

ALMA MATER STUDIORUM · UNIVERSITÀ DI BOLOGNA

---

Scuola di Scienze  
Dipartimento di Matematica  
Corso di Laurea in Matematica

**METODI DI ALGEBRA  
LINEARE NUMERICA  
PER L'ANALISI FATTORIALE**

**Relatore:**  
**Prof. Valeria Simoncini**

**Presentata da:**  
**Luca Bortolan**

Anno Accademico 2015/2016

*”Ma chi l’avrebbe detto che la vita  
mi sorprendevo come hai fatto TU.*

*Tu m’hai aperto come una ferita  
sto sanguinando ma non ti lascio più.”*

*(Lorenzo Cherubini)*

*Alla mia famiglia [Mt25, 14-30],  
ai ragazzacci della C.U.S.L.  
e alla People of Science.*

# Indice

<b>1</b>	<b>Il Modello Fattoriale Ortogonale</b>	<b>2</b>
1.1	Vettori media e matrici di covarianza . . . . .	2
1.2	Struttura del modello fattoriale . . . . .	3
1.3	Metodo di stima del fattore principale . . . . .	6
1.4	Scelta di $m$ . . . . .	7
1.5	Rotazione dei fattori . . . . .	8
1.6	Iterazione QR per il calcolo di autovalori . . . . .	9
<b>2</b>	<b>Applicazioni</b>	<b>13</b>
2.1	Dettagli implementativi . . . . .	13
2.2	Stock-price data . . . . .	14
2.3	Considerazioni sugli stock-price data . . . . .	19
2.4	Secondo dataset . . . . .	22
2.5	Vivibilità nelle città degli USA . . . . .	28

# Introduzione

L'**analisi fattoriale** è una tecnica statistica il cui principale scopo è di descrivere, dove possibile, le relazioni di covarianza tra alcune variabili di un determinato studio in termini di alcune quantità non osservabili chiamate *fattori*. La tecnica suppone che le variabili possano essere suddivise in gruppi grazie ad alcune particolari relazioni di correlazione, cioè ammette che tutte le variabili di uno stesso gruppo siano altamente correlate tra loro e abbiano correlazione relativamente bassa con quelle di un gruppo diverso. A questo punto è possibile pensare che ogni gruppo rappresenti un singolo *fattore*, responsabile della correlazione osservata.

L'elaborato propone uno studio sui metodi dell'algebra lineare utilizzati per la realizzazione dell'analisi fattoriale, in particolare per il metodo del fattore principale, e l'applicazione di tale analisi su due dataset reali. Il primo consiste negli stock-price data dei titoli di 5 società riferiti a 100 settimane, il secondo nello studio di 329 città degli USA attraverso dei punteggi assegnati a 9 criteri di vivibilità.

# Capitolo 1

## Il Modello Fattoriale Ortogonale

### 1.1 Vettori media e matrici di covarianza

Per trattare l'argomento dell'analisi fattoriale abbiamo prima bisogno di dare la definizione di alcune grandezze matematiche necessarie per il nostro lavoro.

Sia  $\mathbf{X}' = [X_1, X_2, \dots, X_p]$  un vettore qualsiasi  $1 \times p$ , supponendo che  $X_i$  abbia funzione densità di probabilità  $f_i$  se è continua, o funzione di probabilità  $p_i$  se è discreta, definiamo le medie  $\mu_i$  e le varianze  $\sigma_{ii}$  come:

$$\mu_i = \begin{cases} \int_{-\infty}^{\infty} x_i f_i(x_i) dx_i, & \text{se } X_i \text{ è una variabile continua} \\ \sum_i x_i p_i(x_i), & \text{se } X_i \text{ è una variabile discreta} \end{cases}$$

$$\sigma_{ii} = \begin{cases} \int_{-\infty}^{\infty} (x_i - \mu_i)^2 f_i(x_i) dx_i, & \text{se } X_i \text{ è una variabile continua} \\ \sum_i (x_i - \mu_i)^2 p_i(x_i), & \text{se } X_i \text{ è una variabile discreta.} \end{cases}$$

Presa invece una coppia di variabili qualsiasi  $X_i$  e  $X_k$ , con funzione densità congiunta  $f_{ik}(x_i, x_k)$  se sono continue, o funzione di probabilità congiunta  $p_{ik}(x_i, x_k)$  se sono discrete, definiamo la covarianza  $\sigma_{ik}$  come:

$$\sigma_{ik} = \begin{cases} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_i - \mu_i)(x_k - \mu_k) f_{ik}(x_i, x_k) dx_i dx_k, & \text{se } X_i, X_k \text{ sono variabili continue} \\ \sum_i \sum_k (x_i - \mu_i)(x_k - \mu_k) p_{ik}(x_i, x_k), & \text{se } X_i, X_k \text{ sono variabili discrete} \end{cases}$$

e  $\mu_i$  e  $\mu_k$  sono le medie marginali.

Quando  $i = k$  la covarianza corrisponde alla varianza marginale. Se  $\text{Cov}(X_i, X_k) = 0$  allora diciamo che  $X_i$  e  $X_k$  sono indipendenti.

Le medie e le covarianze del vettore  $\mathbf{X}$  possono essere scritte usando vettori e matrici. I valori attesi di ogni elemento sono contenuti nel vettore delle medie  $\mu = E(X)$  mentre le varianze  $\sigma_{ii}$  e le covarianze  $\sigma_{ik}$  sono contenute nella matrice simmetrica di varianza:

$$S = Cov(X) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \quad (1.1)$$

È spesso utile separare le informazioni contenute nelle varianze  $\sigma_{ii}$  da quelle contenute in una particolare misura chiamata coefficiente di correlazione  $\rho_{ik}$  che dipende dalle covarianze in questo modo:

$$\rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{kk}}}$$

Definiamo quindi matrice di correlazione  $R$  la matrice simmetrica:

$$R = \begin{bmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} \\ \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} & \cdots & \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{pp}}} \end{bmatrix} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix}. \quad (1.2)$$

Possiamo allora verificare che:

$$V^{1/2}RV^{1/2} = S, \quad (1.3)$$

per una matrice  $V^{1/2}$  chiamata deviazione standard

$$V^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix} \quad (1.4)$$

## 1.2 Struttura del modello fattoriale

Il modello fattoriale suppone che un qualsiasi vettore osservabile  $\mathbf{X}$  sia dipendente da alcune variabili non osservabili  $F_1, F_2, \dots, F_m$ , chiamate **fattori comuni**, e da  $p$  errori,

chiamati **fattori specifici**  $\epsilon_1, \epsilon_2, \dots, \epsilon_p$ , in questo modo:

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \dots + l_{1m}F_m + \epsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \dots + l_{2m}F_m + \epsilon_2 \\ &\vdots \\ X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \dots + l_{pm}F_m + \epsilon_p \end{aligned} \quad (1.5)$$

I termini  $l_{ij}$  sono chiamati *carichi* e sono rappresentati nella rispettiva matrice dei carichi **L**.

In termini matriciali tale scrittura diventa:

$$\underbrace{\mathbf{X}}_{p \times 1} = \underbrace{\boldsymbol{\mu}}_{p \times 1} + \underbrace{\mathbf{L}}_{p \times m} \underbrace{\mathbf{F}}_{m \times 1} + \underbrace{\boldsymbol{\epsilon}}_{p \times 1} \quad (1.6)$$

Il significato di tale modello è quello di sottolineare che ogni grandezza  $X_i$  può essere descritta in maniera diversa da ciascuno dei fattori  $F_i$ , cioè il ruolo dei carichi è che il loro valore mi dice quanto il rispettivo fattore incide sulla grandezza considerata. Questa scrittura ha senso per  $1 \leq m \leq p$ ; in particolare però, come scritto nell'introduzione, quello che cerchiamo è di descrivere  $p$  variabili in funzione di alcuni fattori latenti che siano minori del numero di variabili stesse, quindi il modello ha una valenza su una particolare ricerca se utilizzato con  $m < p$ .

In ogni caso con questo modello le  $p$  deviazioni  $X_i - \mu_i$  sono rappresentate con  $p + m$  variabili  $F_i, \epsilon_j$  ed è quindi difficile poter fare una verifica diretta della correttezza, per questo assumiamo alcune proprietà che i vettori  $\mathbf{F}$  e  $\boldsymbol{\epsilon}$  devono rispettare perché lo studio possa essere semplificato e per poter avere dei risultati interessanti.

In particolare chiediamo che:

$$E(\mathbf{F}) = \mathbf{0}, \quad Cov(\mathbf{F}) = \mathbf{I} \quad (1.7)$$

$$E(\boldsymbol{\epsilon}) = \mathbf{0}, \quad Cov(\boldsymbol{\epsilon}) = \boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \psi_m \end{bmatrix} \quad (1.8)$$

$$Cov(\boldsymbol{\epsilon}, \mathbf{F}) = E(\boldsymbol{\epsilon}\mathbf{F}') = \mathbf{0} \quad (1.9)$$

Così da un modello come quello di partenza e che soddisfa queste assunzioni abbiamo che:

$$\begin{aligned}(X - \mu)(X - \mu)' &= (LF + \epsilon)(LF + \epsilon)' \\ &= (LF + \epsilon)((LF)' + \epsilon') \\ &= LF(LF)' + \epsilon(LF)' + LF\epsilon' + \epsilon\epsilon'\end{aligned}\tag{1.10}$$

quindi, un particolare, risulta

$$\begin{aligned}S = Cov(X) &= E(X - \mu)(X - \mu)' = (LF + \epsilon)(LF + \epsilon)' \\ &= LE(FF')L' + E(\epsilon F')L' + LE(F\epsilon') + E(\epsilon\epsilon') \\ &= LL' + \Psi\end{aligned}\tag{1.11}$$

e, usando le equazioni (1.7) e (1.9):

$$Cov(X, F) = E(X - \mu)F' = LE(FF') + E(\epsilon F') = L\tag{1.12}$$

La porzione di varianza dell'*i*-esima variabile a cui contribuiscono i fattori comuni è chiamata *i*-esima **comunanza**, mentre la porzione a cui contribuiscono i fattori specifici è detta **varianza specifica**. Risulta quindi

$$\sigma_{ii} = l_{i1}^2 + l_{i2}^2 + \dots + l_{im}^2 + \psi_i$$

definiamo  $h$  la comunanza, risulta:

$$h_i^2 = l_{i1}^2 + l_{i2}^2 + \dots + l_{im}^2$$

avremo, in forma ridotta,

$$\sigma_{ii} = h_i^2 + \Psi_i, \quad i = 1, 2, \dots, p.\tag{1.13}$$

Quando  $m = p$  la matrice di covarianza  $S$  può essere scritta esattamente come  $LL'$ , ossia  $\Psi$  può essere la matrice nulla. Nonostante ciò, come detto precedentemente, l'analisi fattoriale risulta utile se, trovando particolari relazioni nella matrice di covarianza, possiamo ricondurre il nostro studio ad alcuni *fattori* di numero minore rispetto a quello dei dati iniziali, riducendo quindi le dimensioni della matrice di partenza; in termini algebrici questo significa che ci interessa il caso in cui  $m$  è minore di  $p$ .

