo punto la domanda è: **quale criterio dovrebbe essere usato per aggregare le modalità?**

2.2.1 Strategie e problematiche nella ricodifica delle variabili

Per approfondire questo importantissimo aspetto, si consideri la tabella 2.2, dove vengono riportate alcune possibili modalità della variabile professione. Si può immaginare che questi risultati provengano da un archivio, oppure che siano le possibili risposte che un ricercatore ha in mente di predisporre alla domanda di un questionario. Si supponga ora di volere, per i motivi precedentemente citati, ridurre il loro numero [1]. Il procedimento abitualmente utilizzato consiste

[1] In questo caso probabilmente non vi è necessità di una riduzione del numero di modalità. L'esempio è puramente "didattico", ma è facile intuire che il problema esposto è estendibile a qualunque situazione.

Tabella 2.2: Modalità della variabile professione, contributi assoluti, sul primo asse fattoriale ($\mathbf{CTA}_1$) e sul secondo asse fattoriale ($\mathbf{CTA}_2$)

Modalità	$\mathbf{CTA}_1$	$\mathbf{CTA}_2$
Ingegnere elettrico	?,??	?,??
Ingegnere edile	?,??	?,??
Commercialista	?,??	?,??
Infermiere	?,??	?,??
Medico	?,??	?,??
Geometra	?,??	?,??
Elettricista	?,??	?,??
Ragionerie	?,??	?,??
Altro	?,??	?,??

nel cercare tra le modalità a disposizione quelle tra loro più simili accorpandole. Ovviamente, la discussione sul concetto di similitudine è aperta ed il risultato sarà determinato in modo decisivo dalle conoscenze e convinzioni di chi si trova a dover decidere. Ad ogni modo, probabilmente in base alla tabella 2.2, si potrebbe decidere di accorpare, per esempio, *Ingegnere edile* e *Geometra* creando ***Professioni edili*** e *Medico* con *Infermiere* creando ***Professioni mediche***. In questo caso si è creata una nuova modalità utilizzando come criterio di accorpamento l'affinità tecnica dei settori lavorativi, tabella 2.3

Tabella 2.3: Possibile procedura di accorpamento per la variabile professione, contributi assoluti, sul primo asse fattoriale ($\mathbf{CTA}_1$) e sul secondo asse fattoriale ($\mathbf{CTA}_2$)

Modalità	$\mathbf{CTA}_1$	$\mathbf{CTA}_2$
Ingegnere elettrico	0.10	0.49
Professioni edili	0.16	0.34
Commercialista	1.02	0.31
Professioni mediche	1.94	0.03
Elettricista	0.05	0.74
Ragionerie	0.45	0.03
Altro	0.43	0.00
Totale	**4.15**	**1.94**

Un'altra strada potrebbe essere quella di accorpare le modalità secondo il criterio del grado di istruzione. In tal caso si potrebbero accorpare, *Ingegnere edile* e

Medico, ottenendo come nuova modalità ***Libero professionista*** e *Geometra* ed *Elettricista* ottenendo ***Tecnico diplomato***, tabella 2.4. Sia le considerazioni seguite per decidere il primo tipo di ricodifica che quelle seguite per il secondo tipo, appaiono altrettanto valide. Si potrebbe anche decidere di accorpare *Ingegnere edile* e *Ingegnere elettrico* creando come nuova modalità ***Ingegnere***, da contrapporre a ***Tecnico diplomato*** e a ***Professioni mediche***.

Tabella 2.4: Possibile procedura di accorpamento per la variabile professione, contributi assoluti, sul primo asse fattoriale ($\mathbf{CTA}_1$) e sul secondo asse fattoriale ($\mathbf{CTA}_2$)

Modalità	$\mathbf{CTA}_1$	$\mathbf{CTA}_2$
Ingegnere elettrico	0.14	0.41
Libero professionista	3.26	0.71
Commercialista	0.70	0.08
Infermiere	5.57	0.13
Tecnico diplomato	0.49	4.11
Ragionerie	0.32	0.09
Altro	0.27	0.05
Totale	**10.72**	**5.57**

E' possibile seguire tanti, apparentemente, validi criteri per decidere quale sia la migliore ricodifica, ed è altrettanto evidente come diversi tipi di ricodifiche determinino risultati differenti, influenzando pesantemente i risultati dell'analisi. Le considerazioni seguite per decidere il tipo di ricodifica da seguire, seppure apparentemente ineccepibili, soffrono dello stesso tipo di errore: non tengono in considerazione le altre relazioni sottostanti il fenomeno che si sta studiando. In altre parole sono avulse dal contesto e fatte a priori. E' possibile determinare un criterio in base al quale giudicare quale sia il miglior tipo di ricodifica svincolandosi da giudizi soggettivi? In assoluto probabilmente no! E' però sicuramente possibile fornire delle indicazioni basandosi su alcune considerazioni desunte da approcci epistemologici concretizzatisi nel tempo in metodologie statistiche. In maniera intuitiva si potrebbe argomentare dicendo che se il tema studiato riguardasse l'opinione sul proprio settore lavorativo, la prima ricodifica sarebbe la più opportuna. Se, invece, l'indagine si riferisse, ad esempio, ad opinioni sull'organizzazione scolastica ed universitaria, sembrerebbe più opportuna la seconda. All'aumentare della complessità del fenomeno però, la mente umana non riuscirebbe più a valutare tutte le possibili interrelazioni tra tutti i possibili aspetti

ed una simile strategia sarebbe inattuabile. Per ovviare a questa impossibilità si può utilizzare proprio l'Analisi delle Corrispondenze Multiple.

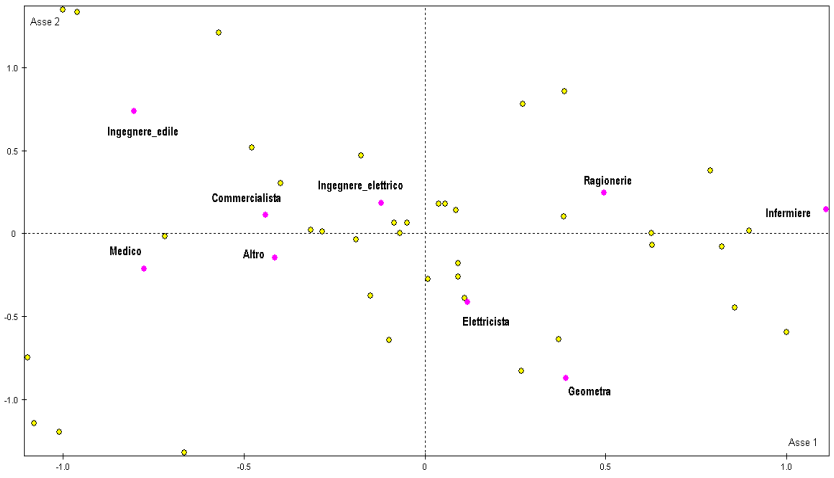


Figura 2.3: Rappresentazione fattoriale della variabile Professione.

L'ACM, attraverso la visualizzazione dei profili colonna individua esattamente quali modalità dello stesso carattere mostrano, rispetto a tutte le altre variabili considerate, un comportamento simile. Dall'analisi della figura 2.3 si evince che le modalità *Ingegnere elettrico*, *Ingegnere edile*, *Medico*, si trovano relativamente vicini sul piano fattoriale. Questo significa che queste tre professioni tendono ad avere, rispetto a tutte le altre variabili considerate, un comportamento simile. Stesso discorso si può fare per *Elettricist*a, e *Geometra*. Da queste considerazioni la ricodifica migliore appare quella basata sul grado di istruzione e non sull' affinità del settore lavorativo. Circostanza ancora più importante da sottolineare è che ai fini della ricodifica, seguendo questo approccio, appare di secondaria importanza il perché si verifica. Chiaramente è il frutto dell'interrelazione di tutte le variabili, per cui al limite si potrebbe coglierne un aspetto, che è appunto ciò che fanno i metodi fattoriali. Ignorare queste considerazioni e ricodificare le modalità secondo altri criteri, può portare a delle conseguenze talmente gravi da invalidare parte dell'analisi. Anzitutto una ricodifica basata sulle convinzioni personali del ricercatore, seppure in molti casi porti a degli ottimi risultati, lascia alcune perplessità. In primo luogo non appare indicata nei metodi fattoriali

per il semplice motivo che viola uno dei principi su cui tali analisi si basano. Secondo Benzécri i modelli devono seguire i dati e non viceversa. In questo contesto seguire i dati significa ricodificare le modalità sulla base delle risultanze del piano fattoriale e non sulla base di considerazioni di natura soggettiva. In questo secondo caso, infatti, sarebbero i dati a seguire il modello. Il modello mentale del ricercatore.

Tabella 2.5: Modalità originarie per la variabile professione, contributi assoluti, sul primo asse fattoriale ($\mathbf{CTA}_1$) e sul secondo asse fattoriale ($\mathbf{CTA}_2$)

Modalità	$\mathbf{CTA}_1$	$\mathbf{CTA}_2$
Ingegnere elettrico	0.14	0.40
Ingegnere edile	1.89	1.94
Commercialista	0.69	0.06
Infermiere	5.58	0.12
Medico	1.37	0.13
Geometra	0.59	3.62
Elettricista	0.04	0.62
Ragionerie	0.28	0.04
Altro	0.32	0.09
Totale	**10.90**	**7.01**

Quella che può apparire come una semplice dissertazione filosofica, in realtà ha delle conseguenze immediate, facilmente verificabili e soprattutto facilmente misurabili. La tabella 2.5 riporta le modalità originarie della variabile professione ed i relativi contributi assoluti sul primo e secondo asse del piano fattoriale riportato in figura 2.3. Per poter pienamente comprendere le conseguenze di una ricodifica che non tenga conto della relazione tra tutte le variabili, che sia cioè fatta a priori e non contestualizzata, si confrontino i contributi riportati nella tabella 2.5, ossia prima della ricodifica, con i contributi successivi alla prima ricodifica e riportati nella tabella 2.3. I contributi assoluti passano da 17.91 a 6.09. diminuiscono di oltre il 50 percento. Una variabile che in un primo momento portava un importante contributo alla costruzione degli assi, passa ad un ruolo decisamente secondario. Questa affermazione viene rafforzata dell'ispezione grafica delle nuove coordinate sul piano fattoriale. In figura 2.3 sono visualizzate le coordinate originarie della variabile. Le modalità appaiono ben distanziate e lontane dal baricentro a testimoniare l'importanza nell'analisi. Nella figura 2.4 sono visualizzate le nuove modalità. Come si evince, esse

appaiono molto più vicine al baricentro e molto meno distanziate tra loro. Un'obiezione che si potrebbe muovere a queste considerazioni, basandosi su quanto detto nella sezione 1.4.3 e ribadito nella sezione 2.2, è che l'inerzia sia calata a causa della diminuzione del numero delle modalità.

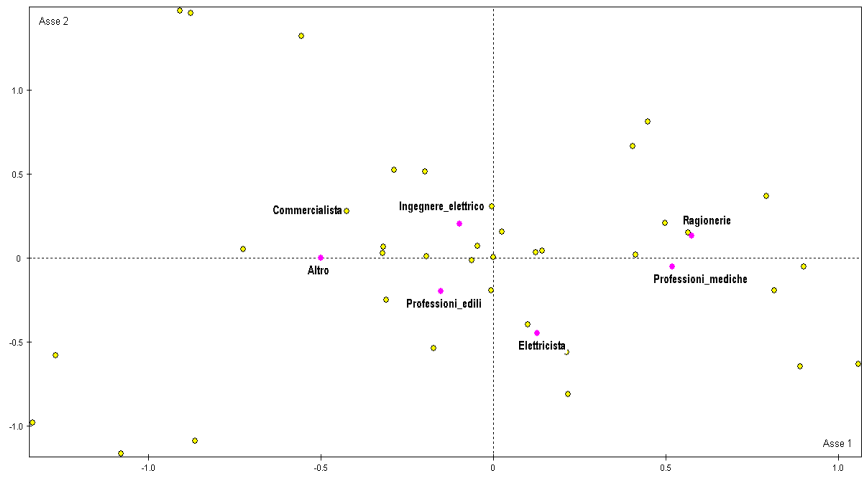


Figura 2.4: Rappresentazione fattoriale della variabile Professione, prima ricodifica

E' immediato verificare come la diminuzione dell'inerzia non sia imputabile semplicemente ad una diminuzione del numero di modalità. La tabella 2.4 riporta i contributi assoluti per la seconda ricodifica precedentemente illustrata e basata, seppur grossolanamente, sui risultati del piano fattoriale. I contributi sono si diminuiti, ma si è passato da un contributo originario, sui primi due assi, di 17.91 a 16.29. Una diminuzione irrisoria. Si ribadisce che l'accorpamento è stato fatto solo sulla base di una ispezione visiva del piano fattoriale. Come si vedrà più avanti quando si procede ad un calcolo esatto delle distanze, e le ricodifiche avvengono su questi presupposti, i contributi rimangono quasi identici. Anche la proiezione delle modalità sul piano fattoriale mette in evidenza la minore distorsione della nuova ricodifica. Esse appaiono ben distanti dal baricentro e ben distanziate tra loro, figura 2.5. Date tutte le considerazioni precedentemente fatte, si possono trarre alcune conclusioni ed indicare le linee guida da seguire per una procedura di ricodifica ottimale. Gli accorpamenti delle mo-

dalità possono portare, se non contestualizzati, ad una riduzione dei contributi assoluti di ciascuna modalità e di ciascuna variabile. Le motivazioni di questo fenomeno vanno ricercate nei motivi stessi che hanno ispirato la nascita dell'A-MD e cioè che il modello deve seguire i dati e non viceversa. La riduzione dei contributi assoluti, si verifica a causa del fatto che una procedura soggettiva di ricodifica viola il principio dell'equivalenza distributiva. Se i dati provengono da databases, ed il numero di modalità è eccessivo, sarà sufficiente verificare la loro posizione nel piano e comportarsi di conseguenza. Questo procedimento è però di più difficile attuazione quando si trattano questionari. In questo caso, l'accorpamento, come precedentemente affermato, avviene prima che l'analisi cominci, nella mente del ricercatore. La situazione è allora come in tabella 2.2, ossia non si conoscono i contributi e tanto meno le coordinate delle variabili, per cui è impossibile identificare le più simili. Si può seguire allora un altro tipo di procedimento. Nel caso si presentasse il problema del numero eccessivo di modalità, indicarne un elenco esteso e ridurre, solo successivamente il loro numero attraverso il procedimento indicato.

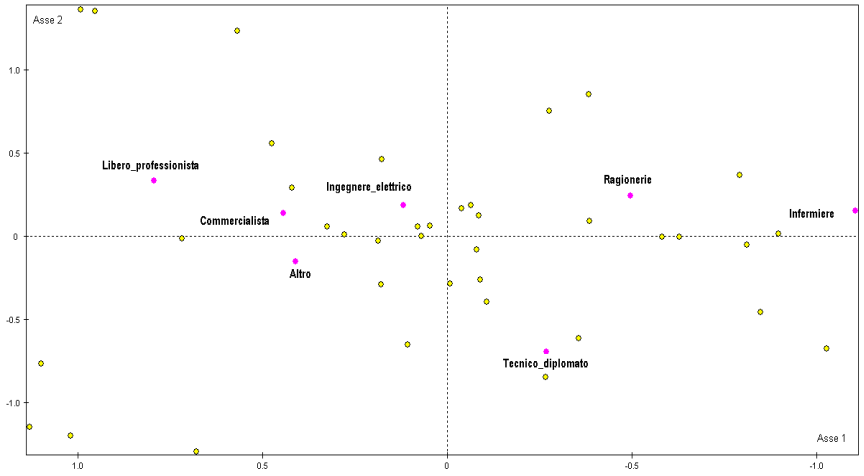


Figura 2.5: Rappresentazione fattoriale della variabile Professione, seconda ricodifica

Questo linea di comportamento, che come già detto, è uno dei principi ispiratori dell'Analisi Multidimensionale dei Dati, ha in realtà radici ancora più antiche.

Questo criterio riflette un importante e ben conosciuto punto di vista epistemologico. Attraverso le parole di Jules Henri Poincaré, fondatore della topologia algebrica (Poincaré 1905):

Lo scopo della scienza non sono le cose in loro stesse, come i dogmatici nella loro semplicità immaginano, ma le relazioni tra le cose; al di fuori di quelle relazioni non c'è nessuna realtà conoscibile.

Nelle situazioni fin qui affrontate, questo significa che le caratteristiche intrinseche delle modalità non devono influenzare la ricodifica, ma solo le relazioni tra loro esistenti dovrebbero essere considerate in questa procedura.

2.3 La Ricodifica Sequenziale Automatica (SAR)

2.3.1 Inroduzione

Tutti i problemi precedentemente affrontati possono essere cosi riassunti:

- la presenza di un gran numero di modalità provoca problemi di visualizzazione a causa della sovrapposizione dei punti e delle traiettorie sul piano fattoriale;
- è fortemente raccomandato che il numero delle modalità sia compreso tra 3 ed 8 anche in modo da evitare modalità con frequenze troppo basse (Lebart et al. 1997);
- l'aggregazione di modalità implica un'enorme mole di lavoro prima che l'analisi statistica vera e propria addirittura cominci (Greenacre 1984);
- l'aggregazione delle modalità dovrebbe essere fatta sulla base del fenomeno esaminato e non sulla base di preconcetti giudizi del ricercatore (Poincaré 1905);
- l'esplosiva crescita di dati ha generato un' urgente richiesta di nuove tecniche e di strumenti automatizzati che possano intelligentemente assistere il ricercatore nel trasformare il vasto ammontare di dati in informazioni utili e conoscenza (Han et al. 2001).

Allo scopo di superare questi problemi, si propone una Ricodifica Sequenziale Automatica (SAR), (Mascia et al. 2006), che ha come obiettivo principale la

riduzione delle modalità. SAR garantisce un'aggregazione automatica delle modalità, indipendente da giudizi soggettivi ed in grado di evitare un'enorme mole di lavoro in fase di ricodifica. E' importante sottolineare che SAR è totalmente svincolata da giudizi soggettivi e totalmente basata sulle relazioni risultanti dalla metodologia scelta. Questo è una delle ragioni che giustificano l'uso della Ricodifica Automatica Sequenziale come uno strumento che possa assistere il ricercatore nella trasformazione dei dati in conoscenza (Han et al. 2001). Finora si è affrontato il problema della complessità computazionale, scomponendola nei tre sottoaspetti fondamentali: elevato numero di osservazioni; elevato numero di variabili; eccessivo numero di modalità. Si è segnalato come la complessità computazionale affligga vari aspetti della statistica e varie metodologie. Si è anche mostrato come una riduzione delle modalità possa essere utile a superare alcuni di questi problemi e come una ricodifica sbagliata possa falsare i risultati dell'analisi. Si è infine proposta una Ricodifica Sequenziale Automatica per superare quest'ultimo aspetto. In linea di principio adattandola alle varie situazioni, SAR potrebbe essere applicata a qualunque metodologia statistica che soffra della presenza di un numero eccessivo di modalità. Nonostante qualche passo, ed anche con buoni risultati, sia stato fatto, (Mola, Mascia 2006), la maggior parte della metodologia è adattata all'Analisi delle Corrispondenze Multiple. Pertanto nella sezione seguente si descriverà SAR nei suoi elementi fondamentali, mostrando poi nel capitolo successivo come possa essere ulteriormente adattata per risolvere numerose problematiche nell'ambito dell'ACM.

2.3.2 La Ricodifica Sequenziale Automatica

Descrizione dei passi dell'algoritmo della SAR

La Ricodifica Sequenziale Automatica inizia con l'applicazione di un'Analisi delle Corrispondenze Multiple classica, utilizzando tutte le variabili presenti nella matrice che s'intende esaminare. Tra i risultati forniti dall'ACM, si focalizza l'attenzione su:

- percentuale di inerzia spiegata da ogni fattore;
- coordinate delle modalità per ogni variabile;
- contributo assoluto per ogni modalità.

SAR può allora essere cosi riassunta. Si consideri una matrice $\mathbf{X}$ con n righe e p variabili.

- ***Passo 1*** Le p variabili $(X_1, X_2, X_3, ...X_p)$ sono ordinate sulla base di un criterio ottenendo la matrice ordinata $\left(X_1^*, X_2^*, X_3^*, ...X_p^*\right)$. Si possono utilizzare diversi criteri di ordinamento: contributi assoluti delle variabili; contributi relativi, o, nel caso si voglia migliorare unicamente la leggibilità del piano fattoriale, nessun ordine. In questo caso si migliora la leggibilità del piano ma non si impone una gerarchia alle variabili e l'aggregazione avviene simultaneamente. Nel seguito, se non indicato diversamente si utilizzerà come criterio il contributo assoluto di ogni variabile, ottenuto come le somma dei contributi assoluti delle singole modalità. Questo passo permette di fornire una gerarchia alle variabili, dalla più importane a quella meno importante per la costruzione degli assi.

- ***Passo 2*** Viene selezionata la variabile col contributo assoluto più alto (X_1^*) e su di essa viene condotta una analisi dei gruppi gerarchica. Le modalità della variabile selezionata (X_1^*) rappresentano le osservazioni, mentre le coordinate sui prima K assi, opportunamente scelti, rappresentano le variabili.

- ***Passo 3*** Le modalità della prima variabile selezionata (X_1^*), accorpate sulla base dei risultati dell'analisi cluster, sono sostituite nella matrice originale da nuove modalità. Le nuove modalità possono assumere diverse forme e dipendono sia dal tipo di carattere utilizzato, da alcune scelte da compiere nella fase d' implementazione dell'algoritmo e da alcune circostanze contingenti. Si potranno quindi avere: classi nel caso il carattere sia numerico; nuove modalità nel caso le modalità unite presentino un'unitarietà di significato come nella sezione 2.2.1, oppure semplicemente unendo i termini delle modalità originali.

