Tecniche di analisi statistica multivariata per l'individuazione dei "sistemi agricoli territoriali" in Italia

Agostino Tarsitano - Giovanni Anania*
Università degli studi della Calabria
Dipartimento di Economia e Statistica
87030 Arcavacata di Rende (Cs)
agotar@unical.it

Lavoro apparso in "I sistemi territoriali agricoli italiani degli anni '90. Contributi metodologici" A cura di Giovanni Cannata. Rubbettino Editore, Soveria Mannelli (Cz). pp.105-242, 1995.

Riassunto

L'esigenza di partizioni in agricoltura è molto avvertita in relazione alla migliore specificazione della politica agraria ed al suo tentativo di modellamento sulle specificità territoriali. In questo lavoro si procede ad una ricognizione delle tecniche statistiche di analisi multivariata utilizzabili per lo studio della multiregionalità e per i sistemi agicoli territoriali. Dopo aver dedicato attenzione alla scelta degli indicatori ed alle varie trasformazioni da adottare in presenza di dati misurati su scale eterogenee. Lo studio prende in considerazione la riduzione della dimensionalità. Preliminarmente dal lato delgi indicatori usando le componenti principali e, in particolare, si approfondisce la dicotomia tra una analisi degli indicatori effettuata in blocco ed una effettuata per blocchi di indicatori. In tale contesto si adoperano i teoremi di Perron-Frobenius e di Gantmacher per spiegare la ricorrenza di certi risultati in tanti e diverse analisi delle componenti principali. Si passa poi alla riduzione del numero di unità con le tecniche della analisi dei gruppi. Tutti i passi dell'analisi sono stati speirmentati con simulazioni prima di essere applicati ai dati reali.

(*) Il capitolo è frutlo dellavoro e della riflessione comune dei due autori che ne condividono la responsabilità. AGovanni Anania va attribuita la stesura materiale del paragrafo 3.1.2 e del sottoparagrafo introduttivo di quello 3.2.6-Gli altri tre sotlopar'grafi del 3.2.6 sono stali redatti dai due aotori assieme. Le restanti parti sono state scritte da Agostino Tarsitano.

3.1. Analisi multivariata ed applicazioni spaziali

Le analisi statistiche territoriali a fini conoscitivi richiedono complesse indagini investigative che coinvolgono una molteplicità di indicatori e, spesso, un alto numero di entità territoriali. In tale situazione si comprende l'importanza di opportune tecniche di analisi multivariata in grado di semplificare la struttura dei dati, di evidenziare le variabili (reali o latenti) importanti e di individuare relazioni e comportamenti tra di esse.

In questo capitolo vengono discussi i passaggi dell'iter metodologico attraverso i quali è necessario passare al fme di definire il progetto esecutivo di una ricerca quale quella che vogliamo realizzare.

3.1.1 L'unità statistica e l'ambito di rilevazione

In generale, l'unità di rilevazione o unità statistica è il soggetto elementare cui l'indagine si rivolge. Può trattarsi di una persona fisica, di un oggetto, di un'azienda, di uno Stato oppure di un gruppo di queste entità o di altre che, dal punto di vista dell'indagine, formino un tutt'uno. Le unità devono essere obiettivamente distinguibili le une dalle altre e deve pure essere stabilito quali siano quelle che interessa considerare e quali debbano invece tralasciarsi. Le unità sono inserite in un sistema di identificazione all'interno del quale è garantita la loro distinguibilità e la conseguente corretta attribuzione della mo-dalità. In genere, le analisi statistiche trattano l'unità in modo anonimoo trascurandone la localizzazione rispetto alle altre. E' considerata sufficiente un'accurata definizione dell'unità che precisi il luogo ed il periodo di tempo in cui deve essere effettuata la rilevazione; il fatto che una unità sia esaminata prima di un'altra o dopo (ovvero che sia vicina o lontana, in senso temporale e/o spaziale) può anche non essere rilevante ai fini dell'analisi.

Nelle indagini territoriali è di fondamentale interesse la distribuzione spaziale delle modalità delle variabili nel presupposto ovvio che il livello o lo status da esse raggiunto sia, almeno in parte, determinato dal fatto che l'unità abbia una certa collocazione e non un'altra. In questo è determinante il tipo di unità geografica che si considera: areale, reticolare,puntuale. Le unità di tipo reticolare (canalizzazioni, fiumi, reti di distribuzione, rotte di navigazione) e quelle di tipo puntuale (centro-città, dogane, porti, miniere, etc.) non sono considerati dalla nostra ricerca che si concentra invece sulle unità di tipo areale.

L'unità di questo tipo è rappresentata da una poligonale chiusa e può sia essere un'entità fisica: un'isola, un lago, un continente; oppure far parte della suddivisione in zone (zoning) di un territorio. Tale suddivisione può obbedire a principi diversi: amministrativi (nazioni, comuni, quartieri, etc.); funzionali (distretti telefonici, scolastici, compartimenti ferroviari, etc.); ecologici (bacini idrografici, aree caratterizzate da microclimi omogenei, etc.); socioeconomiche (aree di diffusione di un dato dialetto, aree omogenee dal punto di vista della espressione politica, bacini commerciali, etc.).

l dati rilevati sulle unità di tipo areale hanno caratteristiche speciali che si riverberano sull'applicabilità di diverse tecniche statistiche:

l) Le unità sono considerate, rispetto alle variabili studiate, del tutto omogenee al loro interno: le misurazioni, cioè, potrebbero anche essere effettuate in punti diversi della stessa unità areale, ma questo non traspare ed all'unità è assegnata la modalità prevalente nel caso di variabile qualitativa o altra misura di sintesi (ad esempio il totale) nel caso di variabili quantitative.

2) La loro selezione non è quasi mai casuale poiché ci sono tra di esse relazioni di contiguità all'interno del territorio studiato poiché, di solito, compongono il mosaico di una intera zona.

3) l valori asseguati alle unità dipendono dai contomi della stessa unità areale; cambiando questi cambierebbero anche quelli. Ci sono perciò problemi di variabilità del dato rispetto alla scala territoriale e al livello di aggregazione prescelto.
4) Le osservazioni presentano sempre un certo grado di autocorrelazione spaziale che può rendere poco plausibile l'ipotesi di indipendenza campionaria.
5) L'esperienza indica che la distribuzione degli indicatori su unità areali non è gaussiana, ovvero Io è molto di rado.

3.1.2 Scelta delle variabili

L'analisi dei Sistemi Agricoli Territoriali in Italia (SAn) necessariamente si basa sull'uso di informazioni di natura molto diversa tra loro. l dati possono essere costituiti da informazioni singole o composite, grezze od elaborate, soggette in modo diverso ad errori o arbitrarietà, tutte ottenibili dalla osservazione - diretta o indiretta - di un qualche fenomeno soggetto a variazioni. L'unico vincolo è che gli indicatori siano di tipo quantitativo, continuo o discreto; sono ammessi anche i ranghi o le dicotomie, purché le variabili di questo tipo non siano numerose. La scelta delle variabili assume quindi una importanza particolare e ad essa vanno destinate attenzione e risorse adeguate.

In generale, il primo passo non può che essere la puntualizzazione delle macro-determinanti che "a priori" si ritengono rilevanti per caratterizzare SATI tra loro diversi. Lo schema presentato in Anania, Bonetti e Cannata (1984) e quello, più articolato, proposto in Cannata costituiscono utili punti di partenza. La ricerca realizzata nell'ambito del progetto finalizzato IPRA del CNR ela successiva sedimentazione dei risultati raggiunti consentono di raggiungere oggi un maggiore livello di articolazione dei nessi causali che concorrono a "spiegare" i differenti SATI rispetto a quelli delineati in quella ricerca. E' bene richiamare, tra l'altro, che, se in Cannata e in Anania, Bonetti e Cannata l'obiettivo, almeno iniziale, era quello di analizzare le emergenze spaziali di marginalità, qui l'obietti vo e esplicitamente diverso, ed è quello di individuare i differenti SAn e la loro distribuzione sul territorio.

Individuate "a priori" le macro-determinanti dei SATI, il passo successivo è dato dalla ricerca delle informazioni ad esse relative effettivamente disponibili al livello di disaggregazione prescelto, nel nostro caso, il comune. A partire dal risultato di questo lavoro di ricerca sarà necessaria una ulteriore riflessione da parte dei ricercatori sugli indicatori che, sulla base delle informazioni elemeutari disponibili, è possibile costruire (ad esempio, in alcuni ambiti, come in quello demografico, esistono in letteratura delle proposte di indicatori che, a partire sostanzialmente da informazioni elementari simili a quelle utilizzate nell'ambito della ricerca IPRA, portano ad indicatori assai più ricchi di informazioni).

A questo punto, completata la raccolta delle informazioni e la produzione degli indicatori "di base", sarebbe utile realizzare un'analisi pilota di tipo esplorativo delle variabili cosl ottenute. L'obiettivo di questo passaggio è di scandagliare le relazioni esistenti tra gli indicatori di base. Questa procedura, in verità assai rapida, consentirà tra l'altro di verificare l'esistenza di inutili duplicazioni di informazioni. L'analisi esplorativa può basarsi sull'uso di una o più delle seguenti tecniche, analisi delle componenti principali, analisi della matrice dei coefficienti di correlazione, *clustering* delle variabili, e *multidimensional scaling*.

Il risultato di questa fase del percorso metodologico è l'individuazione della matrice "definitiva" dei dati, $\boldsymbol{X}$, che costi- tuirà l'oggetto dell'analisi. Si tratterà di una matrice

rettangolare di dimensioni (n x m) , in cui le n righe saranno costituite dai comuni oggetto di indagine, e le m colonne rappresenteranno i valori degli "indicatori di base" che sono stati osservati in ciascun comune. Indichiamo con Xij il valore che lo j-esimo "indicatore di base" assume nel comune i-esimo.

	Indicatori					
Comuni	X_1	X_2	.....	X_j	.....	X_m
I_1	x_{11}	x_{12}	.....	x_{1j}	.....	x_{1m}
I_2	x_{21}	x_{22}	.....	x_{2j}	.....	x_{2m}
.....	.....	.....	.....	.....	.....	.....
I_i	x_{i1}	x_{i2}	.....	x_{ij}	.....	x_{im}
.....	.....	.....	.....	.....	.....	.....
I_n	x_{n1}	x_{n2}	.....	x_{nj}	.....	x_{nm}

Le relazioni esistenti tra le informazioni costituite dagli indicatori sono considerate nella matrice di devianze-codevianzeglobali **T**:

$$\boldsymbol{T} = \sum_{i=1}^{n} (\boldsymbol{x_i} - \mu)(\boldsymbol{x_i} - \mu)^t = (n-1)\boldsymbol{S} \quad con \quad \mu = \left(\frac{1}{n}\right)\sum_{i=1}^{n} \boldsymbol{x_i}$$

dove ***S*** è la matrice di varianze-covarianze, $\mathbf{x}_i$ è il vettore deivalori assunti dagli*m* indicatori nell'i-esimo comune e **μ** è il vettore delle medie degli "m" indicatori. Potrà trattarsi di variabili originarie, di rapporti di variabili (ad esempio delle inisure pro-capite), di percentuali, di scarti, etc.

E' su questa matrice che si baseranno le tecniche di analisimultivariata da noi scelte per effettuare la ricerca: le componenti principali e la cluster analysis.

3.13 Trasformazioni preliminari

Una volta definita la matrice X bisogna valutare l'opportunità di realizzare o meno delle trasformazioni preliminari sugli indicatori e questo, in buona misura, dipenderà dalla procedura di sintesi che verrà prescelta, Poiché i caratteri originari potrebbero essere misurati in scale eterogenee (età in anni, reddito in lire, rapporti in percentuale, etc.) oppure presentare livelli medi molto diversi, od ancora avere campi di variazione più o meno limitati, è pratica comune misurare gli indicatori in unità standard, cioè espresse come scarto dalla media aritmetica diviso per la deviazione standard:

$$Z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad con \qquad E(Z_j) = 0 \quad e\ \sigma^2(z_j)$$

Cosl facendo ad esempio, l'analisi delle componenti principali, si baserà sulla matrice dei coefficienti di correlazione invece che sulla matrice delle varianze-covarianze. Infatti:

$$Cov(Z_i, Z_j) = \frac{1}{n}\sum_{k=1}^{n}\left[\frac{X_{ki}-\mu_i}{\sigma_i}\right]*\left[\frac{X_{kj}-\mu_j}{\sigma_j}\right] = Cor(Z_i, Z_j)$$

La varianza degli indicatori determina il peso attribuito a ciascuno di essi nell'analisi delle componenti principali. Mentre le differenze di variabilità dovute alle differenti scale sono nel nostro caso sicuramente da considerare come non desiderabili, le differenze nelle variabilità dovute a fattori diversi dalla scala costituiscono una informazione "per sé" che, se possibile, sarebbe meglio non perdere, In generale, sarebbe auspicabile definire degli "indicatori di base" espressi in scala comparabile, onde, evitare la standardizzazione e il conseguente appiattimento delle varianze che ne deriva.

Le variabili standardizzate sono un caso particolare di una classe di trasformazioni basate sulla relazione lineare:

$$\text{A) } Y_{ij} = \frac{X_{ij}-c_j}{v_j} \quad dove \begin{cases} c_j = \text{ misura di centralità} \\ v_j = \text{misura di variabilità} \end{cases}$$

In tale classe rientrano, oltre alla già citata standardizzazione, le seguenti trasformazioni:

$$\text{A1) } Y_{ij} = \frac{X_{ij} - min(X_j)}{max(X_j) - min(X_j)};\ \text{A2) } Y_{ij} = \frac{X_{ij} - M_e(X_j)}{Q_3(X_j) - Q_1(X_j)};\ \text{A3) } Y_{ij} = \frac{X_{ij} - \mu(X_j)}{S_e(X_j)}$$

dove Q_j è l'i-esimo quartile, M_e la mediana e S_e è Io scostamen- to semplice medio dell'indicatore j-esimo dal parametro θ,

$$S_e(X_j) = \frac{\sum_{i=1}^{n}|X_{ij}-\theta_j|}{n} \quad con \quad min(X_j) \le \theta_j \le max(X_j)$$

Le trasformazioni della classe "A" sono caratterizzate dal fatto che medie e varianze

$$\mu(Y_j) = \frac{\mu(X_j) - c_j}{v_j};\quad \sigma^2(Y_j) = \frac{\sigma^2(X_j)}{v_j^2}$$

sono costanti rispetto a cambiamenti di scala, sia additivi che moltiplicativi: cioè sommare e/o moltiplicare per una costante non nulla le variabili "X" non altera le medie e varianze delle risultanti "Y",

La Al porta il campo di variazione dei dati osservati tra zero e uno, ma mantiene la diversità tra medie e varianze dei vari indicatori coinvolti nell'analisi, Lo stesso succede con la AZ, tranne che ora metà dei valori osservati di ogni indicatore è negativa e l'altra metà positiva. La A3 porta la trasformata ad avere media zero come nelle standardizzate,

La classe di trasformazioni" A" mira ad un confronto della variabilità non in base ad una misura assoluta, ma a partire dal rapporto tra due misure di variabilità calcolate sullo stesso indicatore. Se questo non appare del tutto soddisfacente, si può pensare di sostituire la misura di variabilità al denominatore con un generico parametro di scala ed utilizzare come riferimento del numeratore un qualsiasi parametro di traslazione di livello:

$$\text{B) } Y_{ij} = \frac{X_{ij} - c_{j1}}{c_{j2}} \quad dove \begin{cases} c_{j1} = \ parametro \ \ di \ \ livello(traslazione) \\ c_{j2} = parametro \ \ di \ \ scala \end{cases}$$