### 1.3 Metodo di stima del fattore principale

La scomposizione spettrale ci assicura che è possibile una fattorizzazione della matrice di covarianza  $S$ . Chiamiamo  $(\lambda_i, e_i)$  le coppie formate dagli autovalori e i rispettivi autovettori di  $S$ , ordinate in modo che risulti  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ .

Allora

$$\begin{aligned}
 S &= \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \dots + \lambda_p e_p e_p' \\
 &= \begin{bmatrix} \sqrt{\lambda_1} e_1 & \sqrt{\lambda_2} e_2 & \dots & \sqrt{\lambda_p} e_p \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} e_1' \\ \sqrt{\lambda_2} e_2' \\ \vdots \\ \sqrt{\lambda_p} e_p' \end{bmatrix}
 \end{aligned} \tag{1.14}$$

Utilizzando questa decomposizione per l'analisi fattoriale siamo nel caso  $m = p$ , ossia avremmo  $\psi_i = 0 \forall i$  e come abbiamo osservato una rappresentazione di  $S$  di questo tipo non è particolarmente utile, vorremmo un modello che spieghi la matrice di covarianza con un numero basso di fattori comuni.

Una modalità per ridurre il numero di variabili è quella di trascurare gli autovalori più bassi, cioè eliminare il contributo di  $\lambda_{m+1} e_{m+1} e_{m+1}' + \dots + \lambda_p e_p e_p'$ . Questa rappresentazione assume che gli errori specifici  $\epsilon_i$  siano trascurabili, mentre vengono inclusi i fattori specifici le cui varianze possono essere prese dagli elementi diagonali di  $S-LL'$ .

L'approssimazione è quindi rappresentabile così

$$\begin{aligned}
 S &= LL' + \Psi \\
 &\approx \begin{bmatrix} \sqrt{\lambda_1} e_1 & \sqrt{\lambda_2} e_2 & \dots & \sqrt{\lambda_m} e_m \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} e_1' \\ \sqrt{\lambda_2} e_2' \\ \vdots \\ \sqrt{\lambda_m} e_m' \end{bmatrix} + \begin{bmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \psi_p \end{bmatrix}
 \end{aligned} \tag{1.15}$$

Dove  $\psi_i = \sigma_{ii} - \sum_{j=1}^m l_{ij}^2$ .

Questo tipo di rappresentazione è noto come *metodo del fattore principale* e può essere applicato alla matrice di covarianza o anche alla matrice di correlazione.

## 1.4 Scelta di $m$

Per il metodo del fattore principale il valore di  $m$  corrisponde quindi alle coppie autovalore-autovettore che si prendono in considerazione e ogni *carico* non cambia se il numero di fattori viene aumentato, lo scopo dello studio allora diventa scegliere il numero di fattori  $m$ .

Se il numero di fattori non è deciso da considerazioni a priori la scelta di  $m$  può essere fatta a partire dalla matrice dei residui che, considerando  $\tilde{L}$  la matrice dei carichi stimati  $\tilde{l}_{ij}$ , e  $\tilde{\Psi}$  la rispettiva matrice delle varianze specifiche, sarà:

$$S - (\tilde{L}\tilde{L}' + \tilde{\Psi}) \quad (1.16)$$

Gli elementi diagonali di questa matrice saranno zero, e la scelta di  $m$  può essere considerata accettabile se anche gli altri elementi sono vicini allo zero.

Analiticamente possiamo stimare questa grandezza partendo da una considerazione generale; per una qualsiasi matrice  $A$  di dimensione  $n \times n$  se chiamiamo  $\tilde{A}$  la sua stima calcolata utilizzando solo i primi  $m$  autovalori avremo che:

$$\|A - \tilde{A}\|_F^2 = \left\| \sum_{i=m+1}^n x_i \lambda_i x_i^T \right\|_F^2 = \|X_2 \Lambda_2 X_2^T\|_F^2 = \|\Lambda_2 X_2^T\|_F^2 = \|\Lambda_2\|_F^2 = (\lambda_{m+1}^2 + \dots + \lambda_n^2)$$

In particolare nel nostro caso questa proprietà implica che:

$$\|S - (\tilde{L}\tilde{L}' + \tilde{\Psi})\|_F^2 \leq (\lambda_{m+1}^2 + \dots + \lambda_p^2) \quad (1.17)$$

Questo mostra che se è piccola la somma dei quadrati degli autovalori trascurati è piccolo anche l'errore di approssimazione. Per calcolare quindi il contributo di ogni fattore rispetto alla varianza (o covarianza) totale dobbiamo rapportarlo con la traccia della matrice  $S$  (o  $R$ ). Il contributo di ognuno degli autovalori sarà:

$$\begin{cases} \frac{\hat{\lambda}_j}{\text{tr}(S)} & \text{per l'analisi fattoriale di } S \\ \frac{\hat{\lambda}_j}{p} & \text{per l'analisi fattoriale di } R \end{cases} \quad (1.18)$$

La somma del contributo di ogni autovalore scelto rappresenta quindi la precisione dello studio, cioè ci dice quanto ogni  $\lambda_i$  contribuisce a rappresentare la variabilità dei dati. Tale quantità e sarà quindi il criterio per decidere il numero di autovalori che vogliamo prendere in considerazione.

## 1.5 Rotazione dei fattori

Vediamo ora una sorta di ambiguità inerente al modello fattoriale.

Prendiamo una matrice ortogonale  $\mathbf{T} \in R^{m \times m}$ , quindi tale che  $\mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$ ; riprendendo il modello fattoriale avremo:

$$X - \mu = LF + \epsilon = LTT'F + \epsilon = L^*F^* + \epsilon \quad (1.19)$$

dove  $\mathbf{L}^* = \mathbf{L}\mathbf{T}$  e  $\mathbf{F}^* = \mathbf{T}'\mathbf{F}$ .

Questo implica anche che  $\mathbf{E}(\mathbf{F}^*) = \mathbf{T}'\mathbf{E}(\mathbf{F}) = \mathbf{0}$  e  $\mathbf{Cov}(\mathbf{F}^*) = \mathbf{T}'\mathbf{Cov}(\mathbf{F})\mathbf{T} = \mathbf{T}'\mathbf{T} = \mathbf{I}$ .

Ciò ci indica che è impossibile, soltanto dall'osservazione di  $X$ , distinguere i carichi  $L$  da i carichi  $L^*$  nonostante siano differenti, cioè i fattori  $F$  e  $F^*$  hanno le stesse proprietà statistiche e generano la stessa matrice di covarianza:

$$S = LL' + \Psi = LTT'L' + \Psi = (L^*)(L^*)' + \Psi \quad (1.20)$$

Questa ambiguità è spiegata dal fatto che una matrice ortogonale corrisponde ad una rotazione o ad una riflessione del sistema di coordinate di  $X$ ; per questa ragione una trasformazione di questo tipo applicata ai carichi è detta **rotazione dei fattori**.

In particolare, chiamata  $\hat{L}$  la matrice dei carichi stimati, anche la matrice dei residui  $S_n - \hat{L}\hat{L}' - \hat{\Psi} = S_n - \hat{L} * \hat{L}' * - \hat{\Psi}$  rimane invariata, come le varianze specifiche  $\hat{\psi}_i$  e le comunanze  $\hat{h}_i^2$ .

La rotazione viene applicata perché i fattori principali potrebbero non essere facilmente interpretabili, ruotarli può essere utile ad avere una struttura più semplice. Idealmente quello che vorremmo raggiungere è che una variabile abbia un carico alto in un singolo fattore e basso negli altri, ma questo non è sempre possibile; risulta comunque utile applicare la rotazione per aggiungere elementi che confermino o modifichino le conclusioni tratte precedentemente.

Vediamo i metodi grafici e analitici per determinare una rotazione ortogonale utile a raggiungere una struttura più semplice. Quando  $m=2$  la struttura semplice può essere trovata facilmente. I fattori comuni non correlati vengono considerati come vettori unitari lungo gli assi coordinati perpendicolari. Un grafico delle coppie di carichi  $(\hat{l}_{i1}, \hat{l}_{i2})$  produce  $p$  punti, ognuno corrispondente ad una variabile. Gli assi coordinati possono allora essere visualizzati ruotati di un angolo  $\phi$  e i nuovi carichi ruotati  $\hat{l}_{ij}^*$  sono determinati dalla

relazione:

$$\underbrace{\hat{\mathbf{L}}^*}_{(p \times 2)} = \underbrace{\hat{\mathbf{L}}}_{(p \times 2)} \underbrace{\mathbf{T}}_{(2 \times 2)} \quad (1.21)$$

$$\text{dove } \begin{cases} \mathbf{T} = \begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix} & \text{rotazione oraria} \\ \mathbf{T} = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} & \text{rotazione antioraria.} \end{cases}$$

Tale relazione viene spesso anche implementata con un'analisi grafica in due dimensioni. In questo caso, gruppi di variabili a volte sono visibili a occhio, e questi gruppi permettono di identificare i fattori comuni senza uno studio delle magnitudini dei carichi.

Nel caso  $m > 2$  si applica una particolare tecnica chiamata **rotazione Varimax** che tende a massimizzare la somma dei quadrati dei carichi. Per prima cosa pesa i carichi dividendoli per la comunanza:

$$\tilde{l}_{ij}^* = \frac{\hat{l}_{ij}^*}{\hat{h}_i}$$

Quello che vogliamo dopo la rotazione è massimizzare questa quantità.