- ***Passo 4*** I dati vengono ri-processati attraverso una ACM. Si ottengono cosi, i nuovi contributi assoluti e le nuove coordinate per ogni modalità. Viene selezionata la variabile con il contributo assoluto più alto (X_2^*), escludendo ovviamente (X_1^*), e su di essa viene condotta un' analisi dei gruppi gerarchica. Ancora una volta le modalità di (X_2^*) rappresentano le osservazioni, mentre le coordinate sui primi K assi rappresentano le variabili. Le modalità della variabile selezionata (X_2^*), accorpate sulla base dei risultati dell'analisi dei gruppi, sono sostituite nella matrice originale da nuove modalità.

- ***Passo 5*** I dati vengono ancora una volta ri-processati dall' ACM (sul-

la matrice modificata) ottenendo i nuovi parametri dell'analisi. I passi dall'uno al cinque vengono ripetuti tenendo conto dei risultati precedenti, finche tutte le variabili originali non siano state ricodificate.

I risultati dell'applicazione dell'algoritmo sono due. Da un lato si ha una riduzione del numero delle modalità e dall'altro che, se si è scelto di ordinare le variabili in base ad uno dei criteri precedentemente esposti, questa gerarchia condiziona il processo di aggregazione delle modalità. Infatti, ad ogni ricodifica si ridefiniscono gli assi e conseguentemente si ottengono nuove coordinate per tutte le variabili, ovviamente anche per quelle non ancora ricodificate. Per cui il risultato dell'aggregazione delle ultime variabili, in base all'ordine scelto, dipende dai risultati delle aggregazioni delle variabili precedentemente ricodificate. In questo modo se alcune variabili presentano un numero eccessivo di modalità non apportando al contempo contributi importanti all'analisi, viene ridimensionata la loro influenza e conseguentemente il rumore causato. Inoltre, nel caso di numerose variabili causanti problemi di visualizzazione, queste verranno visualizzate per ultime 4.4.

Parametri dell'algoritmo

I risultati finali dipendono da alcune scelte preliminari che devono essere fatte al momento di implementare l'algoritmo. Per misurare la distanza tra due modalità l'algoritmo proposto considera la distanza euclidea pesata per i primi K assi:

$$d_{ij} = \left[(x_i - x_j)' \, \mathbf{W_K} \, (x_i - x_j)\right]^{1/2} \tag{2.6}$$

dove $\mathbf{W_K}$ è una matrice diagonale $K \times K$:

$$\mathbf{W_K} = \begin{pmatrix} \frac{1}{\lambda_1} & 0 & 0 & 0 \\ ... & ... & ... & ... \\ 0 & 0 & 0 & 0 \\ ... & ... & ... & ... \\ 0 & 0 & 0 & \frac{1}{\lambda_K} \end{pmatrix}$$

e contiene nella diagonale principale l'inverso dei primi K autovalori $\lambda_1, \lambda_i, ...\lambda_K$. In questo modo si attribuisce più importanza alle distanze tra due modalità nei primi assi rispetto ai sucessivi, (Zani 2000). Si considerino solo due assi, la

distanza totale tra due modalità, può essere scomposta come la somma della distanza calcolata sul primo asse più la distanza calcolata sul secondo asse. Si ipotizzi l'uguaglianza di queste due sottocomponenti, dividendole per i rispettivi autovalori, ed essendo il primo autovalore sempre maggiore del secondo, si ottiene che la distanza derivante dal primo asse diventa minore di quella derivante dal secondo. In questo modo, al momento del raggruppamento, due modalità risulteranno più vicine a causa dell'importanza maggiore attribuita al primo asse e verranno accorpate prima di due modalità che hanno, per ipotesi, la stessa distanza derivante però solo dal secondo asse. Per la scelta del numero di assi si seguono i principi usualmente utilizzati per la scelta del numero ottimo di fattori nell'analisi fattoriale, come per esempio lo scree test. L'aggregazione delle modalità avviene attraverso una classificazione gerarchica. L'algoritmo procede nel modo seguente: si fissa un numero minimo di modalità in modo tale che se una variabile ha un numero uguale od inferiore alla soglia stabilità non avvenga nessuna aggregazione. Sia H questo numero; se la variabile possiede più di H modalità, viene implementata una procedura di accorpamento attraverso una classificazione automaica, altrimenti la variabile non viene ricodificata.

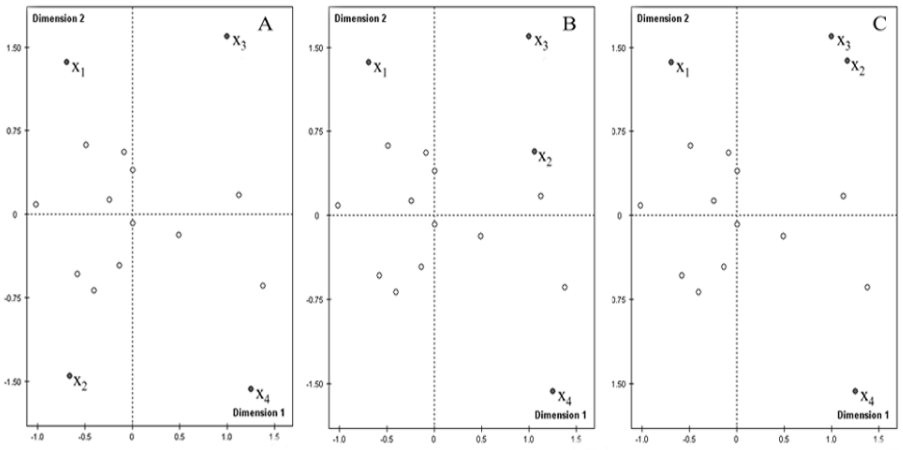


Figura 2.6: Rappresentazione fattoriale della distanza tra modalità: tre diverse situazioni

A questo punto si pone il problema del numero ottimo di modalità, ossia a quale livello tagliare il dendrogramma. Da un punto di vista empirico, a seguito di ripetute applicazioni, si è visto che che tagliare il dendrogramma al livello del

massimo salto di distanza produce risultati insoddisfacenti (Mascia, Mola 2006). Inoltre seguendo i criteri classici, non si può ottenere una sintesi graduale del fenomeno ma solo un unico grado di accorpamento. Si consideri la figura 2.6 (pannello A), se la procedura di clusterizzazione, individua tre gruppi è possibile ottenere l'accorpamento di due modalità tra loro molto diverse. In questa situazione, la riduzione delle modalità ottenuta è inconsistente. Emerge la necessità di poter decidere quando due modalità sono abbastanza vicine da poter essere accorpate e quando non lo sono. Per esempio si consideri una variabile con quattro modalità, x_1, x_2, x_3, x_4. Si potrebbe decidere che la distanza tra x_2 and x_3 nella figura 2.6 pannello C è abbastanza piccola da permettere la loro aggregazione. La decisione opposta potrebbe essere presa nella situazione riportata in figura 2.6 pannello B. Questo significa che l'analista fisserà una soglia per la distanza tra due modalità. L'aggregazione di due punti avviene unicamente se la distanza osservata tra questi due punti è al di sotto della soglia prefissata. La specificazione di questo tipo di soglia, supera due tipi di problemi: in primo luogo saranno permesse solo le aggregazioni significative tra due modalità; come secondo risultato, si rende possibile il raggiungimento di una sintesi graduale del fenomeno. Si riscontra, infatti, un *trade-off* tra soglia minima e numero di modalità accorpate. Riducendo la soglia aumenta il numero di modalità finali e viceversa. I risultati empirici sembrano dimostrare che un buon compromesso tra grado di sintesi, numero di modalità finali e percezione visiva è ottenuto con una soglia uguale al 30 percento della più grande distanza tra due modalità.

Capitolo 3

Applicazioni della Ricodifica Sequenziale delle Modalità

3.1 Introduzione

La Ricodifica Sequenziale Automatica, è una procedura generale che ha come finalità principale quella di ridurre la dimensionalità di una matrice di dati (si veda la sezione 2.3.1). Sempre nella stessa sezione, si è anche affermato come essa possa essere adattata per risolvere problematiche di diversa natura. Nelle sezioni che seguono saranno illustrate alcune di queste applicazioni.

3.2 La Ricodifica per la riduzione del numero di modalità

3.2.1 Per variabli di qualsiasi natura

Come ampiamente illustrato nei precedenti capitoli, quando si utilizza l'ACM in presenza di variabili con un numero eccessivo di modalità si incorre in problemi di visualizzazione a causa della sovrapposizione dei punti sul piano fattoriale e alla sovrapposizione delle traiettorie delle variabili ordinali. Attraverso l'applicazione di SAR, si ottiene un piano fattoriale più leggibile e la possibilità di

lavorare con variabili con un elevato numero di modalità. La procedura ricalca quasi totalmente quella generale, per cui verranno applicati i passi 1-2-3-4-5 elencati nella sezione 2.3.2. Gli unici cambiamenti riguardano la scelta di alcuni parametri inerenti l'algoritmo di aggregazione:

1. K=2; H=3
2. Metodo del legame medio
3. Soglia = 27.5 % della distanza massima tra le modalità di ciascuna variabile

Viene scelto H=3 in conformità a quanto sostenuto da Lebart, secondo il quale il giusto numero di modalità per ciascuna variabile dovrebbe essere compreso tra 3 ed 8, (Lebart et al. 1997). I motivi per cui si sono scelti solo i primi due assi, sono ampiamente discussi nelle conclusioni, 4.5. Per illustrare i risultati dell'applicazione si analizza un dataset raccolto nel contesto di un'indagine tendente a mettere in luce le caratteristiche di un gruppo di compagnie e dei loro proprietari. Il dataset consiste in 11 variabili e 72 modalità osservate su 200 compagnie e riportate nella prime due colonne della tabella 4.25.

Tabella 3.1: **Variabili**, numero di modalità prima dell'aggregazione (**NMPA**), Numero di modalità dopo l'aggregazione (**NMDA**)

Variabili	**NMPA**	**NMDA**
Legal organization	4	3
Profits	6	4
Trend market	3	3
Market	4	3
Kind of market	4	4
Generation of firm	6	3
Idea	5	3
Former occupation	10	5
Industrial category	9	5
Motivations	9	4
Barriers	12	6
Totale	**72**	**43**

Il dataset analizzato è tutto sommato di modeste dimensioni, ma è gia possibile notare, analizzando il piano fattoriale riportato nella figura 3.1 come esso ap-

paia poco leggibile. I primi 4 assi spiegano il 17.35 % dell'inerzia totale, come riportato nella quarta colonna della tabella 4.39 (a).

Tabella 3.2: Risultati numerici del'ACM prima dell'applicazione della SAR (a) e dopo l'applicazione della SAR (b): Autovalori, percentuale di inerzia spiegata e percentuale cumulata di inerzia spiegata

(a)				(b)			
N	Autovalori	Inerzia	Cum	N	Autovalori	Inerzia	Cum
1	0.284	5.13	5.13	1	0.276	9.20	9.20
2	0.250	4.50	9.63	2	0.242	8.07	17.27
3	0.219	3.95	13.58	3	0.172	5.75	23.02
4	0.209	3.77	17.35	4	0.159	5.31	28.33
5	0.203	3.66	21.02	5	0.156	5.19	33.52
6	0.199	3.58	24.60	6	0.152	5.05	38.57
7	0.195	3.51	28.11	7	0.139	4.64	43.20
8	0.190	3.43	31.53	8	0.131	4.36	47.57
9	0.181	3.27	34.80	9	0.120	4.01	51.57
10	0.175	3.16	37.96	10	0.118	3.94	55.51
..	.....	....	.....	..	.....	....	.....
61	0.005	0.10	100.00	33	0.018	0.61	100.00

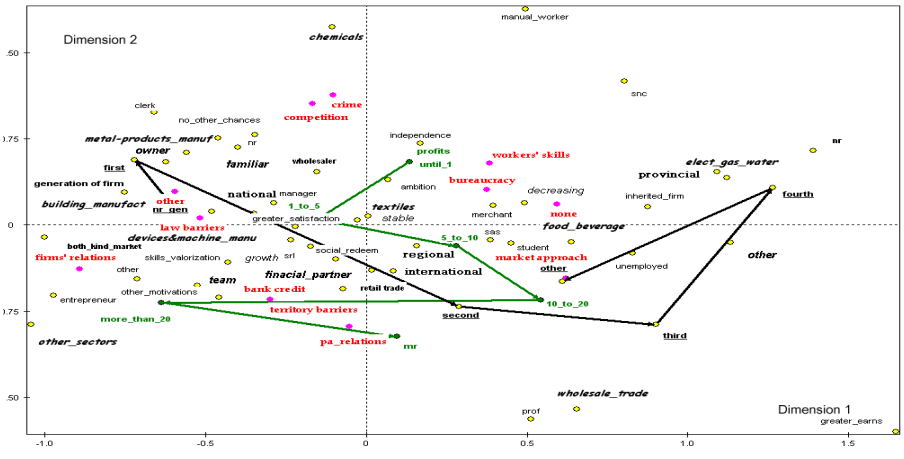


Figura 3.1: Visualizzazione dei profili colonna rispetto al primo piano fattoriale prima della ricodifica

A seguito dall'applicazione della metodologia proposta, il numero di modalità viene ridotto da 72 a 43, come riportato nella sezione (b) della tabella 4.39, mentre la percentuale di inerzia spiegata dai primi 4 fattori sale al 28.33%. Le categorie "ridotte"ammontano a 29 mentre il guadagno di inerzia spiegata è del 10.98% nei primi quattro assi. Le figure 3.1 e 3.2 riportano il primo piano fattoriale prima e dopo la procedura di ricodifica. Il confronto tra i due piani fattoriali mostra, nel secondo caso, una maggiore leggibilità permettendo una più facile lettura dei risultati.

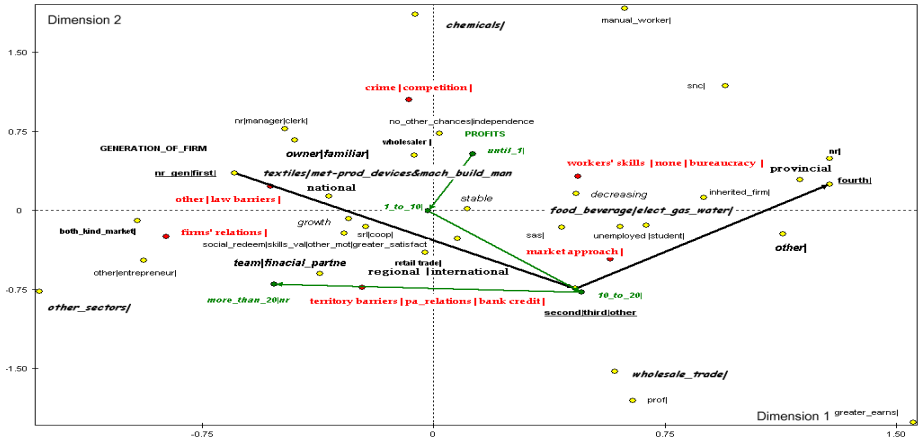


Figura 3.2: Visualizzazione dei profili colonna rispetto al primo piano fttoriale dopo la prima ricodifica

Allo scopo di migliorare ulteriormente la leggibilità del piano fattoriale, è possibile compiere un ulteriore passo. Si consideri la variabile *Motivazione*. Le modalità prima della ricodifica sono:
Ambizione; Maggiori guadagni; Maggiori soddisfazioni; Indipendenza; Società ereditata; Valorizzazione delle proprie competenze; Riscatto sociale; Assenza di alternative; Altre motivazioni. A seguito della ricodifica, esse diventano:
Maggiori guadagni;
Società ereditata;
Assenza di alternative-Indipendenza-Ambizione;
Riscatto sociale- Valorizzazione delle proprie competenze- Altre motivazioni-Maggiori soddisfazioni;
La modalità *Riscatto sociale-Valorizzazione delle proprie competenze- Altre motivazioni-*

Maggiori soddisfazioni, rappresenta motivazioni legate alla sfera emozionale, opposta per esempio a *Maggiori guadagni* rappresentante aspetti materiali. In situazioni simili, è possibile rinominare questa categoria come *Motivazioni emozionali* e sostituirla nella matrice originale. Quest'ultima procedura è stata implementata per alcune modalità ed i risultati sono riportati nella figura 3.3.

Figura 3.3: Visualizzazione dei profili colonna rispetto al primo piano fattoriale dopo la seconda ricodifica

L'analisi del piano fattoriale mostra un'ulteriore leggibilità ed una ancora più facile interpretazione.

3.2.2 Per variabli ordinabili

La procedura appena esposta può essere applicata a variabili misurate su ogni tipo di scala, ossia scala nominale, scala ordinale, scala ad intervalli e scala a rapporti. In questo caso però non si terrebbe conto della specificità del tipo di variabile. Per tenere nel giusto conto questi aspetti, la procedura precedentemente illustrata deve essere leggermente modificata. L'aspetto peculiare, quando si trattano variabili ordinali, consiste nel fatto che non ha molto significato l'aggregazione di due modalità non contigue. Si consideri la variabile grado di istruzione, con modalità: *NT* (nessun titolo); *LE* (licenza elementare); *LM* (Licenza media); *D* (Diploma); *L* (laurea). Una ricodifica del tipo: *LM-L*, *LE-D*; *NT* non avrebbe molto significato. Si impone allora un vincolo d'ordine in

modo tale che possano avvenire aggregazioni solo di modalità vicine. In realtà le uniche modifice che si apportano, riguardano il tipo di legame e la matrice delle distanze. Si utilizza il legame singolo al posto del legame medio, mentre per quanto concerne la matrice delle distanze, una normale matrice delle distanze contiene le distanze tra tutte le possibili coppie di modalità. La matrice delle distanze nel caso di variabili ordinali con vincolo d'ordine, assume invece la forma riportata nella figura 3.3, ossia vengono considerate solo le distanze tra modalità attigue.

	A	B	C	D
A	0	d_{ab}	0	0
B	0	0	d_{bc}	0
C	0	0	0	d_{cd}
D	0	0	0	0

Tabella 3.3: *Matrice delle distanze*

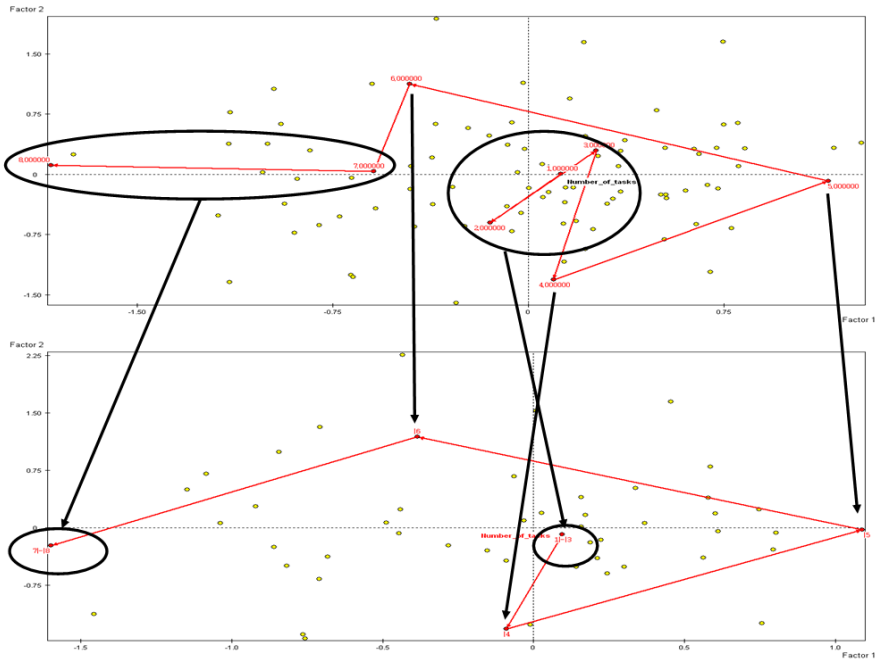
Al momento di aggregare le modalità, l'algoritmo considererà solo le distanze diverse da zero. La prima riduzione avverrà pertanto tra le due modalità che: a) sono contigue; b) hanno la distanza minore rispetto a tutte le altre. Ovviamente si può rinunciare a tale opzione e considerare la variabile come non regolata da nessun ordine.

3.2.3 Per variabli numeriche

Per le variabili numeriche si pongono le stesse problematiche illustrate nel caso di variabili ordinali. Si consideri una ipotetica variabile numerica **X** con modilità: *1,2,3,4,5,6,7,8*. Una ricodifica del tipo: *1-5*; *3-7*; *4-6*, ossia classi sovrapposte, non ha nessun significato statistico. Ancora una volta si impone un vincolo di contiguità come illustrato nella figura 3.3, ed il principio di aggregazione in questo caso è identico a quello precedentemente illustrato. La differenza principale, riguarda le etichette da assegnare alle nuove modalità. Si consideri ancora la variabile **X**, e si supponga che la ricodifica effettuata dall'algoritmo sia: $\{1,2,3\}$; $\{4\}$; $\{5\}$; $\{6\}$; $\{7,8\}$. In questo caso appare inopportuno e superfluo elencare tutte le modalità, apparendo più idoneo indicare unicamente gli estremi. Le nuove etichette saranno pertanto :$\{1-3\}\,;\{4\}\,;\{5\}\,;\{6\}\,;\{7-8\}$. Si ottiene in questo modo una ricodifica automatica in classi, si veda la figura 3.4. Questo tipo di ricodifica, ha un' importante proprietà: non è una ricodifica a priori ma totalmente basata sui risultati dell'ACM. Nella prossima sezione verrà illustrata

la procedura per la ricodifica di variabili continue. Per non generare confusione tra la procedura appena illustrata e la successiva, è opportuno indicare quale criterio si è utilizzato per distinguere una variabile numerica da una continua. Si sono considerate numeriche quelle variabili composte da numeri e che abbiano un numero di modalità ristretto. Questa distinzione è importante perché impone un'ulteriore modifica alla procedura che verrà qui di seguito illustrata.