Ecco Ire esempi della classe "B ":

$$\text{B1) } Y_{ij} = \frac{X_{ij} - min\left(X_j\right)}{\mu\left(X_j\right)};\ \text{B2) } Y_{ij} = \frac{X_{ij} - M_e\left(X_j\right)}{\mu\left(X_j\right)};\ \text{B3) } Y_{ij} = \frac{X_{ij} - \mu\left(X_j\right)}{\mu\left(X_j\right)}$$

Le trasformazioni di questa classe, come quelle in "A", sono invarianti rispetto a trasformazioni moltiplicative, ma sono deformate da trasformazioni additive. Le medie e le varianze delle trasformazioni della classe "B" sono:

$$\mu\left(Y_j\right) = 1 - \frac{c_{j1}}{c_{j2}}; \quad \sigma^2\left(Y_j\right) = \frac{\sigma^2\left(X_j\right)}{c_{j2}^2}$$

Anche in questo caso sono preservate le differenze tra medie e varianze; in particolare, per la B 1 essendo

$$\mu\left(Y_j\right) = 1 - \frac{min\left(X_j\right)}{\mu\left(X_j\right)}; \quad \sigma^2\left(Y_j\right) = \frac{\sigma^2\left(X_j\right)}{\mu_j^2}$$

il confrouto della variabilità passa per i coefficienti di variazione che, oltre ad essere costanti rispetto a variazioui proporzionali, sono comparabili con il massimo raggiungibile sugli "n" dati e cioè:

$$\sigma^2\left(Y_j\right) \le (n-1) \quad se \ \ c_{j1} \le min\left(X_j\right)$$

Una scelta molto interessante del parametro di scala è la norma euclidea del vettore formato dalle osservazioni dell'indicatore j-esimo:

$$\text{B4) } Y_{ij} = \frac{X_{ij}}{\left\|X_j\right\|} con \quad \left\|X_j\right\| = \sqrt{\sum_{i=1}^{n} X_{ij}^2}$$

In questo caso gli indicatori avrebbero le seguenti caratteristiche:

$$\mu(Y_j) = \frac{\mu}{\|X_j\|}; \quad \sigma^2(Y_j) = \frac{\sigma^2(X_j)}{\|X_j\|^2}$$

che preserva le differenze tra medie e varianze pur assicurando l'indipendenza rispetto a variazioni proporzionali. La matrice di varianre covarianre delle trasformate "B4" ha caratteristiche molto simili alla matrice di correlazione (o, se si vuole, della matrice di varianze-covarianze delle standardizzate). Infatti

$$Cov(Y_i, Y_j) = \sum_{k=1}^{n} \frac{X_{kj} X_{ki}}{\|X_j\| \|X_i\|} = \frac{X_j{}^t X_i}{\|X_j\| \|X_i\|} = \cos(\phi_{ij})$$

dove ϕ_{ij} è l'angolo fra i due indicatori visti come vettori n-dimensionali. Quindi gli elementi sono tutti tra -1 e + l. Inoltre, sulla diagonale si ha

$$\sigma^2(Y_j) = \frac{\sigma^2(X_j)}{\|X_j\|^2} \le 1 \quad dato \ che \quad \sigma^2(X_j) = \frac{\|X_j\|^2}{n} - (\mu_j)^2$$

E' ovvio che se le variabili sono misurate come scarti dalla media, cioé $c_{jl} = \mu_j$ la B4 coincide (a parte la divisione per *n*)con la standardizzazione.

L'uso delle variabili trasformate linearmente elimina molti problemi relativi alle unità di misura (ma ne aggiunge altri, come vedremo). Talvolta però può valere la pena di considerare delle trasformazioni non lineari, quali ad esempio la Box-Cox

$$Y_j(\lambda) = \begin{cases} \dfrac{X_j^{\lambda} - 1}{\lambda} & per \quad \lambda \neq 0 \\ Ln(X_j) & per \quad \lambda = 0 \end{cases}$$

Per λ=*1/2* si ha la trasformazione: $Y_j = 2(\sqrt{X_j}-1)$; per λ=*-1* si ha la reciproca: $Y = 1 - l/X_j$

Lo scopo delle Box-Cox è di rendere la distribuzione delle trasformate più vicine al modello gaussiano (complicando però le interpretazioni delle componenti che risulterebbero espresse in variabili fittizie, anche se collegate a quelle originarie da una relazione monotona). Poiché sia la analisi delle componenti principali che la *cluster analysis*, per la loro natura esplorativa, possono prescindere da considerazioni inferenziali (che ci permettono di mettere da parte la necessità di assumere una distribuzione gaussiana degli indicatori) le trasformazioni a questo finalizzate non sono nel nostro caso necessarie Una qualche possibilità andrebbe forse lasciata alla trasformazione logaritmica che ha il merito di rendere uguale la varlanza di indicatori misurati su scale tra di loro proporzionali: *Var(Log(aX$_j$)) = Var(Log(X$_i$))* ovvero ad altre trasformazioni stabilizzatrici della varianza, quali quella cubica o quella quadratica.

3.2 L'analisi delle componenti principali

La tecnica delle componenti principali è uno strumento di sintesi particolarmente prezioso. Essa, contribuisce in maniera determinante a far chiarezza nelle relazioni lineari, più o meno latenti, tra gli indicatori e suggerisce le linee strategiche più appropriate per confermare o smentire il quadro delle ipotesi fondamentali della ricerca

La procedura ha per scopo la trasformazione di m indicatori $X_1, X_2,\ldots,X_m$ in un nuovo insieme di variabili ortogonali $Y_1, Y_2,\ldots,Y_p$ tali che

l. è molto più piccolo di m (realisticamente: $[m/6] \le p \le [m/3]$)

2. Ogni Y_j è una combinazione lineare delle X_j

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{im}X_m \quad i = 1,2,\ldots,p$$

3. La norma dei vettori $\boldsymbol{a}_i$ è unitaria: $\boldsymbol{a}_i^t\boldsymbol{a}_i = 1$. La metrica scelta, la matrice identità I, è quella euclidea, ma si potrebbe usare una qualsiasi altra matrice simmetrica positiva definita e vincolare la componente alla relazione $\boldsymbol{a}_i^t\mathbf{B}\boldsymbol{a}_i = 1$, con il risultato però di aumentare in questo caso l'arbitrarietà, dato che si dovrà fissare non solo il valore della norma (che non necessariamente deve essere unitaria), ma anche la matrice $\boldsymbol{B}$.

4. La prima componente Y_1 è la combinazione lineare $\mathbf{a}_1^t\mathbf{X}$ che rende massima $Var(\boldsymbol{a}_1^t\boldsymbol{X})=\boldsymbol{a}_1^t\boldsymbol{S}\boldsymbol{a}_1$ sotto il vincolo $\boldsymbol{a}_1^t\boldsymbol{a}_1 = 1$. Questo implica che $\boldsymbol{a}_1$ sia l'autovettore normalizzato associato all'autovalore massimo $\lambda_{(1)}$ di $\boldsymbol{S}$.

5. Le componenti successive alla prima si determinano, solitamente, come la combinazione lineare che rende massima $Var(\boldsymbol{a}_i^t\boldsymbol{X})=\boldsymbol{a}_i^t\boldsymbol{S}\boldsymbol{a}_i$ sotto i vincoli $\boldsymbol{a}_i^t\boldsymbol{a}_i = 1$ e $\mathrm{cov}(\boldsymbol{a}_i^t\boldsymbol{X},\boldsymbol{a}_{i-j}^t\boldsymbol{X})=0$ ovvero $\boldsymbol{a}_i^t\boldsymbol{a}_{i-j}=0$ per $i=2,3,\ldots,p$ e per $j=1,\ldots,i-1$.

6. La matrice di varianze-covarianze delle componenti Y è la matrice diagonale composta dagli autovalori $\lambda_{(i)}$ della matrice **S** (la matrice di varianze-covarianze):

$$E\left(Y_iY_j^t\right) = \boldsymbol{L} = Diag\left(\lambda_{(1)},\lambda_{(2)},\ldots,\lambda_{(p)}\right) \quad con \quad \lambda_{(1)} \ge \lambda_{(2)} \ge \ldots \ge \lambda_{(p)}$$

Tra questa e la matrice S esiste l'importante ~Iazione: $\boldsymbol{L} = \boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^t$ e, poiché la matrice **A** è ortogonale cioè $\boldsymbol{A}\boldsymbol{A}^t = \boldsymbol{I}$, abbiamo $Tr(\boldsymbol{L})=Tr(\boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^t)=Tr(\boldsymbol{A}\boldsymbol{A}^t\boldsymbol{S})=Tr(\boldsymbol{S})$ cosicché la somma delle varianze di tutte le componenti coincide con la somma delle varianze degli indicatori originari.

7. La matrice **S** può essere ricostruita a partire dai suoi autovalori e dagli autovettori ad essi collegati. Infatti:

$$S = \sum_{i=1}^{m}\lambda_{(i)}\boldsymbol{a}_i\boldsymbol{a}_i^t \quad con \quad \sum_{i=1}^{m}\boldsymbol{a}_i\boldsymbol{a}_i^t = \boldsymbol{I}$$

altro non è che la decomposizione spettrale della matrice $\boldsymbol{S}$. E' ovvio che il numero di componenti p~scelte sarà tale che le differenze tra la $\boldsymbol{S}$ e la sua ricostruzione in base ai primi p autovettori siano trascurabili.

Quando si opera in un contesto territoriale a fini di ricerca socioeconomica si considera, di solito, una gamma molto ampia di variabili, per cui non sono infrequenti i

casi di sovrapposizione di informazione in cui, pur senza arrivare al punto limite di un indicatore *de facto* duplicato di un altro, si verifica un elevato grado di correlazione lineare tra di essi. Peraltro, gli effetti di una forte multicollinearità non sono, nel nostro contesto, da considerare con preoccupazione, in quanto essi si tradurranno nell'accorpamento in una nuova variabile unica degli effetti dovuti agli indicatori legati dal rapporto di multicollinearità.

3.2.1 Confronto con l'analisifattoriale

Prima di passare a discutere a fondo le peculiarità della procedura di analisi delle componenti principali è necessario un confronto di massima con la procedura rivale: l'analisi fattoriale. Cominceremo la discussione su questo punto cruciale delle analisi territoriali richiamando una affermazione contenuta nel manuale del "Systat" (uno dei packages statistici più diffusi):

Nonostante qualcuno tra gli utilizzatori dell' analisi fattoriale sia molto suscettibile su questo punto e voglia riservare il termine "analisi fattoriale" per il solo modello dei fattori comuni, i risultati ottenuti utilizzando su dati reali l'analisifattoriale e l'analisi delle componenti principali raramente presentano differenze significative, (Wilkinson, 1987, pag. F-2)

La differenza tra analisi fattoriale e analisi delle componenti principali non è tanto nei risultati o nei metodi di calcolo, quanto nello schema teorico da cui muovono. L'idea di fondo del modello di analisi fattoriale è che gli "m" indicatori $X_1, X_2,\ldots, X_m$ possano essere espressi, al netto di un termine di errore, come combinazioni lineari di p superindicatori ortogonali (fattori comuni) $Y_1, Y_2,\ldots, Y_p$ tali che

$$X_i = a_{i1}Y_1 + a_{i2}Y_2 + \ldots + a_{im}Y_p + e_i \quad i = 1,2,\ldots,m$$

ovvero $\mathbf{X=AY+e}$ dove gli a_{ij} sono i pesi fattoriali ed e_i è un termine di errore, detto fattore specifico, perché1egato al solo indice i.

Sia questa procedura che quella delle componenti principali hanno lo stesso obiettivo: la riduzione della dimensionalità dal lato degli indicatori. La analisi fattoriale però tenta di raggiungerlo postulando un modello di relazione lineare tra variabili originarie e fattori comuni costruito su di una serie di ipotesi:

l) $E(\boldsymbol{e})=E(\boldsymbol{Y})=\boldsymbol{0}$; 2) $E(\boldsymbol{ee}^t) = \Psi$ *(diagonale);*
3) $E(\boldsymbol{Ye}^t) = \boldsymbol{0}$; 4) $E(\boldsymbol{YY}^t) = \boldsymbol{I}$ *(matrice identità)*

La formulazione è simile a quella di un modello di regressione multiequazionale lineare, ma con una sostanziale eccezione. Nel modello di analisi fattoriale la matrice $\boldsymbol{A}$ è incognita come lo sono i fattori specifici $\boldsymbol{e}$. Tirando le somme si vede come questa procedura dovrebbe stimare un numero di parametri superiore al numero di osservazioni, con evidenti e insuperabili problemi di indeterminatezza.

L'analisi delle componenti principali si basa su ipotesi molto meno specifiche e quindi non porta a verifiche di particolari ipotesi sulla struttura delle relazioni tra gli indi-

catori. Tale procedUra costituisce solo un modo diverso, più semplice, di rappresentare con una trasformazione le stesse informazioni.

Il confronto Ira analisi fattoriale e analisi delle componenti è esaurientemente condotto in Jolliffe (1986) dove sono messe in risalto analogie e difformità tra le due tecniche e che qui riprendiamo sommariamente. Innanzitutto, l'analisi fattoriale produce in genere meno fattori comuni di quanti non ne suggerisca l'analisi delle componenti e questo perché laddove in questa sono possibili delle componenti separate (connesse ad una sola variabile), nessun fattore comune può invece essere legato a meno di due variabili (se cosl non fosse il corrispondente fattore comune Y confluirebbe, confondendosi, nel fattore specifico ***e***). Ai fini della riduzione della dimensionalità sembrerebbe quindipiù efficiente l'analisi fattoriale, anche se "ingabbiare" in un fattore unico due o più indicatori altrimenti distinti potrebbe portare ad interpretazioni forzose.

Un altro punto rilevante è che entrambe le tecniche hanno come oggetro la matrice di varianze-covarianze ***S*** (di indicatori variamente trasformati), ma, mentre l'analisi delle componenti principali si concentra soprattutto sugli elementi della diagonale in quanto cerca di massimizzare la varianza delle componenti, l'analisi fattori aie cerca di massimizzare la rappresentazione degli elementi di S fuori della diagonale. Infatti, la ***S*** è considerata come la somma di altre due matrici $\boldsymbol{S = A*A' + \Psi}$ e poiché Ψ è diagonale il termine dei fattori comuni ***AY*** è più influenzato dalle covarianze che non dalle varianze. Da notare poi che i "punteggi fattoriali" usati in entrambe le tecniche sono ottenibili in maniera esatta dall 'analisi delle componenti principali dato che queste sono delle funzioni lineari deterministiche delle variabili originarie. Lo stesso non è possibile per l'analisi fattoriale in quanto, nella relazione che lega le *X* e le *Y* compare il vettore incognito "e" ed i punteggi debbono essere stimati. Infine, c'è la diversa reattività delle due tecniche alla alterazione del numero di componenti utilizzate. Se p passa da p_1 a p_2 con $p_2 > p_1$ si introducono $(p_2 - p_1)$ nuove componenti, ma le prime p_1 non scompaiono e non si modificano. Nell'analisi fattoriale, l'aumento del numero di fattori, porta alla completa ridefinizione di tutti i fattori e fra i nuovi p_2 potrebbero non comparire uno o più dei fattori ottenuti per $p=p_1$. In conclusione si può dire che la scelta tra le due procedure dovrà essere guidata dalle finalità dell'analisi: se si è interessati ad una esplorazione dei dati che prescinda da particolari modelli allora la "semplicità" dell'analisi delle componenti principali è preferibile. Se invece i dati si prestano alla formulazione in termini di variabili endogene e fattori esogeni diviene allora appropriata l'analisi fattoriale. Le due tecniche non sono in competizione, anzi possono essere utilmente impiegate sullo stesso insieme di dati.