Definiamo procedura Varimax quella che seleziona la rotazione per trovare il massimo della seguente funzione:

$$\mathbf{V} = \frac{1}{p} \sum_{j=1}^m \left\{ \sum_{i=1}^p (\tilde{l}_{ij}^*)^4 - \frac{1}{p} \left( \sum_{i=1}^p (\tilde{l}_{ij}^*)^2 \right)^2 \right\} \quad (1.22)$$

## 1.6 Iterazione QR per il calcolo di autovalori

Nel prossimo capitolo utilizzeremo, per il calcolo degli autovalori della matrice di covarianza o correlazione, necessari per la decomposizione spettrale, la funzione **eig** di **MATLAB**. Questa funzione sfrutta una particolare iterazione utilizzata per fattorizzare le matrici, chiamata **iterazione QR**.

Più precisamente, data una generica matrice  $\mathbf{A} \in \mathbb{R}^{n \times m}$  con  $n \geq m$ ,  $\mathbf{A}$  può essere scritta come prodotto di due matrici come segue:

$$\mathbf{A} = \mathbf{QR} = [\mathbf{Q}_1, \mathbf{Q}_2] [\mathbf{R}'_1, 0]' = \mathbf{Q}_1 \mathbf{R}_1$$

dove  $\mathbf{R}_1 \in \mathbb{R}^{m \times m}$ ,  $\mathbf{Q}$  è una matrice  $n \times n$  reale ortogonale,  $\mathbf{Q} = [\mathbf{Q}_1, \mathbf{Q}_2]$  con  $\mathbf{Q}_1 \in \mathbb{R}^{n \times m}$  e  $\mathbf{Q}_2 \in \mathbb{R}^{n \times (n-m)}$ , dove le colonne di  $\mathbf{Q}_1, \mathbf{Q}_2$  sono ortonormali. Inoltre,  $\mathbf{R}_1$  è triangolare

superiore ed  $\mathbf{R}$  ha le stesse dimensioni di  $\mathbf{A}$ ; la versione  $A = Q_1 R_1$  prende il nome di fattorizzazione QR ridotta, perché non genera né memorizza la matrice  $Q_2$ . Da questo procedimento si ha anche che  $Q' A = R$  ( $Q^H A = R$  nel caso complesso), cioè  $Q$  è la matrice che trasforma  $A$  in una matrice triangolare superiore. Vediamo così il primo indizio sulla modalità che possiamo utilizzare per i problemi agli autovalori nel caso in cui  $n = m$ .

Quest'ultima osservazione è proprio quella che sta alla base del comando `eig` di MATLAB il cui algoritmo determina una successione di matrici  $\{\mathbf{T}_k\}_{k \in N}$  tale che:

$$\mathbf{T}_k := \mathbf{U}_k^H \mathbf{A} \mathbf{U}_k \rightarrow \mathbf{T}, \quad k \rightarrow +\infty,$$

dove  $\mathbf{U}_k^H$  è il prodotto delle matrici  $Q_k^H$  determinate dalla fattorizzazione QR delle matrici  $T_k$  e  $\mathbf{T}$  è una matrice triangolare superiore (diagonale se  $A$  è normale) avente come elementi diagonali gli autovalori di  $A$ .

Come vedremo, questa iterazione tende ad ottenere la decomposizione di Schur della matrice  $A$ , cioè:

$$\mathbf{A} = \mathbf{U}_\infty \mathbf{T} \mathbf{U}_\infty^H. \quad (1.23)$$

La successione  $\{\mathbf{T}_k\}_{k \in N}$ , può essere determinata come segue:

$$\mathbf{T}_0 = \mathbf{A}$$

Per  $k = 0, 1, \dots$ ,

$$\mathbf{T}_k = \mathbf{Q}_k \mathbf{R}_k \text{ (fattorizzazione QR di } \mathbf{T}_k \text{)}$$

$$\mathbf{T}_{k+1} := \mathbf{R}_k \mathbf{Q}_k$$

end

Nella prima operazione viene effettuata una fattorizzazione QR della matrice quadrata  $T_k$ . Nella seconda operazione, viene creata una nuova matrice  $T_{k+1}$  come prodotto di due fattori  $R_k, Q_k$  in ordine scambiato; in questo modo, dato che dalla fattorizzazione segue  $Q_k^H T_k = R_k$ , si ha  $T_{k+1} = Q_k^H T_k Q_k$ . Essendo le matrici  $Q_k$  unitarie per ogni  $k$ , si ha che tutte le matrici  $T_k$  sono simili tra loro. Inoltre,

$$\mathbf{T}_{k+1} = \mathbf{Q}_k^H \mathbf{Q}_{k-1}^H \mathbf{T}_{k-1} \mathbf{Q}_{k-1} \mathbf{Q}_k = \dots = \underbrace{\mathbf{Q}_k^H \mathbf{Q}_{k-1}^H \dots \mathbf{Q}_0^H}_{\mathbf{U}_k^H} \mathbf{T}_0 \mathbf{Q}_0 \dots \mathbf{Q}_{k-1} \mathbf{Q}_k =: \mathbf{U}_k^H \mathbf{A} \mathbf{U}_k,$$

con  $\mathbf{U}_k$  unitaria quindi gli autovalori di  $A$  vengono mantenuti.

Il costo computazionale di questo algoritmo di base è alquanto elevato, infatti la fattorizzazione QR costa un  $O(n^3)$  flops se  $T_k$  è piena, così come il costo del prodotto per  $T_k$  costa  $O(n^3)$  flops.

Nel nostro caso però la matrice  $A$  è una matrice di correlazione quindi in particolare sarà una matrice simmetrica; questo permette di avere un costo minore per la fattorizzazione QR perchè possiamo dapprima ridurla, mediante trasformazioni ortogonali di similitudine, alla forma tridiagonale. Questo implica anche che la rispettiva decomposizione di Schur, che abbiamo visto, avrà la matrice  $T$  diagonale con elementi diagonali gli autovalori di  $A$ . Per realizzare questo obiettivo costruiamo una prima matrice simmetrica  $U_1$  che azzeri gli elementi  $a_{3,1}, a_{4,1}, \dots, a_{n,1}$  della prima colonna di  $A$ , e formiamo la nuova matrice, simile ad  $A$ :

$$A_1 = U_1 A U_1$$

abbiamo:

$$\begin{aligned} U_1 A U_1 &= \begin{bmatrix} 1 & 0 \\ 0 & U_1^{(n-1)} \end{bmatrix} \begin{bmatrix} a_{11} & a_1^T \\ a_1 & A_{11} \end{bmatrix} \begin{bmatrix} 1 & o \\ o & U_1^{(n-1)} \end{bmatrix} = \\ &= \begin{bmatrix} a_{11} & a_1^T U_1^{(n-1)} \\ U_1^{(n-1)} a_1 & U_1^{(n-1)} A_{11} U_1^{(n-1)} \end{bmatrix} \end{aligned} \quad (1.24)$$

ma poichè  $U_1^{(n-1)}$  è stato costruito proprio per produrre la trasformazione  $U_1^{(n-1)} a_1 = (x, 0, \dots, 0)^T$ , avremo anche  $a_1^T U_1^{(n-1)} = (U_1^{(n-1)} a_1)^T = (x, 0, \dots, 0)$ . Pertanto  $A_1$  assumerà la forma (simmetrica):

$$A_1 = \begin{bmatrix} * & * & 0 & 0 & \dots & 0 \\ * & * & * & * & \dots & * \\ 0 & * & * & * & \dots & * \\ 0 & * & * & * & \dots & * \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & * & * & * & \dots & * \end{bmatrix}$$

Successivamente determiniamo la matrice  $U_2$  in modo che:

$$A_2 = U_2 A_1 U_2$$

sia di forma (simmetrica):

$$\mathbf{A}_3 = \begin{bmatrix} * & * & 0 & 0 & \cdots & 0 \\ * & * & * & 0 & \cdots & 0 \\ 0 & * & * & * & \cdots & * \\ 0 & 0 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & * & * & \cdots & * \end{bmatrix}$$

e così continuiamo sino a raggiungere la matrice finale

$$\mathbf{A}_{n-2} = \mathbf{U}_{n-2} \mathbf{A}_{n-3} \mathbf{U}_{n-2}$$

simmetrica, tridiagonale e simile alla matrice iniziale  $\mathbf{A}$ .

Una volta che si è ottenuta una matrice tridiagonale il costo della fattorizzazione QR sarà molto inferiore, in quanto devono essere azzerati solo gli  $n-1$  elementi della sottodiagonale. Questo può essere fatto applicando una procedura che non approfondiremo in questo elaborato: le rotazioni di Givens. Tale procedimento sfrutta le proprietà di particolari matrici che, se moltiplicate alla matrice di partenza, ne eliminano alcuni elementi, con un limitato costo computazionale e mantenendo la forma tridiagonale di essa. Questo metodo si applica fino a convergenza, cioè fino a che gli elementi sotto la diagonale principale non saranno abbastanza piccoli. Il costo di ogni iterazione è di un ordine di grandezza inferiore al caso di una matrice  $\mathbf{A}$  generica.

Riassumendo l'iterazione QR più avanzata è della forma:

Data  $\mathbf{T}_0 = \text{Tridiag}(\mathbf{A})$

Per  $k = 0, 1, \dots$ ,

$T_k = Q_k R_k$  (QR con rotazioni di Givens)

$T_{k+1} = R_k Q_k$

se  $\max_j \frac{|(T_{k+1})_{j+1,j}|}{|(T_{k+1})_{j+1,j+1}| + |(T_{k+1})_{j,j}|} < \epsilon$  Stop, altrimenti Continua

end

dove con  $\text{Tridiag}$  si indica la procedura per trasformare  $\mathbf{A}$  in forma tridiagonale.

# Capitolo 2

## Applicazioni

Il seguente capitolo propone due diverse applicazioni del metodo del fattore principale studiato. La prima, che riguarda lo studio sui tassi di rendimento di 100 settimane dei titoli di 5 società americane, ha lo scopo di vedere se attraverso la ricerca dei fattori latenti si trovano delle particolari relazioni tra alcune di queste aziende. La seconda riguarda lo studio di 9 criteri di vivibilità di 329 città degli USA e ha come scopo quello di trovare dei fattori che accomunino alcuni di questi criteri. Per entrambi gli esempi applicheremo prima il modello con l'algoritmo base, poi proveremo a dare un'interpretazione dei dati che verificheremo eseguendo anche la rotazione dei fattori.

### 2.1 Dettagli implementativi

Per l'applicazione dell'analisi fattoriale sui dati in possesso è stato scritto un codice **MATLAB** che sfruttasse alcune funzioni già implementate nel programma per la ricerca di una soluzione del metodo del fattore principale.