Figura 3.4: Esempio di ricodifica di una variabile numerica.



Le variabili continue, presentano un numero enorme di modalità, verosimilmente tutte, o quasi, con frequenze minori della soglia solitamente imposta per rendere l'analisi robusta. Se venissero inserite direttamente nella matrice dei dati, queste modalità peserebbero in modo eccessivo sull'analisi. Si rende pertanto necessaria una variazione metodologica.

3.3 La Ricodifica di variabili continue

E' noto che per il trattamento delle variabili continue nell'ACM, si possono seguire due strade (Lebart et al. 1997). La prima consiste nell'utilizzare le variabili continue come illustrative mentre la seconda consiste nel rendere discreta

la variabile attraverso una suddivisione in classi. L'importanza della variabile continua ai fini dell'analisi è ottenuta attraverso la correlazione della stessa con gli assi fattoriali. Gli inconvenienti consistono nella possibilità di quantificare solo il grado di correlazione lineare e soprattutto di poter utilizzare la variabile solo come illustrativa. Il secondo procedimento invece, risente delle scelte soggettive del ricercatore, questo comporta che diverse suddivisioni possono portare a diversi risultati dell'analisi (Greenacre, 1984). Un altro inconveniente consiste nella circostanza che la suddivisione in classi è effettuata a priori e non tiene in considerazione il fenomeno oggetto di studio. Da ultimo, se le variabili continue sono numerose, questo procedimento richiede un'enorme mole di lavoro. La ricodifica sequenziale automatica delle modalità (Mascia and Mola, 2006), inizia con l'applicazione di una ACM classica utilizzando tutte le variabili presenti nel dataset da analizzare. Tra i risultati forniti dall'analisi, l'attenzione è focalizzata sui seguenti indicatori:

- Percentuale d'inerzia spiegata da ogni fattore;
- Coordinate delle modalità per ogni variabile;
- Contributi assoluti delle modalità;

La ricodifica automatica può essere cosi riassunta. Si consideri una matrice $\mathbf{X}$ con n righe e p variabili e per brevità si consideri il caso di una sola variabile continua $\mathbf{X}_c$ con k modalità $(x_1, x_2, ...x_k)$.

- ***passo 1*** Si conduce un'ACM classica sulle variabili nominali e si proiettano le modalità della variabile continua, considerate anch'esse come modalità di una variabile nominale, in supplementare sul piano fattoriale. Successivamente, si conduce un'analisi di raggruppamento (*cluster analysis*) sulla variabile. Le modalità sono considerate come osservazioni mentre le coordinate sugli assi rappresentano le variabili. Di ciascun gruppo, si calcolerà il valore medio delle modalità appartenenti al gruppo stesso. La variabile continua originaria è sostituita da una nuova variabile nominale con numero di modalità uguale al numero di gruppi scelti e costituita dalle medie dei gruppi. Nell'ipotesi che si ottengano θ gruppi si avrà la seguente sostituzione:

$$\underbrace{x_1, x_2, x_3}_{\overline{x1}} \quad \underbrace{x_4, x_5, x_6}_{\overline{x2}} .... \underbrace{x_7, x_8, ...x_k}_{\overline{x\theta}} \tag{3.1}$$

Al posto di $\mathbf{X}_c$ originariamente con k modalità $(x_1, x_2, ...x_k)$ si sostituisce $\mathbf{X}_c^{\circ}$ con θ modalità con $\theta < k$ Successivamente $\mathbf{X}_c^*$ sarà utilizzata come una variabile nominale.

- ***Passo 2*** Le p variabili $\left(X_1, X_2, X_c^{\circ}, ...X_k\right)$ sono ordinate sulla base dei contributi assoluti di ciascuna variabile; dalla variabile con il contributo più alto a quella con il contributo più basso, ottenendo il dataset $\left(X_1^*, X_2^*, X_c^{\circ *}, ...X_k^*\right)$.

- ***Passo 3*** Viene presa in considerazione la variabile con il più alto contributo assoluto (X_1^*) e su di essa viene condotta una analisi dei gruppi. Le modalità della variabile (X_1^*) rappresentano le osservazioni mentre le coordinate sui primi K assi, opportunamente scelti, rappresentano le variabili.

- ***Passo 4*** Le modalità della prima variabile selezionata (X_1^*), aggregate in base ai risultati dell'analisi, sono sostituite nel dataset originario con nuove modalità, rimpiazzando quelle di (X_1). Come primo risultato è importante evidenziare che le nuove modalità, sono in numero inferiore rispetto a quelle originarie.

- ***Passo 5*** Un'ACM classica viene condotta sul dataset modificato a seguito della sostituzione delle modalità originarie, ottenendo i nuovi contributi assoluti per ciascuna variabile. I passi dal secondo al quinto sono ripetuti fino a che tutte le variabili sono state ricodificate. Come risultato si ottiene un numero inferiore di modalità e che le coordinate delle variabili meno importanti sono forzate da quelle delle variabili più importanti nell'analisi.

I risultati dipendono da alcune scelte che il ricercatore deve compiere come per esempio il numero di assi, la distanza o la procedura di raggruppamento. Come misura di distanza tra due categorie, l'algoritmo proposto utilizza la distanza euclidea classica pesata con gli autovalori corrispondenti ai primi 2 assi:

$$d_{ij} = \left[(x_i - x_j)' \mathbf{W}_2 (x_i - x_j)\right]^{\frac{1}{2}} \tag{3.2}$$

dove

$$\mathbf{W}_2 = \left| \begin{array}{cc} \frac{1}{\lambda_1} & 0 \\ 0 & \frac{1}{\lambda_2} \end{array} \right| \tag{3.3}$$

contiene sulla diagonale principale l'inverso dei primi due autovalori $\frac{1}{\lambda_1}$ e $\frac{1}{\lambda_2}$. L'aggregazione delle modalità è ottenuta attraverso una classificazione gerarchica basata sul metodo del legame medio.

La procedura proposta permette di inserire variabili continue nel processo di ricodifica automatico tendente alla riduzione delle modalità nell'ACM. I vantaggi della procedura proposta consistono nel poter trattare anche variabili continue, di poterle utilizzare come attive, nell'eliminazione della soggettività della riduzione in classi e dalla totale automatizzazione della procedura che snellisce il lavoro del ricercatore nel caso di numerose variabili continue.

3.4 La Ricodifica per le modalità con bassa frequenza

3.4.1 Introduzione

Un altro grande problema, come ampiamente evidenziato nei paragrafi precedenti, è dovuto alla presenza di modalità con frequenze eccessivamente basse. Questa situazione si risolve generalmente grazie all'assegnazione casuale delle osservazioni appartenenti a quelle modalità, privando però in questo modo queste categorie del loro ruolo attivo e utilizzandole come supplementari. Per ovviare a questo inconveniente, si può utilizare ancora una volta la Ricodifica Sequenziale Automatica adattandola alla specificità del problema, (Mascia 2006). In questo caso lo scopo è quello di evitare l'assegnazione casuale e di mantenere le modalità come attive, attraverso l'assegnazione delle modalità a più bassa frequenza a quelle di frequenza maggiore (Bolasco 1999). In realtà, come si vedrà più avanti, le modalità sono meglio definibili come *Semi-Attive*.

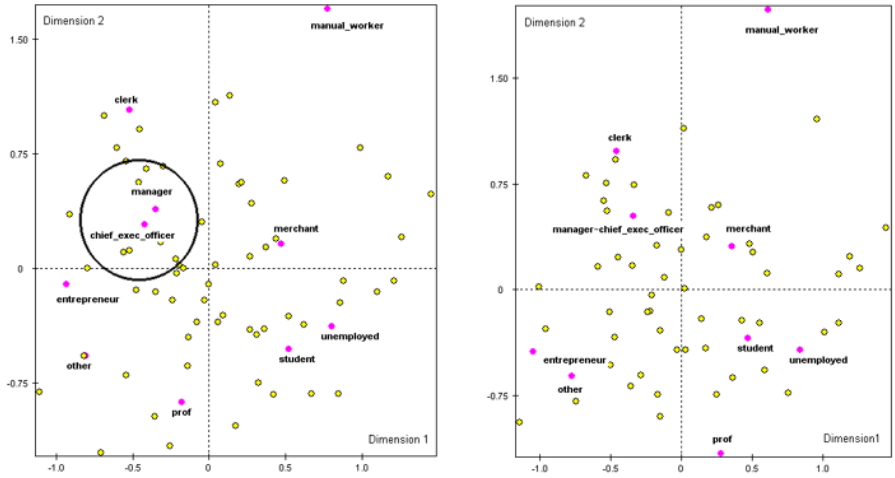
Per meglio comprendere il problema, si consideri un'indagine su un non meglio specificato argomento. Tra le variabili a disposizione si consideri, la professione degli intervistati, tabella 3.4. La modalità *Chief executive officer* ha una frequenza eccessivamente bassa, cosi per ottenere un'analisi più robusta, questa categoria diventa supplementare e le sue osservazioni sono casualmente assegnate alle altre modalità.

La terza colonna della tabella 3.4, riporta le frequenze della modalità considerata susseguente all'assegnazione casuale. Originariamente c'erano quattro *Chief executive officers*, ora un'osservazione viene assegnata alla modalità *Professor*, una alla modalità *Entrepreneur* e due alla modalità *Students*. E' importante

Tabella 3.4: Statistiche per la variabile *Professione*: modalità, numero di osservazioni prima dell'assegnazione casuale (NOPAC), modalità, numero di osservazioni dopo l'assegnazione casuale (NODAC).

Former Occupation	NOPAC	NODAC
Merchant	11	11
Clerk	12	12
Chief executive officer	4	Random Assigned
Unemployed	19	19
Manager	5	5
Professor	5	6
Manual worker	6	6
Entrepreneur	13	14
Student	14	16
Other	12	12

Figura 3.5: Visualizzazione grafica della variabile *Former Occupation* rispetto al primo piano fattoriale.



sottolineare che questa procedura è in buona parte casuale. Al posto di una assegnazione casuale, l'adattamento della SAR considera la distanza delle modalità a bassa frequenza da tutte le altre modalità con frequenza elevata della stessa variabile. Successivamente le modalità a bassa frequenza vengono aggregate alla modalità più vicina. Nella situazione precedentemente illustrata, essendo *Manager* la modalità più vicina a *Chief executive officer*, verrà a questa

aggregata, si veda la parte sinistra della figura 3.5, ottenendo una nuova modalità chiamata *Manager* ∼*Chief executive officer*, parte destra della figura 3.5. Questo criterio riflette il già accennato punto di vista secondo cui il criterio di decisione debba essere la relazione tra le variabili...e non il caso.

3.4.2 Passi di SAR per il trattamento di modalità di bassa frequenza

In ogni variabile, si sostituiscono le modalità con frequenza inferiore alla soglia (usualmente il 2%) con la moda di ogni specifica variabile. In questo modo, gli assi ottenuti e le coordinate delle modalità non sono influenzate dalle modalità con frequenze eccessivamente basse. Si conduce una ACM e tra i risultati ottenuti, si focalizza l'attenzione sui seguenti:

- Percentuale d'inerzia spiegata da ogni fattore;
- Coordinate delle modalità per ogni variabile;
- Contributi assoluti delle modalità;

Si consideri una matrice $\mathbf{X}$ con n righe e p variabili.

- ***Passo 1*** Le p variabili, $(X_1, X_2, ...X_p)$ vengono ordinate sulla base dei contributi assoluti ottenendo la matrice ordinata $\left(X_1^*, X_2^*, ...X_p^*\right)$.
- ***Passo 2*** Viene selezionata la variabile con il contributo assoluto più alto (X_1^*) che viene divisa in due vettori (X_1^{*a}) e (X_1^{*s}). Il vettore (X_1^{*a}) è il vettore delle modalità attive, mentre (X_1^{*s}) è il vettore delle modalità supplementari: modalità con frequenze eccessivamente basse, solitamente meno del 2 percento.
- ***Passo 3*** Il vettore (X_1^{*s}) viene rappresentato come supplementare sul piano fattoriale.
- ***Passo 4*** Le coordinate del vettore supplementare (X_1^{*s}) vengono unite in unico vettore insieme con le coordinate delle modalità attive della stessa variabile.
- ***Passo 5*** Ogni modalità supplementare è aggregata alla modalità attiva che si trova alla minima distanza. Per misurare la distanza tra due

modalità, l'algoritmo proposto considera la distanza euclidea pesata per l'inverso dei primi due autovalori. Ossia, come nella procedura generale si ha:

$$d_{ij} = \left[(x_i - x_j)' \mathbf{W_2} (x_i - x_j)\right]^{1/2} \tag{3.4}$$

dove $\mathbf{W_2}$ è una 2×2 matrice diagonale che ha nella diagonale l'inverso dei primi due autovalori λ_1 and λ_2.

$$\mathbf{W_2} = \begin{pmatrix} \frac{1}{\lambda_1} & 0 \\ 0 & \frac{1}{\lambda_2} \end{pmatrix} \tag{3.5}$$

- ***Passo 6*** Susseguentemente al quinto passo, vengono costruite delle nuove modalità. Le precedenti modalità vengono successivamente sostituite nella matrice di partenza da queste ultime.
- ***Passo 7*** Una ACM classica viene ricondotta sulla matrice modificata, ottenendo i nuovi contributi assoluti e le nuove coordinate. I passi dall'uno al cinque vengono ripetuti fino a che tutte le variabili originali, in cui erano presenti modalità con basse frequenze, siano state ricodificate. Tenendo, ovviamente, in considerazione i risultati precedenti.

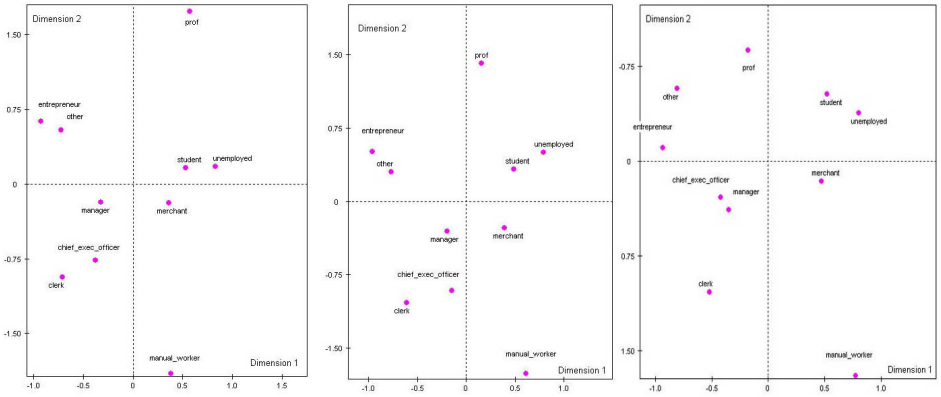
Con questo metodo, queste nuove modalità potrebbero essere definite come "Semi-Attive "in quanto sono state ottenute dalla fusione di modalità attive e modalità che all'inizio sono illustrative e solo successivamente diventano attive. C'è una sostanziale differenza tra il tradizionale metodo dell'assegnazione casuale e la nuova procedura proposta. Entrambe le metodologie assegnano le osservazioni delle modalità a bassa frequenza ad altre modalità. Questa assegnazione è fatta in entrambi i casi nella matrice originale. Nell'approccio tradizionale l'assegnazione avviene in modo casuale, in quella proposta le osservazioni sono assegnate alla modalità che mostra il comportamento più simile, riducendo cosi l'arbitrarietà dell'assegnazione.

3.4.3 Valutazione comparativa dei risultati

Per la valutazione dei risultati si utilizzerà il dataset presentato nella sezione 3.2 le cui variabili sono riportate nella prima e seconda colonna della tabella 3.5.

La presenza di modalità con basse frequenze, rende l'ACM veramente instabile. Solitamente la soglia minima suggerita è del 2%. La figura 3.6, riporta per la variabile *Former Occupation*, la visualizzazione delle proiezioni delle modalità rispetto tre differenti soglie, (2%, 3%, 4%). Nonostante bassi cambiamenti nella soglia scelta è possibile notare rappresentazioni fattoriali piuttosto differenti. E' perciò facile immaginare l'entità dei cambiamenti quando si considerano tutte le variabili simultaneamente. Per migliorare la robustezza si dovrebbe aumentare la soglia oltre il 2 %, ma cosi facendo si presenta un altro grosso problema. Si consideri la tabella 3.5, che riporta il numero di modalità assegnate casualmente, (NMAC) ed il numero di osservazioni, assegnate casualmente, (NOAC). In questa circostanza si è scelta come soglia il 4%.

Figura 3.6: Visualizzazione grafica della variabile *Occupazione* nel primo piano fattorial per differenti soglie (2%, 3%, 4%).



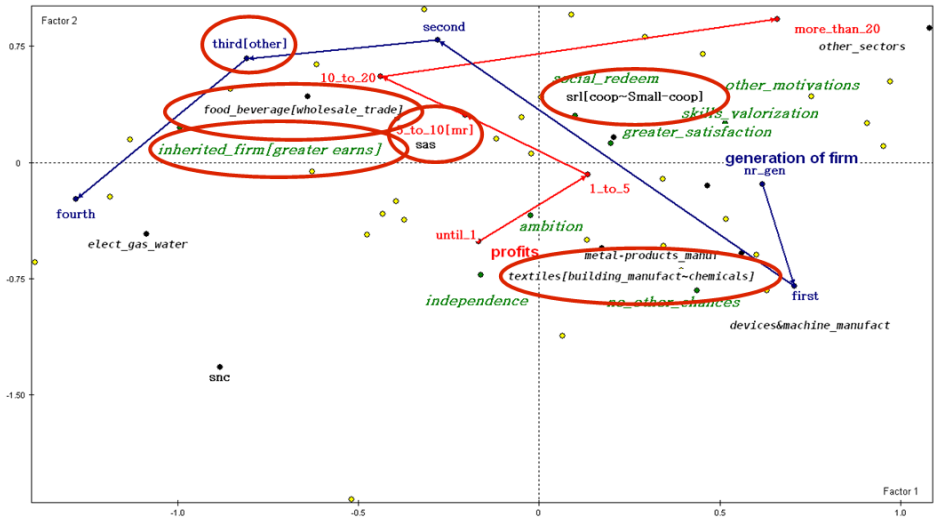
Soglie del 2%, 3%, 4% potrebbero essere valutati valori abbastanza bassi che non dovrebbero cambiare sostanzialmente i risultati dell'analisi. In realtà è possibile vedere che in questa applicazione ci sono 38 osservazioni (37.6 %) assegnate ad altre modalità e 11 modalità (15.2%) utilizzate come supplementari. Questi sono valori decisamente elevati che influenzano pesantemente l'intera analisi.
Grazie all'adattamento della SAR per evitare l'assegnazione casuale, 38 osservazioni non sono più assegnate casualmente ma sono aggregate alla modalità più simile. Undici modalità non sono punti supplementari nel piano fattoriale ma sono utilizzate come variabili Semi-Attive. In altre parole, adesso, queste modalità influenzano l'orientamento degli assi fattoriali, influenzando cosi le coordinate delle altre modalità e non sono più un mero ausilio all'interpretazione del fenomeno. Anche l'incremento della soglia è adesso, in termini di *trade-off*

Tabella 3.5: Variabili, Numero di modalità prima dell'assegnazione casuale (NMPA), Numero di modalità assegnate casualmente (NMAC), Numero di osservazioni assegnate casualmente (NOAC).

Variabili	NMPA	NMCA	NOAC
Legal organization	5	2	5
Profits	6	1	3
Trend market	3	0	0
Market	4	0	0
Kind of market	4	0	0
Generation of firm	6	1	3
Idea	5	0	0
Former occupation	10	1	4
Industrial category	9	3	12
Motivations	9	1	3
Barriers	12	2	8
Total	72	11	38

robustezza capacita esplicativa del modello, più bassa.

Figura 3.7: Visualizzazione grafica della variabile *Occupazione* nel primo piano fattorial per differenti soglie (2%, 3%, 4%).



In figura 3.7 si riporta la rappresentazione fattoriale delle variabili Semi-Attive. L'etichetta delle variabili Semi-Attive è composta di due parti. La parte sini-

stra rappresenta la modalità attiva, caratterizzata dal possedere una frequenza non inferiore alla soglia scelta. La parte destra, racchiusa tra parentesi quadre, riporta una o più modalità con frequenza iniziale inferiore alla soglia scelta. Si ricorda che la soglia usualmente scelta ammonta al 2%. Ad esempio nell'etichetta *Srl[coop∼coop-small]*, *srl* costituisce la modalità originariamente attiva mentre *coop* e *small-coop*, le modalità originariamente illustrative. Si noti che: in questo stadio esiste una sola modalità attiva chiamata *Srl[coop∼small-coop]*; questa modalità potrebbe essere ulteriormente rinominata e semplificata come *Srl[coop]*; a seguito dell'accorpamento *coop∼small-coop*, la frequenza di questa nuova modalità diventa maggiore della soglia prefissata, potrebbe essere quindi disaggregata dalla modalità *Srl[coop∼coop-small]* formando una nuova modalità e risolvendo del tutto il problema sia della robustezza dell'analisi, sia l'obbligo dell'uso di alcune modalità come illustrative.