3.2.2 Componenti principali ed unità di misura

Come abbiamo già avuto modo di dire è piuttosto frequente calcolare le componenti principali dopo aver standardizzato gli indicatori. Questa operazione si realizza con il prodotto di matrici ***Z=CXD*** dove ***X*** è la matrice *(nxm)* degli indicatori nella scala originale, ***C*** la matrice di centramento (simmetrica e idempotente) tale che ***CX*** sia la matrice di scarti degli indicatori dalle rispettive medie, cioè:

$$\boldsymbol{C} = \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{u}\boldsymbol{u}^t \right)$$

(con **u** vettore (nxI) formato da soli uno); **D** è una matrice diagonale formata dai reciproci degli scarti quadratici medi:

$$\boldsymbol{D} = diag\left(\frac{1}{s_1}, \frac{1}{s_2}, \ldots, \frac{1}{s_m}\right), \quad dove \quad s_j = \sqrt{\frac{\sum_{i=1}^{n}\left(X_{ij} - \mu_j\right)^2}{n-1}}; \; j = 1,2,\ldots,m$$

La standardizzazione porta ad estrarre le componenti dalla matrice di correlazione ***R*** invece che dalla matrice di varianzecovarianze ***S***.

$$\boldsymbol{R} = \boldsymbol{Z'Z} = \boldsymbol{DX'C'CXD} = \boldsymbol{DX'CXD} = \boldsymbol{KS}\text{K} \qquad con \quad \boldsymbol{K} = diag\left(\frac{\sqrt{n-1}}{s_j}\right)$$

La procedura di calcolo è la stessa. ma i risultati sono molto diversi e non c'è alcuna relazione immediata che permetta di ottenere le componenti principali della ***S*** una volta calcolati quelli della ***R***.

Chatfield e Collins (1980) dimostrano che le componenti principali di ***S*** sono diverse da quelle di ***R*** (o di qualsiasi altra matrice ottenibile con la pre- e post-moltiplicazione per una matrice diagonale), tranne che nei casi seguenti:

I. La matrice di varianze-covarianze coincide con quella di correlazione. Il caso è improbabile, ma si ha, ad esempio, se

$$\boldsymbol{S} = \left(s_{ij}\right) = \begin{cases} 1 & per \;\; i = j \\ -1 & per \;\; i \neq j \end{cases}$$

2, Gli elementi sulla diagonale di ***K*** sono tutti uguali, di modo che $\boldsymbol{K}=\alpha\boldsymbol{I}$ dove α è la costante in diagonale, Questo significherebbe che le variabili sono scalate allo stesso modo e ciò ha senso solo se gli indicatori hanno la stessa varianza, il che renderebbe superfluo Il ricorso alla standardizzazione,

3, Nel caso che sulla diagonale di ***K*** ci siano elementi uguali a gruppi, gli indicatori collegati agli elementi uguali possono essere correlati tra di loro, ma debbono essere incorrelati con gli altri, Se tutti gli elementi sono diversi allora ***S*** deve essere una matrice diagonale, ma ciò renderebbe superflua l'analisi delle componenti principali visto che ogni indicatore si identificherebbe con una e una sola specifica componente.

Estrazione delle componenti da indicatori trasformati

La sensibilità delle componenti principali al tipo di *scaling* effettuato sugli indicatori le rende aspetti non univoci della matrice dei dati, Se in un gruppo di indicatori ci fosse una chiara gerarchia di variabilità, le componenti principali riporteranno fedelmente tale gerarchia senza badare troppo alla struttura delle covarianze e saranno incuranti di quegli aspetti latenti e trasversali di cui si è alla ricerca,

Per confermare questa asserzione si è studiato un campione di 2000 entità da una distribuzione multinormale di ordine 5 avente

$$\mu = [2 \;\; 4 \;\; 6 \;\; 8 \;\; 10]; \qquad \Sigma = \begin{vmatrix} 2 & 2 & 2 & 2 & 2 \\ 2 & 4 & 2 & 2 & 2 \\ 2 & 2 & 6 & 2 & 2 \\ 2 & 2 & 2 & 8 & 2 \\ 2 & 2 & 2 & 2 & 10 \end{vmatrix}$$

Una matrice di varianze-covarianze con elementi tutti positivi è abbastanza comune in molti contesti applicativi: socio-economici, biometrici, psicometrici, Da tale matrice affiora una prima (e molto spesso la sola significativa) componente che è media ponderata (con pesi tutti dello stesso segno) degli indicatori e che coglie l'aspetto dimensionale del fenomeno, cioè quel fattore che induce l'accrescimento congiunto di tutti gli aspetti considerati dai vari indicatori (tomeremo su questo punto nel paragrafo 3.2.6).

L'analisi delle componenti principali condotta sulla matrice ***R*** dei dati campionari dà luogo ai seguenti risultati essenziali:

$$\lambda = [2.31 \; 0.93 \; 0.84 \; 0.66 \; 0.26]$$
$$PVS = [46.3 \; 18.5 \; 16.8 \; 13.2 \; 5.2]$$
$$A = \begin{vmatrix} 0.90 & 0.05 & 0.09 & -0.16 & -0.39 \\ 0.81 & 0.07 & 0.15 & -0.47 & 0.30 \\ 0.67 & 0.12 & 0.40 & 0.61 & 0.10 \\ 0.51 & 0.31 & -0.78 & 0.18 & 0.06 \\ 0.38 & -0.90 & -0.20 & 0.10 & 0.04 \end{vmatrix}$$

dove PVS è la percentuale di variabilità spiegata dalla componente. La prima componente è quel fattore dimensionale di cui si diceva (la matrice ***A*** ha come colonne le componenti; per righe si leggono i pesi che le variabili assumono nelle componenti stesse). Le componenti successive presentano ciascuna un peso molto più rilevante rispetto agli altri denotando il loro collegamento ad un singolo indicatore: 2ª/v5, 3ª/v4, 4ª/v3 e 5ª/vl. La gerarchia delle varianze è in buona sostanza rispettata, anche se v2, con varianza maggiore di vI, spartisce la sua influenza su tutte le componenti tranne la seconda.

Se l'analisi delle componenti principali è condotta sulla ***S*** campionaria con gli indicatori in scala naturale si hanno invece i risultati:

$$\lambda = [32.88 \; 16.81 \; 8.27 \; 3.39 \; 0.65]$$
$$PVS = [53.0 \; 27.1 \; 13.3 \; 5.5 \; 1.1]$$
$$A = \begin{vmatrix} 0.49 & 0.63 & 0.67 & -0.62 & -0.72 \\ 0.52 & 0.72 & 0.89 & -1.55 & 0.35 \\ 0.60 & 0.98 & 2.45 & 0.81 & 0.09 \\ 0.85 & 3.78 & -0.97 & 0.20 & 0.04 \\ 5.59 & -0.80 & -0.20 & 0.08 & 0.02 \end{vmatrix}$$

Anche in questo caso sono confermate le attese: l'indicatore con la varianza più alta è associato alla componente più rilevante e la caratterizza in modo esclusivo. Si profila inoltre una graduatoria di componenti/indicatori che segue rigorosamente l'ordine stabilito dalle varianze degli indicatori stessi. E' per questo che è diffusa e forte la raccomandazione che l'analisi delle componentiprincipali debba applicarsi solo a dei groppi di indicatori che già in scala origioale abbiano variabilità sostanzialmente uniforme. La scelta cade spesso sulla preliminare standardizzazione:

$$Z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad j = 1,2,\ldots,m$$

in quanto elimina le differenze rispetto alla media e rispetto alla varianza che diventano uguali per tutti gli indicatori: zero la prima e uno la seconda (questo implica che l'analisi delle componenti sia effettuata sulla matrice **R**).

Se per la media il problema non è cruciale dato che gli indicatori entrerebbero nell'analisi come scarti dalle rispettive medie originarie, per la varianza la questione è più seria. Non c'è infatti alcuna ragione di pensate che tutti gli indicatori stiano sullo stesso piano dal punto di vista della variabilità: la standardizzazione rende uguali le varianze e dunque annulla, oltre alle differenze dovute alla diversità di scala, anche quelle proprie, dovute alla maggiore dinamica di certi indicatori rispetto ad altri. L'uso delle variabili standardizzate dovrebbe limitarsi solo a quelle situazioni in cui non esista, almeno a priori, una gerarchia di variabilità tra gli indicatori.

Forme alternative di standardizzazione

La discussione fatta nel paragrafo 3.1.3 ha mostrato come l'indipendenza dall'unità di misura possa essere raggiunta da trasformazioni tipo *Z*, che preservano le differenze tra medie e variabilità, por rimanendo invarianti rispetto ai cambiamenti di scala.

Applichiamo al campione già considerato nel precedente sottoparagrafo la trasformazione unitaria AI:

$$U_{ij} = \frac{X_{ij} - min\left(X_j\right)}{max\left(X_j\right) - min\left(X_j\right)}$$

Per i soliti parametri di riferimento si ottiene:

$$Medie = [0.48 \;\; 0.49 \;\; 0.47 \;\; 0.50 \;\; 0.51]; \quad dev.std. = [0.15 \;\; 0.13 \;\; 0.15 \;\; 0.16 \;\; 0.16]$$

con una sostanziale uniformità tra medie e varianze senza arrivare alla identità perfetta. L'analisi effettuata sulla "S" degli indicatori unitarizzati comporta i seguenti risultati:

$$\lambda = [0.05 \;\; 0.02 \;\; 0.02 \;\; 0.02 \;\; 0.01]$$
$$PVS = [44.5 \;\; 20.1 \;\; 18.1 \;\; 12.7 \;\; 4.6]$$

$$A = \begin{bmatrix} 0.13 & 0.00 & 0.03 & -0.02 & -0.66 \\ 0.12 & 0.01 & 0.04 & -0.07 & 0.04 \\ 0.11 & 0.01 & 0.04 & 0.10 & 0.01 \\ 0.08 & 0.01 & -0.12 & 0.00 & 0.01 \\ 0.06 & -0.14 & -0.06 & 0.00 & 0.00 \end{bmatrix}$$

Se non si tiene conto dei livelli numerici dei pesi delle componenti, non si intravvedono differenze apprezzabili rispetto ai risultati ottenuti partendo dalla matrice di correlazione, se non una più chiara distinzione del legame tra la quinta componente e V1. Rimane peraltro la confusione sulla V2 che è ancora collegata, pur con segni discordanti, a tutte le componenti. Analizziamo ora alcune trasformazioni della classe B, segnatamente la divisione degli indicatori per la rispettiva media aritmetica che è suggerita da Rizzi (1985)

$$U_{ij} = \frac{X_{ij}}{\mu_j}$$

Le medie dei nuovi indicatori sono tutte pari ad uno con deviazioni standard: [0.70 0.51 0.47 0.49 0.57], che coincidono con i coefficienti di variazione:

$$\sigma^2(u_j) = \sqrt{\frac{\sigma_j^2}{\mu_j^2}} = \frac{\sigma_j}{|\mu|_j}; \quad j = 1,2,\ldots,m$$

Questa trasformazione, come si è visto nel paragrafo 3.1.3, annulla le differenze di variabilità attribuibili a dei cambiamenti di scala, ma mantiene quelle dovute a traslazioni di livello. La variabilità maggiore si riscontra ora su V1 e V5 mentre le altre rimangono su di un piano di parità.

L'analisi delle componenti principali sulla ***S*** degli indicatori *U* implica

$$\lambda = [0.78 \;\; 0.30 \;\; 0.28 \;\; 0.15 \;\; 0.09]$$
$$PVS = [51.0 \;\; 19.7 \;\; 13.9 \;\; 9.6 \;\; 5.6]$$

$$A = \begin{bmatrix} 0.67 & 0.08 & 0.07 & 0.04 & -0.18 \\ 0.40 & 0.08 & 0.11 & 0.16 & 0.22 \\ 0.29 & 0.04 & 0.02 & -0.35 & 0.08 \\ 0.21 & 0.06 & -0.44 & 0.03 & 0.03 \\ 0.20 & -0.53 & -0.03 & -0.01 & 0.02 \end{bmatrix}$$

Questi risultati sono piuttosto coerenti con quanto ci si aspettava: c'è il fattore dimensionale e l'abbinamento componente/indicatore segue l'ordine di importanza sulla base del coefficiente di variazione; V1 e V2 si combinano nella prima e si contrastano nell'ultima

componente. I1 tutto avviene con contorni più netti rispetto a quanto ottenuto con la unitarizzazione e peraltro in modo simile all'esito della standardizzazione. Poiché i "fattori" che si ottengono dall'analisi delle componenti principali sono solitamente espressi con media zero, si tende ad adottare trasformazoni che diano questa garanzia. A tal fine basta adottare (si veda più avanti il relativo sottoparagrafo) trasformazioni che inducano media zero negli indicatori trasformati: ad esempio per la A3:

$$v_{ij} = \frac{x_{ij} - \mu_j}{\mu_j}$$

I risultati ovviamente non cambiano visto che la covarianza è calcolata comunque in termini di scarti dalla media, cioè $Cov(X_i,X_j) = Cov(X_i-\mu_i,X_j-\mu_j)$. Esaminiamo ora la "normalizzazione degli indicatori", cioè la divisione delle variabili per la loro norma euclidea:

$$V_{ij} = 100 \frac{X_{ij}}{\|X_j\|}$$

in cui la moltiplicazione per la costante 100 ha il solo scopo di espandere l'ordine di grandezza numerica e limitare gli errori di approssimazione. L'analisi delle componenti principali sulla nuova $\boldsymbol{S}$ dà luogo a:

$$\lambda = [2.85 \;\; 1.15 \;\; 0.86 \;\; 0.61 \;\; 0.32]$$
$$PVS = [49.3 \;\; 19.8 \;\; 14.8 \;\; 10.5 \;\; 5.6]$$

$$A = \begin{vmatrix} 1.21 & 0.14 & 0.14 & 0.10 & -0.37 \\ 0.81 & 0.16 & 0.24 & 0.34 & 0.40 \\ 0.61 & 0.10 & 0.06 & -0.70 & 0.15 \\ 0.43 & 0.15 & -0.88 & 0.06 & 0.06 \\ 0.40 & -1.03 & -0.07 & 0.01 & 0.03 \end{vmatrix}$$

I risultati non differiscono nella sostanza da quelli ottenuti con standardizzazione o con la unitarizzazione. Tuttavia, si ha l'impressione che questi ottenuti dalla normalizzazione siano più nitidi: c'è infatti una maggiore differenza tra i pesi rilevanti e quelli no; è più regolare l'abbinamento componente variabile: C2/V5, C3/V4, C4/V3 e C5/V2; è meglio evidente il contrasto della V1 e della V2 nella quinta componente.

Esaminiamo ora anche una trasformazione non lineare:

$$L_{ij} = Ln[X_{ij} + c_j] \; ; \qquad j=I,2,...,m$$

dove cj è una costante che rende positivo, per ogni unità *i*, l'argomento del logaritmo (la costante aggiunta può essere omessa nel caso l'indicatore sia positivo per costruzione). Ad esempio $c_j = \alpha - min(X_j)$ con $\alpha > 0$. Questa scelta rende la trasformazione costante rispetto a traslazioni di livello perché annullati dallo scarto rispetto al minimo; l'effetto dei cambiamenti di scala moltiplicativi sulla variabilità può essere ammortizzato scegliendo la α abbastanza piccola: ad esempio $\alpha=0.0001$.