In particolare si è scritta una funzione `fact.m` che prende in input una matrice qualsiasi  $\mathbf{A} \in \mathbb{R}^{p \times n}$  e un valore  $\mathbf{m}$  e restituisce in output 3 valori: la matrice  $n \times m$  che rappresenta i carichi dei fattori stimati; un vettore  $n \times 1$  che descrive la porzione di varianza dovuta ad ogni autovalore; la porzione totale di varianza spiegata, cioè la somma degli elementi del vettore precedente.

Nel caso di  $m=2$  la funzione disegna anche il grafico che pone la prima colonna della matrice in output nell'asse delle ascisse e la seconda colonna nell'asse delle ordinate.

```

function [L] = fact_m(a,m)
    C=corrcoef(a);
    p=size(C,1);
    [V,D]=eig(C);
    [d1,i1]=sort(diag(D),'descend');
    V1=V(:,i1); % V1, d1 autocopie ordinate
    L=V1(:,1:m)*diag(sqrt(d1(1:m)));
    proporzione=d1/sum(d1)
    prop_m=sum(d1(1:m))/sum(d1)
    if m==2
        plot(L(:,1),L(:,2),'md')
        text(L(:,1),L(:,2),num2str((1:p)'))
    end
    %cumproporzione=cumsum(d1)/sum(d1);
end

```

Le funzioni principali già presenti in **MATLAB** e utilizzate sono: `corrcoef` che calcola la matrice dei coefficienti di correlazione di un qualsiasi vettore; `eig` che calcola le matrici  $[V,D]$  degli autovettori e degli autovalori della matrice in input, e di cui abbiamo visto l'algoritmo principale; `sort` che utilizziamo per ordinare in maniera discendente gli elementi di un vettore.

Per applicare invece la rotazione dei fattori ai nostri dataset abbiamo usato la funzione di **MATLAB** `rotatefactors`. Tale funzione riceve in input una matrice  $A$  a cui ruota i carichi in modo da massimizzare la somma delle varianze dei loro quadrati e restituisce in output una matrice  $B$ . Coerentemente con la matrice in input, le righe di  $B$  corrispondono alle variabili analizzate mentre le colonne rappresentano i fattori ruotati; ossia l'elemento  $b_{ij}$  rappresenterà il coefficiente ruotato della  $i$ -esima variabile rispetto il  $j$ -esimo fattore.

## 2.2 Stock-price data

Gli stock-price data che consideriamo nel primo esempio sono i rendimenti dei titoli delle società americane *Allied Chemical*, *Du Pont*, *Union Carbide*, *Exxon*, *Texaco* per un tempo di 100 settimane.

Allied Chemical	Du Pont	Union Carbide	Exxon	Texaco
0.000000	0.000000	0.000000	0.039473	-0.000000
0.027027	-0.044855	-0.003030	-0.014466	0.043478
0.122807	0.060773	0.088146	0.086238	0.078124
0.057031	0.029948	0.066808	0.013513	0.019512
0.063670	-0.003793	-0.039788	-0.018644	-0.024154
0.003521	0.050761	0.082873	0.074265	0.049504
-0.045614	-0.033007	0.002551	-0.009646	-0.028301
0.058823	0.041719	0.081425	-0.014610	0.014563
0.000000	-0.019417	0.002353	0.001647	-0.028708
0.006944	-0.025990	0.007042	-0.041118	-0.024630
0.010345	0.006353	0.083916	0.010291	-0.000000
-0.030717	0.020202	-0.040860	-0.039049	-0.050505
-0.003521	0.118812	0.089686	0.060070	0.021276
0.060071	0.079646	0.028807	0.036666	0.026041
-0.003333	-0.001025	0.028000	0.028938	-0.010152
0.055596	0.091282	0.042759	0.059375	-0.015812
0.051282	-0.007519	-0.041431	-0.016269	0.058510
-0.060976	-0.043561	0.023576	0.004566	-0.015075
-0.035714	0.018170	-0.021113	-0.007575	-0.010204
0.000000	-0.021569	-0.007843	0.088549	0.082474
-0.006734	-0.015030	-0.086956	-0.021037	-0.019047
0.000000	-0.017294	0.017316	0.054441	0.033980
0.030508	0.047619	0.055319	-0.008152	0.032863
0.023026	0.012846	-0.002016	0.013698	-0.031518
-0.061093	-0.043902	-0.042424	-0.029729	-0.014084
0.041096	0.016326	0.048523	0.018105	0.071428
-0.013158	-0.004016	-0.038229	-0.042407	-0.048888
0.003333	-0.008065	-0.014992	0.000000	-0.028037
-0.056478	-0.014228	-0.038627	-0.005714	-0.019607
0.051899	0.018557	0.066964	0.020302	-0.015000
-0.013559	-0.029352	0.012552	-0.008571	-0.010152

---

-0.037801	0.003252	-0.012397	-0.020172	-0.025641
-0.021429	0.031466	0.039749	0.016176	0.005263
-0.014599	-0.024390	-0.010060	0.004341	-0.005235
-0.014815	-0.020833	-0.091463	-0.007204	-0.015789
0.011278	-0.017021	0.064877	0.065312	0.026737
-0.096654	-0.075758	-0.073529	-0.053133	-0.026041
0.020576	0.058548	0.018141	0.063309	0.016042
0.088710	0.046460	0.022272	0.004059	-0.000000
0.007407	0.019027	0.045752	-0.008086	0.052631
-0.022059	0.002075	-0.017272	-0.021739	-0.045000
-0.031579	0.010352	0.012848	-0.013888	0.010695
0.039370	0.054303	-0.014799	0.011428	-0.005291
0.015151	0.029154	-0.021459	-0.009887	-0.021276
0.000000	-0.010466	0.035088	-0.014265	0.038043
-0.037313	-0.024038	-0.019068	-0.024602	-0.010471
0.015504	-0.027586	0.006479	0.022255	-0.026455
0.034351	0.024316	0.034335	0.020319	0.005434
-0.036900	0.011869	0.014523	0.007112	0.016216
0.068965	0.014663	0.016360	0.038135	0.063829
0.089606	0.079961	0.102616	0.002721	0.020000
0.000000	0.016949	0.029197	0.002713	0.004901
0.059210	0.077193	0.019504	-0.012178	0.039024
0.027950	0.009772	0.000000	-0.000000	-0.000265
-0.004196	0.014516	-0.031696	-0.004445	-0.014354
0.018405	-0.046900	0.061594	-0.043235	-0.029126
0.069277	0.056888	0.040956	0.040816	0.020000
-0.016901	-0.018268	-0.008197	-0.005602	-0.019607
-0.017192	-0.001618	-0.001653	-0.016901	0.005000
-0.040816	-0.035656	0.000000	0.014326	0.004975
-0.018237	-0.003361	-0.028146	0.035310	0.014851
-0.003096	-0.021922	-0.027257	0.005457	0.039024
0.018634	0.025862	-0.017513	0.018995	-0.004694
-0.057927	-0.018487	0.000000	-0.023968	-0.037735

---

0.087379	0.049657	0.033868	0.047748	0.039215
0.000000	-0.011419	-0.010345	-0.005208	0.028301
-0.019367	-0.011551	-0.022817	0.007853	0.013761
-0.046012	0.035893	0.044964	0.040612	0.004608
-0.077170	-0.004029	-0.003442	0.003797	-0.027522
0.034843	-0.008157	-0.018998	0.008827	-0.014151
-0.006734	-0.019737	-0.026408	0.023749	0.014354
-0.023729	-0.019295	-0.032550	-0.001221	0.023584
0.065972	0.024807	0.057944	0.020782	0.004608
0.000000	-0.036728	-0.014134	-0.007185	0.004587
-0.052117	-0.058925	-0.069892	0.009650	0.009132
0.054983	-0.003683	0.026975	-0.002389	0.009049
-0.003257	-0.009242	-0.022514	0.005988	-0.013452
0.022876	0.033582	0.001919	0.026190	0.004545
-0.003195	-0.005415	-0.003831	-0.013921	-0.000165
0.043590	-0.014519	-0.013385	0.021176	0.013824
-0.009317	0.013812	0.021654	-0.014927	-0.009090
-0.056426	-0.005557	-0.003854	-0.023696	-0.018348
0.003322	-0.041475	-0.029014	-0.002427	-0.004672
0.016556	0.017308	0.033864	0.034063	0.009389
-0.009772	-0.016068	-0.003854	0.014117	0.013953
0.026316	-0.016330	-0.009671	0.032482	0.027522
0.009615	0.009766	0.017578	0.016247	0.017857
-0.047619	-0.027079	-0.051823	-0.045468	-0.021929
-0.026667	-0.061630	-0.056680	-0.013452	-0.040358
0.010274	0.023305	0.034335	-0.018181	-0.004672
-0.044068	0.020704	-0.006224	-0.018518	0.004694
0.039007	0.038540	0.024988	-0.028301	0.032710
-0.039457	-0.029297	-0.065844	-0.015837	-0.045758
0.039568	0.024145	-0.006608	0.028423	-0.009661
-0.031142	-0.007941	0.011080	0.007537	0.014634
0.000000	-0.020080	-0.006579	0.029925	-0.004807
0.021429	0.049180	0.006622	-0.002421	0.028985

0.045454	0.046375	0.074561	0.014563	0.018779
0.050167	0.036380	0.004082	-0.011961	0.009216
0.019108	-0.033303	0.008362	0.033898	0.004566

Tabella 2.1: Stock-price data

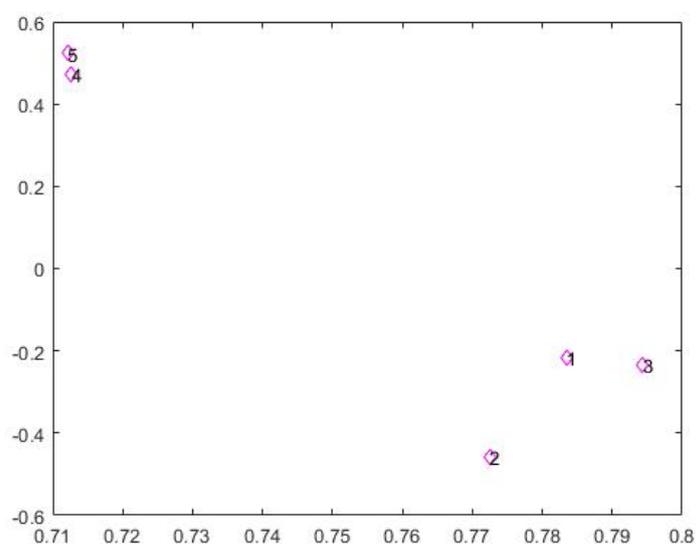
Quello che vogliamo fare è trovare la soluzione del metodo del fattore principale (1.6) calcolando i carichi, cioè trovando i coefficienti della combinazione lineare espressa nell'equazione (1.5), e attraverso questi poi stimare i fattori. Quello che vorremmo é trovare dei fattori per cui ogni variabile ha carico alto solo per uno di questi e carico vicino allo zero per gli altri. Per fare questo si è applicata la funzione `fact.m` prima descritta e si sono raccolti i risultati in una tabella che, coerentemente a quanto visto in (1.13) e (1.18), riporta i carichi stimati  $l_{ij}$ , le varianze specifiche di ognuna delle società  $\tilde{\psi}_i = 1 - l_{i1}^2 - l_{i2}^2 - \dots - l_{im}^2$  e la porzione di varianza spiegata  $\frac{\hat{\lambda}_1 + \hat{\lambda}_2 + \dots + \hat{\lambda}_m}{p}$ .