Capitolo 4

Applicazioni su Datasets reali

4.1 Introduzione

Nel capitolo precedente sono state illustrate diverse applicazioni della SAR finalizzate alla risoluzione di specifici problemi generalmente incontrati nell'applicazione della ACM. Nella sezione 3.2, la SAR è stata utilizzata per la riduzione del numero di modalità, quando si è in presenza di problemi di visualizzazione o per evitare che nella fase di ricodifica si abbia una perdita dei contributi delle modalità dovuta ad accorpamenti che non rispettino il principio dell'equivalenza distributiva. Nella sezione 3.3, si è invece illustrato coma la SAR possa essere utilizzata per la ricodifica automatica di variabili continue, illustrandone potenzialità e limiti. Infine nella sezione 3.4, si è proposto un procedimento alternativo all'assegnazione casuale per evitare che le modalità con frequenze troppo basse influenzino eccessivamente i risultati dell'ACM. Nonostante queste applicazioni della SAR siano state presentate, ed in effetti lo siano, come autonome, esse offrono i migliori risultati ed esplicano appieno le loro potenzialità quando unite in un unico procedimento. Ovviamente, seppur i risultati non cambino di molto, la diversa combinazione degli algoritmi può portare a diversi contesti di applicazione e a diverse interpretazioni della metodologia proposta. Da un lato, infatti, essa può essere vista come una variazione metodologica dell'Analisi delle Corrispondenze Multiple, ma anche come uno strumento di ausilio nell'interpretazione dei risultati al momento della lettura del piano fattoriale od

ancora come uno strumento di consulto al momento della ricodifica di alcune variabili. Allo scopo di illustrare queste diverse interpretazioni, si illustreranno diverse combinazioni delle procedure proposte nelle sezioni 3.2, 3.3, 3.4, attraverso l'analisi di una matrice di dati che, per le sue caratteristiche, può essere considerata al limite delle diverse interpretazioni proposte.

4.2 Descrizione della matrice dei dati

Il dataset analizzato è il risultato delle risposte di un panel di utilizzatori internet di un noto Internet Service Provider. Lo scopo dell'indagine è la comprensione della tipologia della propria utenza, con la speranza di poter definire alcune figure tipiche, attraverso la conoscenza delle loro caratteristiche, abitudini ed esigenze al momento della navigazione. La prima colonna della tabella 4.1, riporta le variabili presenti nel dataset, mentre nella seconda colonna, sono riportate le rispettive modalità.

Tabella 4.1: **Variabili**, numero di modalità prima dell'aggregazione (**NMPA**), Numero di modalità dopo l'aggregazione (**NMDA**)

Variabili	**NMPA**	**NMDA**
Professione	10	4
Titolo di Studio	6	4
Numero Prodotti Tecnologici Posseduti	14	3
Luogo di Collegamento	4	3
Tecnologia Principalmente Usata	13	7
Tecnologia di Connessione	8	4
Anzianita Internet	8	3
Acquisti Online	2	2
Frequenza di collegamento	4	3
Provider	9	5
Sesso	2	2
Dimensione della Famiglia	5	3
Interessi	13	7
Regione	20	9
Donazione	3	3
Reddito	Cont	Cont
Età	Cont	Cont
Totale	**121**	**62**

Questo dataset è particolarmente adatto a mettere in luce diverse problematiche

affrontate nei capitoli precedenti. Innanzi tutto, la presenza di variabili con un numero di modalità che va da un minimo di due ad un massimo di venti, permette di apprezzare i vantaggi di poter decidere attraverso la fissazione di una soglia, si veda la sezione 2.3.1, il grado di sintesi più opportuno per la migliore descrizione possibile dei dati. Il dataset è composto da 17 variabili di cui due continue per un totale di 121 modalità su 6552 osservazioni. Diciassette variabili con 121 modalità rappresentano probabilmente le tipiche dimensioni di un dataset su cui si applica l'ACM. Più avanti, si veda la figura 4.1, sarà immediatamente percettibile come una lettura del piano fattoriale sia piuttosto difficoltosa e come incrementando ulteriormente le dimensioni possa essere quasi impossibile la corretta individuazione delle posizioni, e la stessa lettura, delle etichette delle modalità sul piano. L'utilità del dataset consiste quindi proprio nel fatto che permette ancora un confronto tra i risultati di una ACM classica con la metodologia qui proposta. D'altra parte, come si vedrà nella sezione 4.4, oltre un certo limite anche la SAR illustrata nella sezione 4.3 mostra i suoi limiti se impiegata come strumento per il Data Mining, si rende necessaria pertanto una ulteriore variazione per renderla utilizzabile anche con enormi moli di dati. Da ultimo, la presenza di variabili piuttosto comuni come l'*Età* o la *Professione* e di altre invece meno note come il tipo di *Internet Service Provider* (ISP), o l'*Anno di primo utilizzo d'internet*, permetteranno di valutare i risultati della SAR in contesti noti, confrontando le differenze tra un accorpamento soggettivo ed uno oggettivo, e l'utilità di un supporto in situazioni del tutto sconosciute.

Tabella 4.2: Intervistati per grado di istruzione; frequenze assolute; frequenze percentuali.

Titolo di studio	**Intervistati**	**Percentuali**
Nessuna scuola	26	0.40
Licenza Elementare	43	0.66
Licenza Media	3063	46.75
Diploma	532	8.12
Studente Universitario	745	11.37
Laurea	2143	32.71
Totale	**6552**	**100.00**

Il grado d'istruzione prevalente è la licenza media, in linea con l'andamento nazionale. E' da sottolineare tuttavia la forte presenza di laureati (32.71%) e la presenza di un titolo di studio atipico come *Studente universitario* con una

percentuale del 11.37%. Il campione è formato prevalentemente da uomini, tabella 4.3 e solo il 20% circa sono donne.

Tabella 4.3: Intervistati per sesso; frequenze assolute; frequenze percentuali.

Sesso	**Intervistati**	**Percentuali**
Uomo	5208	79.49
Donna	1344	20.51
Totale	**6552**	**100.00**

Le due professioni prevalenti, tabella 4.4, sono *Impiegato* e *Libero professionista*. Da sole queste due professioni rappresentano quasi metà del campione.

Tabella 4.4: Intervistati per professione; frequenze assolute; frequenze percentuali.

Professione	**Intervistati**	**Percentuali**
Impiegato	1755	26.79
Libero Professionista	1413	21.57
Studente	689	10.52
Dirigente / Quadro	680	10.38
Imprenditore	571	8.71
Pensionato	275	4.20
Operaio	206	3.14
Non Occupato	193	2.95
Casalinga	63	0.96
Altro	707	10.79
Totale	**6552**	**100.00**

Percentuali di rilievo mostrano anche le modalità *Studente* e *Dirigente/Quadro* formanti un altro 20% del campione. Percentuali decisamente basse invece per le casalinghe; avendo una percentuale inferiore al 2%, questa modalità, se non opportunamente trattata, potrebbe pesare eccessivamente nell'analisi, per cui, nell'applicazione della ACM classica, le unità appartenenti a questa modalità verranno assegnate casualmente alle altre modalità, mentre nella SAR, la modalità *Casalinga* verrà accorpata alla modalità ad essa più vicina. La tabella 4.5 riporta le principali caratteristiche delle due variabili continue presenti nel

dataset: *Reddito* ed *Età*. In una prima fase queste due variabili sono state ricodificate al fine di eliminare quei dati palesemente errati o incongruenti.

Tabella 4.5: Statistiche sommarie per le variabili Reddito ed Età: Media (Med), Scostamento quadratico medio (Sm), Coefficiente di variazione (Cv), Minimo (Min) e Massimo (Max).

Variabile	**Med**	**SM**	**CV**	**Min**	**Max**
Reddito	24048.90	10823.90	45.01	5500	45500
Età	38.85	11.59	29.83	17	75

Il *Reddito* mostra una variabilità piuttosto elevata con un coefficiente di variazione pari al 45%; variabilità più modesta si ha invece per l'*età*.

Tabella 4.6: Intervistati per regione di residenza; frequenze assolute; frequenze percentuali.

Regione	**Intervistati**	**Percentuali**
Lombardia	1392	21.25
Sardegna	913	13.39
Lazio	912	13.92
Piemonte	454	6.93
Sicilia	435	6.64
Veneto	397	6.06
Campania	327	4.99
Emilia Romagna	303	4.62
Toscana	301	4.59
Puglia	296	4.42
Calabria	144	2.20
Marche	137	2.09
Liguria	125	1.91
Friuli Venezia Giulia	90	1.37
Umbria	87	1.33
Abruzzo	84	1.28
Basilicata	75	1.14
Trentino Alto Adige	41	0.63
Valle D'Aosta	21	0.32
Molise	18	0.27
Totale	**6552**	**100.00**

Dall'analisi della tabella 4.6, si nota la presenza di tre regioni ''forti''. La Lom-

bardia, il Lazio e la Sardegna; da sole queste tre regioni rappresentano il 48.56% del campione. All'estremo opposto si trovano una serie di regioni con percentuali non superiori al 2%, come: Liguria, Friuli Venezia Giulia, Umbria, Abruzzo, Basilicata, Trentino Alto Adige, Valle D'Aosta, Molise, più altre due regioni con percentuali di poco superiori al 2% come Marche e Calabria. Quel che emerge è che il campione è fortemente distorto non rappresentando i pesi reali delle singole regioni dal punto di vista demografico. Infatti, se da un lato sono giustificate le basse frequenze d'alcune regioni come la Valle D'Aosta non lo sono sicuramente quelle di altre come per esempio la Sicilia. D'altra parte si ignora la reale composizione dell'utenza di questo provider. Per quanto concerne più specificatamente l'ACM un numero cosi alto di modalità crea i ben noti problemi evidenziati nei capitoli precedenti.

Tabella 4.7: Intervistati per numero di componenti la famiglia; frequenze assolute; frequenze percentuali.

Dimensione	**Intervistati**	**Percentuali**
1 Persona	735	11.22
2	1293	19.73
3	1790	27.32
4	1984	30.28
5 o più	750	11.45
Totale	**6552**	**100.00**

L'ultima tabella inerente le caratteristiche strutturali degli utenti che hanno risposto al questionario 4.7, riporta la dimensione della famiglia. La famiglia maggiormente rappresentata è la tipica famiglia composta da tre o quattro componenti, 57.60% del campione. Le altre tipologie sono rappresentate in modo bilanciato e la frequenza più bassa non scende al di sotto dell' 11%.

Tabella 4.8: Intervistati per attitudine ad aquistare on line; frequenze assolute; frequenze percentuali.

Acquisti	**Intervistati**	**Percentuali**
Si	4363	66.59
No	2189	33.41
Totale	**6552**	**100.00**

Le altre variabili del dataset riportano alcune caratteristiche comportamentali tendenti ad investigare alcune abitudini degli utenti, le motivazioni di utilizzo, tipo di tecnologia utilizzata e luogo di collegamento. Dalla tabella 4.8 si evince che oltre due terzi dei rispondenti acquistano in internet.

Tabella 4.9: Intervistati per anno di primo utilizzo di internet; frequenze assolute; frequenze percentuali.

Anzianità	**Intervistati**	**Percentuali**
Prima del 1997	1864	28.45
Durante 1997	775	11.83
Durante 1998	1041	15.89
Durante 1999	978	14.93
Durante 2000	1003	15.31
Durante 2001	520	7.94
Durante 2002	297	4.53
Durante 2003	74	1.13
Totale	**6552**	**100.00**

Che oltre il 40% di coloro che hanno risposto sono utilizzatori "storici" di internet, tabella 4.9, dichiarando di utilizzarlo gia dal 1997 o addirittura da prima.

Tabella 4.10: Intervistati tipo di tecnologia utilizzata per la connessione ad internet; frequenze assolute; frequenze percentuali.

Tecnologia di connessione	**Intervistati**	**Percentuali**
Modem Standard	3072	46.89
ADSL	2455	37.47
ISDN	635	9.69
Fibra Ottica	198	3.02
Rete Locale	99	1.51
Satellite	16	0.24
Altro	35	0.53
Non so	42	0.64
Totale	**6552**	**100.00**

La quasi totalità (84.36%) del campione analizzato utilizza come tecnologia

di connessione o il Modem standard o l'ADSL, tabella 4.10; Poco più del 9% utilizza l'ISDN, mentre le altre modalità si presentano con modalità bassissime.

Tabella 4.11: Intervistati per provider utilizzato; frequenze assolute; frequenze percentuali.

Provider	**Intervistati**	**Percentuali**
Tiscali	3925	59.91
Libero	759	11.58
Virgilio	553	8.44
AliceTi	520	7.94
Fastweb	242	3.69
Tele2	155	2.37
Kataweb	8	0.12
Altro	353	5.39
Non so	37	0.56
Totale	**6552**	**100.00**

I due fornitori *Tiscali* e *Libero*, dominano nelle preferenze degli intervistati. Oltre il 70%, infatti, dichiara di utilizzare l'uno o l'altro. Fanalino di coda *Kataweb* con soli 8 utenti, tabella 4.11.

Tabella 4.12: Intervistati per Luogo di collegamento; frequenze assolute; frequenze percentuali.

Luogo di collegamento	**Intervistati**	**Percentuali**
Casa	2551	38.93
Lavoro	788	12.03
Casa-Lavoro	3168	48.35
Altro	45	0.69
Totale	**6552**	**100.00**

la tabella 4.12 riporta il luogo di collegamento prevalentemente utilizzato dagli utenti. In questo caso le risposte fornite sono, ovviamente, abbastanza scontate. Il 48.35% dichiara di collegarsi sia da casa che dal lavoro. Al secondo posto si trovano coloro che si collegano da casa con il 38.93%, mentre solo il 12.03% dichiara di collegarsi esclusivamente dal lavoro. La categoria *Altro*, comprende coloro che si collegano dagli internet point, da casa di amici o dall'università. Come riportato in tabella 4.13, ben il 77.40%, dichiara di collegarsi tutti i giorni.

Tabella 4.13: Intervistati per numero di giorni di connessione abituale; frequenze assolute; frequenze percentuali.

Numero di giorni	**Intervistati**	**Percentuali**
1-2 al Mese	189	2.88
1-2 a Settimana	503	7.68
3-5 a Settimana	789	12.04
Tutti i Giorni	5071	77.40
Totale	**6552**	**100.00**

Al diminuire della frequenza di connessione, diminuiscono anche gli utenti, fino ad arrivare ad un 2.88% che si connette solo una o due volte al mese. Come si vedrà più avanti la frequenza di utilizzo rappresenta una vera discriminante per i comportamenti degli utenti.

Tabella 4.14: Intervistati per tipologia di interesse in internet; frequenze assolute; frequenze percentuali.

Interessi	**Intervistati**	**Percentuali**
Sport	1797	27.43
Tecnologia	314	4.79
Arte	1225	18.70
Radio Tv	417	6.36
Bricolage	756	11.54
Natura	152	2.32
Auto	405	6.18
Moda	141	2.15
Economia	273	4.17
Benessere	257	3.92
Cucina	208	3.17
Lettura	397	6.06
Altri interessi	210	3.21
Totale	**6552**	**100.00**

Le tabelle 4.14, 4.15, 4.16, riportano informazioni che sono state considerate utili al gestore internet per comprendere se gli interessi e gli hobby praticati dagli utenti siano in qualche modo correlati alle abitudini di navigazione e soprattutto all'attitudine di acquistare o meno on line. Si può cosi notare dalla tabella 4.14, che l'interesse dominante è senza dubbio lo Sport con il 27.43%, seguito a

debita distanza dall'Arte (18.70%) e successivamente dal Bricolage con l'11.54 %. Decisamente basse tutte le altre modalità.

Tabella 4.15: Intervistati per tipologia di tecnologia principalmente utilizzata frequenze assolute; frequenze percentuali.

Tecnologia	**Intervistati**	**Percentuali**
Pc	3138	47.89
Fotocamera	794	12.12
Videocamera	707	10.79
Cellulare	396	6.04
Vcr	375	5.72
Pc Portatile	322	4.91
PayTv	246	3.75
webcam	241	3.68
Home Cinema	156	2.38
Dvd	98	1.50
Stampante	34	0.52
Scanner	27	0.41
Masterizzatore	18	0.27
Totale	**6552**	**100.00**

La tabella, 4.15, riporta lo strumento tecnologico che gli utenti dichiarano di utilizzare in modo prevalente rispetto agli altri. Il computer fisso è senza dubbio lo strumento più utilizzato con il 47.89%, seguito dalla fotocamera digitale con il 12.12%. Anche in questa variabile, sono presenti alcune modalità con frequenza piuttosto basse. Oltre a chiedere agli intervistati, qual è lo strumento tecnologico principalmente utilizzato, si è anche chiesto quanti fossero gli strumenti complessivamente posseduti. Anche in questo caso lo scopo era quello di verificare l'eventuale correlazione tra l'interesse per la tecnologia e le altre caratteristiche precedentemente menzionate. La tabella 4.16 riporta i risultati. Il 30% circa dichiara di possedere tra gli 8 e i 9 prodotti tecnologici. Il fenomeno mostra un andamento piuttosto normale e solo l'1.54% dichiara di possedere solo un prodotto tecnologico, verosimilmente il personal computer senza nessuna periferica. Da segnalare inoltre una percentuale, seppur bassa, che dichiara di possedere 13 o più prodotti tecnologici! Praticamente tutti quelli elencati in tabella 4.15 più qualcun altro ancora.

Tabella 4.16: Intervistati per numero di prodotti tecnologici posseduti; frequenze assolute; frequenze percentuali.

Numero prodotti	**Intervistati**	**Percentuali**
1	101	1.54
2	67	1.02
3	141	2.15
4	344	5.25
5	517	7.89
6	722	11.02
7	883	13.48
8	975	14.88
9	927	14.15
10	773	11.80
11	546	8.33
12	364	5.56
13	146	2.23
14	46	0.70
Totale	**6552**	**100.00**

Tabella 4.17: Intervistati per destinazione della donazione; frequenze assolute; frequenze percentuali.

Donazione	**Intervistati**	**Percentuali**
Alleanza di Misericordia	2784	42.49
SolidAfrica	2057	31.40
Scegliete voi	1711	26.11
Totale	**6552**	**100.00**

Per incentivare la compilazione del questionario da parte degli utenti, si è fatto leva sullo spirito di solidarietà destinando un euro, per ogni questionario pervenuto, a due associazioni umanitarie. All'intervistato è stata lasciata la possibilità di scegliere il destinatario tra i due proposti dal gestore. Per amor di completezza, e per l'utilità di quel che si dirà in seguito, si è deciso di lasciare nel dataset anche questa variabile. Nella tabella 4.17, sono riportati i risultati delle preferenze degli intervistati.

4.3 La SAR come strumento per la riduzione delle modalità

In questa sezione si illustrano nei dettagli i risultati della SAR come strumento per la riduzione delle modalità, l'attenzione sarà pertanto concentrata quasi esclusivamente sulla migliore leggibilità del piano fattoriale. Per apprezzare al meglio i risultati, e mettere in luce le principali differenze, della ricodifica automatica delle variabili, si eseguirà preliminarmente una Analisi delle Corrispondenze Multiple classica sulla matrice originaria dei dati. In seguito, si confronteranno i piani fattoriali, la ricodifica effettuata e le coordinate di alcune variabili allo scopo di poter criticamente confrontare le differenze e le analogie dei risultati e suffragare le considerazioni fatte nelle sezioni precedenti.

4.3.1 I risultati della ACM classica

Tabella 4.18: Risultati numerici del'ACM prima dell'applicazione della SAR (a) e dopo l'applicazione della SAR (b): Autovalori, percentuale di inerzia spiegata e percentuale cumulata di inerzia spiegata

(A)				(B)			
N	Autovalori	Inerzia	Cum	N	Autovalori	Inerzia	Cum
1	0.185	2.61	2.61	1	0.201	6.20	6.20
2	0.131	1.85	4.47	2	0.142	4.39	10.59
3	0.124	1.75	6.22	3	0.103	3.20	13.79
4	0.115	1.63	7.85	4	0.095	2.95	16.74
5	0.105	1.49	9.33	5	0.088	2.71	19.45
6	0.102	1.44	10.77	6	0.077	2.37	21.82
7	0.097	1.37	12.15	7	0.075	2.31	24.13
8	0.090	1.28	13.42	8	0.074	2.29	26.42
9	0.088	1.24	14.66	9	0.071	2.20	28.62
10	0.087	1.23	15.90	10	0.070	2.16	30.79
..	. ...	. ..	.. ..	..	. ....	. ..	. ..
106	0.019	0.26	100.00	55	0.006	0.18	100.00

La parte A della tabella 4.18, riporta, gli autovalori, l'inerzia e l'inerzia cumulata per i primi 10 assi fattoriali. I primi 2 assi, spiegano appena il 4.47% dell'inerzia totale, mentre i primi 4 spiegano il 7.85%. Come sempre si conferma la bassa percentuale di inerzia spiegata dall'ACM. Una bassa percentuale di

inerzia spiegata nel caso dell'ACM è un risultato pessimistico della reale capacità esplicativa del modello. Nel caso si voglia rivalutare il tasso di inerzia, si può utilizzare la correzione di Benzécri.