L'analisi delle componenti principali sulla $\boldsymbol{S}$ degli indicatori L produce:

$$\lambda = [0.36 \;\; 0.16 \;\; 0.15 \;\; 0.13 \;\; 0.06]$$
$$PVS = [41.5 \;\; 18.7 \;\; 17.7 \;\; 14.7 \;\; 7.3]$$
$$A = \begin{vmatrix} 0.35 & 0.04 & 0.05 & 0.08 & -0.19 \\ 0.30 & 0.09 & 0.09 & 0.19 & 0.15 \\ 0.32 & 0.01 & 0.07 & -0.29 & 0.05 \\ 0.18 & -0.35 & -0.13 & 0.04 & 0.02 \\ 0.12 & -0.17 & -0.35 & -0.02 & 0.02 \end{vmatrix}$$

Al solito, è riprodotta al primo posto la componente dimensionale, anche se qui il peso di V3 è maggiore di quello di V2; in aggiunta, la variabilità spiegata dalla prima componente è la minore rispetto a quella che spiegavano le altre prime componenti in tutte le prove condotte. E' anche scambiata la graduatoria tra V4 e V5 dato che ora V4 è associata alla 2a componente e V5 alla 3a. La quarta e la quinta componente hanno una struttura simile a quelle ottenute con le altre trasformazioni; in particolare, non è confermata la maggiore presenza della V2 sulla quinta componente che ha infatti un peso, in valore assolato, minore della VI, peraltro già presente sulla prima componente.

Le prove effettuate non hanno certo pretesa di proporre conclusioni sul problema dell'unità di misura nell'analisi delle componenti principali. Piuttosto, confermano che tale problema diventa centrale se gli indicatori sono espressi in scale eterogenee. La standardizzazione in questo caso, pur semplificando i tennini della questione, provoca un uso inefficiente delle informazioni contenute nei dati originali, in cui le differenze nei livelli assoluti e nella variabilità sono impananti di per sé, magari perché legati alla particolare definizione delle unità di osservazione.

In altemativa alle variabili standardizzate si può fare ricorso a diverse formule di trasformazione tra cui scegliere quella che risulti più neutrale rispetto agli indicatori di cui si dispone e agli obiettivi dello studio. La nostra sperimentazione dà dei suggerimenti in questo senso (ad esempio la normalizzazione si candida natura1mente a sostituire la standardizzazione) senza ovviamente poter fornire, per le caratteristiche proprie della simulazione realizzata, indicazioni di natura generale.

3.2.3 Alcune noIe sull' analisi del/e componenti principali

In questo paragrafo presentiamo alcuni risultati utili per realizzare ed interpretare l'analisi delle componenti principali.

Normalizzazione

Le componenti principali sono determinate a meno di una costante scalare, cioè se a; è una componente principale Io è anche ba;, con b scalare non nullo. Questa indeterminatezza è risolta vincolando la norma delle componenti ad un livello arbitrario, ma prefissato (in genere si sceglie $\boldsymbol{a}_j^t \boldsymbol{a}_j = 1$; questo permette di comparare i pesi degli indicatori su componenti diverse in quanto ora variano tutti nell'intervallo *[-l,l]*. La normalizzazione può anche farsi in un modo più articolato con la trasformazione: $b_j = \boldsymbol{a}_i \sqrt{\lambda_i}$ per $i=l, 2, ..., p$. In tal caso la norma delle componenti è pari a:

$$\boldsymbol{b}_i^t \boldsymbol{b}_i = \left(\sqrt{\lambda_i}\right)\boldsymbol{a}_i^t \boldsymbol{a}_i \sqrt{\lambda_i} = \lambda_i \boldsymbol{a}_i^t \boldsymbol{a}_i = \lambda_i$$

Gli elementi dell'autovettore bj sono tali che i pesi relativi alle componenti più importanti (cioé con maggiore variabilità) sono più grandi di quelli associati alle componenti meno rilevanti e questo dovrebbe facilitare l'interpretazione delle componenti principali.

Relazioni tra componenti e indicatori
La relazione tra le componenti *y* e gli indicatori *x* è sintetizzata dal prodotto scalare:

$$Y_i = \boldsymbol{a}_i^t \boldsymbol{X}, \quad i = 1, 2, \ldots, p$$

dove $\boldsymbol{a}_i$ è il vettore dei pesi che descrive il modo in cui ogni variabile entra nella componente. Il vettore $\boldsymbol{Y}_i$ è una delle "supervariabili" che sostituirà gli indicatori originali nelle analisi successive. Per meglio comprendere la natura di tali supervariabili vediamone le caratteristiche salienti. Fermo restando che, per costruzione, la varianza di $\boldsymbol{Y}_i$ è uguale all'autovalore i-esimo della matrice $\boldsymbol{S}$, la sua media è

$$E(Y_i) = E\left(\boldsymbol{a}_i^t \boldsymbol{X}\right) = \boldsymbol{a}_i^t \mu, \quad i = 1, 2, \ldots, p$$

che dipende perciò dalla trasformazione adottata (si veda il paragrafo 3.1.3). Ad esempio, nel caso della trasformazione B3 (che implica $\mu = 1$) la media di ogni componente sarebbe parialla somma dei suoi pesi. Invece, nel caso di traslazioni di livello tali che $\mu = 0$, sarebbe nulla anche la media dell'indicatore. Per la correlazione lineare tra componenti ed indicatori abbiamo:

$$Cor\left(X_i, Y_j\right) = \frac{E\left(X_j, Y_i\right) - E(X_i)E\left(Y_j\right)}{\sigma\left(X_i\right)\sigma\left(Y_j\right)} = \frac{E\left(X_j, \boldsymbol{X}^t \boldsymbol{a}_i\right) - \mu_j E\left(\mu^t \boldsymbol{a}_i\right)}{\sigma_j \sqrt{\lambda_i}} = \frac{\boldsymbol{S}_j^t \boldsymbol{a}_i - \mu_j E\left(\mu^t \boldsymbol{a}_i\right)}{\sigma_j \sqrt{\lambda_i}}$$

dove $\boldsymbol{S}_j$ è la j-esima riga della $\boldsymbol{S}$. Poichè per costruzione si ha: $\boldsymbol{S}_j^t \boldsymbol{a}_i = \lambda_i \boldsymbol{a}_i$ si avrà anche: $\boldsymbol{S}_j^t \boldsymbol{a}_i = \lambda_i a_{ij}$ e quindi

$$Cor\left(X_i, Y_j\right) = \frac{\lambda_i a_{ij} - \mu_j E\left(\mu^t \boldsymbol{a}_i\right)}{\sigma_j \sqrt{\lambda_i}}$$

La lettura di questa formula non è immediata. Nel caso di variabili standardizzate ($\boldsymbol{S}_j = 1$; $\mu = 0$) si ha $Cor(X_j, Y_i) = \sqrt{\lambda_i} a_{ij} = b_{ij}$ per cui il coefficiente normalizzato nel senso del sottoparagrafo precedente è pari al coefficiente di correlazione tra componente ed indicatore quando questi è misurato in unità standard. Se si ha solo $\mu_j = 0$ la correlazione diventa:

$$Cor\left(X_i, Y_j\right) = a_{ij}\sqrt{\frac{\lambda_j}{\sigma_j^2}}$$

e cioé proporzionale al peso con un coefficiente di proporzionalità che è pari al rapporto tra la deviazione standard della componente e quella dell'indicatore.

Presenza di autovalori eguali

La simmetria della matrice di varianze-covarianze garantisce che gli autovalori da essa ricavati siano tutti reali e che le componenti siano pure reali e tra di loro ortogonali. Non c'è pero garanzia che gli autovalori siano distinti. Se $l_{(i)}$ è l'autovalore i-esimo in ordine decrescente di grandezza della matrice di varianzecovarianze $\sum$ è possibile che $l_{(q+l)} = l_{(q+2)} = ... = l_{(q+k)}$ Gli autovettori associati ad autovalori multipli non possono essere determinati univocamente (ci sono uno o più gradi di libertà, a secondo del valore assunto da *k*) e la varianza ad essi associata è la stessa. E' chiaro che essi andranno inseriti o esclusi in blocco dalle componenti da trattenere per le analisi successive. Ad esempio, se gli indicatori dessero luogo ad una matrice di varianze-covarianze "egualizzata" del tipo menzionato da Motrison (1967)

$$\boldsymbol{S} = \sigma^2 \begin{vmatrix} 1 & \rho & \mathbf{L} & \rho & \rho \\ \rho & 1 & \mathbf{L} & \rho & \rho \\ \mathbb{M} & \mathbb{M} & \mathbf{O} & \mathbb{M} & \mathbb{M} \\ \rho & \rho & \mathbf{L} & 1 & \rho \\ \rho & \rho & \mathbf{L} & \rho & 1 \end{vmatrix}; \quad con \quad 0 < \rho \leq 1$$

all'autovalore massimo $\lambda_{(l)} = \sigma^2[1+(m-l)\rho]$ è associato l'autovettore normalizzato: $\boldsymbol{a}_1 = \left[m^{-0.5}, m^{-0.5}, \ldots, m^{-0.5}\right]$ che spiega il $l00[1+(m-l)\rho/m$ di variabilità. Gli altri autovalori sono tutti uguali tra di loro (infatti: $\lambda_{(i)} = \sigma^2(I-\rho)$ *per i = 2,3,...,*m e gli altri autovettori sono una qualsiasi delle *(m-l)* soluzioni dell'equazione $\sigma^2\rho(a_{12} + a_{22} + ... + a_{m2})=0$.

In questo tipo di matrici c'è una componente dominante che assorbe buona parte della variabilità totale (la quota aumenta all'aumentare di "p") e nella quale tutti gli indicatori sono rappresentati in modo paritario. Motrison afferma che tale componente ha un orientamento equiangolare nello spazio degli indicatori originali. Le altre componenti si dispongono simmetricamente rispetto alla componente dominante spiegando ciascuna un'eguale quota di variabilità totale: $l00(1-\rho)/m$. Se volessimo aumentare il grado di copertura della variabilità spiegata non sapremmo quale componente scegliere e si rimarrebbe in posizione di stallo.

La molteplicità degli autovalori è comunque molto improbabile per dati campionari Anzi, si pone spesso il problema contrario, e cioè che differenti campioni dalla stessa popolazione diano stime molto diverse degli autovalori. Se ad apparire uguali (almeno a livello di dati campionari) sono gli ultimi k autovalori si parla di "sfericità" nelle ultime

componenti, che non appaiono legate a particolari variabili o gruppi ristretti di variabili. Piuttosto sono componenti indifferenziate irrilevanti nella spiegazione della variabilità dei dati che è invece assorbita dalle prime $p=m-k$ componenti.

Presenza di autovalori nulli

Anche questa è una situazione improbabile per dati cam pionari perché un autovalore nullo implicherebbe l'esistenza di una perfetta relazione lineare (o una perfetta sovrapposizione) tra due o più indicatori. Tuttavia, a meno di errori nella definizione degli indicatori, è vinualmente impossibile trovare relazioni lineari esatte in dati reali. Più realistico è il problema di individuare relazioni di quasi dipendenza. In questi casi è facile trovare degli autovalori molto grandi (conseguenza delle alta collinearità fra gli indicatori) e contemporanea presenza di autovalori molto piccoli (la somma degli autovalori, come sappiamo è fissa). Uno o più autovalori quasi nulli significa che esistono indicatori per i quali le entità del collettivo sono poco differenziate, ovvero che $\boldsymbol{a}_{p+1}{}^{t}\boldsymbol{X}$ è quasi costante e che perciò possono essere utilizzate, senza perdita sostanziale di informazione, le sole prime **p** componenti.

Dipendenza dai rapporti tra le correlazioni

Supponiamo che gli indicatori siano tali che la loro varianza campionaria sia la stessa: σ^2. Se si moltiplicano gli elementi esterni alla diagonale principale di $\boldsymbol{S}$ per la stessa costante $0<r<l$ gli autovalori cambiano, ma gli autovettori rimangono gli stessi. In termini matriciali: $\mathbf{S}^* = [\sigma^2(1-r)\boldsymbol{I} + r\mathbf{S}]$. Se $\boldsymbol{a}$ è un autovettore di $\boldsymbol{S}$ sarà anche un autovettore di $\boldsymbol{S}^*$. Infatti, la relazione $\boldsymbol{Sa}=\boldsymbol{Ia}$, moltiplicando entrambi i membri per **r** e sommando per entrambi il vettore $\sigma^2\boldsymbol{(I-a)Ia}$, può essere scritta come

$$\boldsymbol{aSa} + \sigma^2\boldsymbol{(I-a)Ia}=\boldsymbol{aIa} + \sigma^2\boldsymbol{(I-a)a}$$
$$[\mathbf{aS} + \sigma^2\boldsymbol{(I-a)I}]\mathbf{a}=[\mathbf{aI} +\sigma^2\boldsymbol{(I-a)}]\mathbf{a}$$
$$\mathbf{S^*a = I^*a}$$

con $\mathbf{I}^* =\sigma^2\mathbf{(I-a) + aI}$ ovvero una trasformata degli autovalori originari e con "a" autovettore sia di **S** che di $\mathbf{S}^*$ Questo siguifica che matrici di varianze-covarianze molto diverse possono dar luogo alle stesse compouenti per cui gli autovalori debbono essere sempre considerati con attenzione quando si interpretanoi risultati. La normalizzazione degli indicatori elimina del tutto questo problema.

Il segno dei pesi

Alcuni packages controllano che nei pesi delle componenti non ci siano più valori negativi che positivi, se questo accade ne invertono la direzione (moltiplicando tutti i pesi per -1). E' evidente che il cambio di segno non modifica ne la varianza spiegata dalla componente ne la sua ortogonalità rispetto alle altre componenti, ne l'interpretazione attribuibile alla componente.

Componenti isolate
Se uno degli indicatori è del tutto incorrelato con gli altri questo si rifletterà in una componente il cui autovettore avrà un solo elemento non nullo al posto corrispondente alla variabile isolata e con tutti gli altri elementi pari a zero; quindi ci sarà la perfetta identità indicatore=componente Se gli indicatori fossero incorrelati le componenti non farebbero altro che riprodurre, in ordine di varianza, gli indicatori originari. Nel caso delle variabili standardizzate questo implicherebbe autovalori tutti pari ad uno. Le prime *p* componenti spiegherebbero ancora la stessa percentuale della variabilità totale: le prime otto componenti di dieci indicatori spiegano 1'80% della variabilità totale, ma come decidere quali siano i due indicatori da trascurare?

3.2.4 Scelta del numero di componenti
L'obiettivo dell'analisi delle componenti principali è la determinazione di *p* supervariabili che possano essere validamente sostituite alle "m" variabili originarie in tutte le analisi successive. Peraltro, l'obiettivo si considera pienamente raggiunto se p è molto piccolo rispeno ad m e se il contenuto informativo dei due insiemi di indicatori, originali e latenti, non differisce in maniera apprezzabile. Si capisce perciò l'importanza di una accurata scelta del numero di componenti principali, ovvero di decidere quali siano gli autovalori "grandi" e quali quelli "piccoli". I metodi disponibili per effettuare questa scelta sono diversi, nessuno dei quali privo di soggettività. Spesso, anzi, essi vanno usati in modo ongiunto per non dar luogo a proposte arbitrarie: solo un valore di p su cui ci sia ampia convergenza nelle indicazioni cheprovengano da metodi diversi può essere accettato.

Percentuale cumulata di variabilità spiegata
Il criterio più immediato per la determinazione del numero di componenti è l'ammontare di variabilità spiegata dalle prime *p* componenti. In genere l'ammontare di variabilità spiegata complessiva *b* che si vuole raggiungere è legato al tipo di ricerca, ma raramente è inferiore al 75% o superiore al 95%. Se così è *p* può essere dato dalla formula

$$p = \underset{1\le k\le m}{Min}\left\{ t_k = 100\left(\frac{\sum_{i=1}^{k}\lambda_{(i)}}{\sum_{i=1}^{m}\sigma_i^2}\right) \ge \beta \right\}, \quad 75 \le \beta \le 95$$

Secondo questo criterio si continuano ad estrarre componenti fino a che non si sia raggiunta la percentuale *b* di variabilitàspiegata. Poiché le componenti sono determinate in ordine di importanza, l'aggiunta di variabilità spiegata diminuisce aggiungendo una componente, fino a diventare trascurabile.