Vediamo, per esempio il risultato di tale funzione applicata alla matrice con  $m=1$ :

Società	Carichi stimati	Varianze specifiche
1.Allied Chemical	0.783	0.39
2.Du Pon	0.773	0.40
3.Union Carbide	0.794	0.37
4.Exxon	0.713	0.49
5.Texaco	0.712	0.49
<b>Porzione di varianza spiegata</b>	0.571	

Nel caso di  $m=2$ , oltre ad avere una precisione migliore, la funzione disegna anche il grafico, questo aiuta a vedere con più immediatezza le relazioni tra le diverse variabili:

Società	Carichi stimati		Varianze specifiche
1.Allied Chemical	0.783	-0.217	0.34
2.Du Pon	0.773	-0.458	0.19
3.Union Carbide	0.794	-0.234	0.31
4.Exxon	0.713	0.472	0.27
5.Texaco	0.712	0.524	0.22
<b>Porzione di varianza spiegata</b>	0.733		



## 2.3 Considerazioni sugli stock-price data

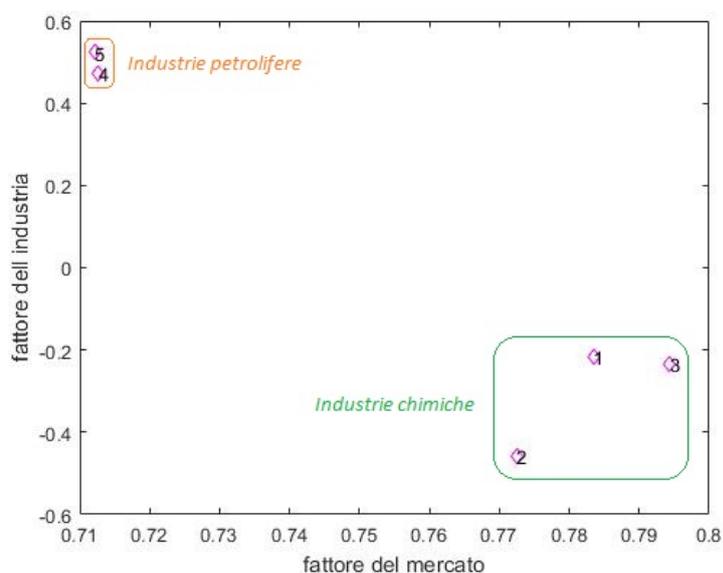
Nella soluzione con due fattori abbiamo spiegato già una porzione adeguata della varianza totale, il 73.3%, quindi da questa soluzione possiamo trarre le prime conclusioni. La prima colonna di carichi  $L_1$  risulta alta per tutte e 5 le società, cioè, in termini matematici, il coefficiente del primo fattore nella combinazione lineare (1.5) è grande per tutte le variabili. Questo significa che tutte le società sono altamente descritte dal primo fattore e quindi non si chiarisce particolarmente una divisione in gruppi diversi di queste. Il fattore sembra quindi rappresentare la condizione economica delle aziende, positiva per tutte, può essere chiamato il *fattore del mercato*.

La colonna di carichi  $L_2$  invece fornisce dei risultati interessanti perché mette fortemente in contrasto due gruppi di aziende: il primo formato da 1., 2. e 3. che hanno un valore compreso tra -0.458 e -0.217, e il secondo formato da 4. e 5. che hanno i valori 0.472 e 0.524. Questa è una prima prova della presenza di fattori non osservabili che inducono alta correlazione tra alcune variabili; questi due gruppi infatti possono essere riconoscibili perché le società appartenenti al primo insieme sono industrie chimiche, mentre quelle appartenenti al secondo insieme sono industrie petrolifere. Il secondo fattore differenzia insomma le aziende per il loro ambito e può essere chiamato il *fattore dell'industria*.

Vediamo come l'analisi in questo caso è utile perché raggiunge lo scopo di ridurre il nu-

mero di dati da studiare: dalle 5 aziende iniziali abbiamo trovato 2 fattori che descrivono esaurientemente l'oggetto della nostra ricerca, senza perdere un'alta percentuale di varianza.

Queste considerazioni sono state tutte raggiunte con un'analisi strettamente numerica ma, come abbiamo visto, anche graficamente possono essere colte le stesse conclusioni. Vediamo infatti che nel grafico le aziende si dispongono proprio in due gruppi distinti:



Vediamo ora cosa si ottiene applicando la rotazione dei fattori all'esempio; eseguendo la funzione *rotatefactors* sui carichi del caso  $m=2$  osserviamo che vengono così ruotati:

Società	Carichi stimati		Carichi ruotati	
	1	2	1	2
1.Allied Chemical	0.783	-0.217	<b>0.746</b>	0.323
2.Du Pon	0.773	-0.458	<b>0.889</b>	0.128
3.Union Carbide	0.794	-0.234	<b>0.766</b>	0.316
4.Exxon	0.713	0.472	0.258	<b>0.815</b>
5.Texaco	0.712	0.524	0.226	<b>0.855</b>

La rotazione rende molto chiaro ciò che, grazie al grafico, avevamo precedentemente concluso: le aziende chimiche hanno carichi alti nel primo fattore mentre quelle petrolifere hanno carichi alti nel secondo. Come detto questo è proprio lo scopo dell'analisi, trovare dei fattori tali che ogni variabile ha carico alto solo per uno di essi. È quindi rafforzata

l'idea che ci sia un'alta correlazione all'interno di questi due differenti gruppi. I fattori così ruotati possono essere ridefiniti: il primo rappresenta quei criteri che riguardano la crescita delle industrie chimiche, il secondo rappresenta quelli che riguardano invece le industrie petrolifere.

## 2.4 Secondo dataset

Per la seconda applicazione è stato usato un dataset preso dal *Places Rated Almanac* di *Richard Boyer* e *David Savageau*; si tratta dello studio su 329 città a cui sono stati assegnati dei punteggi per ognuno dei seguenti 9 criteri: *Clima*, *Alloggi*, *Salute*, *Criminalità*, *Trasporti*, *Educazione*, *Arte*, *Divertimento*, *Economia*.

Città	Clima	Alloggi	Salute	Criminalità	Trasporti	Educazione	Arte	Divertimento	Economia
Abilene	521	6200	237	923	4031	2757	996	1405	7633
Akron	575	8138	1656	886	4883	2438	5564	2632	4350
Albany	468	7339	618	970	2531	2560	237	859	5250
Albany-Schenectady-Troy	476	7908	1431	610	6883	3399	4655	1617	5864
Albuquerque	659	8393	1853	1483	6558	3026	4496	2612	5727
Alexandria	520	5819	640	727	2444	2972	334	1018	5254
Allentown, Bethlehem	559	8288	621	514	2881	3144	2333	1117	5097
Alton, Granite City	537	6487	965	706	4975	2945	1487	1280	5795
Altoona	561	6191	432	399	4246	2778	256	1210	4230
Amarillo	609	6546	669	1073	4902	2852	1235	1109	6241
Anaheim-Santa Ana	885	16047	2025	983	3954	2843	5632	3156	6220
Anchorage	195	12175	601	1223	5091	2414	2346	3000	7668
Anderson, IN	530	5704	580	878	2865	2469	430	838	3370
Anderson, SC	591	5725	820	975	2707	2772	169	613	4262
Ann Arbor	546	11014	2508	1067	3433	3346	7559	2288	4579
Anniston	560	5530	598	1125	3051	2189	268	1165	4730
Appleton-Oshkosh-Neenah	396	7877	833	525	3298	2844	1166	2315	5275
Asheville	694	6722	1204	566	5086	2990	1391	1542	5196
Athens	601	6691	605	933	1866	2646	3546	1001	5193
Atlanta	696	8316	3195	1308	8409	3057	7559	1362	6315
Atlantic City	615	11074	637	1878	3556	2929	621	2711	8107
Augusta	534	6292	1798	872	2523	2915	1047	913	5431
Aurora-Elgin	474	10384	1203	821	3943	2208	3857	1800	5097
Austin	435	8831	782	1049	3670	3063	5355	1063	7439
Bakersfield	560	8068	420	1561	3725	2564	1222	1568	6056
Baltimore	567	9148	3562	1730	7405	3471	9788	2925	5503
Bangor	451	7277	780	651	5613	2934	1995	2148	5172
Baton Rouge	427	8083	342	1565	3329	2635	4237	1413	6308
Battle Creek	527	6342	900	1031	4652	2483	354	1648	4008
Beaumont-Port Arthur	423	6288	616	1313	2782	2745	1795	1813	6019
Beaver County	586	7866	861	310	2960	2535	1284	1480	3119
Bellingham	772	8329	240	825	3776	2778	1302	3200	4247
Benton Harbor	566	6761	570	1190	2989	2545	79	1477	3635
Bergen-Passaic	559	14607	2661	857	3511	3653	9304	1918	6016
Billings	452	8315	479	810	6285	3008	778	2046	5913
Biloxi-Gulfport	584	6458	441	810	2516	2592	679	2106	5801
Binghampton	550	8257	1007	415	4529	3052	1599	1722	5614
Birmingham	612	6811	1692	1123	5177	2851	3958	1234	4843
Bismarck	149	8365	804	413	4303	2686	1211	1630	6019
Bloomington	558	7056	731	657	1746	2873	2152	1990	4829
Bloomington-Normal	487	8654	815	673	5889	2854	1470	1605	5863
Boise City	592	8221	453	880	6575	2391	2385	1672	4633
Boston	623	11609	5301	1215	6801	3479	21042	3066	6363
Boulder-Longmont	459	11914	962	1088	7108	2587	3663	4012	7127
Bradenton	440	8242	333	1093	3805	2712	154	1349	7437
Brazoria	423	8394	438	768	2391	2718	1506	1512	6020
Bremerton	808	9060	310	651	1670	2544	382	1973	5671
Bridgeport-Milford	648	13429	2550	943	3197	3029	8368	1913	7197
Bristol	516	10041	975	545	4495	2628	514	777	6527
Brockton	575	8263	916	1336	3810	2729	2001	1217	6900
Brownsville-Harlington	440	5376	91	974	3119	2413	162	3000	4968
Bryan-College Station	383	8228	640	1016	2530	2973	2002	1413	8040
Buffalo	571	8064	2465	971	5384	3121	8567	2441	5047
Burlington, NC	637	6179	994	707	1910	2519	131	701	5680
Burlington, VT	383	9673	1809	494	7146	3323	1741	3357	6726
Canton	575	7332	443	650	4279	2754	989	1157	4847
Casper	401	9839	345	989	4410	2453	303	1435	6303