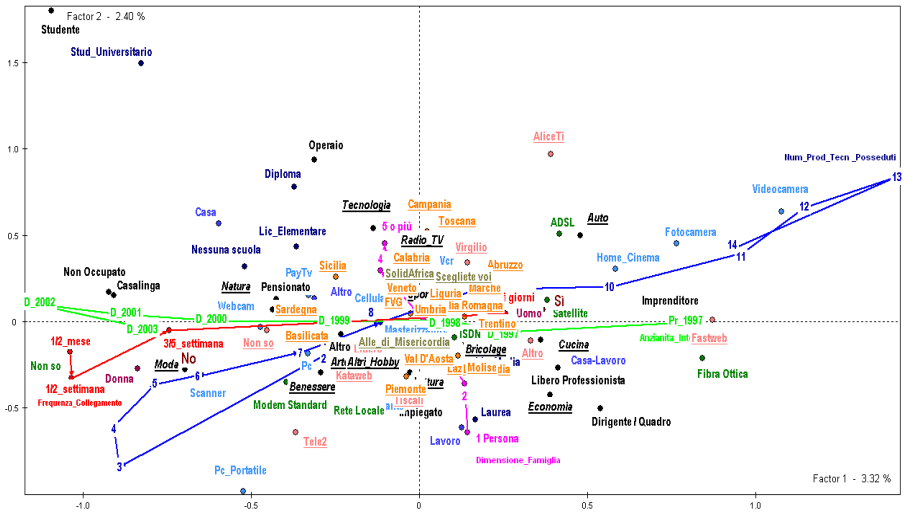


Figura 4.1: Visualizzazione dei profili colonna rispetto al primo piano fattoriale prima della ricodifica.

Ad ogni modo è fuor di dubbio che il vero potenziale di questa metodologia sia la rappresentazione grafica dei profili colonna, per cui, è su questa che ora ci si concentrerà. L'analisi della figura 4.1, evidenzia una forte sovrapposizione delle etichette, rendendo, soprattutto nella parte centrale, impossibile l'interpretazione dei risultati. Le traiettorie delle variabili ordinali, importantissime e spessissimo utilizzate per dare un orientamento agli assi fattoriali, risultano in alcuni casi coperte dalle etichette a tal punto de non essere distinguibili. Si veda in particolare la traiettoria della variabile *Dimensione della famiglia*. Va peraltro rimarcato, che la nuvola dei punti variabile è schiacciata dalla presenza di due modalità anomale come *Studente universitario*, per quanto riguarda la variabile *Professione*, e *Studente*, per quanto riguarda la variabile *Titolo di studio*. Per rendere il piano fattoriale più leggibile, alcuni software permettono di ingrandire parti del piano, in questo modo si può in parte ovviare a questo inconveniente. In figura 4.21, si veda l'Appendice C per un'immagine di maggiori dimensioni, viene riportato un ingrandimento della parte centrale. Per

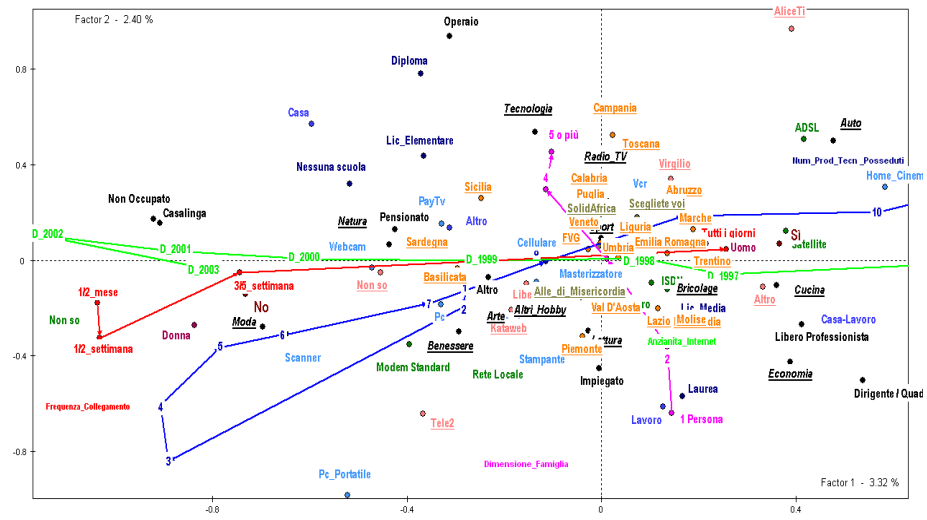


Figura 4.2: Visualizzazione dei profili colonna rispetto al primo piano fattoriale prima della ricodifica: ingrandimento della parte centrale.

quanto, se confrontato con il piano originale, quest'ultimo risulti più leggibile, la leggibilità resta problematica ed è facile notare ancora la sovrapposizione di molti punti. Per quanto si tenti di spostare le etichette, oltre un certo limite non è materialmente possibile avere sott'occhio tutti i punti.

Tabella 4.19: Correlazione delle variabili Reddito ed Età nei primi 5 assi

Variabile	**Asse 1**	**Asse 2**	**Asse 3**	**Asse 4**	**Asse 5**
Reddito	-0.01	0.00	0.00	-0.02	0.00
Età	0.16	-0.28	-0.08	-0.24	0.19

Come illustrato nella sezione 3.3, uno dei possibili modi per l'analisi delle variabili continue nell'ACM, consiste nel calcolare la correlazione tra gli assi fattoriale con queste ultime. La tabella 4.19, riporta le correlazione delle variabili *Reddito* ed *Età* con i primi 5 assi fattoriali. Il reddito non risulta correlato con nessuna delle prime 5 dimensioni individuate dall'ACM. Per quanto riguarda l'età, invece, si può notare una debolissima correlazione con il secondo e quarto asse fattoriale. Complessivamente si può comunque affermare che le due va-

riabili continue presenti nella matrice iniziale non apportino nessun contributo alla comprensione del fenomeno, non risultando significativamente correlate con nessuna delle nuove dimensioni individuate.

4.3.2 I risultati della ricodifica

La presenza di numerose modalità con frequenze inferiori alla soglia del 2%, impone un primo pre-trattamento della matrice onde evitare che queste modalità influenzino in modo eccessivo le direzioni degli assi fattoriali. Nell'analisi classica, come più volte rimarcato nelle sezioni precedenti, questo problema viene superato grazie all'assegnazione casuale delle osservazioni appartenenti a queste modalità, ad altre modalità della stessa variabile, con frequenze superiori alla soglia predeterminata. Allo scopo di ridurre la casualità insista in questo processo, qui si adotterà la procedura proposta nella sezione 3.4, che porterà alla definizione delle gia menzionate modalità Semi-Attive, si veda ancora la sezione 3.4.

Tabella 4.20: Modalità della variabile Tecnologia di connessione prima dell'applicazione della SAR e dopo l'applicazione della SAR

Modalità originarie	**Modalità ricodificate**
Modem Standard Rete Locale Non so	**modem standard[non so∼rete locale]**
ADSL Altro	**adsl[altro]**
ISDN Satellite	**isdn[satellite]**
Fibra Ottica	**fibra Ottica**

La tabella 4.20, riporta i risultati per la variabile *Tecnologia di connessione*. La parte della nuova modalità racchiusa tra parentesi quadre, rappresenta la modalità con frequenza originaria minore della soglia, mentre la parte alla sua sinistra, la modalità con frequenza superiore alla soglia prefissata, ad essa più vicina. Cosi ad esempio, nella modalità *modem standard[non so∼rete locale]*, le modalità originariamente con frequenza inferiore al 2% sono: *[non so∼rete locale]*, mentre *modem standard* rappresenta la modalità attiva ad esse più vicina. Dovendo interpretare il significato di questa nuova modalità, si potrebbe

facilmente affermare che chi possiede un modem standard ha un comportamento molto simile a coloro che ignorano perfino quale sia la tecnologia che utilizzano. Questa prima parte appare piuttosto chiara: da una parte del piano, si trovano coloro che probabilmente scelgono attentamente quale tipo di tecnologia sia più adatta alle loro esigenze, ed ovviamente sanno anche quale sia. Un gruppo di utilizzatori ''attenti''. Dall'altra parte del piano, rappresentati da questa modalità, un gruppo che probabilmente ignora il ventaglio di scelte a disposizione accontentandosi della tecnologia base proposta al momento dell'acquisto del PC. All'interno di questo gruppo, una parte ne ignora persino il nome. Date queste premesse, appare del tutto illogico l'accorpamento con *Rete locale*. In realtà questo risultato non è il frutto di un accorpamento azzardato, già al momento della presentazione dei risultati in sede al committente, ancor prima di un qualunque accorpamento, si era constatato che la modalità *Rete locale* presentava un comportamento piuttosto anomalo. Gli accorpamenti delle restanti modalità, sono riportate nell'appendice **A**.

Tabella 4.21: Ricodifica in classi della variabile Reddito; frequenze assolute; frequenze percentuali.

Reddito	**Intervistati**	**Percentuali**
$5500\vert - \vert 12000$	897	13.69
$12500\vert - \vert 14000$	247	3.77
$14500\vert - \vert 34000$	3664	55.92
$34500\vert - \vert 45500$	1744	26.62
Totale	**6552**	**100.00**

Tabella 4.22: Ricodifica in classi della variabile Età; frequenze assolute; frequenze percentuali.

Età	**Intervistati**	**Percentuali**
$17\vert - \vert 27$	1103	16.83
$28\vert - \vert 58$	5010	76.46
$59\vert - \vert 63$	214	3.26
$64\vert - \vert 75$	225	3.43
Totale	**6552**	**100.00**

Il secondo passo consiste nel ricodificare le variabili continue attraverso la pro-

cedura proposta nella sezione 3.3, ed i cui risultati sono riportati nelle tabelle 4.21 e 4.22. Dall'analisi del reddito, si nota una ricodifica atipica, con classi decisamente lontane da quelle che si sarebbero ottenute se si fosse ricodificato il reddito seguendo, seppur approssimatamente, il criterio delle classi equi-ampie o equi-frequenti. Un analogo discorso può essere fatto analizzando le classi di età. Come prima conclusione, si può quindi affermare che le classi ottenute con la SAR sono decisamente diverse da quelle che si sarebbero ottenute seguendo i criteri classici di suddivisione in classi.

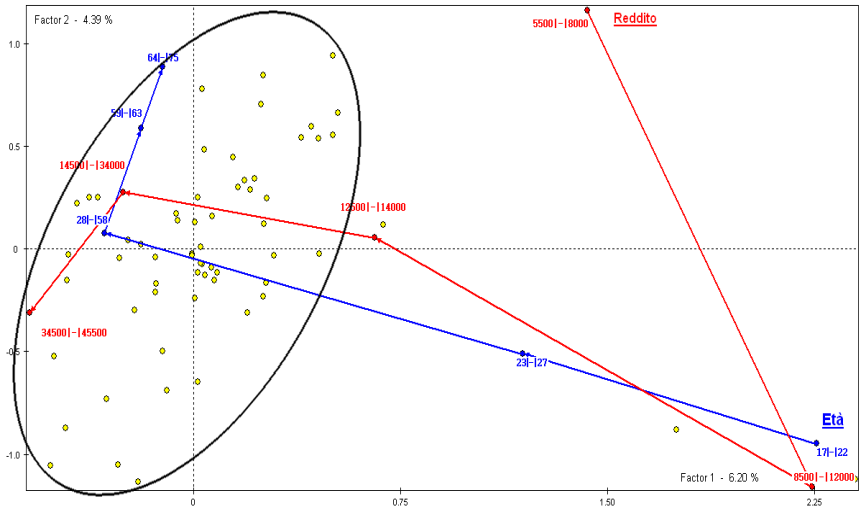


Figura 4.3: Visualizzazione della nuvola dei profili colonna rispetto al primo piano fattoriale: particolare delle variabili Reddito ed Età.

La figura 4.3 riporta le modalità delle variabili *Età* e *Reddito* sul piano fattoriale formato dai primi due assi. L'analisi delle coordinate delle variabili sul piano fattoriale, permette importanti considerazioni. La prima nasce dal forte legame esistente tra reddito ed età. Sulla parte destra del piano fattoriale si trovano coloro che dichiarano un basso reddito. Seguendo le traiettorie delle due variabili lungo il piano fattoriale, si desume che, entro certi limiti, all'aumentare dell'età aumenta anche il reddito dichiarato. Superata la parte centrale del piano, si assiste ad una netta divaricazione, all'aumentare dell'età il reddito diminuisce. Queste sono le considerazioni derivanti dall'analisi delle traiettorie

delle due variabili. Che dire delle peculiarità della ricodifica proposta? La prima, e più importante, circostanza da sottolineare è che le due variabili hanno un'influenza fortissima nell'interpretazione dei risultati. La seconda paradossale considerazione deriva dal fatto che mentre attraverso la ricodifica in classi le due variabili assumono un'importanza persino eccessiva, utilizzandole come supplementari, appaiono del tutto inutili alla comprensione del fenomeno con delle correlazioni non significative. La spiegazione di questo ricorrente fenomeno, è di facile spiegazione. In primo luogo, quando utilizzate come supplementari, le variabili non influenzano la direzione degli assi, e di conseguenza non apportano nessun contributo alla loro costruzione. In secondo, l'unica correlazione che viene misurata è quella lineare. Se, come accade spesso, si è in presenza di correlazioni non lineari tra gli assi e le variabili, queste non verranno rilevate dal coefficiente di correlazione e pertanto anche variabili importanti verranno trascurate. Queste prime considerazioni fanno propendere per una ricodifica in classi delle variabili continue. A questo punto si ripropone il problema di quale sia la migliore ricodifica. I motivi presentati nella sezione 3.3, fanno propendere per una ricodifica basata sui principi della SAR. L'ultima considerazione, forse altrettanto paradossale della seconda, nasce dalle stesse motivazioni che hanno giustificato la SAR, ossia una ricodifica basata esclusivamente sui risultati dell'ACM. La ricodifica ottenuta è talmente sovra-adattata ai dati da schiacciare la restante nuvola dei punti e risultare preponderante rispetto a tutte le altre variabili. Essendo ormai chiaro il ruolo delle due variabili continue, allo scopo di meglio investigare le relazioni tra le restanti variabili, esse saranno nel proseguo utilizzate come illustrative. Per quanto riguarda i parametri dell'algoritmo di ricodifica, si è deciso di utilizzare solo i primi due assi e di non ricodificare la variabili che abbiano 3 o meno di tre modalità. La soglia che sembra garantire il miglior compromesso tra riduzione delle modalità e capacità esplicativa è 22.5%. Quindi:

1. K=2
2. H=3
3. Soglia = 22.5 % della distanza massima tra le modalità di ciascuna variabile

A seguito dall'applicazione della metodologia proposta, il numero di modalità viene ridotto da 121 a 62, come riportato nella sezione (B) della tabella 4.1, mentre la percentuale di inerzia spiegata dai primi 4 fattori sale dal 7.85% al

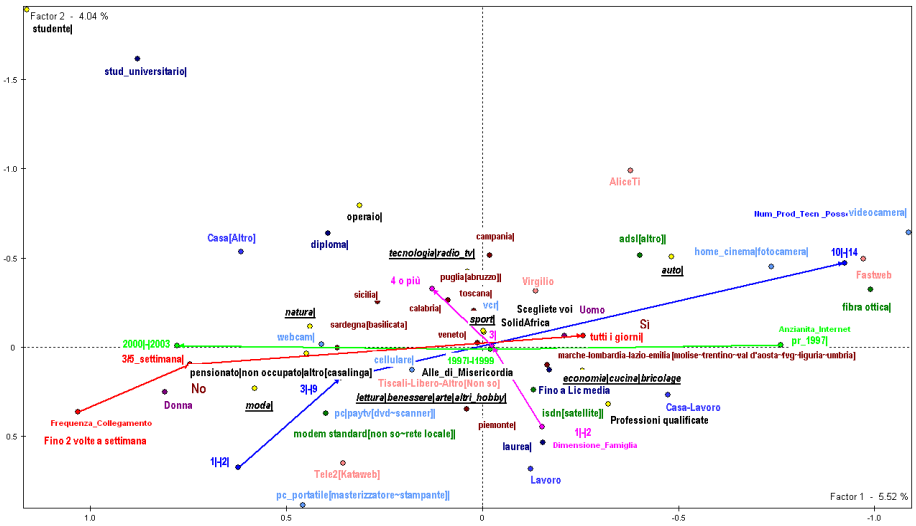


Figura 4.4: Visualizzazione della nuvola dei profili colonna rispetto al primo piano fattoriale dopo la ricodifica.

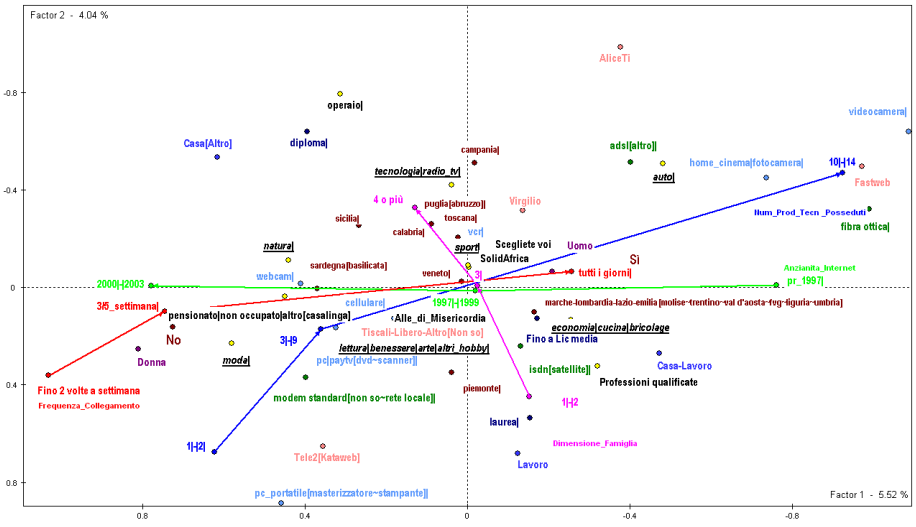


Figura 4.5: Visualizzazione della nuvola dei profili colonna rispetto al primo piano fattoriale dopo la ricodifica: ingrandimento della parte centrale.

16.74%. Le categorie "ridotte"ammontano pertanto a 59 mentre il guadagno di inerzia spiegata nei primi quattro assi è del 8.89%, praticamente raddoppiata! Le figure 4.4 e 4.22, si veda l'Appendice C per un'immagine di maggiori dimensioni, riportano il primo piano fattoriale dopo la procedura di ricodifica ed un ingrandimento della parte centrale identico a quello effettuato per il piano fattoriale delle modalità non ricodificate. Il confronto tra i due piani fattoriali prima della ricodifica, figura 4.21, e dopo la ricodifica, figura 4.22 evidenzia nel secondo caso, una maggiore pulizia permettendo una più facile lettura dei risultati.

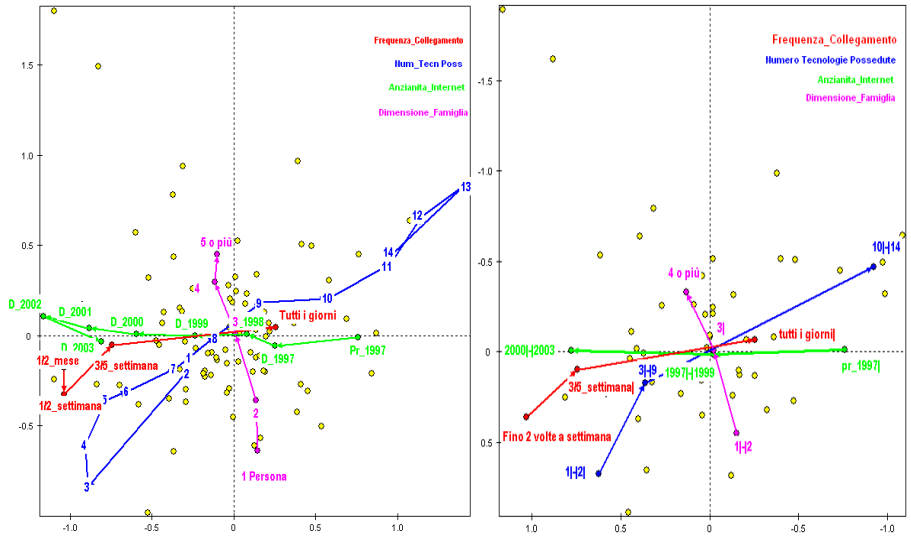


Figura 4.6: Confronto delle traiettorie delle variabili ordinali prima e dopo la ricodifica.

Allo scopo di illustrare ulteriormente alcune caratteristiche della SAR, si analizzeranno nel dettaglio le ricodifiche per le variabili ordinali e per la regione di residenza degli intervistati. La figura 4.23, si veda l'Appendice C per un'immagine di maggiori dimensioni, riporta solo le traiettorie delle variabili ordinali per il primo piano fattoriale. Come si può notare, si ha una fortissima riduzione nel numero di modalità, che scende da 31 a 12, senza che ci sia un cambiamento apprezzabile nella struttura delle traiettorie. Infatti, nonostante la forte riduzione nel numero delle modalità, il verso delle traiettorie rimane immutato e conse-

guentemente non cambia neppure l'interpretazione del piano fattoriale. Questa circostanza, dimostra che una riduzione delle modalità, basata sulle coordinate dell'ACM non toglie informazioni significative all'interpretazione del fenomeno, rendendo al contempo il piano fattoriale più leggibile e l'analisi più robusta a causa della riduzione del numero di dimensioni.