Percentuale di variabilità spiegata residuale
Poiché gli autovalori sono considerati in ordine decrescente di grandezza, la percentuaìe

di variabilità spiegata dalla componente k-esima si riduce man mano che "k" si avvicina ad *m*. Se gli indicatori fossero tra di loro ortogonali, l'analisi delle componenti principali produrrebbe esattamente *m* componenti, ognuna collegabile ad un solo indicatore. Quindi, converrà trattenere le componenti fino a che la percentuale di variabilitàspiegata sia superiore alla varianza di almeno un indicatore (regola di Kaiser)

$$p = \underset{1 \le k \le m}{Max} \left\{ \lambda_{(k)} \ge \underset{1 \le k \le m}{Min} \left\{ \sigma_i^2 \right\} \right\}$$

se tale limite sembra troppo debole si può scegliere il criterio più generale

$$p = \underset{1 \le k \le m}{Max} \left\{ \left(\frac{\lambda_{(k)}}{\sum_{i=1}^{m} \sigma_i^2} \right) \ge 100\gamma \right\}, \quad 0 \le \gamma \le 1$$

cioè ci si fenna non appena la componente (p+l)-esima spiega meno del γ% della variabilità totale degli indicatori originali. Dei valori ragionevoli per la soglia di percentuale sono il 10%, il 5% od anche (l/m)% cioé si ritiene la componente purché la quota di variabilità da essa spiegata è almeno pari alla quota media spiegata dagli indicatori.

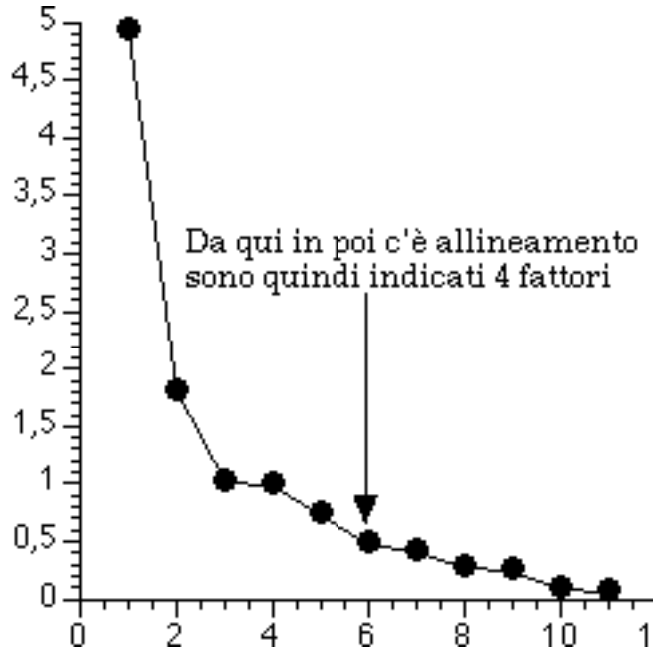
Modello del baslone spezzato.
Supponiamo che un bastone di lunghezza unitaria venga spezzato in *m* pezzi e che le linee di frattura si dispongano casualmente per la sua lunghezza. Si dimostra che, in media, il frammento che, in ordine di lunghezza, ha la k-esima lunghezza, misura

$$t_k = \frac{1}{m} \sum_{j=k}^{m} \frac{1}{j}$$

Un modo per decidere se includere o meno la p-esima componente è quella di confrontare la percentuale di variabilità spiegata con t_p: solo se

$$\left(\frac{\lambda_{(p)}}{\sum_{i=1}^{m} \lambda_{(i)}} \right) \ge t_p$$

potrà valer la pena di portare fino a p il numero delle componenti.



Metodi Scree e LEV

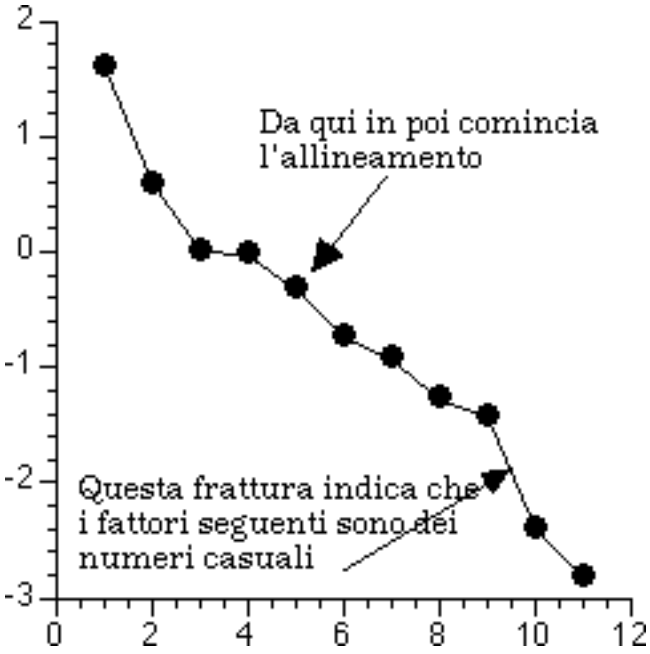
Due metodi grafici possono rivelarsi molto utili. il cosiddetto grafico *scree* pone sulle ascisse l'ordine degli autovalori e questi ultimi sulle ordinate. I punti così definiti sono congiunti con dei segmenti di retta per formare una spezzata. il metodo consiste nell'individuare il punto in cui l'inclinazione dei segmenti diventa quasi costante in modo da presentarsi quasi come facenti parte di una sola retta. il punto di svolta coincide con il punto prima della retta e indica il numero di fattori da selezionare

Analiticamente, il grafico *scree* punta a considerare la differenza tra autovalori successivi $d_k = \lambda_{(k-1)} - \lambda_{(k)}$. Il numero di componenti è detemùnato dal punto 10 cui dk risulta pratica-mente costante per diversi valori successivi di k. Il difetto principale del metodo è che in molti casi o non ci sono punti di svolta chiaramente delineati oppure ve ne sono diversi. In questi casi il grafico scree è di scarsa utilità.

In alternativa si può usare un grafico dove le ordinate sono misurate in scala logaritmica (grafico LEV). Anche in questo caso il numero di componenti sarà indicato da un allineamento degli autovalori laddove una brusca caduta nei valori indicherà che da Il in poi i fattori sono trascurabili. Da notare che nel grafico LEV non ci si basa sulla differenza tra autovalori successivi, bensì sul loro rapporto:

$$d_k = Ln(\lambda_{k-1}) - Ln(\lambda_k) = Ln(\lambda_{k-1}/\lambda_k)$$

cioè si deve stabilizzare non più la differenza tra aulovalori successivi, bensì il loro rapporto. Anche il grafico LEV ha problemi simili al precedente. La figura è illuminante: per gli stessi dati del grafico *scree*, quello LEV non dà chiare indicazione sul numero di componenti ovvero dà indicazioni su un numero diverso di componenti,



Test probabilistici

Vista la debolezza dell'impianto inferenziale in cui si potrebbe collocare l'analisi delle componenti principali e vista anche la scarsa funzionalità di tale impianto per l'efficacia di questa procedura, i metodi di determinazione del numero di componenti basati su tests statistici non sono molto sviluppati. Esistono tuttavia dei test approssimati che, se considerati con prudenza, possono trovare un utile spazio applicativo.

lnnanzitutto si può impiegare un test di tipo generale per valutare se nel blocco delle variabili originarie ci siano intercor relazioni di entità sufficiente a giustificare la stessa analisi delle componenti principali. Nell'ipotesi che le variabili siano distribuite normalmente e siano tra loro incorrelate, la quantità

$$B = -n\sum_{i=1}^{m} Ln\left(\lambda_{(j)}\right)$$

ha una distribuzione ben approssimata dalla χ^2*(m(m+ 1)/2)*. Se il valore di *B* nou è significativo non si potrà rifiutare l'ipotesi di incorrelazione delle variabili originarie, stabilendo quindi un n*olle prosequi* per l'analisi delle componenti principali.

Il test di Bartlett può essere generalizzato per sottoporre a verifica l'ipotesi che le prime *p* componenti della matrice di varianze-covarianze (e quindi non della matrice di correlazione) assorbano tutta la variabilità degli indicatori originari. Infatti, la distribuzione della quantità

$$B_p = -n'(m-p)\left\{ Ln\left[\frac{\sum_{j=p+1}^{m}\lambda_{(j)}}{m-p}\right] - \frac{\sum_{j=p+1}^{m}\lambda_{(j)}}{m-p}\right\} \quad con \quad n' = n - \frac{2m+11}{6}$$

nell'ipotesi che da *k* in poi gli autovalori siano uguali, è ben approssimata dalla χ^2*((m-k-1)(m-k+2)/2)*. E' noto, peraltro, che il test di Bartlett tenda ad includere più componenti del necessario.

3.2.5 Rotazioni

La determinazione di una data componente avviene, come si è visto, secondo i criteri della ortogonalità e della massimizzazione della variabilità complessiva "residuale" che cioè rimane non spiegata dalle componenti principali già estratte. Entrambe questi criteri possono essere tralasciati per ottenere configurazioni più facilmente interpretabili ovvero determinare una configurazione di pesi che preveda:

1) pesi trascurabili per variabili poco rilevanti e pesi elevati per le variabili significative in ciascuna componente;
2) ogni variabile entri in maniera significativa in una sola componente e che nessun peso si collochi in grado intermedio.

fatti comunque salvi il numero di componenti e l'ammontare complessivo di variabilità da esse spiegata.

Il problema della rotazione consiste nel moltiplicare la matrice dei pesi **A** per una matrice di trasformazione: **AT=B**; se **A** è invertibile la determinazione di **T** è semplicemente $\mathbf{T=A^{-1}B}$. Poichè il numero di componenti è minore del numero di variabili **A** non sarà quadrata e **T** non potrà essere calcolata direttamente. Per ottenere una soluzione indiretta si può partire dal fatto che $\mathbf{(A'A)^{-1}(A'A)=I}$ e premoltiplicando per $\mathbf{(A'A)^{-1}A'}$ la **AT=B** si ha $\mathbf{(A'A)^{-1}A'AT=(A'A)^{-1}A'B}$ che implicherebbe la relazione $\mathbf{T=(A'A)^{-1}A'B}$. Il problema è che **B** non è nota e deve essere determinata solo sulla base di principi soggettivi, analitici o grafici, ma che comunque sono legati al particolare metodo di rotazione che si sceglie (ne sono noti in lettaratura circa una ventina).

Il regime di ortogonalità tra le colonne di **A** è usuale. Questa scelta induce a ritenere che i "superindicatori" che collegano trasversalmente le variabili originarie non abbiano però legami di linearità. Una tale imposizione può apparire eccessiva: spesso i superindicatori sono riconducibili a macrofenomeni tra cui sono facilmente ipotizzabili dei rapporti di dipendenza, anche lineare. Una rotazione di tipo "obliquo" che prevedesse perciò la possibilità di ottenere fattori intercorrelati è più realistica. Tuttavia, la ACP è una tecnica numerica che viene forzatamente imposta ai dati e che ha soprattutto finalità di semplificazione senza obbligo di realismo. Le rotazioni oblique pur risultando più flessibili di quelle ortogonali, producono soluzioni più elaborate e molto più complicate da interpretare soprattutto per la difficoltà, in tale contesto, di definire e spiegare la "semplicità" della configurazione.

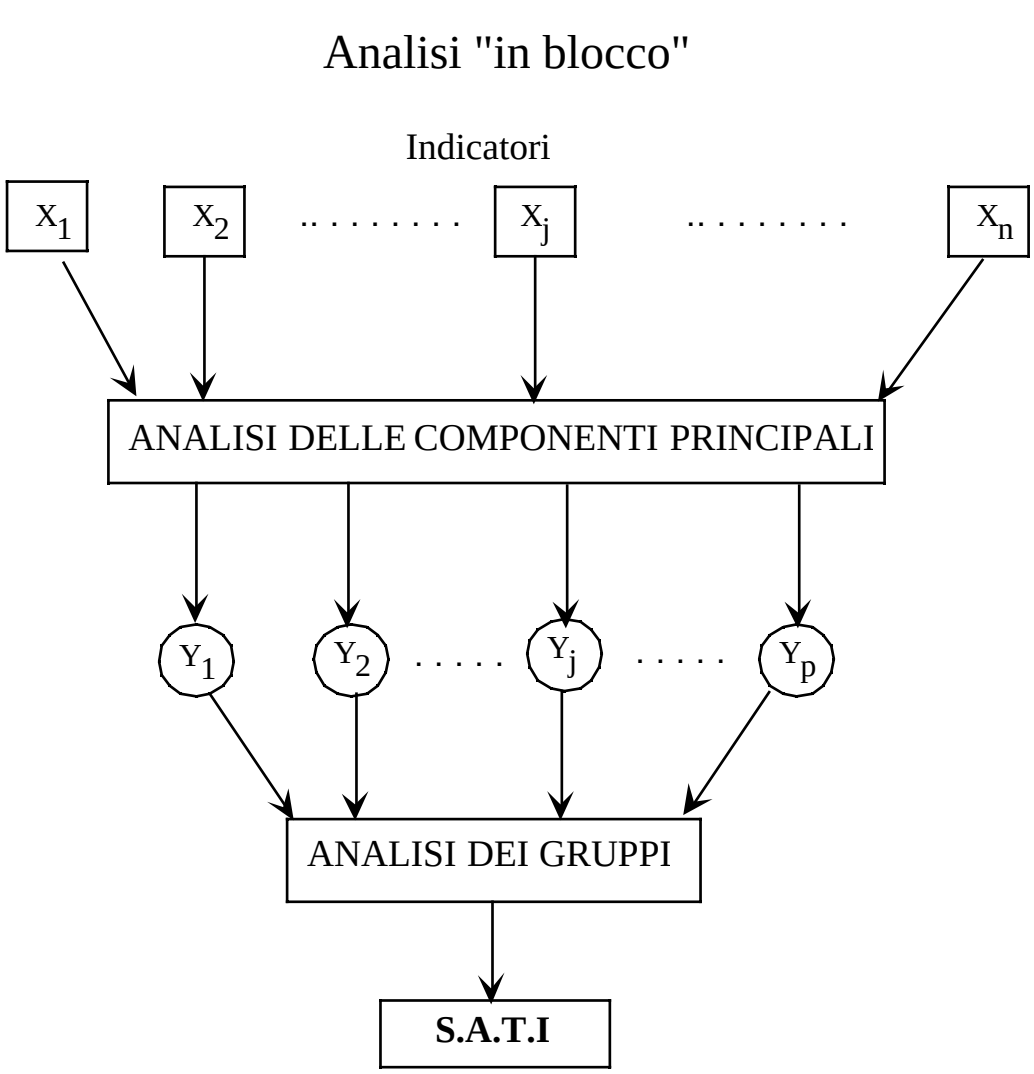
Stabilito dunque che riteniamo preferibile l'uso delle rotazioni ortogonali a quelle oblique resta da decidere quale, tra i tanti metodi di disponibili, sia preferibile. In pratica, la configurazione ideale sarebbe una matrice dei pesi **A** composta da colonne di zeri tranne che per un solo valore in corrispondenza di una specifica variabile e che questo valore non nullo non si ritrovi mai nella stessa posizione. Va detto subito che i due punti attraverso cui abbiamo articolato la nostra idea di "configurazione semplice" non sono facilmente conciliabili. Ci si può muovere cercando una soluzione che punti a ridurre la complessità di riga in modo cioè che sia facilitata l'interpretazione, in termini di componenti principali, delle variabili. In questo caso si opta per il metodo "Quartimax" che porta a massimizzare la somma della potenza quarta dei pesi. Il Quartimax riduce la presenza della variabile tra i fattori facendo sì che sia minimo il numero di fattori per cui la singola variabile ottiene pesi significativi. Spesso si ottiene una soluzione in cui la prima componente è un fattore generale con pesi moderati o piccoli su tutte le variabili.