Cedar Rapids	434	7774	837	714	5270	2619	904	1501	5009
Champaign-Urbana-Rantoul	525	8627	672	1022	7447	3147	2203	1700	5485
Charleston, SC	569	7402	1463	1495	4207	3164	2993	2561	5153
Charleston, WV	627	7789	708	721	5470	2894	2605	844	5257
Charlotte-Gastonia	644	7169	999	1273	6099	3031	4313	1236	5671
Charlottesville	618	9531	1348	756	6041	3489	1422	1704	6055
Chattanooga	576	6189	564	946	3401	2415	2483	1238	4487
Chicago	514	10913	5766	1034	7742	3486	24846	2856	5205
Chico	603	8587	243	947	4067	3126	1647	1543	5307
Cincinnati	584	8143	2138	978	5748	2918	9688	2451	5270
Clarksville-Hopkinsville	544	6007	446	736	2226	2654	111	2219	4880
Cleveland	579	9168	3167	1138	7333	2972	12679	3300	4879
Colorado Springs	526	8509	721	1086	3389	2754	1749	2375	7699
Columbia, MO	541	7702	1951	1065	3893	2377	2882	1331	5147
Columbia, SC	526	7519	1421	1524	5859	2908	2489	1484	5279
Columbus, GA-AL	517	5817	833	820	2995	2665	1861	1214	4812
Columbus, OH	558	8093	1837	1092	4364	2928	6648	2020	5165
Corpus Christi	362	6929	458	1335	3626	2840	1992	2037	6690
Cumberland	591	6054	760	337	3709	3363	373	1036	4741
Dallas	544	9318	2825	1529	6213	3269	10438	2310	7710
Danbury	569	14420	2350	548	2715	3029	7415	1572	7060
Danville	545	5709	593	379	3161	2943	85	501	4491
Davenport-Rock Island	440	8083	1113	834	3907	2901	1017	1920	4997
Dayton-Springfield	544	7635	2253	1151	4775	2772	6935	1122	4532
Daytona Beach	561	7203	723	1347	4117	2612	809	3967	6592
Decatur	480	7395	732	897	3867	2683	298	1222	4274
Denver	521	10789	2533	1365	8145	3145	8477	2324	7164
Des Moines	444	8028	1256	1044	5521	2613	1857	1802	5346
Detroit	536	8525	4142	1587	4808	3064	10389	2483	3904
Dothgan	336	5708	593	930	2232	2230	117	714	5453
Dubuque	419	7993	640	571	3668	2701	340	1587	3949
Duluth	193	6040	1159	488	5205	2619	2377	3107	3922
East St. Luis-Belleville	537	6501	444	1096	6539	2630	904	1610	6113
Eau Claire	257	7078	798	433	3197	2960	1807	1397	5348
El Paso	592	7343	528	1323	3705	2479	3800	1101	5080
Elkhart-Goshen	521	6573	596	524	4168	2537	353	1023	4214
Elmira	467	7078	562	582	3324	3000	1048	1600	4813
Enid	461	6829	626	845	2312	2764	215	1200	8268
Erie	605	7715	529	635	5754	2641	2032	1340	4299
Eugene-Springfield	741	9370	539	874	5293	3118	2631	3400	3045
Evansville	550	6743	783	864	3496	2797	1876	1622	5206
Fall River	643	9017	900	861	4602	2439	749	2005	4884
Fargo-Moorhead	148	8168	920	503	6325	2506	2111	1414	5594
Fayetteville	561	6274	872	1150	4402	3051	844	709	5255
Fayetteville-Sprindale	549	6686	594	545	3581	2334	1915	1695	4631
Fitchburg-Leominster	507	8252	655	655	2244	2799	270	790	5098
Flint	540	7204	724	1671	4912	2511	2163	1355	3724
Florence, AL	546	5962	607	516	1454	2427	1021	994	4492
Florence, SC	552	6508	818	1334	4963	3109	628	800	4842
Fort Collins	490	9951	731	744	2637	2413	1609	4200	6631
Fort Lauderdale-Hollywood	572	10810	1252	1536	4186	2734	2027	2455	7136
Fort Meyers	342	9298	546	787	4583	2729	380	4005	7166
Fort Pierce	602	8842	527	1422	2143	3154	368	2058	7973
Fort Smith	482	5784	466	663	3092	2927	145	1736	4849
Fort Walton Beach	536	7554	484	544	2886	2809	87	2092	6342
Forth Wayne	509	6733	1060	710	5416	2772	2846	1711	4195
Forth-Arlington	528	7956	1038	1348	4472	2627	6466	2366	6862
Fesno	559	9291	369	1483	4388	2407	3596	2984	5746
Gadsden	526	5382	622	749	2174	2299	153	300	4220
Gainesville	402	7388	1731	1658	3527	3094	3335	1237	5739
Galveston-Texas City	727	7767	1437	1213	3423	2809	1756	3000	6026
Gary-Hammond	483	7641	1364	996	5855	2526	4115	1940	3826
Glens Falls	476	7120	43	568	2241	2674	603	1883	5166
Grand Forks	105	7898	1109	401	5587	2721	1921	1304	5646
Grand Rapids	513	7780	1274	952	3454	2705	3255	1909	4848
Great Falls	410	7143	667	792	3747	2737	401	2176	4697
Greeley	490	8218	706	994	1641	2854	1254	739	5443

Green Bay	367	8401	916	583	3793	2622	2547	1925	5650
Greensboro-Winston-Salem	626	7064	1694	967	4453	3090	4188	1651	5204
Greenville-Spartanburg	655	6336	1260	1185	3950	3236	2569	1410	5012
Hagerstown	568	7763	818	627	3431	2990	825	1491	4477
Hamilton-Middletown	542	8227	1135	892	3338	2747	2316	1604	4618
Harrisburg-Lebanon-Carlisle	556	7891	2087	629	6164	3224	3083	1532	5322
Hartford	516	11652	2521	1279	7120	3628	3616	1790	6307
Hickory	623	6760	1006	765	2703	2726	188	797	4728
Honolulu	717	17021	1298	891	5911	2502	7168	3703	5187
Houma-Thibodaux	427	7094	583	400	1145	1995	725	2700	6662
Houston	424	9760	2467	1499	4626	3271	11073	1825	7464
Huntington-Ashland	636	6632	875	665	4001	2525	2195	840	4383
Huntsville	600	6283	685	924	2661	2257	1921	1075	6412
Indianapolis	557	7012	2243	1000	5804	2690	6348	1906	5082
Iowa City	434	9429	2437	830	2770	2842	2255	1506	5165
Jackson, MI	518	6794	679	1021	4800	2654	323	1933	3822
Jackson, MS	412	7245	1792	1091	4917	3130	3209	1427	5186
Jacksonville, FL	457	6626	1181	1211	5611	3048	2162	2884	6139
Jacksonville, NC	564	6111	740	967	1780	2646	567	1177	6386
Janesville-Beloit	466	7447	700	858	3092	2532	1092	1615	4371
Jersey City	601	8810	1759	1434	4982	2574	7420	1001	4889
Johnson City-Kingsport	663	6119	1152	424	2532	2925	1925	2155	4903
Johnstown	547	6524	731	353	4343	2691	666	903	4181
Joliet	479	9327	1058	837	4645	2868	3177	1636	4631
Joplin	580	5159	500	628	2335	2871	80	801	5324
Kalamazoo	527	7919	1043	1120	5419	2896	2071	2163	4794
Kankakee	483	7230	609	976	3444	2855	75	1119	5579
Kansas City, KS	549	8126	1711	1142	5006	3028	1167	1045	6166
Kansas City, MO	549	7076	1939	1468	5869	2949	5553	2043	4865
Kenosha	496	8516	1067	911	4473	2918	3844	2224	5176
Kileen-Temple	365	6463	398	733	2862	3167	920	931	6331
Knoxville	670	6692	960	622	4273	2761	3309	2514	5537
Kokomo	512	6616	596	413	1817	2904	285	1000	3429
La Crosse	352	8310	686	676	6096	3027	1466	1953	5648
Lafayette, IN	494	7778	655	465	4956	2945	2235	1814	4333
Lafayette, LA	429	8572	548	1030	5268	2305	1772	1734	9702
Lake Charles	469	6921	314	1093	3549	2336	1456	1855	5872
Lake County	514	13282	1237	822	3422	2607	3746	2435	5755
Lakeland-Winter Haven	307	6680	323	1373	3412	2998	309	2513	5594
Lancaster	559	8631	1111	413	3908	3097	1015	1147	5120
Lansing-East Lansing	480	7907	1371	894	5557	2891	4206	1609	4747
Laredo	424	6152	465	1050	3322	2827	150	702	5264
Las Cruces	552	6962	588	1457	2989	2736	2804	1609	5341
Las Vegas	556	9906	412	1913	5900	2241	1586	3996	6035
Lawrence	513	7497	621	1018	2931	2700	3150	1752	4573
Lawrence-Haverhill	548	10414	1202	909	3575	2479	2111	1879	6527
Lawton	479	5850	477	1156	2366	2375	1280	1757	6105
Lewiston-Auburn	490	6876	759	764	2941	2694	736	1853	4444
Lexington-Fayette	635	8340	1860	1055	4080	2861	3596	1403	6245
Lima	522	6986	741	855	4084	2629	1352	1428	4313
Lincoln	398	8256	775	789	5618	2878	4523	1804	4908
Little Rock	497	7270	1861	1328	4186	2581	2180	1462	5273
Longview-Marshall	500	6608	509	976	2680	2816	334	834	6898
Lorain-Elyria	579	8309	1105	609	3629	2582	2565	1602	3301
Los Angeles, Long Beach	885	13868	5153	1960	4345	3195	23567	3948	5316
Louisville	616	6812	2111	937	5420	3028	4916	1942	5402
Lowell	526	9640	1083	819	3820	2479	3057	1129	6651
Lubbock	604	6990	900	1608	4158	2545	3402	1702	5923
Lynchburg	642	6934	732	643	4909	2803	1079	1439	4926
Macon, Warner Robbins	447	6235	593	783	3144	2651	1435	1204	5659
Madison	378	9897	2168	779	6084	3047	5123	1944	5448
Manchester	404	9860	737	633	4595	2728	475	837	7101
Mansfield	558	6881	303	1072	2876	2871	554	1133	4386
McAllen-Edinburg-Mission	238	5345	372	836	2117	2644	1231	1059	5739
Medford	611	9008	256	728	3512	2797	1856	1416	3692
Melbourne-Titusville-Palm Bay	582	8721	517	1039	2560	2814	437	3800	7089
Memphis	514	7015	2043	1488	6247	2804	4486	1994	5160