Tabella 4.23: Modalità della variabile Regione prima dell'applicazione della SAR e dopo l'applicazione della SAR

Modalità originarie	**Modalità ricodificate**
Sardegna Basilicata	**sardegna**[**basilicata**]
Sicilia	**Sicilia**
Campania	**Campania**
Veneto	**Veneto**
Marche Lombardia Lazio Emilia Romagna Molise Trentino Val D'Aosta FVG Liguria Umbria	**marche-lombardia-lazio-emilia** [**molise∼ trentino∼ val d'aosta∼ fvg∼ liguria∼ umbria**]
Puglia Abruzzo	**puglia**[**abruzzo**]
Piemonte	**piemonte**
Calabria	**calabria**
Toscana	**toscania**

L'ultima variabile che si analizzerà, è la regione di residenza. La tabella 4.23, riporta le modalità di questa variabile prima della ricodifica, coincidenti con le singole regioni, e le modalità dopo la ricodifica. La modalità che maggiormente attira l'attenzione è senza dubbio:
marche-lombardia-lazio-emilia
[molise∼trentino∼val d'aosta∼fvg∼liguria∼umbria].
Questa ricodifica è senza dubbio piuttosto anomala e diventa veramente difficile coglierne il significato. Sicuramente dovendo fare una ricodifica a priori,

e tenendo in considerazione i diversi aspetti e le diverse e note caratteristiche socio-economiche delle regioni, questa ricodifica non avrebbe mai avuto luogo. Il punto centrale, è però, che se attraverso la SAR si è arrivati ad una simile aggregazione, questo significa che le regioni occupavano sul piano una posizione simile che ne denota la similarità di comportamento rispetto a tutte le altre regioni. Inoltre, avendo fissato una soglia oltre la quale l'aggregazione non avrebbe avuto luogo, se tale aggregazione è avvenuta, questo significa che queste regioni occupavano delle posizioni estremamente ravvicinate nel piano.

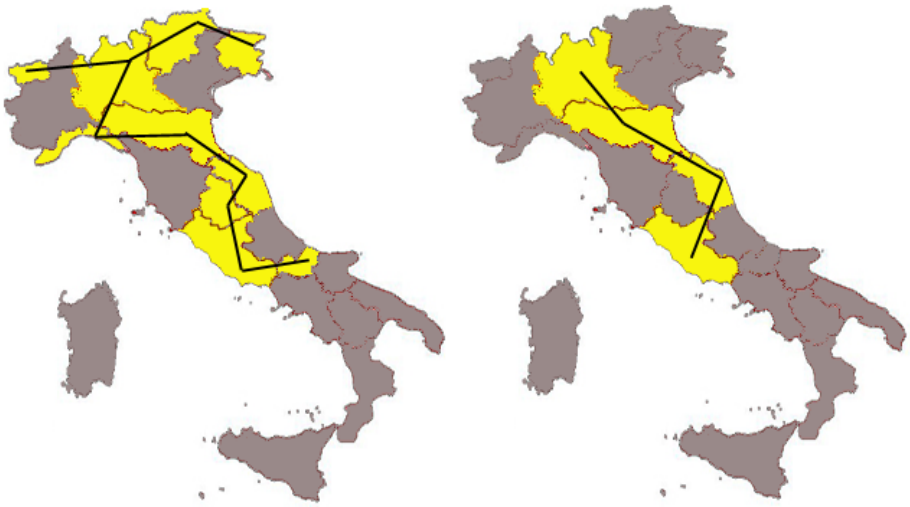


Figura 4.7: Rappresentazione grafica di una modalità della variabile Regione dopo la ricodifica.

La figura 4.7 riporta il cartogramma relativo alla modalità sopraccitata, è permette di compiere ulteriori considerazioni. Analizzando la parte sinistra del piano, si nota immediatamente che, ad esclusione della Valle D'Aosta, le regioni sono territorialmente contigue ed è come se fossero unite tra loro da un'asse immaginario. Se si considera solo la parte attiva della variabile, *marche-lombardia-lazio-emilia*, questa circostanza è ancora più evidente, come mostra il lato destro della figura 4.7. Che cosa si può concludere da quanto detto? Probabilmente queste regioni sono tra loro collegate da qualche fattore che va oltre l'analisi diretta della matrice dei dati. Essendo il tema dell'analisi la comprensione delle

caratteristiche degli utenti, si potrebbe azzardare che queste regioni siano state servite prima di altre da una particolare tecnologia, che ne giustificherebbe la vicinanza geografica. Il fatto che la Lombardia il Lazio e l'Emilia siano regioni palesemente più importanti, geograficamente, economicamente e politicamente, non fa che rafforzare questa ipotesi. Inoltre basandosi sui più noti modelli territoriali, si potrebbe affermare che le Marche, essendo geograficamente vicina a queste ne abbia subito l'influsso. Ovviamente allo stato attuale delle conoscenze disponibili, queste sono solo ipotesi. Il fatto più importante è però senza dubbio che una simile ricodifica, porta a delle riflessioni ed ad ulteriori spunti di analisi. Da ultimo è chiaro che una classica ricodifica, basata magari sulla sola vicinanza territoriale o su altre considerazioni a priori, non avrebbe permesso di cogliere questo, eventuale, aspetto nascosto. In conclusione, questa variabile racchiude in se lo spirito sottostante alla Ricodifica Sequenziale delle Modalità: non potendo conoscere tutti gli aspetti sottostanti il fenomeno, è meglio nelle procedure di ricodifica basarsi sulle risultanti relazioni tra le variabili che procedere soggettivamente.

4.4 La SAR come strumento per il Data Mining

A seguito della diminuzione del numero delle modalità, il piano fattoriale diventa chiaramente più leggibile. Il dataset utilizzato nella sezione precedente, mette in luce i limiti della ACM classica ed i vantaggi della SAR quando si analizzano matrici contenenti variabili con un numero eccessivo di modalità. Ricordando quanto detto nelle sezioni precedenti, un numero eccessivo di modalità può creare problemi per vari motivi. Infatti, un numero eccessivo di modalità:

1. può rendere instabile l'analisi a causa della eccessiva frammentazione delle variabili che può portare a modalità con basse frequenze;

2. rende difficoltosa l'interpretazione del piano fattoriale a causa del numero eccessivo di etichette da visualizzare;

3. obbliga a ricodifiche che possono violare il principio dell'equivalenza distributiva e far perdere importanza ad una variabile.

D'altra parte, l'analisi della figura 4.4, evidenzia chiaramente che anche dopo la ricodifica si possono avere problemi di visualizzazione. Nonostante con un leggero ingrandimento della parte centrale, si veda la figura figura 4.22, il problema venga del tutto risolto, problema che invece appare irrisolvibile prima della ricodifica, è chiaro che oltre un certo limite qualunque ricodifica diviene inefficace. Chiaramente non è possibile definire a priori un numero di modalità massimo oltre il quale diventa impossibile la visualizzazione. La fissazione a priori non è possibile in quanto i problemi di visualizzazione dipendono oltre che dal numero dei punti, anche dalla loro posizione nel piano. Ad ogni modo, superato tale numero la SAR diventa inefficace. Ovviamente continua a mantenere la sua utilità in quanto continua ad essere un valido supporto per i punti 1 e 3. Ad ogni modo, ancora una volta la SAR può essere adattata per risolvere problemi di visualizzazione anche con un numero elevatissimo di variabili.

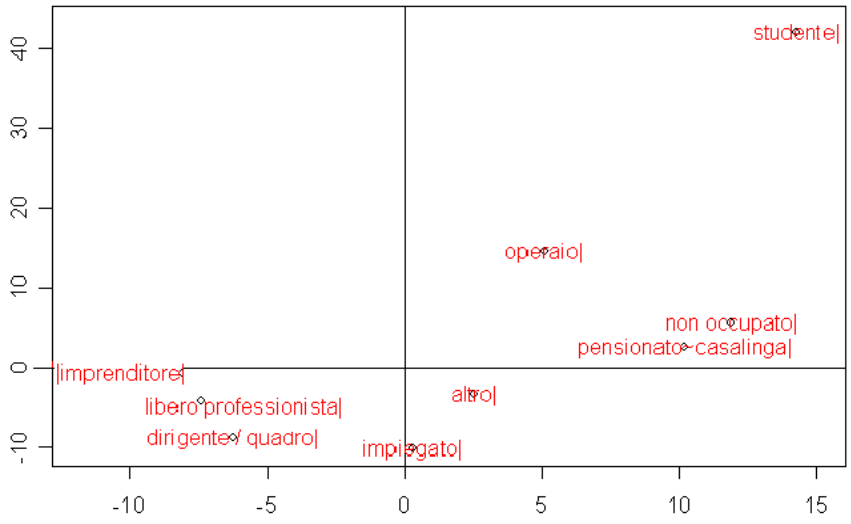


Figura 4.8: Rappresentazione grafica della prima variabile con il contributo assoluto più elevato: nessuna ricodifica.

Si supponga di essere in un contesto di Data Mining. Essendo lo scopo quello di trovare le relazioni più importanti, SAR procede per passi alla visualizzazione delle variabili, proiettandole in base al contributo che forniscono all'analisi. Il primo passo consiste nel selezionare la variabile col contributo più elevato, e proiettare le sue modalità sul piano fattoriale. Si supponga che la variabile

più importante sia la *Professione*, si avrà allora la proiezione delle modalità di questa variabile come mostrato in figura 4.8. Successivamente, si procede con la ricodifica scegliendo una valore basso per la soglia. Con una soglia uguale al 15% della distanza massima, si ottiene l'aggregazione dei *Pensionati*, *Casalinghe*, *Non occupati* da una parte e *Liberi professionisti*, *Dirigenti/quadri* dall'altra, come mostrato in figura 4.9.

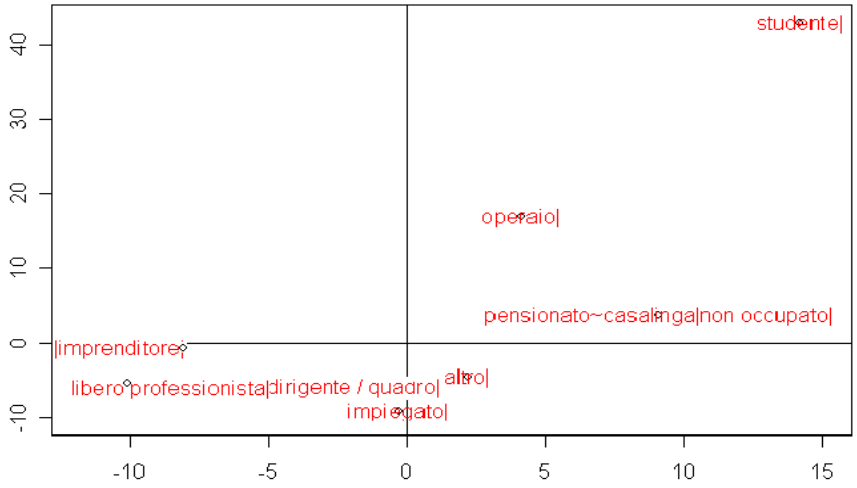


Figura 4.9: Rappresentazione grafica della prima variabile con il contributo assoluto più elevato: soglia=15%.

Si procede scegliendo una soglia più elevata: 25%. A seguito dell'incremento della soglia ai *Liberi professionisti* e *Dirigenti/quadri*, vengono aggiunti gli *Impiegati*. Ci si potrebbe fermare qui in quanto si sono identificate due categorie ben distinte: da una parte una serie di occupazioni mediamente o altamente qualificate e dall'altra persone non attive nel mercato del lavoro. Si supponga ad ogni modo di incrementare ulteriormente la soglia al 30%. Ai *Pensionati*, *Casalinghe*, *Non occupati* si aggiungono coloro che hanno dichiarato *Altre professioni*. Dall'analisi del piano, figura 4.10, ci si rende conto che qualunque altra ricodifica accorperebbe modalità troppo diverse tra loro per cui ci si ferma qui. Ovviamente il punto di arresto della ricodifica è abbastanza soggettivo. Questo permette però di unire le conoscenze del ricercatore sull'argomento con l'oggettività della ricodifica. E' importante sottolineare che il ricercatore decide solo il punto di arresto, ma non può in alcun modo entrare nel merito delle aggrega-

zioni. In altre parole egli potrà solo decidere il grado di sintesi del fenomeno ma non il tipo di sintesi. E' altrettanto chiaro che la scelta del miglior compromesso tra sintesi e capacità informativa è lasciata alla sensibilità dell'analista.

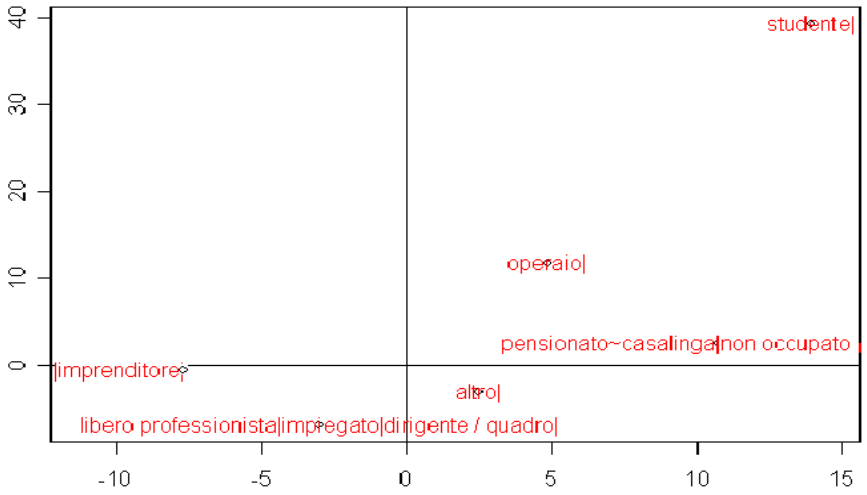


Figura 4.10: Rappresentazione grafica della prima variabile con il contributo assoluto più elevato: soglia=25%.

Una volta definita la migliore ricodifica, se possibile, si rinomineranno le nuove modalità in modo da ridurre ulteriormente lo spazio occupato nel piano. I risultati di quest'ultimo passo, sono riportati nella figura 4.12. Ricodificata la variabile col maggior contributo assoluto, si passa alla seconda in graduatoria. Si supponga che la seconda variabile in ordine di importanza sia *Numero di prodotti tecnologici posseduti*. Il risultato della proiezione delle modalità di questa variabile sul piano sono riportate nella figura 4.13. In questo caso si sta proiettando una variabile che ha gia subito una ricodifica con una soglia uguale al 20%.

Si prosegue con la ricodifica scegliendo un valore leggermente più elevato per la soglia ad esempio il 30%. Il risultati dell'aggregazione per questo valore di soglia, sono riportati in figura 4.14. E' di immediata constatazione che nonostante la riduzione del numero delle modalità, il significato attribuito al paino fattoriale non muta. Successivamente si incrementa la soglia, riducendo ulteriormente le modalità, finche si ritiene che il significato del piano non muti. Terminata la ricodifica della seconda variabile, si passa alla terza e cosi via. E' chiaro che

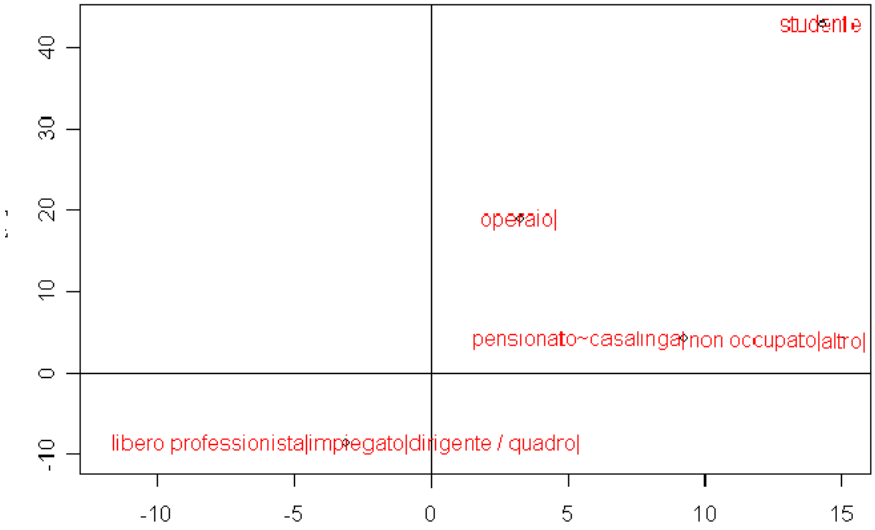


Figura 4.11: Rappresentazione grafica della prima variabile con il contributo assoluto più elevato: soglia=30%.

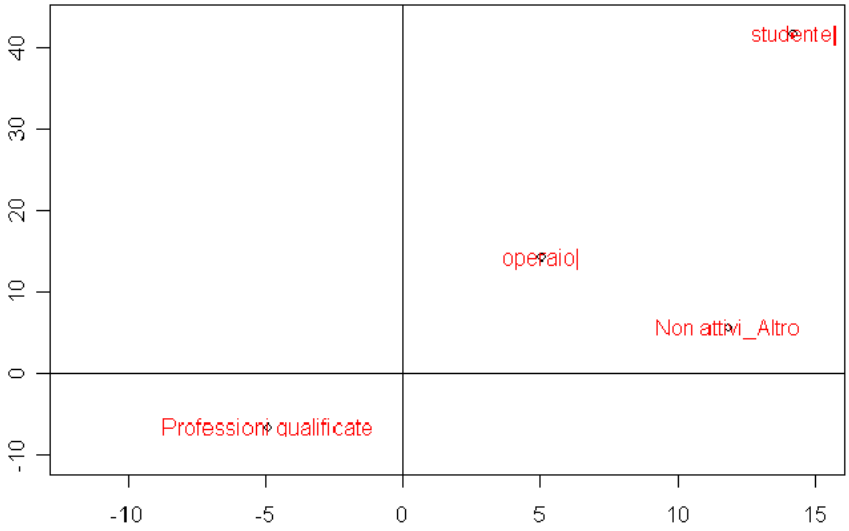


Figura 4.12: Rappresentazione grafica della prima variabile con il contributo assoluto più elevato: ridefinizione delle etichette.

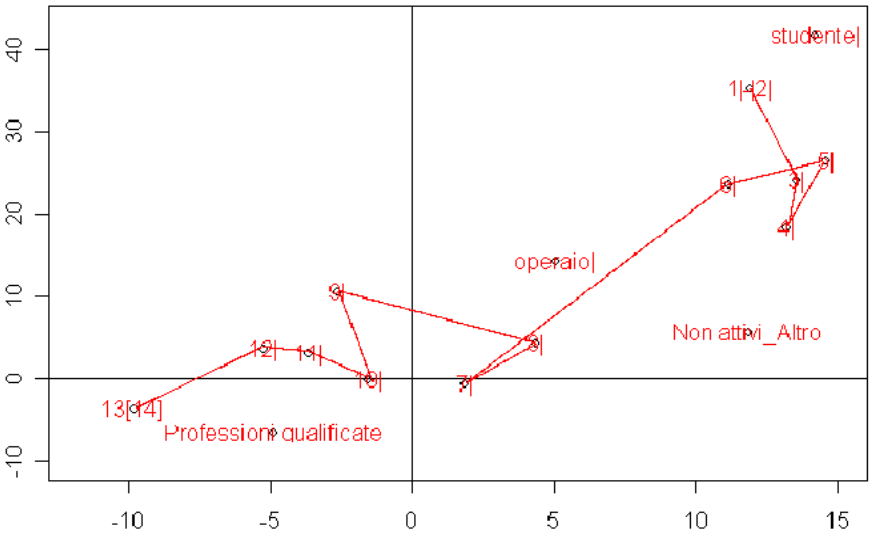


Figura 4.13: Rappresentazione grafica della prime due variabili con il contributo assoluto più elevato: soglia=20%

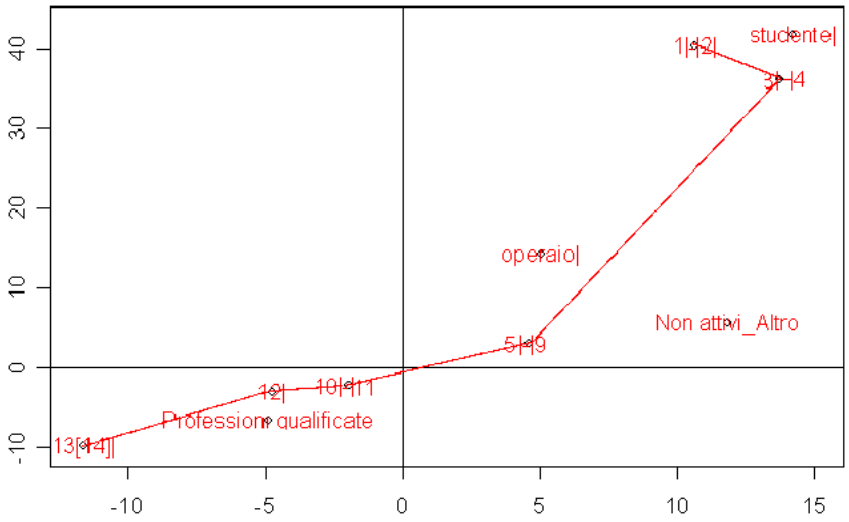


Figura 4.14: Rappresentazione grafica della prime due variabili con i contributi assoluti più elevati: soglia=30%.

in questo modo si avrà la possibilità di esplorare le relazioni più importanti all'interno di una matrice di dati riducendo al minimo la saturazione del piano fattoriale. Questo processo continua finché:
a) le variabili mostrano dei contributi significativi;
b) si raggiunge la saturazione dello schermo;
c) il ricercatore ha un'idea abbastanza chiara del fenomeno e preferisce proseguire autonomamente.

La differenza fondamentale rispetto alla procedura esposta nella sezione precedente, consiste nella gerarchia delle variabili. Mentre in un processo di ricodifica generico, teso principalmente a migliorare la leggibilità del piano, le variabili possono anche essere ricodificate simultaneamente, in questo caso, è imprescindibile la fissazione di un ordine che determini da quale variabile debba iniziare la procedura di ricodifica.