In alternativa si può optare per la riduzione della complessità di colonna in modo da facilitare l'interpretazione delle componenti in termini delle variabili originarie. Questo equivale a massimizzare la varianza del quadrato dei pesi in ciascuna colonna e orientarsi a soluzioni in cui solo poche variabile hanno peso significato sulla componente. La soluzione Varimax è la più diffusa perchè sembra più rispondente alle finalità proprie della ACP.

Tenuto conto che i due precedenti criteri puntano a due diverse angolature di semplificazione è d'obbligo considerare anche metodi che tengano conto di entrambe. Fra i tanti si può considerare l'Equimax che pone sullo stesso piano il criterio usato per il Quartimax e quelle del Varimax

3.2.6 Analisi in blocco o blocchi di analisi?

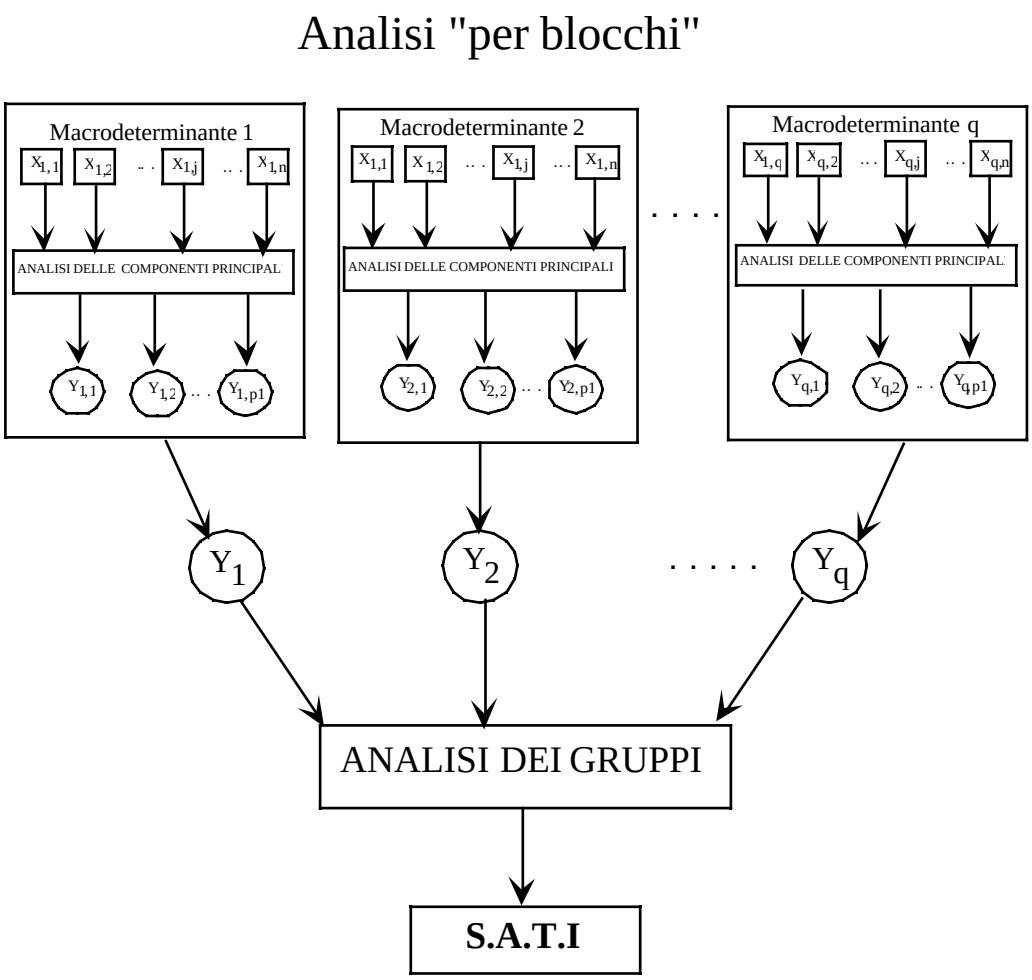
Le impostazioni delle analisi territoriali sono varie e molto articolate. Prevale spesso l'idea di raccogliere quanti più possibili indicatori sul numero massimo di aspetti rilevanti: sociali, economici, amministrativi, demografici, culturali, geomorfologici, ambientali ed usare poi l'analisi delle componenti principali per eliminare le inevitabili ridondanze e duplicazioni. L'approccio della "analisi in blocco", con lo studio in contemporanea di tutte le variabili di base cerca di raggiungere una visione della realtà territoriale più completa e globale di quanto non sia possibile considerando isolatamente le caratteristiche delle diverse entità.



Un possibile limite di tale approccio è costituito dalla difficoltà di calcolo numerico che intervengono quanto il numero delle variabili di base è molto elevato, diciamo oltre 200 o 250 od anche 500 (un numero che potrebbe essere superiore alla capacità di elaborazione di molti PC e delle *work stastion* più diffuse).

La situazione ottimale sarebbe quella in cui tutte le caratteristiche rimaste dopo la selezione iniziale collassassero in un numero ridotto di superindicatori senza che si perda un ammontare di informazione significativo.

Una procedura alternativa si basa sull' idea che un insieme di dati, ampio e articolato, è più facilmente trattabile da un punto di vista statistico, e da un punto di vista concettuale se si individua uno sfondo teorico, un sistema di ipotesi, su cui proiettare i risultati e rispetto al quale condurre l'analisi della documentazione statistica. Il quadro di ipotesi in cui ci muoviamo permette di scomporre l'intero fenomeno in angolature o meglio su piani autonomi o almeno analizzabili separatamente.



Il presupposto di questo approccio è che si possano individuare variabili "tipiche" di ciascun aspetto caratteristico dei SATI talmente specifiche che in generale sia possibile non assegnare una data variabile di base a più di una determinante dei SATI, che cioé fra i gruppi non vi siano sovrapposizioni. Nell'approccio "per blocchi", una volta definiti i gruppi di variabili, si procede all'analisi delle componenti principali per ciascun gruppo di indicatori, individuando per ciascun gruppo i superindicatori che faranno da base per il *clustering* delle entità. Ogni superindicatoree è, una combinazione lineare degli indicatori inclusi nel blocco considerato. Se il numero di superindicatori che complessivamente si

rpoducono nelle sétté-analisi risulta ancora elevato è possibile, in teoria, effettuare una analisi delle componenti principali di secondo livello: le componenti già ottenute diventano delle nuove "variabili di base" e forniscono l'input pere una nuova analisi delle componenti principali. Anche al secondo livello si può procedere in locco o per blocchi separati in dipendenza del numero di nuove variabili di base e della possibilità di individuare tra di esse dei gruppi che ha senso considerare come blocchi. Lo schema è ovviamente iterativo: si pruò procedere all'analisi delle componenti principali di terzo livello, quarto livello, etc. Riteniamo però che molto difficilmente si possa superare il secondo livello ed anche questo sarà attivato solo nel caso in cui si maturasse la convinzione che le macrodeterminanti non sono in realzione lineare diretta, ma solo attraverso le loro componenti principali.

La linea di demarcazione fra i due approcci passa per il momento in cui si applica l'analisi delle componenti principali: su tutti i dati "in blocco" nel primo, separatamente "per ogni blocco" nel secondo, con eventuali iterazioni. I due approcci coincidono se se si possono individuare dei gruppi di variabili fortemente correlate all'interno del gruppo, ma prive di relazioni lineari significative con le variabili di altri gruppi (in pratica, le componenti principali estratte in ogni sottinsieme devono essere incorrelate con le variabili di altri gruppi e conseguentemente con le loro componenti). Se però la matrice di correlazione è indecomponibile ovvero le variabili escluse da un blocco perchè non legate alle componenti principali del blocco stesso, hanno legami lineari significativi con le componenti principali di altri blocchi, i risultati che si ottengono con i due approcci possono essere molto diversi perché nella "analisi per blocch" risulterebbe messo da parte un pezzo dell'informazione complessivamente presente nelle variabili.

Teoria dei due approcci

Sia **X** la matricedegli indicatori (eventualmente trasformati) e sia **S** la matrice delle loro varianze-covarianze. Sia inoltre **A** la matrice con colonne formate dagli autovettori normalizzati di **S**. Suddividiamo gli "m" indicatori in due macrodeterminanti distinte: m_1 ed m_2 e riordiniamo le colonne della **X** in modo che **S** possa essere partizionata in quattro blocchi:

$$S = \left[\begin{array}{c|c} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \hline \mathbf{S}_{21} & \mathbf{S}_{22} \end{array}\right]$$

Con $\mathbf{S_{ij}}$ matrice quadrata di varianze-covarianze degli indicator in "i" e in "j".

Nell'approccio "per blocchi" l'analisi delle componenti principali è applicata aid due gruppi di indicatori fattorizzando separatamente le matrici $\mathbf{S_{11}}$ ed $\mathbf{S_{22}}$ ed ottenendo le matrici di autovettori $\mathbf{A_1}$ ed $\mathbf{A_2}$. Quindi $\mathbf{S_{11}}=\gamma_i\mathbf{a}_{1j}$, i=1,2,…,$m_1$; $\mathbf{S_{22}}=\delta_i\mathbf{a}_{2j}$, i=1,2,…,$m_2$ dove a_{1j} e a_{2j} sono le colonne i-esime di $\mathbf{A_1}$ ed $\mathbf{A_2}$, γ_i e δ_i autovalori di $\mathbf{S_{11}}$ e $\mathbf{S_{22}}$ rispettivamente. E' evidente che se tra le due macrodeterminanti non sussistono relazioni lineari e cioè se $\mathbf{S_{21}}=\mathbf{S_{12}}=\mathbf{0}$ il vettore formato impilando $\mathbf{a}_{1j}$ e i$\mathbf{a}_{2j}$ è l'autovettore i-esimo di **S** cioè a_i. Infatti,

$$\mathbf{S}\mathbf{a}_i = \left[\begin{array}{c|c} \mathbf{S}_{11} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{S}_{22} \end{array}\right]\mathbf{a}_i = \lambda_i\mathbf{a}_i \Rightarrow \left[\begin{array}{c|c} \mathbf{S}_{11} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{S}_{22} \end{array}\right]\left[\begin{array}{c} \mathbf{a}_{1i} \\ \hline \mathbf{a}_{2i} \end{array}\right] = \lambda_i\mathbf{a}_i \quad \textit{ovvero} \quad \begin{cases} \mathbf{S}_{11}\mathbf{a}_{1i} = \lambda_i\mathbf{a}_{1i} \\ \mathbf{S}_{22}\mathbf{a}_{2i} = \lambda_i\mathbf{a}_{2i} \end{cases}$$

Le ultime due relazioni altro non sono che le equazioni caratteristiche gerneratrici degli autovalori ed autovettori di $\mathbf{S_{11}}$ ed $\mathbf{S_{22}}$. Ne consegue che γ_i e δ_i sono anche autovalori della matrice aggregata **S** e perciò l'operare "in blocco" o "per blocchi" è in questo caso del tutto indifferente (a meno di problemi computistici e numerici che potrebbero insorgere nei calcoli.

Se gli indicatori dei due gruppi presentano relazioni lineari più o meno intense allor ala scelta tra le due alternative diventa importante. Infatti, alla base dell'approccio "in b blocco" c'è l'idea di applicare l'analisi delle componenti principali alla matrice di varianze-covarianze totali **S** e di ridurre la dimensionalità del problema scegliendo le prime "k" componenti per formare la matrice **A** degli autovettori. Nell'approccio "per blocchi", dopo aver classificato gli "m" indicatori in macrodeterminanti, si effettua l'analisi delle componenti principali per ognuna di esse. Da ogni sotto-analisi si conservano le componenti che garantiscono la copertura di un'elevata percentuale di variabilità totale, diciamo del 90%, e si fanno confluire nella matrice **B** degli autovettori normalizzati. Nel caso di due gruppi se $\mathbf{S_{12}}\neq\mathbf{0}$ si ha anche $\mathbf{A}\neq\mathbf{B}$. Supponiamo che $\mathbf{B_1}$ e $\mathbf{B_2}$ inglobino gli autovettori di i $\mathbf{S_{11}}$ ed $\mathbf{S_{22}}$ con $\mathbf{S_{12}}\neq\mathbf{0}$. Affinché il vettore $(\mathbf{b}_{1i}, \mathbf{b}_{2i})^t$ formato impilando gli autovettori i -esimi di $\mathbf{S_{11}}$ ed $\mathbf{S_{22}}$ sia anche un autovettore di **S** deve aversi

$$\left[\begin{array}{c|c} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \hline \mathbf{S}_{21} & \mathbf{S}_{22} \end{array}\right]\left[\begin{array}{c} b_{1i} \\ \hline b_{2i} \end{array}\right]=\lambda_i\left[\begin{array}{c} b_{1i} \\ \hline b_{2i} \end{array}\right]\Rightarrow\begin{cases}\mathbf{S}_{11}\mathbf{a}_{1i}+\mathbf{S}_{12}\mathbf{a}_{2i}=\lambda_i b_{1i}\\ \mathbf{S}_{21}\mathbf{a}_{1i}+\mathbf{S}_{22}\mathbf{a}_{2i}=\lambda_i b_{2i}\end{cases}\Rightarrow\begin{cases}\lambda_i b_{1i}+\mathbf{S}_{12}\mathbf{a}_{2i}=\lambda_i b_{1i}\\ \mathbf{S}_{21}\mathbf{a}_{1i}+\lambda_i b_{2i}=\lambda_i b_{2i}\end{cases}$$

ovvero

$$\begin{cases}\mathbf{S}_{12}b_{1i}=\mathbf{0}\\ \mathbf{S}_{21}b_{2i}=\mathbf{0}\end{cases}$$

e queste ultime due relazioni sono impossibili, sia perché $\mathbf{S_{12}}\neq\mathbf{0}$ e sia perché gli autovettori sono non nulli per costruzione.

Le covarianze riscontrabili empiricamente tra macrodeterminanti non sono mai nulle a causa delle fluttuazioni campionarie e degli effetti di legami lineari più o meno spuri. Ma quale ordine di grandezza è compatibile con l'ipotesi $\mathbf{S_{12}}=\mathbf{S_{21}}=\mathbf{0}$? Analizzeremo ora un problema concreto per valutare quali siano le conseguenze della scelta di operare "in blocco" o "per blocchi".