Miami-Hialeah	634	10267	2314	2459	5202	2879	4837	4300	5840
Middlesex-Somerset, Hunterdon	559	12135	2589	691	4198	3539	8058	1596	6324
Middletown	593	11652	884	646	4636	3128	730	1682	6307
Midland	603	8672	97	1166	5310	2416	438	1502	9980
Milwaukee	460	10176	3053	826	4945	3044	8766	2902	4982
Minneapolis-St. Paul	293	9559	3934	906	5606	3013	11714	2158	5843
Mobile	442	6704	1469	1511	3345	2779	1764	2164	4565
Modesto	639	8630	347	1154	2000	2616	631	833	5107
Monmouth-Ocean	615	11660	2482	819	2690	2787	7563	3544	6154
Monroe	455	6190	331	957	3606	2453	1528	1541	5537
Montgomery	483	6754	832	815	3509	2388	1374	1112	4892
Muncie	530	5800	949	783	4325	2965	2498	1428	3980
Muskegon	580	6391	699	1537	3353	2630	529	2666	3708
Nashua	538	10757	853	452	3320	2728	2122	1523	6962
Nashville	600	7800	1850	984	5030	2763	4342	1849	5938
Nassua-Suffolk	656	11138	3919	566	2119	3234	8640	3705	7371
New Bedford	643	8087	519	1012	3219	2439	766	1450	4937
New Britain	516	10509	1245	903	4900	3128	1360	1217	6470
New Haven-Meriden	583	11460	2068	893	5938	3495	7852	1604	5478
New London-Norwich	583	10218	556	633	4505	3244	1164	2281	6672
New Orleans	498	8515	2586	1604	4579	2995	7978	3500	6453
New York	638	13358	7850	2498	8625	2984	56745	3579	5338
Newark	601	14220	4106	1461	3514	3362	14224	1818	5690
Niagara Falls	554	7686	507	775	5561	2538	966	1873	4463
Norfolk-Virginia Beach	632	8568	1932	997	3215	3014	7087	2964	5866
Norwalk	648	20151	2530	625	3536	3029	7273	2268	6432
Oakland	910	13135	2362	1533	6430	2646	6162	2394	5457
Ocala	333	6750	489	1327	3798	2864	266	3095	7060
Odessa	603	6689	384	1698	1944	2721	259	900	7565
Oklahoma City	554	7186	1623	1297	4459	2908	4843	1742	8119
Olympia	726	8263	338	752	4083	2625	708	1451	4912
Omaha	440	7128	2559	1008	5806	3069	3787	1977	5853
Orange County	509	10173	1574	804	3299	2754	6987	1866	5822
Orlando	457	8196	765	1671	5887	2976	2681	2881	7413
Owensboro	524	6760	812	568	2273	2491	804	852	6062
Oxnard-Ventura	890	14000	1106	791	2238	2155	2769	2135	5514
Panama City	536	6373	201	1344	2778	2500	755	2089	6083
Parkerburg-Marietta	617	6657	665	488	4399	2503	91	1148	5187
Pascagoula	584	6248	593	591	1750	1701	155	1956	4491
Pawtucket-Woonsocket-Attleboro	586	9462	1117	744	4738	3058	480	1513	5154
Pensacola	536	6479	563	1472	2918	2914	1954	2160	6029
Peoria	491	8388	1184	921	2967	2452	2294	1688	4073
Philadelphia	630	8310	5158	1059	5903	3781	17270	1979	5638
Phoenix	536	8921	1584	1268	4729	2942	4573	2472	6415
Pine Bluff	463	5674	617	1169	1671	2554	373	793	4247
Pittsburgh	586	8099	3413	687	5616	3544	11069	2145	5261
Pittsfield	482	7807	694	638	3759	3264	228	1420	5483
Portland, ME	483	8100	834	823	5185	2973	2351	3366	6186
Portland, OR	768	9912	1590	1504	5947	3343	5160	2532	4535
Portsmouth-Dover-Rochester	469	9966	596	475	2321	3026	165	1390	8367
Poughkeepsie	488	9981	355	633	4166	2898	785	1670	6746
Providence	586	9274	2467	998	5474	3558	6152	2263	5154
Provo-Orem	500	9321	198	485	4546	2618	1985	3300	3459
Pueblo	497	6637	468	1181	3501	2653	1307	1619	4646
Racine	496	8943	931	1055	3558	2732	1171	2016	4415
Raleigh-Durham	647	8230	3476	981	6544	3455	5730	1606	6405
Reading	614	7614	1154	522	3120	3028	1108	1549	5587
Redding	664	8584	274	892	5727	2471	845	2424	4459
Reno	535	12449	615	1116	6767	2529	2210	2386	5677
Richland-Kinnewick-Pasco	664	8461	300	779	4714	1728	529	1204	5326
Richmond-Petersburg	585	8343	2448	1076	6680	2940	5697	1943	5870
Riverside-San Bernardino	615	9754	2201	1475	3141	2596	5327	1918	4923
Roanoke	652	7476	1036	784	3872	2723	1263	2036	5287
Rochester, MN	308	9193	2966	437	4399	2134	769	1503	6099
Rochester, NY	536	8609	1969	894	5165	3582	6956	2659	6304
Rockford	466	7584	969	1156	2987	2680	1026	933	4592
Sacramento	576	9855	1027	1363	5097	2793	4483	2306	5309

Saginaw-Bay City-Midland	515	7368	1022	1068	3186	2772	1708	2059	3709
St. Cloud	195	7235	603	343	4565	2502	1871	1572	5488
St. Joseph	475	5589	223	969	2689	2927	879	1265	5991
St. Louis	537	7605	2850	1306	7119	3530	8896	2243	5800
Salem	716	8378	749	1014	4732	3278	691	1873	3835
Salem-Glouster	644	11622	1232	490	3459	2729	3276	2234	6309
Salinas-Seaside-Monterey	843	13838	352	1107	4160	2439	1004	3179	5656
Salt Lake City-Ogden	541	9466	1631	969	6228	2340	5528	3900	4942
San Angelo	488	6321	236	1032	2938	2707	1301	1136	7720
San Antonio	398	6898	1337	1197	5387	2938	4295	1509	6873
San Diego	903	14465	2416	1099	5489	2794	8818	3347	5489
San Francisco	910	17158	3726	1619	8299	3371	14226	4600	6063
San Jose	850	16048	2117	1065	5224	2709	6446	1964	7270
Santa Barbara-Santa Maria	855	15547	532	1026	5662	2719	2684	3300	5821
Santa Cruz	843	14303	1035	964	5010	2611	3748	1703	5335
Santa Rosa-Petaluma	732	12931	1052	912	3313	2722	3457	2255	5703
Sarasota	391	9560	801	939	3742	2626	817	2535	7715
Savannah	542	6896	1084	1614	5958	2456	2262	2237	5591
Scranton-Wilkes Barre	575	6697	1219	372	3683	3230	1832	1386	4907
Seattle	808	10183	2715	1170	6634	2710	9577	4800	5901
Sharon	570	6697	700	384	2017	3022	52	1100	4055
Sheboygan	442	8121	593	450	3458	2557	268	1316	4765
Sherman-Denison	524	5722	394	1035	1922	2652	68	937	6213
Shreveport	508	6534	1445	1197	4401	2858	2826	1389	6585
Sioux City	385	6528	846	759	4316	2673	1393	1359	4648
Sioux Falls	276	7983	1041	556	6271	2651	465	1324	5204
South Bend-Mishawaka	545	5938	830	1038	5634	2874	2672	1819	5056
Spokane	574	6927	497	869	5534	2774	2988	1517	4722
Springfield, IL	524	7882	1877	1225	6172	3078	1983	1536	5384
Springfield, MA	453	8039	710	1212	6159	3525	3466	1514	5289
Springfield, MO	544	6343	577	892	3828	2709	1634	1737	5932
Stamford	648	23640	2610	835	3110	3029	7865	1729	6158
State College	575	8405	612	540	2740	3169	1271	1200	4677
Steubenville-Weirton	542	6578	505	418	1532	2672	147	1460	3744
Stockton	625	8474	342	1395	4427	2155	1579	1630	5672
Syracuse	548	7670	1040	689	6951	3144	5080	2851	4474
Tacoma	808	7770	539	1162	4730	2546	4297	4000	4887
Tallahassee	404	8029	370	1161	5530	2790	2181	1936	6021
Tampa-St. Petersburg	440	7442	1189	1493	5588	3044	5040	2943	7256
Terre Haute	557	5527	453	630	3550	3012	1226	1401	4353
Texarkana	467	5717	343	822	2537	2899	63	669	4772
Toledo	518	7767	1738	998	5323	2852	4389	1952	4534
Topeka	501	7110	1148	999	5348	2795	1632	1141	5464
Trenton	636	10616	1372	1181	4786	3311	5029	1646	5772
Tuscon	589	8548	1259	1400	4397	2685	4889	3131	6147
Tulsa	530	7498	1581	1080	3758	2628	4248	2024	7115
Tuscaloosa	470	6464	674	1014	4723	2390	1432	1090	4900
Tyler	500	7298	672	955	3460	3283	404	631	7327
Utica-Rome	548	6744	391	400	4592	2970	858	1750	5226
Vallejo-Fairfield-Napa	821	10503	1079	964	4153	2498	2962	1559	5819
Vancouver	768	9015	517	752	3817	2332	1557	1464	4571
Victoria	336	7143	260	1092	2407	2696	87	1410	7599
Vineland-Millville-Bridgeton	615	7295	807	1135	4133	2747	2097	1474	5023
Visalia-Tulare-Porterville	543	7778	210	1132	3094	2128	511	2800	5563
Waco	412	6106	538	1166	3018	2867	1141	1248	6259
Washington	631	13724	4361	1317	8236	3635	21701	1578	6072
Waterbury	569	10024	1218	789	2434	2995	318	946	5656
Waterloo-Cedar Falls	347	7881	925	700	3351	2889	3000	1900	4407
Wasau	308	7642	818	442	3496	2749	761	1654	4300
West Palm-Boca Raton-Delray	509	10512	375	1783	5201	3224	2888	3772	7992
Wheeling	542	6576	791	308	2450	3002	422	1271	4740
Wichita	494	7061	806	1164	3933	2981	2987	1508	6036
Wichita Falls	456	6404	549	1179	2793	2747	599	1126	6805
Williamsport	558	7284	860	464	3097	2906	196	726	3288
Wilmington, DE-NJ-MD	597	7927	1445	1115	4532	3112	4545	1923	6174
Wilmington, NC	564	6858	1099	1423	2904	2876	1077	2668	5390
Worcester	562	8715	1805	680	3643	3299	1784	910	5040