4.5 La SAR come strumento di supporto per le decisoni

La Ricodifica Sequenziale Automatica, può anche essere impiegata solo come strumento utile per decidere che tipo di accorpamenti da fare e non unicamente come strumento per ridurre le modalità. In altre parole anche quando il numero di modalità rende facilmente interpretabile il piano fattoriale, può essere desiderabile o necessario accorpare alcune modalità. L'esempio più classico è la ricodifica in classi. Si è visto nella sezione 4.3.2, che la ricodifica in classi ottenuta tramite SAR, ha delle caratteristiche peculiari e risulta talmente sovra-adattata ai dati da mettere spesso in secondo piano le altre variabili. Se questo è da un lato un inconveniente, dall'altro permette di individuare gruppi particolari d'osservazioni che possono apportare informazioni ulteriori sul fenomeno investigato. L'analisi di una modalità anomala nel caso della ricodifica delle regioni, sempre nella sezione 4.3.2, ha sollevato interrogativi ed ulteriori spunti di analisi portando all'attenzione una similarità che sarebbe passata sicuramente inosservata con una ricodifica classica. In conclusione, è pur sempre interessante confrontare una ricodifica soggettiva, basata sulle conoscenze e convinzioni del ricercatore, con una totalmente asettica ottenuta automaticamente, ma pur sempre, basata sui principi ispiratori dell'Analisi dei Dati.

Conclusioni e ulteriori sviluppi

L'analisi dei dati viene spesso connotata con aggettivi tendenti a metterne in luce la modernità in contrasto alla scuola classica. La Moderna Analisi Multidimensionale dei Dati, si è distinta rispetto all'impostazione classica grazie alla capacità di dare risposta alla crescente disponibilità di dati e rendere possibile il trattamento simultaneo di numerose variabili. Se la connotazione moderna è derivata dall'incremento di dati e dall'implementazione di strumenti adatti a trattarli, la situazione odierna dovrebbe portare alla definizione di Statistica Contemporanea. Infatti, metodologie che qualche decennio fa erano perfettamente in grado di trattare in modo adeguato il tipo di dati a disposizione, oggi cominciano a mostrare i propri limiti. L'Analisi delle Corrispondenze Multiple in particolare, soffre per varie ragioni l'eccessivo numero di modalità. Nel presente lavoro si è cercato di proporre alcune modifiche metodologiche allo scopo si superare questo tipo di problemi. La Ricodifica Automatica Sequenziale, si è mostrato uno strumento flessibile e capace di dare risposta a numerosi problemi causati principalmente dall'eccessivo numero di modalità. La varietà di problemi affrontati, rende difficile definire se la SAR sia una variazione metodologica o più semplicemente un tool per migliorare la leggibilità del piano fattoriale. I metodi fattoriali prendono connotazioni diverse a seconda delle discipline in cui sono utilizzati. Tendenzialmente essi sono comunque visti come metodi di riduzione della dimensionalità o riduzione del rumore. La SAR tende ad accorpare le modalità che, rispetto agli assi individuati dai metodi stessi, non apportano, dal punto di vista puramente geometrico, nessuna informazione significativa. L'eliminazione di una modalità equivale, sempre dal punto di vista geometrico, equivale alla riduzione di una dimensione. Seguendo questa impostazione, la SAR può essere vista come una variazione metodologica dell'ACM. Uno dei

motivi per cui la percentuale di inerzia spiegata dall'ACM non viene considerata attendibile, e che essa diminuisce all'aumentare delle dimensioni, per cui se si aggiungessero variabili senza nessun significato (rumore) la percentuale di variabilità spiegata diminuirebbe automaticamente. Ma questo non significa, ovviamente, che il sistema di assi trovato spieghi meno del fenomeno analizzato (Gherghi, Lauro 2000). Spesso si associa la esistenza di rumore dei dati alla presenza di variabili che nulla hanno a che vedere con i dati d'interesse. Più raramente si associa il rumore alla presenza di modalità superflue, ma questo non esclude la presenza di modalità inutili che non apportano nessuna informazione aggiuntiva se non rumore. Per meglio spiegare questo concetto, e meglio giustificare la metodologia finora proposta, si segua un ragionamento inverso a quello seguito finora.

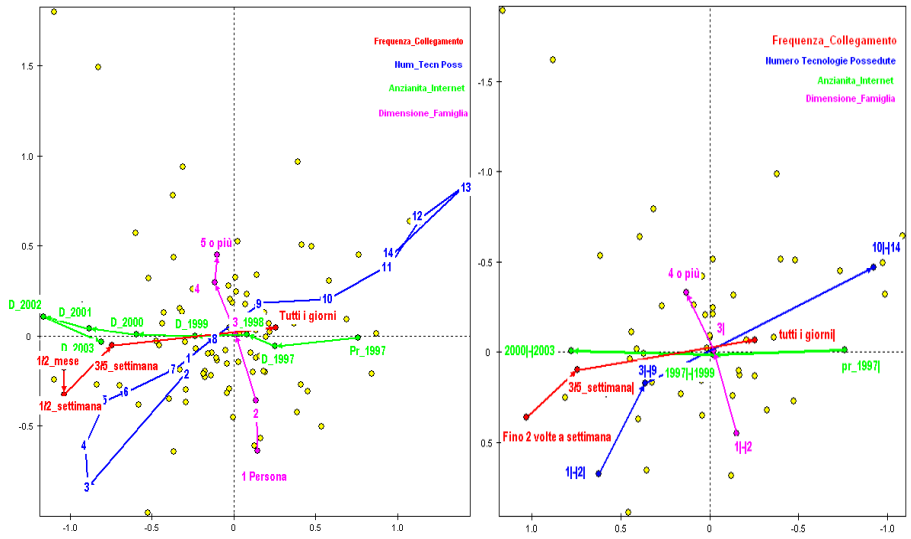


Figura 4.15: Confronto delle traiettorie delle variabili ordinali prima e dopo la ricodifica.

Si consideri come situazione iniziale una serie di variabili con modalità uguali a quelle riportate nel lato destro della figura 4.15. Queste variabili hanno complessivamente 12 modalità. Si supponga ora che venga proposto, di incrementare il dettaglio dell'informazione aumentando il numero di modalità per ciascuna variabile. Si porti il numero totale delle modalità a 31. Il risultato

che si otterrebbe, è riportato nel lato sinistro della stessa figura. E' chiaro che non si è apportata nessuna informazione significativa, in quanto il significato del piano non cambia e quello che si è aggiunto è solo inutile rumore. La SAR, permette di passare dalla situazione descritta nella parte sinistra della figura, a quella decritta nella parte destra. Si Ancora una volta quindi SAR appare come una variazione metodologica che permette una riduzione del rumore dovuto ad un eccessivo numero di modalità. D'altra parte come mostrato nelle sezioni 4.4 e 4.5 la SAR può essere vista anche ed unicamente come uno strumento di ausilio. Tutte le applicazioni presentate in questo lavoro si sono basate sulle coordinate delle modalità sui primi due assi. Nonostante in diverse applicazioni si siano utilizzati più di due assi, si è ritenuto in questa fase privilegiare una metodologia che permettesse un confronto visuale tra i risultati della ricodifica ed i dati originali. Infatti, anche considerando solo due assi il confronto visuale è assai arduo. Non di rado è sembrato di assistere all'accorpamento di modalità apparentemente lontane nel piano. L'ispezione visiva delle modalità prima e dopo la ricodifica, comporta non pochi problemi. In primo luogo la visualizzazione rettangolare non è idonea a rappresentare le distanze, inoltre la distanza tra due modalità è pesata per l'importanza dell'asse: maggiore è la varianza dell'asse minore sarà, a parità di altri fattori, la distanza tra due modalità in quell' asse. La gerarchia imposta fa si che le coordinate cambino dopo ogni ricodifica, l'ispezione visiva, permette di vedere solo l'inizio e la fine del processo ma non i passi intermedi. Quando si passa a un numero di assi superiore a due tutte queste ed altre problematiche devono essere affrontate (Lauro, Decarli 1982). Naturalmente aumentano anche le potenzialità dell'analisi, e questo è uno dei primi passi da compiere in futuro. Gli algoritmi presentati nelle sezioni precedenti, sono piuttosto autonomi ed ognuno di essi può essere utilizzato per risolvere i problemi specifici per cui è stato implementato. In un'ottica di ottimizzazione delle procedure una strada da seguire, è senza dubbio quella di unire alcuni di essi in modo da ridurre lo sforzo computazionale. Soprattutto la procedura delle modalità Semi-Attive può essere combinata con le altre in modo da ottenere direttamente l'aggregazione della modalità a bassa frequenza evitando al contempo la loro eccessiva influenza e permettendo di sfruttare tutti i vantaggi di poter fissare una soglia di sensibilità dell'aggregazione. La riduzione delle modalità avviene esclusivamente grazie ad una variazione metodologica. Lo sviluppo di strumenti grafici interattivi permetterebbe di migliorare ulteriormente la leggibilità del piano fattoriale e soprattutto permettere all'analista una più facile ed immediata navigazione in presenza di enormi moli di dati. Spesso ci

si trova in presenza di variabili, soprattutto ordinali o numeriche, che seguono traiettorie identiche o che comunque sono la rilevazione in forme diverse dello stesso fenomeno. In casi simili, oltre alla riduzione delle modalità, si potrebbero ridurre anche le variabili allo scopo di rendere ulteriormente sintetica l'analisi. Partendo dall'esempio delle regioni e delle variabili ordinali, si nota come l'inserimento di vincoli, possa migliorare la ricodifica delle modalità. Sarebbe dunque di enorme utilità l'implementazione di algoritmi che permettano all'analista di inserire diversi tipi di vincoli relativamente alle diverse situazioni che si possono presentare. Il numero eccessivo di modalità non è, ovviamente, un problema che affligge solo l'ACM. La SAR basata sull'ACM, può essere orientata alla risoluzione di problemi relativi ad altre metodologie. Un piccolo passo in questa direzione è gia stato fatto (Mola, Mascia. 2006). Oltre a migliorare l'interazione tra diverse tecniche, il passo più importante da compiere, è quello di adattare la SAR in modo tale che la riduzione delle modalità o delle variabili, avvenga sulla base dei risultati della metodologia su cui deve essere applicata.

Appendice A

Come applicazione della ricodifica in classi di una variabile continua, si consideri come variabile continua l'età degli intervistati. Allo scopo di confrontare la ricodifica della variabile attraverso SAR con altri metodi, si ricodifica la variabile in altri due modi, classi equi-ampie e classi equi-frequenti.

Tabella 4.24: Correlazione della variabile età, sul primo asse fattoriale ($\mathbf{COR}_1$) e sul secondo asse fattoriale ($\mathbf{COR}_2$)

Modalità	$\mathbf{COR}_1$	$\mathbf{COR}_2$
Età	0.25	0.11

Utilizzando la variabile come supplementare, non si riscontra una particolare influenza di questa variabile, infatti, le correlazioni sia sul primo che sul secondo asse sono piuttosto basse, come riportato in tabella 4.24.

Tabella 4.25: Modalità della variabile età ricodificate attraverso SAR, contributi assoluti, sul primo asse fattoriale ($\mathbf{CTA}_1$) e sul secondo asse fattoriale ($\mathbf{CTA}_2$)

Modalità	$\mathbf{CTA}_1$	$\mathbf{CTA}_2$
18-24	0.00	6.25
25-38	0.38	0.00
39-44	2.13	0.01
45-62	4.81	2.02
Totale	**7.31**	**8.38**

Allo scopo di utilizzare la variabile come attiva e non supplementare, la si può ricodificare in classi. La tabella 4.25, riporta i contributi della variabile ricodificata attraverso la procedura SAR per i primi due assi. Dall'analisi della

tabella e della figura 4.16 si può notare chiaramente che attraverso la ricodifica e l'utilizzo della variabile come attiva, essa acquista importanza nell'analisi.

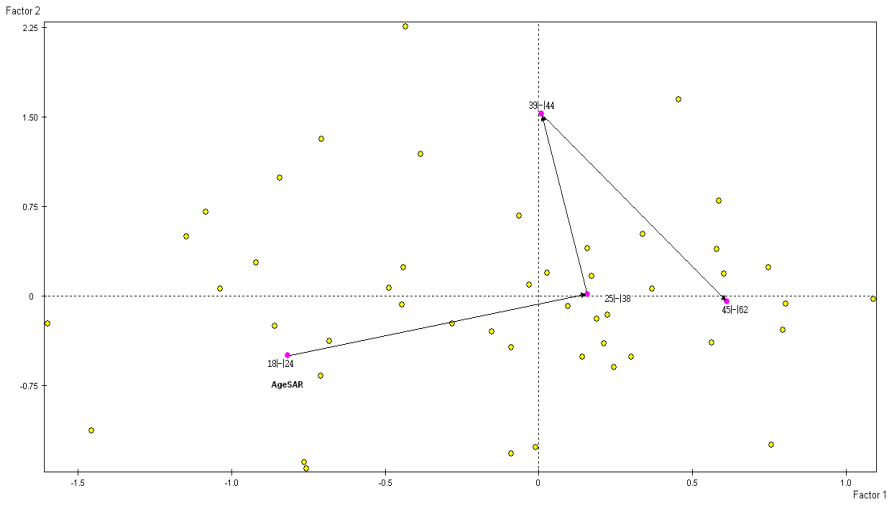


Figura 4.16: Rappresentazione fattoriale della variabile età (SAR).

Il fenomeno appena descritto è piuttosto comune, ed è dovuto al fatto che in questo secondo modo la variabile influenza la direzione degli assi fattoriali.

Tabella 4.26: Modalità della variabile età ricodificate col metodo delle classi equi-ampie, contributi assoluti, sul primo asse fattoriale ($\mathbf{CTA}_1$) e sul secondo asse fattoriale ($\mathbf{CTA}_2$)

Modalità	$\mathbf{CTA}_1$	$\mathbf{CTA}_2$
18-28	3.80	2.63
29-39	0.96	0.27
40-50	0.07	0.45
51-62	3.91	0.85
Totale	**8.74**	**4.20**

Per un ulteriore confronto, si è ricodificata la variabile in modo da ottenere classi equi-frequenti. I contributi sono ancora rilevanti, si veda la tabella 4.26 e la figura 4.17, ma comunque inferiori a quelli ottenuti attraverso una Ricodifica Sequenziale Automatica. L'utilizzo di classi equi-frequenti porta sicuramente

ad un miglioramento della rappresentazione fattoriale ma ad un peggioramento dei contributi. I risultati per questa ricodifica, sono riportati in tabella 4.27 ed in figura 4.18.

Tabella 4.27: Modalità della variabile età ricodificate col metodo delle classi equi-frequenti, contributi assoluti, sul primo asse fattoriale ($\mathbf{CTA}_1$) e sul secondo asse fattoriale ($\mathbf{CTA}_2$)

Modalità	$\mathbf{CTA}_1$	$\mathbf{CTA}_2$
18-24	2.02	0.68
25-30	2.23	0.01
31-39	0.22	0.12
40-62	4.99	0.71
Totale	**9.45**	**2.52**

Dal confronto delle tre tabelle, si evince che la ricodifica SAR ottiene dei contributi assoluti maggiori delle altre ricodifiche nei primi due assi.

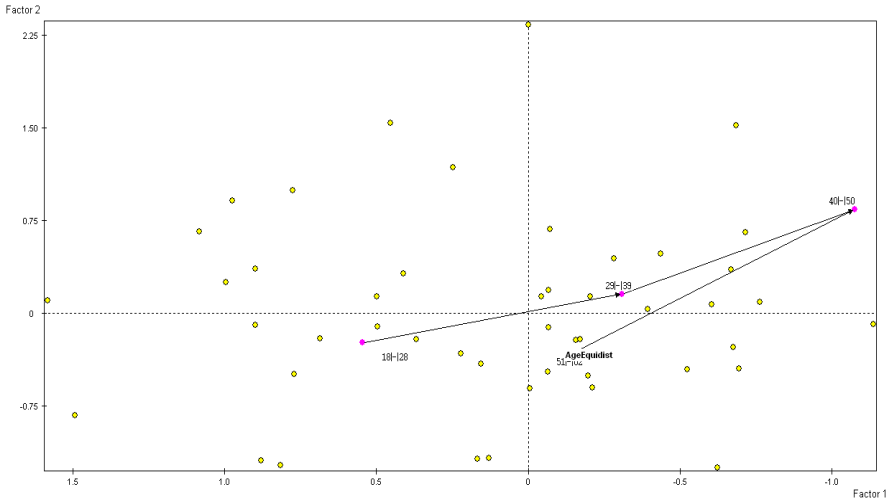


Figura 4.17: Rappresentazione fattoriale della variabile età (equi-ampie).

La figura 4.19, riporta i contributiassoluti per le tre ricodifiche per i primi 10 assi mentre la figura 4.20, riporta i contributi cumulati. La ricodifica SAR mostra contributi assoluti superiori per i primi due assi e tendenzialmente inferiori per

i successivi. Quest'andamento non deve stupire in quanto la ricodifica è basata sui risultati dei primi due assi.

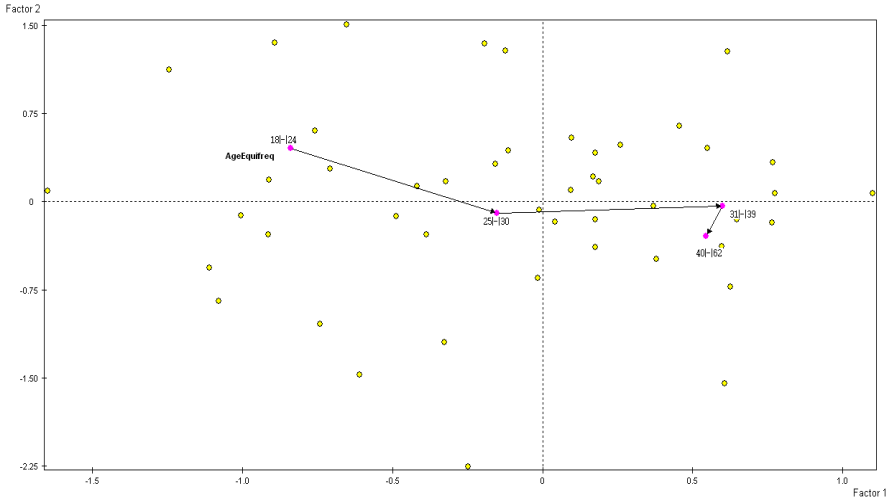


Figura 4.18: Rappresentazione fattoriale della variabile età (equi-frequenti).

Purtroppo non sempre la ricodifica SAR garantisce contributi più elevati. Nel caso in cui la distribuzione presenti delle forti asimmetrie, questo tipo di ricodifica tende ad identificare una classe con una frequenza elevata e più classi con basse frequenze. Questo fenomeno è probabilmente dovuto sia alla circostanza che le coordinate sono ottenute in supplementare sia al fatto stesso che l'ACM tende a mettere in evidenza i comportamenti che si discostano dal profilo medio. Per cui la ricodifica SAR tende a separare una grossa massa, identificabile come comportamento generale, e tante piccole classi con comportamenti peculiari. Se questa caratteristica può apparire desiderabile, bisogna comunque rimarcare i bassi contributi generalmente ottenuti. Probabilmente un compromesso tra le due ricodifiche, sarebbe il risultato più auspicabile.

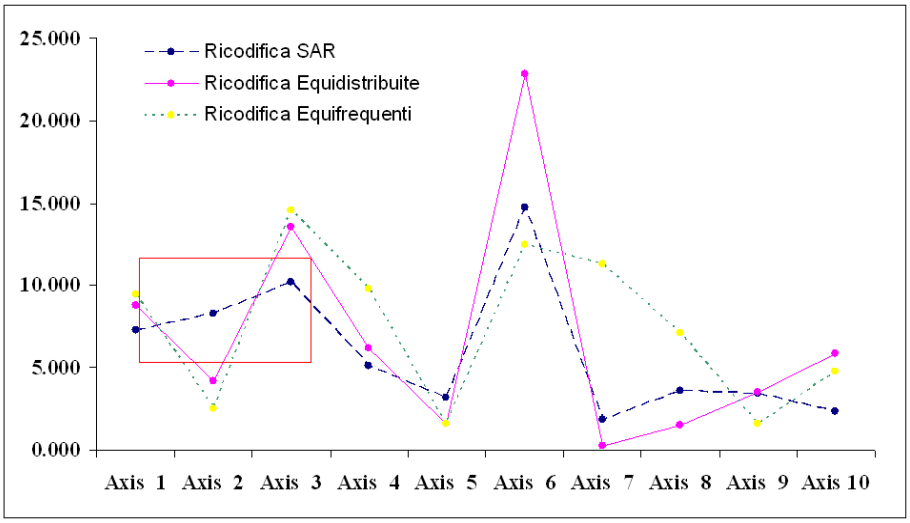


Figura 4.19: Andamento dei contributi assoluti per i primi 10 assi.

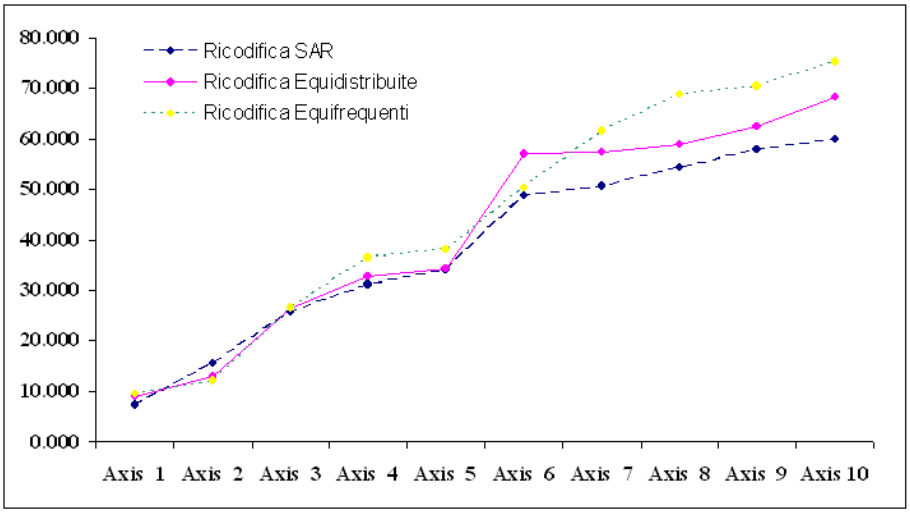


Figura 4.20: Andamento dei contributi assoluti cumulati per i primi 10 assi.