Caso della matrice diagonale a blocchi

La struttura a blocchi diagonali della matrice di correlazione può essere verificata tentando un riordino di righe e colonne $\mathbf{S = P\,S\,P^{-1}}$ dove **P** è una matrice di permutazione (ottenuta cambiando di ordine le righe della matrice di identità) che agisce modificando l'ordine delle righe di **S**, e $\mathbf{P}^{-1}$ è l'inversa di **P** ed agisce sulle colonne di **S**. Se esiste una matrice **P** tale che

$$S=\left[\begin{array}{c|c} \mathbf{S}_{11} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{S}_{22} \end{array}\right]$$

allora l'analisi delle componenti principali, come si è visto, può essere condotta sia sulla matrice aggregata **S** che separatamente per le due sottomatrici $\mathbf{S_{11}}$ ed $\mathbf{S_{22}}$

La ricerca di una possibile decomposizione della matrice **S** è l'essenza stessa dell'approccio "per blocchi". E' però una procedura laboriosa quando il numero di variabili è anche solo moderatamente elevato ed il riordino per gruppi di variabili è piuttosto complicato. Subentrano perciò considerazioni logiche e teoriche che inidcano "ragionevoli" arrangiamenti delle righe e delle colonne di **S** (cioè di disposizioni delle variabili in gruppi correlati al loro interno e non correlati tra loro). La "ragionevolezza" deve essere confermata dai risultati: i coefficienti di correlazione debbono effettivamente essere lontani da zero per le variabili interne ai gruppi e molto vicine a zero per quelle esterne.Consideriamo ad esempio la seguente matrice di correlazione con un numero di indicatori m=10 in cui le prime cinque variabili presentano forti correlazioni posivite tra di loro e correlazioni nulle con le ultime cinque. Lo stesso vale per gli indicatori 6-10.

Esempio di matrice di correlazione diagonale a blocchi

	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
W1	1.00									
W2	0.90	1.00								
W3	0.80	0.70	1.00							
W4	0.70	0.60	0.75	1.00						
W5	0.65	0.50	0.55	0.85	1.00					
W6	0.00	0.00	0.00	0.00	0.00	1.00				
W7	0.00	0.00	0.00	0.00	0.00	0.55	1.00			
W8	0.00	0.00	0.00	0.00	0.00	0.65	0.80	1.00		
W9	0.00	0.00	0.00	0.00	0.00	0.75	0.70	0.95	1.00	
W10	0.00	0.00	0.00	0.00	0.00	0.85	0.60	0.65	0.70	1.00

Nella tabella riportata sotto sono inclusi i risultati dell'analisi delle componenti principali "per blocchi" realizzata separatamente sui prime cinque e sugli ultimi cinque indicatori.

	1	2	3	4	5		1	2	3	4	5
Autvl	3.81	0.69	0.33	0.12	0.05		3.89	0.62	0.32	0.14	0.03
PVS	76.16	13.75	6.59	2.43	1.06		77.76	12.47	6.45	2.8	0.52

	C1A	C2A	C3A		C1B	C2B	C3B
W1	0.93	0.26	0.13	**W6**	0.94	0.13	0.31
W2	0.89	-0.36	-0.12	**W7**	0.92	0.31	0.18
W3	0.87	0.16	-0.45	**W8**	0.86	-0.44	0.03
W4	0.85	0.43	0.24	**W9**	0.86	-0.41	-0.17
W5	0.81	-0.53	0.21	**W10**	0.83	0.39	-0.4

Le due sottomatrici di correlazione propongono entrambe il caso della matrice di correlazione positiva e indecomponibile. Per i teoremi di Perron-Frobebius (si veda Gantmacher, 1974, p.53) a queste matrici è associato un valore masimo unico il cui autovettore ha pesi tutti strettamente diversi da zero e dello stesso segno. Questo significa che nella prima componente principale si riflettono, quale più quale meno, tutte le variabili ed è perciò da considerare un fattore "trasversale" che misura la dimensione complessiva del fenomeno. Sempre in riferimento alla proprietà delle matrici indecomponibili, la percentuale di variabilità speigata (PVS) dalla prima componente è compresa nell'intervallo

$$\frac{\underset{1\le i\le m}{Min}\left\{\sum_{i=1}^{m}\mathbf{S}_{ij}\right\}}{Tr(\mathbf{S})} \le PVS(aut.max) \le \frac{\underset{1\le i\le m}{Max}\left\{\sum_{i=1}^{m}\mathbf{S}_{ij}\right\}}{Tr(\mathbf{S})}$$

che nei due casi particolari descritti dalla matrice precedente implicano i limiti (71%-81%) e (73%-82%) puntualmente confermati dai risultai empirici.

La seconda componente è un fattore "bipolare" in cui c'è la suddivisione per segno e valore quasi paritaria dei pesi significativi e che perciò esprime le contrapposizioni che rimangono tra due gruppi principali di variabili una volta che si sia eliminata l'influenza del fattore dimensionale. Si tratta pure di una presenza "classica" e non sorprendente visto che i pesi della prima componenti sono tutti dello stesso segno (ad esmpio positivi) e la seconda deve essere a questa ortogonale. Manca però una giustificazione teorica analoga a quella del fattore trasversale (si dimostra che nerlle componenti diverse dalla prima deve essere presente almeno un cambiamento di segno nei pes, ma il numero di tali inversioni può essere stabilito solo per matrici particolari.

L'autovalore dominante di ciascuna macrodeterminante risulta l'unico significativo e l' autovettore ad esso associato spiega da solo gran parte della variabilità (circa l'80%) delle sottomatricu e quindi può essere proposto come *leading indicator*. Qualora si rendesse necessariauna maggiore copertura si dovrebbe aggiungere la prima componente bipolare che porterebbe, in entrambi le macrodeterminanti, al 90% la variabilità spiegata.
Analizziamo la matrice aggregata S che ha ora la forma di matrice diagonale a blocchi e perciò perde la caratteristica di indecomponibilità, ma mantiene quella di non negatività (rimanendo simmetrica e definita positiva). Essa ha ancora un autovalore massimo il cui autovettore ha pesi dello stesso segni o nulli (Gantmacher, 1974, p.66). Ecco i risultati

Si confermano le aspettative dettate dalla teoria, ma si evidenzia anche il fatto che ora esistono due fattori dominanti di importanza pressoché uguale: ciascuno spiega il 38% circa della variabilità totale e sintetizza un gruppo di variabili specifico: le prime cinque variabili e le seconde cinque. Analoga distinzione si può fare per i due fattori bipolari 8terza e quarta componente): spiegano ognuno il 6% circa della variabilità totale ed entambi riportano le contrapposizioni tra gli indicagtori del gruppo cui sono legati. La seprazione tra macrodeterminanti è completa, ed è indifferente procedere al calcolo delle componenti principali in modo aggregato "in blocco" o "per blocchi". La loro importanza e la loro interpretabilità sono fedelmente riprodotte nelle due elaborazioni.

	1	2	3	4	5	1	2	3	4	5
Autvl	3.89	3.81	0.69	0.62	0.33	0.32	0.14	0.12	0.05	0.03
PVS	38.88	38.08	6.88	6.24	3.29	3.22	1.40	1.22	0.53	0.26

	C1AB	C2AB	C3AB	C4AB	C5AB	C6AB
W1	0.94	0.00	0.00	0.13	0.00	0.31
W2	0.92	0.00	0.00	0.31	0.00	0.18
W3	0.86	0.00	0.00	-0.44	0.00	0.03
W4	0.86	0.00	0.00	-0.41	0.00	-0.17
W5	0.83	0.00	0.00	0.39	0.00	-0.40
W6	0.00	0.93	-0.26	0.00	-0.13	0.00
W7	0.00	0.89	0.36	0.00	0.12	0.00
W8	0.00	0.87	-0.16	0.00	0.45	0.00
W9	0.00	0.85	-0.43	0.00	-0.24	0.00
W0	0.00	0.81	0.53	0.00	-0.21	0.00

Caso della matrice indecomponibile

Ma che succede se la separazione delle variabili in gruppi non è completa ed un legame, anche se debole, esiste tra macrodeterminanti diverse? A tal fine consideriamo una nuova matrice in cui i valori nulli della matrice precedente sono stati sostituiti con il valore 0.1, una soglia che indica un legame lineare positivo, ma debolissimo tra due variabili inserite in blocchi diversi.

	W1	W2	W3	W4	W5	W6	W7	W8	W9	W0
W1	1.00									
W2	0.90	1.00								
W3	0.80	0.70	1.00							
W4	0.70	0.60	0.75	1.00						
W5	0.65	0.50	0.55	0.85	1.00					
W6	0.10	0.10	0.10	0.10	0.10	1.00				
W7	0.10	0.10	0.10	0.10	0.10	0.55	1.00			
W8	0.10	0.10	0.10	0.10	0.10	0.65	0.80	1.00		
W9	0.10	0.10	0.10	0.10	0.10	0.75	0.70	0.95	1.00	
W0	0.10	0.10	0.10	0.10	0.10	0.85	0.60	0.65	0.70	1.00

La matrice aggregata **S** è ora una matrice irriducibile e ad essa si applicano in pienoi teoremi di Perron-Frobenius già visti per le due sottomatrici: ci si aspetta quindi un autovalore massimo che spiega una quota prefissata entro limiti precisi di variabilità totale ed a cui è asociata una componenti con pesi tutti diversi e dello stesso segno (il fattore trasversale) ed una componente bipolare con almeno un peso in contrasto rispetto agli altri.

	1	2	3	4	5	6	7	8	9	10
Autvl	4.35	3.35	0.69	0.62	0.33	0.32	0.14	0.12	0.05	0.03
PVS	43.49	33.48	6.88	6.24	3.29	3.22	1.40	1.22	0.53	0.26

	C1AB	C2AB	C3AB	C4AB	C5AB	C6AB
W9	0.72	0.59	0.00	0.13	0.01	0.31
W8	0.71	0.59	0.00	0.31	0.01	0.18
W6	0.67	0.54	0.00	-0.44	0.00	0.03
W10	0.67	0.54	0.00	-0.41	-0.01	-0.18
W7	0.67	0.51	0.00	0.39	-0.02	-0.40
W1	0.65	-0.65	-0.26	0.00	-0.13	0.01
W4	0.65	-0.62	0.36	0.00	0.12	0.00
W3	0.63	-0.60	-0.16	0.00	0.44	0.00
W2	0.62	-0.58	-0.43	0.00	-0.24	0.00
W5	0.59	-0.55	0.53	0.00	-0.21	0.00

Rispetto alla matrice diagonale a blocchi del paragrafo precedente i cambiamenti sono pochi, ma significativi: l'importanza del primo autovalore è aumentata; il primo autovalore è diminuita quella del secondo e sono rimasti praticamente inalterati gli altri. Le componenti associate agli autovalori più piccoli (dal terzo in poi) sono pure rimaste le stesse, ma per le prime due componenti la situazione è ben diversa. Nella nuova matrice si ripropone il fattore trasversale dominante che è ora però fortemente correlato a *tutti* gli indicatori senza distinzione alcuna per i due gruppi. Segue poi il fattore bipolare in cui metà dei pesi sono positivi e metà negativi a seconda dei gruppi a cui fanno riferimento (questo è dovuto alla sostanziale uniformita dei pesi mella prima componente). Il legame tra macrodeterminanti, per quanto tenue, ha agito da collante ed ha prodotto una coppia di componenti che spiega circa il 77% della variabilità totale, come del resto facevano le due componenti ottenute dalle analisi separate, ma che rispetto a queste hanno una chiave di lettura alternativa e che porta a conclusioni molto dissimili. Da notare che è stata annullata la contrapposizione all'interno dei due gruppi ed è proposto un contrasto tra macrodeterminantiche non era previsto e che potrebbe essere solo artificiale.

Caso della matrice indecomponibile non positiva

La fusione delle macrodeterminanti potrebbe essere legata al fatto che le relazioni tra i loro indicatori siano dello stesso segno. Conviene quindi vedere quello che succede se le covarianze sono di segno diverso. Ecco quindi un esempio di matrice **S** in cui sono presenti anche delle entrate negative. I segni ed i valori sono disposti in modo tale che la matrice sia definita positiva.

Anche con relazioni di segno diverso tra gli indicatori dei due gruppi, la sostanza dei risultati non cambia: emergono un fattore bipolare ed un fattore trasversale che congiuntamente spiegano, di nuovo, circa il 77% della variabilità totale (rispetto al caso precedente, i due fattori si sono scambiati la posizione d'ordine). A questi si possono aggiungere due altri fattori "misti" che portano la variabilità spiegata al 92%.

	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
W1	1.00									
W2	0.90	1.00								
W3	0.80	0.70	1.00							
W4	0.70	0.60	0.75	1.00						
W5	0.65	0.50	0.55	0.85	1.00					
W6	0.10	0.10	-0.10	0.10	0.10	1.00				
W7	-0.10	-0.10	-0.10	-0.10	-0.10	0.55	1.00			
W8	0.10	0.10	-0.10	-0.10	-0.10	0.65	0.80	1.00		
W9	0.10	0.10	-0.10	-0.10	-0.10	0.75	0.70	0.95	1.00	
W0	-0.10	-0.10	-0.10	0.10	0.10	0.85	0.60	0.65	0.70	1.00

Sono proprio questi fattori misti che possono complicare l'interpretazione. Il fattore trasversale ed il fattore bipolare possono infatti essere facilmente interpretati.Della decifrazione dei fattori misti non si diffiderà mai abbastanza.

	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00	10.00
Autvlò	3.97	3.76	0.94	0.53	0.34	0.20	0.12	0.10	0.04	0.01
PVS	39.66	37.64	9.36	5.26	3.41	2.04	1.17	0.99	0.38	0.09

	C1AB	C2AB	C3AB	C4AB	C5AB	C6AB
W9	-0.78	0.52	0.22	0.09	0.07	0.21
W8	-0.77	0.51	0.28	-0.10	0.14	0.14
W6	-0.67	0.56	-0.24	0.34	-0.09	-0.18
W10	-0.71	0.47	-0.42	0.09	-0.21	0.06
W7	-0.74	0.36	0.09	-0.49	0.03	-0.25
W1	0.49	0.80	0.28	0.11	0.07	0.00
W4	0.52	0.73	-0.35	-0.13	-0.02	0.04
W3	0.58	0.66	0.14	-0.22	-0.38	0.09
W2	0.44	0.73	0.38	0.23	0.03	-0.18
W5	0.47	0.66	-0.45	-0.09	0.34	0.03

Esem pio con le matrici di tipo Jacobi

Un altro caso interessante di differenza profonda di risultati tr approccio “in blocco” e “Per blocchi” è dato da matrici di varianze-covarianze di tipo Jacobi. Si tratta di matrici in cui sono poisitivi gli elementi sulla diagonale principale nonché quelli immediatamente sotto ed immediatamente a destra della diagonale.

$$a_{ii} > 0 \;\; per \;\; i = 1,2,\ldots,m; \qquad a_{i+1,i} > 0, \;\; a_{i,i+1} > 0 \;\; per \;\; i = 1,2,\ldots,m-1;$$

Tutti gli altri elemetni sono nulli. Ad esempio, nel caso di blocchi separati (5x5) e tenuto conto che discutiamo di matrici simmetriche, si avrebbe

$$S = \left[\begin{array}{c|c} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \hline \mathbf{S}_{21} & \mathbf{S}_{22} \end{array}\right];\ \mathbf{S}_{11} = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & 0 & 0 \\ \beta_1 & \alpha_2 & \beta_2 & 0 & 0 \\ 0 & \beta_2 & \alpha_3 & \beta_3 & 0 \\ 0 & 0 & \beta_3 & \alpha_4 & \beta_4 \\ 0 & 0 & 0 & \beta_4 & \alpha_5 \end{bmatrix};\quad \mathbf{S}_{22} = \begin{bmatrix} \phi_1 & \nu_1 & 0 & 0 & 0 \\ \nu_1 & \phi_2 & \nu_2 & 0 & 0 \\ 0 & \nu_2 & \phi_3 & \nu_3 & 0 \\ 0 & 0 & \nu_3 & \phi_4 & \nu_4 \\ 0 & 0 & 0 & \nu_4 & \phi_5 \end{bmatrix}$$

Ciascuna delle sottomatrici sula diagonale di **S** è di tipo Jacobi. Tali matrici, come si deriva immediatamente dal teorema di Gantmacher (Gantmacher 1974, p. 105) hanno autovalori positivi e distinti. La prima componente ha pesi non nulli e delo stesso segno; la componente associata all'autovalore i-esimo (i>1) in ordine di grandezza ha esattametne (i-1) inversioni di segno (è cioè una componente di contrasto) ma solo in nmero limitato e specifico di indicatori. Ad esempio, strutture di segni per le ultime quattro componenti in un gruppo di cinque compatibili con il teorema di Gantmacher sono le seguenti $(-+++)$, $(+-++)$, $(++-+)$, $(-+-+)$. Un approccio "per blocchi" implicherebbe perciò, per ogni macrodeterminante, come prime due componenti un fattore trasversale ed un fatto di contrasto (che può anche essere bipolare) nella prima o nell'ultima componente. Ecco un esempio di matrice di varianze-covarianze che rispecchia la struttura Jacobi.