Appendice B

Tabella 4.28: Modalità della variabile Professione prima dell'applicazione della SAR e dopo l'applicazione della SAR

Modalità originarie	**Modalità ricodificate**
Libero Professionista Impiegato Imprenditore Dirigente/Quadro	**Professioni qualificate**
Operaio	**Operaio**
Studente	**Studente**
Pensionato Non Occupato Casalinga Altro	**Pensionato[casalinga]|non occupato| altro**

Tabella 4.29: Modalità della variabile Luogo di collegamento prima dell'applicazione della SAR e dopo l'applicazione della SAR

Modalità originarie	**Modalità ricodificate**
Lavoro	**Lavoro**
Casa-Lavoro	**Casa-Lavoro**
Casa Altro	**Casa[Altro]**

Tabella 4.30: Modalità della variabile Numero di prodotti tecnologici posseduti prima dell'applicazione della SAR e dopo l'applicazione della SAR

Modalità originarie	**Modalità ricodificate**
1 2	**1| − |2**
3 4 5 6 7 8 9	**3| − |9**
10 11 12 13 14	**10| − |14**

Tabella 4.31: Modalità della variabile Grado di istruzione prima dell'applicazione della SAR e dopo l'applicazione della SAR

Modalità originarie	**Modalità ricodificate**
Nessuna scuola Licenza Elementare Licenza Media	**Fino a Licenza Media**
Diploma	**Diploma**
Studente Universitario	**Studente Universitario**
Laurea	**Laurea**

Tabella 4.32: Modalità della variabile Tecnologia principalmente utilizzata prima dell'applicazione della SAR e dopo l'applicazione della SAR

Modalità originarie	**Modalità ricodificate**
Pc PayTv Scanner Dvd	**pc|paytv[dvd∼scanner]**
Webcam	**Webcam**
Videocamera	**Videocamera**
Cellulare	**Cellulare**
Vcr	**Vcr**
Pc Portatile Masterizzatore Stampante	**pc portatile[masterizzatore ∼stampante]**
Home Cinema Fotocamera	**home cinema|fotocamera**

Tabella 4.33: Modalità della variabile Tecnologia di connessione prima dell'applicazione della SAR e dopo l'applicazione della SAR

Modalità originarie	**Modalità ricodificate**
Modem Standard Rete Locale Non so	**modem standard[non so∼rete locale]**
ADSL Altro	**adsl[altro]**
ISDN Satellite	**isdn[satellite]**
Fibra Ottica	**fibra Ottica**

Tabella 4.34: Modalità della variabile Anno di inizio utilizzo di internet prima dell'applicazione della SAR e dopo l'applicazione della SAR

Modalità originarie	**Modalità ricodificate**
Pr 1997	**Pr 1997**
D 1997 D 1998 D 1999	**1997**| − |**1999**
D 2000 D 2001 D 2002 D 2003	**2000**| − |**2003**

Tabella 4.35: Modalità della variabile Frequenza di collegamento prima dell'applicazione della SAR e dopo l'applicazione della SAR

Modalità originarie	**Modalità ricodificate**
1-2 mese 1 2 settimana	**Fino 2 volte a Settimana**
3-5 settimana	**3-5 settimana**
Tutti i giorni	**Tutti i giorni**

Tabella 4.36: Modalità della variabile Dimensione della famiglia prima dell'applicazione della SAR e dopo l'applicazione della SAR

Modalità originarie	**Modalità ricodificate**
1 Persona 2	**1**| − |**2**
3	**3**
4 5 o più	**4 o più**

Tabella 4.37: Modalità della variabile Tipo di interessi prima dell'applicazione della SAR e dopo l'applicazione della SAR

Modalità originarie	**Modalità ricodificate**
Sport	**Sport**
Tecnologia Radio TV	**tecnologia**|**radio tv**
Economia Cucina Bricolage	**economia**|**cucina**|**bricolage**
Natura	**Natura**
Auto	**Auto**
Moda	**Moda**
Lettura Benessere Arte Altri Hobby	**lettura**|**benessere**|**arte**|**altri hobby**

Tabella 4.38: Modalità della variabile Provider utilizzato prima dell'applicazione della SAR e dopo l'applicazione della SAR

Modalità originarie	**Modalità ricodificate**
Virgilio	**Virgilio**
Tiscali Libero Altro Non so	**tiscali**|**libero**|**altro[non so]**
AliceTi	**aliceTi**
Fastweb	**fastweb**
Tele2 Kataweb	**tele2[kataweb]**

Tabella 4.39: Modalità della variabile Regione prima dell'applicazione della SAR e dopo l'applicazione della SAR

Modalità originarie	**Modalità ricodificate**
Sardegna Basilicata	**sardegna**[**basilicata**]
Sicilia	**Sicilia**
Campania	**Campania**
Veneto	**Veneto**
Marche Lombardia Lazio Emilia Romagna Molise Trentino Val D'Aosta FVG Liguria Umbria	**marche-lombardia-lazio-emilia** [**molise∼ trentino∼ val d'aosta∼ fvg∼ liguria∼ umbria**]
Puglia Abruzzo	**puglia**[**abruzzo**]
Piemonte	**piemonte**
Calabria	**calabria**
Toscana	**toscania**

Appendice C

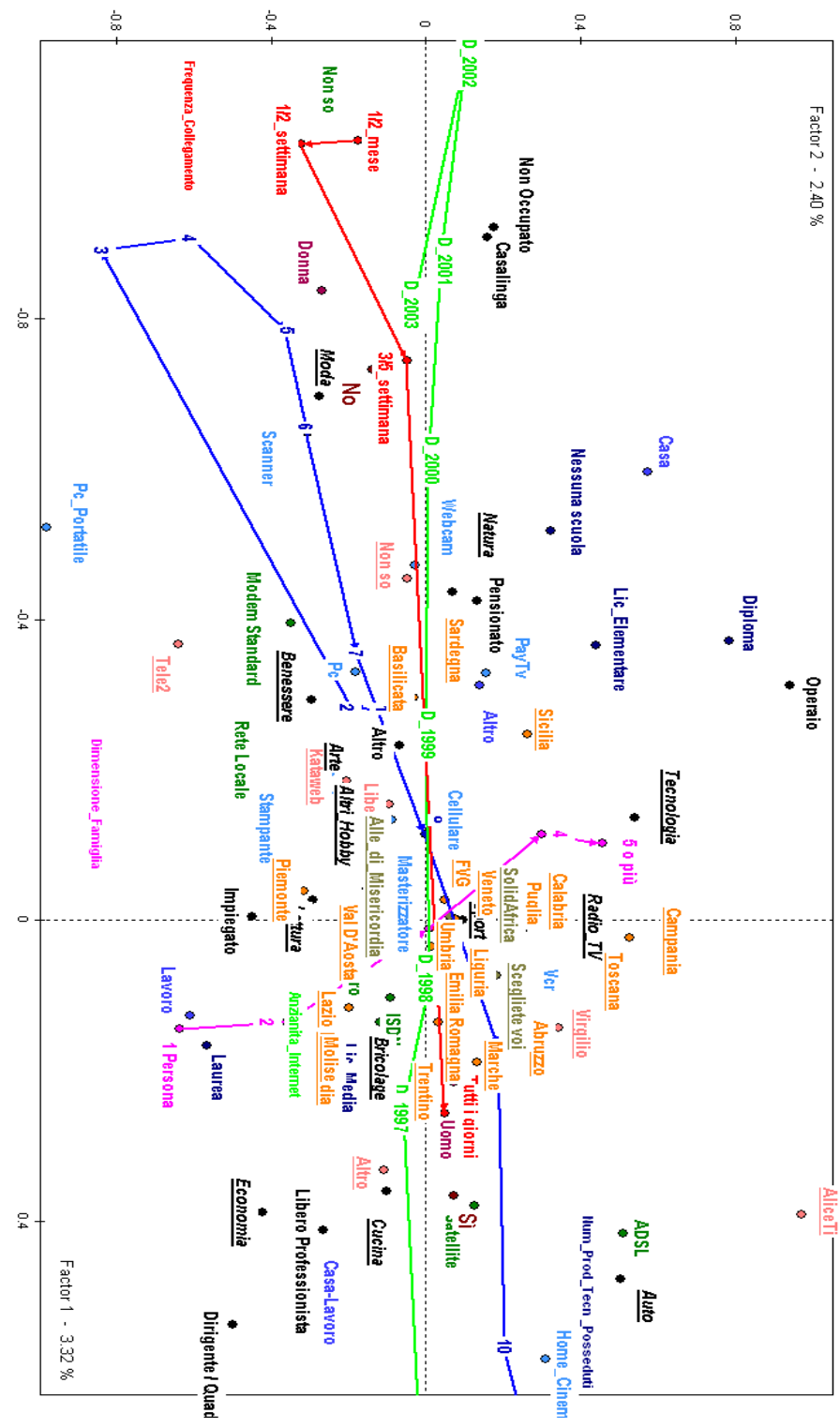


Figura 4.21: Visualizzazione dei profili colonna rispetto al primo piano fattoriale prima della ricodifica: ingrandimento della parte centrale.

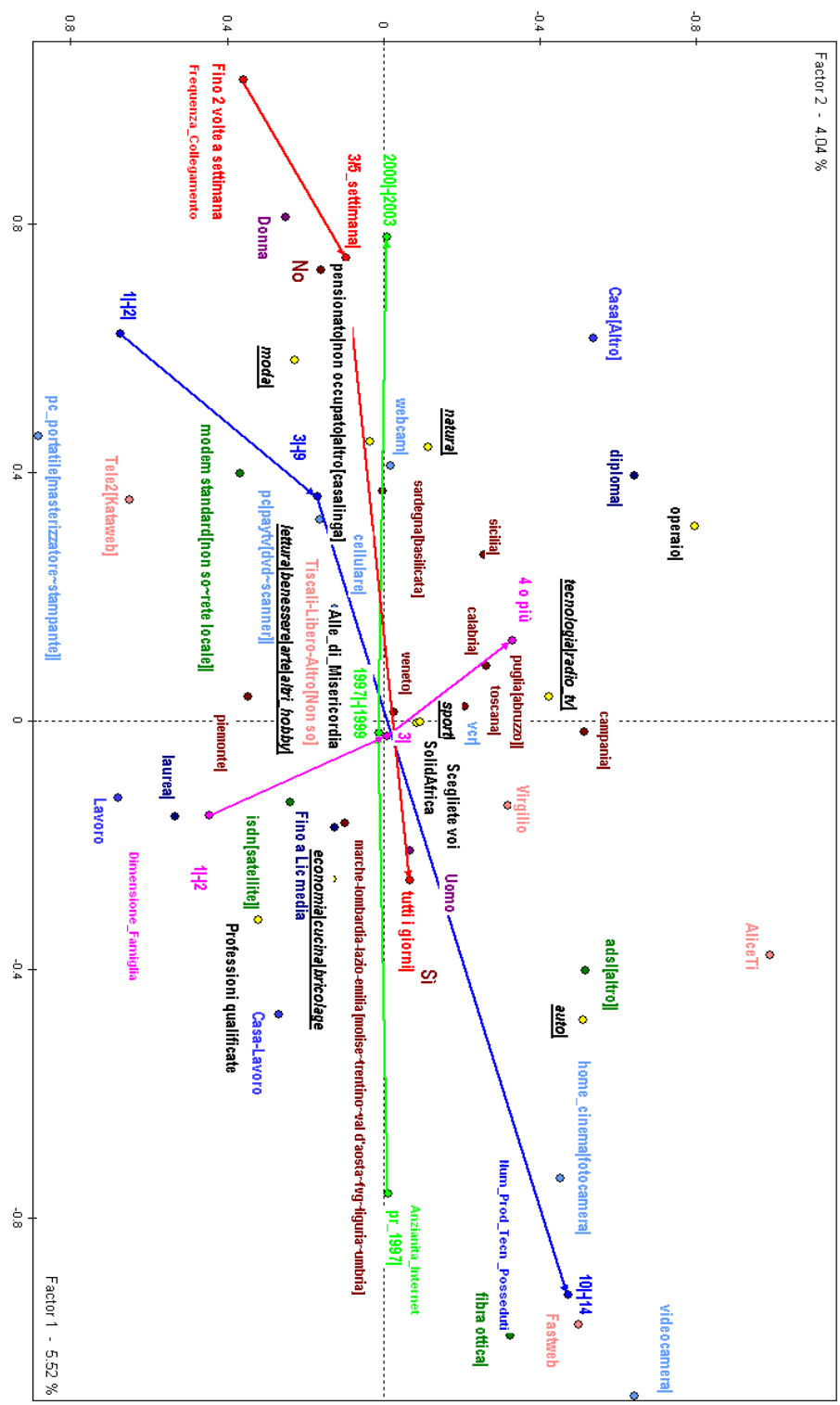


Figura 4.22: Visualizzazione della nuvola dei profili colonna rispetto al primo piano fattoriale dopo la ricodifica: ingrandimento della parte centrale.

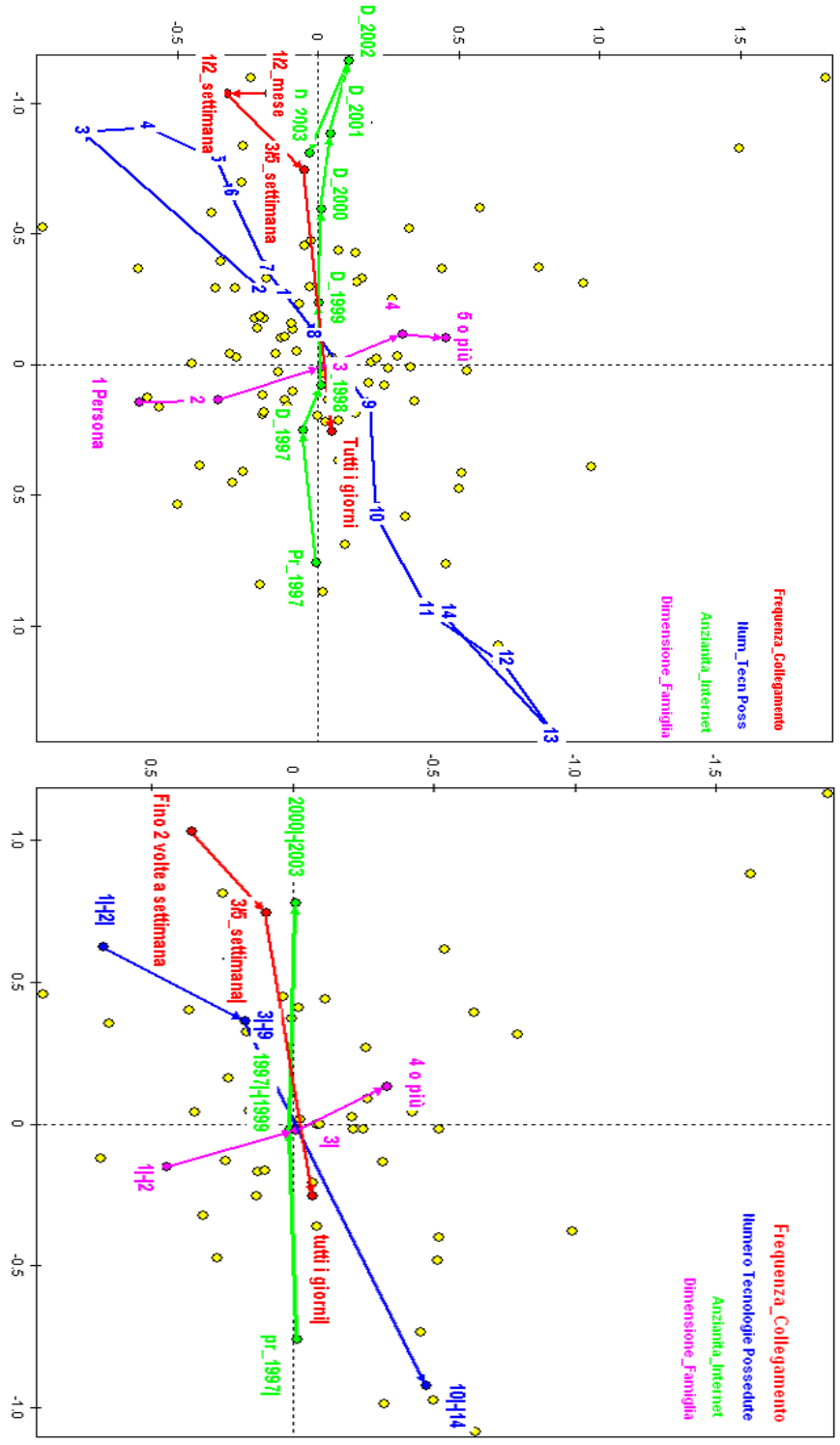


Figura 4.23: Confronto delle traiettorie delle variabili ordinali prima e dopo la ricodifica.

Bibliografia

Agresti, A. (2002), *Categorical Data Analysis*, second edn, Wiley Series in probability and statistics.

Benzécri, JP. (1973), *L'Analyse des Données, (Vol.1)*, Dunod, Paris.

Benzécri, JP. (1973), *La Taxinomie, (Vol.2)*, Dunod, Paris.

Benzécri, JP. (1973), *L'Analyse des Correspondances*, Dunod, Paris.

Bolasco, S. (1999), *Analisi multidimensionale dei dati* , Carocci editore, Roma.

Caridad, j., Espejo, R. & Gallego, A. (1999), Automatic aggregation of categories in multivariate contingency tables using information theory, *in* 'Computational Statistics and Data Analysis', New York.

Fisher, R. A. (1940), 'The precision of discriminant functions', *in*, 'Annals of Eugenics', 10, pp. 422–429.

Gherghi, M., Lauro, C. (2004), *Appunti di Analisi dei Dati Multidimensionali*, RCE edizioni, Napoli

Greenacre, MJ. (1984), *Theory and application of Correspondence Analysis*, Academic Press, London.

Greenacre, MJ. (1984), Clustering the Rows and Columns of a Contingency Table, *in* 'Journal of Classification', New York.

Greenacre, M. (2000), 'Correspondence analysis of square asymmetric matrices', *Applied Statistics* **49** (3), 297–310.

Guttman, L. (1941), 'The Quantification of a Class of Attributes: A Theory and Method of Scale Construction', *in* P. Horst *et al.*, 'The Prediction

of Personal Adjustment', Social Science research Council, New York, pp. 319–348.

Han, J., Kamber, M. (2001), *Data Mining Concepts and Tecniques*, Morgan Kaufmann Publishiers, San Diego.

Hastie, T., Tibshirani, R. & Friedman, J. H. (2000), *The Elements of Statistical Learning*, Springer.

Hayashi, L. (1950), 'On The Quantification of Qualitative Data from the Mathematical-statistical Point of View', *in*, 'Ann. of the Inst. of Stat. Math.', 2, pp. 35–47.

Hayashi, L. (1956), 'On The Quantification of Qualitative Data from the Mathematical-statistical Point of View', *in*, 'Proc. of the Inst. of Stat. Math.', 4, 2, pp. 19–30.

Hirschfeld, H. (1935), 'A Connection Between Correlation and Contingency ', *in*, 'Cambridge Philosophical Soc. Proc. of the Inst. (Math. Proc)', 31, pp. 520–524.

Lauro, NC., Decarli, A. (1982), *Correspondence analysis and log-linear models in multiway contingency tables study. Some remarks on experimental data*, *in* 'Metron n° 1-2', Roma, pp. 213- 234.

Lebart, L., Morineau, A.& Piron, M. (1997), *Statistique exploratorie multidimensionelle* , Dunod, Paris.

Lebart, L., Morineau, A. & Fenelon, J. (1979), *Traitement des données statistiques*, Dunod.

Mascia, P., Mola, F. (2006), Categories Reduction in Multiple Correspondence Analysis, *in* 'Robust Classification and Discrimination With High Dimensional Data', Firenze.

Mascia, P., Miele, R., Mola, F. (2005), Outliers detection in Regression Trees via Forward Search, *in* 'CLADAG', Parma.

Mascia, P. (2006), Categories Reduction for Ordinal or Numeric Variables in Multiple Correspondence Analysis trough Sequential Automatic Recoding *in* '8th Workshop Of The Ercim Workshops On Matrix Computations And Statistics', Salerno.

Mola, F. Mascia, P. (2006), Categories Reduction in Classification Tree through Sequential Automatic Recoding *in* '8th Workshop Of The Ercim Workshops On Matrix Computations And Statistics', Salerno.

Mascia, P., Mola, F. (2006), On The Aggregation of Categories in Multiple Correspondence Analysis: some Proposals, *Submitted.*.

Mascia, P. (2006), Una Rappresentazione Alternativa delle Variabili Continue nell' Analisi delle Corrispondenze Multiple, *in* 'Knowledge Extraction and Modelling', Capri.

Mascia, P. (2006), The Semi-Active Categories in Multiple Correspondence Analysis, *in* 'MTISD', Procida.

Piccolo, D. (1998), *Statistica*, Il Mulino.

Poincaré, JH. (1905), *Science and Hypotesis*, Walter Scott Publishing, London.

Tukey, J. (1977), *Exploratory data analysis*, Addison Wesley.

Zani, S. (2000), *Analisi dei dati statistici II*, Giuffré ed., Milano.