	W1	**W2**	**W3**	**W4**	**W5**	**W6**	**W7**	**W8**	**W9**	**W10**
W1	2.00	1.00								
W2	1.00	4.00	2.00							
W3		2.00	8.00	3.00						
W4			3.00	16.00	3.00					
W5				3.00	32.00					
W6						3.00	2.00			
W7						2.00	6.00	3.00		
W8							3.00	9.00	4.00	
W9								4.00	12.00	5.00
W10									5.00	15.00

Nella tabella che segue sono inclusi i risultati della analisi delle componenti principali "per blocchi" che ricalcano quanto prestabilito dal terorema di Gantmacher.

	1	**2**	**3**	**4**	**5**		**1**	**2**	**3**	**4**	**5**
Autvl	32.96	16.14	7.87	3.56	1.46		19.77	12.25	7.56	4.24	1.57
PVS	53.17	26.04	12.70	5.74	2.36		43.05	27.23	16.79	9.43	3.50

	C1A	**C2A**	**C3A**	**C4A**	**C5A**		**C1B**	**C2B**	**C3B**	**C4B**	**C5B**
w1	0.00	0.02	0.21	0.92	1.05	**w6**	0.03	0.27	0.81	1.16	0.96
w2	0.10	0.23	1.25	1.44	-0.57	**w7**	0.26	1.25	1.83	0.72	-0.69
w3	0.16	1.40	2.32	-0.78	0.19	**w8**	1.15	2.44	0.41	-1.20	0.37
w4	1.34	3.65	-0.93	0.20	-0.04	**w9**	2.79	1.04	-1.52	0.88	0.17
w5	5.58	-0.92	0.15	-0.28	0.01	**w10**	3.19	-1.90	1.02	-0.41	0.06

Accontendosi di una variabilità spiegata del 70%, i risultati indurrebbero a ritenere, in ogni blocco, due fattori: uno dimensionale ed uno di contrasto. Questo non necessariamente succede se si opera "in blocco". Innanzitutto perché la matrice aggregata **S** non è più una matrice di tipo Jacobi (manca la relazione lineare tra w5 e w6) che sarebbe in contraddizione con l'ipotesi che le matrici siano separate. Inoltre, dovrebbero essere nulle le covarianze tra tutte le coppie di indicatori in macrodeterminanti diverse. In ogni caso l'analisi "in blocco" dovrà necessariamente dare risultati diversi da quella " per blocchi". Ad esempio, ipotizzan do Cov(w5,w6)=5 si ottiene:

	1	2	3	4	5	6	7	8	9	10
Autvl	33.74	19.37	16.24	12.25	7.90	7.47	4.00	3.56	1.46	1.03
PVS	31.53	18.10	15.17	11.45	7.38	6.98	3.74	3.32	1.37	0.96

	C1AB	C2AB	C3AB	C4AB	C5AB	C6AB
W1	0.00	0.00	0.02	0.00	0.21	0.04
W2	0.01	0.00	0.23	0.01	1.24	0.20
W3	0.15	0.00	1.39	0.04	2.30	0.32
W4	1.29	-0.01	3.67	0.04	-0.90	-0.19
W5	5.59	-0.01	-0.83	-0.07	0.10	0.16
W6	0.91	0.03	-0.32	0.23	0.22	-0.63
W7	0.07	0.26	-0.07	1.25	0.29	-1.81
W8	0.01	1.15	-0.03	2.44	0.03	-0.46
W9	0.00	2.79	0.01	1.05	-0.22	1.53
W10	0.00	3.19	0.03	-1.90	0.16	-1.02

Rispetto all'analisi separata c'è il fatto nuovo della presenza dell'indicatore w6 nelle varie compponenti che è in fondo l'affioramento del legame lineare stabilito attraverso Cov(w5,w6)=5. E' pure cambiata la struttura dei segni: è comparso un fattore bipolare (il segno cambia in effetti una volta sola) che, sia pure attenuato, potrebbe indurre interpretazioni ardite quanto fantasiose.

A conclusione del paragrafo possiamo dire che, ipotizzati solo legami lineari deboli tra gruppi diversi di indicatori, la differenza tra l'approccio "per blocchi" e "in blocco" non passa né per il numero né per la capacità esplicativa dei fattori. Nel caso della matrice positiva indecomponibile si arriva, in entrambi i casi, a quattro componenti rappresentative dei dieci indicatori che congiuntamente spiegano più del 90% della varaibilità totale di partenza. Nel caso delle matrici Jacobi si arriva pure a quattro componenti ed al 70% di variabilità spiegata seguendo entrambi gli approcci. La grande diversità è nella interpretazione delle componenti principali: i segni ed i valori dei pesi possobo essere molto diversi per i due approcci e dar luogo a supervariabili con significati con coerenti tra loro.

3.2.7 Le componenti di secondo livello

Una importante variante dell'approccio per blocchi e costituita daIle componenti principali di secondo livcllo, c cioe dall'analisi delle componenti principali applicata alle componenti principali ottenute nei vari blocchi in cui sana stati raggruppati gli indicatori.

L'approccio “per blocchi” in realtà non può garantire che i raggruppamenti di indicatori siano del tutto privi di interrelazioni. Anzi, spesso si usa uno stesso indicatore o indicatori motoa simili in macrodeterminanti diverse (ad esempio gli indicatori demografici) ed è perciò naturale aspettarsi che le covarianze tra indicatori, sebbene collocati in blocchi differenti, siano talvolta significativamente diverse da zero. Questo, oltre ad originarele diversità interpretative illustrate nel paragrafo precedentc, implica che le componenû principali ottenute da gruppi diversi di indicarori, non siano ortogonali.

Consideriamo ancora gli indicatori divisi in due gruppi, la matrice di varianze-covarianze delle componenti principali separatamente estratte dai due blocchi è

$$E\left(\mathbf{y}_{i1}\mathbf{y}_{j2}^{t}\right)=\left[\begin{array}{c|c}\mathbf{L}_1 & \mathbf{C}\\ \hline \mathbf{C} & \mathbf{L}_2\end{array}\right] \quad con \quad \begin{cases}\mathbf{y}_{i1} = autovettore\ \ i-esimo\ \ gruppo\ 1\\ \mathbf{y}_{j2} = autovettore\ \ j-esimo\ \ gruppo\ 2\end{cases}$$

$\mathbf{L}_1$ ed $\mathbf{L}_2$ sono matrici diagonali formate con gli autovalori delle matrici $\mathbf{S}_{11}$ ed $\mathbf{S}_{22}$. La matrice $\mathbf{C}$ ingloba le covarianze tra autovettori di gruppi diversi e se non c'è ragione di pensare che le relazioni lineari tra indicatori di gruppi diversi siano nulle è naturale aspettarsi che $\mathbf{C}\neq\mathbf{O}$. Questa è una notevole differenza, rispetto all'approccio “in blocoo” dato che qui le componenti sono ortogonali per costruzione.

Perché ci si preoccupa della ortogonalità delle supervariahili? Il fatto è che è su queste che si baserà la *cluster analysis*. Poiché uno dei criteri che vogliamo utilizzare mira alla mimizzazione del determinante della matrice di varianze-covarianze delle supervariabili, se queste fossero fortemente correlate, il criterio sarebbe già basso di per sé e l'intera procedura di ottimizzazione ne sarebbe danneggiata. Occorre inoltre considerare la possibilità che la scelta degli indicatori inseriti nelle macro-determinanti sia tale da determinare delle relazioni lineari molto forti tra $\mathbf{y}_{i1}$ e $\mathbf{y}_{i2}$ al punto che, pur calcolate all’interrno di macrodeterminanti diverse, siano quasi ol’una il duplicato dell’altra. Tali ridondanze sono discordi con l'obiettivo della riduzione della dimensionalità tipico dell'analisi delle componenti principali. Sopratttto nel caso delle ricerche territoriali in cui tale analisi è un passo intermedio prima della *cluster analysis*. Qui, a causa del numero elevato di entità coinvolte, è essenziale che il numero di indicatori con cui esse sono descritte sia il minore pussibile evitando percia di coinvolgere informazioni non strettamente neeessarie. Questo ha portato Narayanaswamy e Raghavarao (1991) a riapplicare l'analisi delle componenti principali ai fattori estratti dai vari blocchi chiamando componenti di secondo livello i nuovi fattori. Essi pero propongono una procedura sequenziale di inglobazione delle componenti di primo livello in quelle di secondo che non è soddisfacente per la nostra ricerca, per cui proponiamo una variante più consona alla situazione analizzata.

Supponiamo che gli "m" indicatori siano ,stati sutldivisi in "q" macrodetenninanti distinte e che per ognuna siano state detenninate p_i componenti principali (che garantiscono, in tutte le macro-detenninanti, una quota elevata di variabilita spiegata, diciamo 90-95%) e sia $p^*=p_1 + p_2 + ... + p_q$. Consideriamo la matrice $\mathbf{Y}$ Ie cui colonne siano fonnate dagli autovettori normalizzati e con media zero ottenuti nelle sottoanalisi.

$$\mathbf{Y} = \left[\mathbf{y}_{11},\ \mathbf{y}_{21},\ \ldots,\ \mathbf{y}_{p_1 1} \,\middle|\, \mathbf{y}_{12},\ \mathbf{y}_{22},\ \ldots,\ \mathbf{y}_{p_2 2} \,\middle|\, \ldots \,\middle|\, \mathbf{y}_{1q},\ \mathbf{y}_{2q},\ \ldots,\ \mathbf{y}_{p_q q} \right]$$

che ha seguente matrice di varianze-covarianze

$$E\left(\mathbf{y}_{ir}\mathbf{y}_{js}^t\right) = \bar{\mathbf{S}} = \begin{bmatrix} \mathbf{L}_1 & & & & \\ \mathbf{S}_{12} & \mathbf{L}_2 & & & \\ \mathbf{S}_{13} & \mathbf{S}_{23} & \mathbf{L}_3 & & \\ \mathbf{K} & \mathbf{K} & \mathbf{K} & \mathbf{O} & \\ \mathbf{S}_{1q} & \mathbf{S}_{2q} & \mathbf{S}_{3q} & \mathbf{L} & \mathbf{L}_q \end{bmatrix}$$

La metodologia di ricerca prevede di applicare aIle "p*" supervariabili "y" l 'analisi delle componenti principali conservando le prime p<p* componenti di secondo livello che coprono una percentuale soddisfacente di variabilità spiegata. Le componenti di secondo livello sono naturalmenre ortogonali. E' chiaro che se una delle colonne di $\bar{\mathbf{S}}$ è nulla (tranne che per l'elemento sulla diagonate) la componente di primo livello si trasferirà direttamente al secondo senza subire alcuna alterazione (si veda la dimostrazione all'inizio del paragrafo). Applichiamo la metodalogia alIa matrice di varianze-covarianze positiva con separazione debole del paragrafo 2.6. Le sottoanalisi sono quelle effettuate per i due gruppi di cinque indicatori anche se ora conserviarno tre componenti in agni macrodeterminante portando la variabilità spiegata al 96.50 e al 96.68 rispettivamente. Queste sei supervariabili vengono ora considerate degli indicatori di base ai quali applicare nuova-mente l'analisi delle componenti principali. La loro matrice di varianze-covarianze è data da

$$\bar{\mathbf{S}} = E\left(\mathbf{y}_{ir}\mathbf{y}_{js}^t\right) = a_{ir}^t \mathbf{S}_{rs} a_{js}^t \qquad i, j = 1,2,3; \quad r, s = 1,2$$

che nel caso particolare dell'esempio diventa

	$\mathbf{y}_{11}$	$\mathbf{y}_{21}$	$\mathbf{y}_{31}$	$\mathbf{y}_{12}$	$\mathbf{y}_{22}$	$\mathbf{y}_{32}$
$\mathbf{y}_{11}$	3.810					
$\mathbf{y}_{21}$	0.000	0.069				
$\mathbf{y}_{31}$	0.000	0.000	0.330			
$\mathbf{y}_{12}$	1.870	-0.018	0.004	3.890		
$\mathbf{y}_{22}$	-0.002	0.000	0.000	0.000	0.620	
$\mathbf{y}_{32}$	-0.002	0.000	0.000	0.000	0.000	0.000

Come si vede c'è una sola entrata significativa che è affioramento dell'ipotesi

$$\mathbf{S}_{12} = 0.1 \begin{bmatrix} 1 & 1 & 1 & \mathbf{L} & 1 \\ 1 & 1 & 1 & \mathbf{L} & 1 \\ 1 & \mathbf{S}_{23} & \mathbf{L}_3 & \mathbf{L} & 1 \\ \mathbf{K} & \mathbf{K} & \mathbf{K} & \mathbf{O} & 1 \\ 1 & 1 & 1 & \mathbf{L} & 1 \end{bmatrix}$$

Le componenti di secondo livello si ottengono analizzando la nuova matrice $\bar{\mathbf{S}}$ che produce le seguenti componenti

	1	2	3
Autvl	5.72	1.98	0.69
PVS	59.22	20.79	7.14

	CP1	CP2	CP3
$\mathbf{y}_{11}$	1.673	-1.005	0.005
$\mathbf{y}_{21}$	-0.007	-0.025	-0.830
$\mathbf{y}_{31}$	-0.001	0.000	0.000
$\mathbf{y}_{12}$	1.709	0.984	-0.008
$\mathbf{y}_{22}$	0.000	0.000	0.001
$\mathbf{y}_{32}$	0.000	0.001	0.000

Esse, in tre, coprono 1'87% della variabilità delle componenti di primo livello. Rispetto agli indicatori originali si puo ritenere che la variabilità spiegata sia dell'83% cioè quanto si poteva ottenerc fattorizzando la matrice originaria **S**. Questo è coerente con l'idea che le componenti di secondo livello colgano, almeno in parte, le interazioni tra indicatori in macrodeterminanti diverse trascurate dalle sottoanalisi.

Le componenti di secondo livello hanno però due debolezze: una di tipo tecnico ed una logica. Da un lato le componenti tralasciate in una sottoanalisi perché non rilevanti ai fini della variabilità spiegata di quella macro-determinante potrebbcro invece avere legami molto forti con le componenti di altri blocchi. Il calcolo di queste ultime componenti risulterebbe perciò privato di una fonte potenzialmente rilevante di informazione. L'altra carenza riguarda la interpretabilità delle componenti di secondo livello: sono già note le difficoltà e la necessità di forti doti creative per comprendere il senso delle componenti di primo livello, tanto che, spesso, ci si deve contentare di descrizioni vaghe, di larga massima e soggettive. Il problema si dilata per le componenti di secondo livello, che sono labili fili che collegano fjli già piultosto evanescenti. Al primo problema è difficile ovviare: la speranza è che i benefici siano maggiori dei costi connessi. Il seconda invece non è molto grave se si tiene conto che le componenti saranno utilizzate per la *cluster analysis* e che la caratterizzazione dei grnppi avverrà, come suggeriscono Friedman e Rubin (1967), nei tcrmini degli indicatori originali e non sulla base degli indicatori artificiali generati dall'analisi delle componenti principali.