Yakima	535	6440	317	1106	3731	2491	996	2140	4986
York	540	8371	713	440	2267	2903	1022	842	4946
Youngstown-Warren	570	7021	1097	938	3374	2920	2797	1327	3894
Yuba Cityt	608	7875	212	1179	2768	2387	122	918	4694

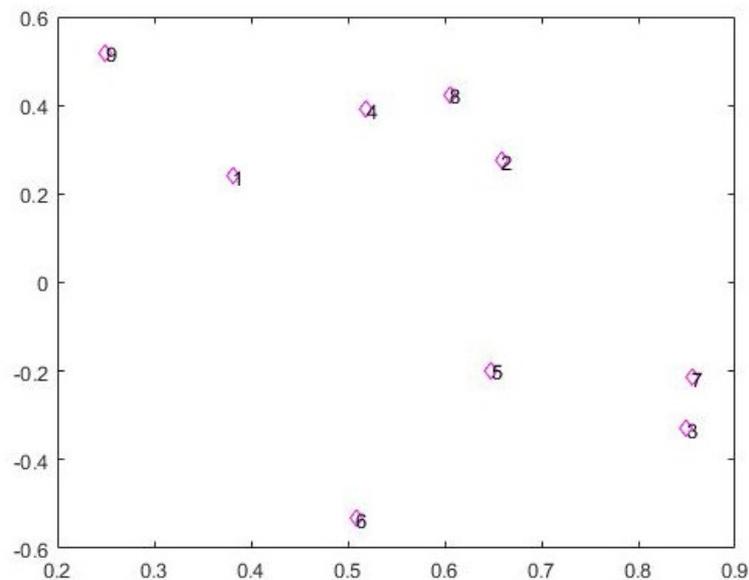
Tabella 2.5: Vivibilità nelle città degli USA

## 2.5 Vivibilità nelle città degli USA

Utilizziamo la stessa funzione scritta per il primo dataset per stimare i carichi e vedere le relazioni che ci sono tra le varianze dei 9 criteri. In questo caso dovremo lavorare con un numero di autovalori più grande per avere dei risultati accettabili in termine di varianza spiegata.

Iniziamo comunque con il caso  $m=2$  che, come spiegato precedentemente, ci può già dire qualcosa perchè possiamo rappresentare i dati anche graficamente:

Criteria	Carichi stimati	
1.Clima	0.381	0.240
2.Alloggi	0.658	0.276
3.Salute	0.850	-0.330
4.Criminalità	0.519	0.392
5.Trasporti	0.648	-0.198
6.Educazione	0.508	-0.533
7.Arte	0.855	-0.215
8.Divertimento	0.605	0.424
9.Economia	0.250	0.519
<b>Porzione di varianza spiegata</b>	0.514	

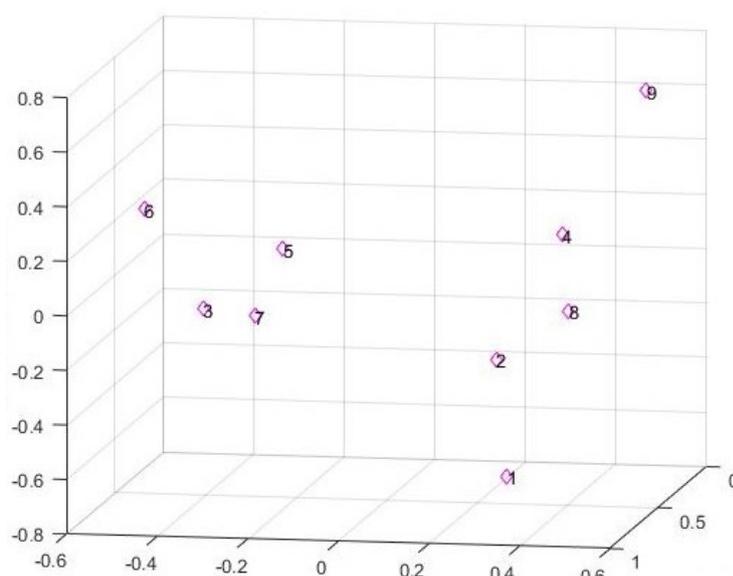


Dalla tabella e dal grafico possiamo notare come il primo fattore non differenzi particolarmente tra loro i diversi criteri, sono infatti tutti positivi e non c'è una differenza sostanziale tra il maggiore e il minore.

Con il secondo fattore vediamo già qualche differenza in più, abbiamo in particolare la distinzione tra i criteri che hanno un carico positivo: clima, alloggi, criminalità, divertimento ed economia; e quelli che hanno un carico negativo: salute, trasporti, educazione e arte. Nel secondo gruppo sono compresi i criteri che indicano tutti i servizi che una città può offrire: quelli ospedalieri come ambulatori pubblici e privati, quelli di trasporto come autostrade, ferrovie e stazione aeroportuali, quelli educativi come scuole e istituti di altro tipo e quelli artistici come teatri, musei e librerie. Possiamo quindi considerare questo come il *fattore dei servizi*.

Vediamo però che con 2 soli fattori la parte di varianza spiegata è pari solo al 51.4%, per avere una precisione più alta possiamo applicare l'analisi nel caso di  $m=3$  e vedere il grafico 3D che ne risulta:

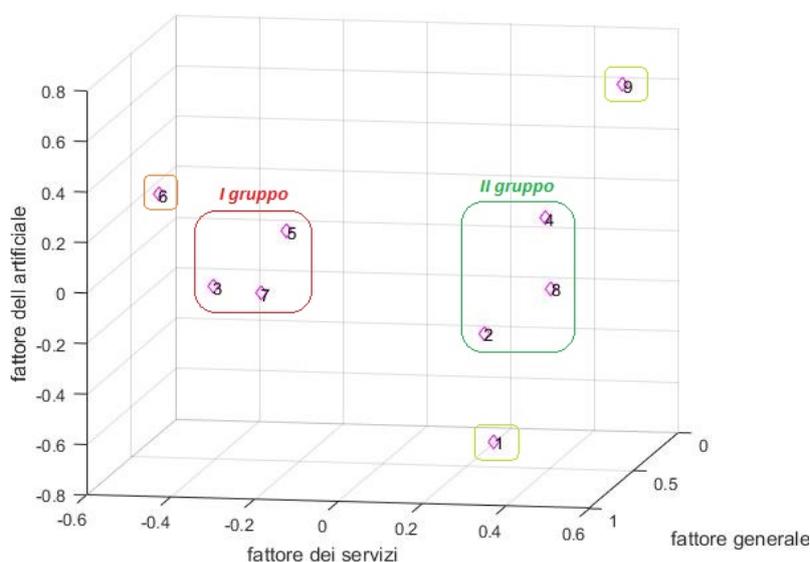
<b>Criteri</b>	<b>Carichi stimati</b>		
1.Clima	0.381	0.240	-0.737
2.Alloggi	0.658	0.276	-0.222
3.Salute	0.850	-0.330	-0.008
4.Criminalità	0.519	0.392	0.198
5.Trasporti	0.648	-0.198	0.156
6.Educazione	0.508	-0.533	0.245
7.Arte	0.855	-0.215	-0.028
8.Divertimento	0.605	0.424	-0.0543
9.Economia	0.250	0.519	0.649
<b>Porzione di varianza spiegata</b>	0.640		



Il terzo fattore è di difficile interpretazione, vediamo che ci sono due valori con un valore rispettivamente molto basso e molto alto, ossia clima ed economia, mentre tutti gli altri si concentrano in valori prossimi allo zero senza particolari differenze. Possiamo considerarlo come un fattore che distingue ciò che riguarda la natura, come il clima, e ciò che è artificiale, come l'economia. In posizione centrale si trovano quei criteri che non sono riconducibili esclusivamente alla natura o all'artificiale, come i trasporti, gli alloggi o la salute, che oltre a descrivere i servizi ospedalieri descrive la qualità dell'aria e dell'acqua della città. Lo chiameremo quindi il *fattore dell'artificiale*.

Quello che osserviamo dal grafico invece ci aiuta a vedere quali sono i gruppi che si formano in seguito all'applicazione dell'analisi; rispetto al primo esempio in questo caso la rappresentazione è necessaria perché numericamente è impossibile individuarli e non sono così netti da poter essere definiti in maniera chiara. Vediamo però che si formano due gruppi principali, uno formato da clima trasporti e arte (3.,5.,7.) e uno da alloggi criminalità e divertimento (2.,4.,8.), i restanti tre criteri sono più isolati ma osserviamo comunque che l'educazione (6.) ha caratteristiche più simili al primo gruppo, mentre il clima (1.) e l'economia (9.) più simili al secondo.

L'applicazione dell'analisi fattoriale a questo esempio sembra quindi indicare che i punteggi che assumono i criteri siano essenzialmente determinati da un fattore generale (ossia in cui tutti hanno un valore alto), dai servizi che offrono e dalla presenza di un azione artificiale in essi.

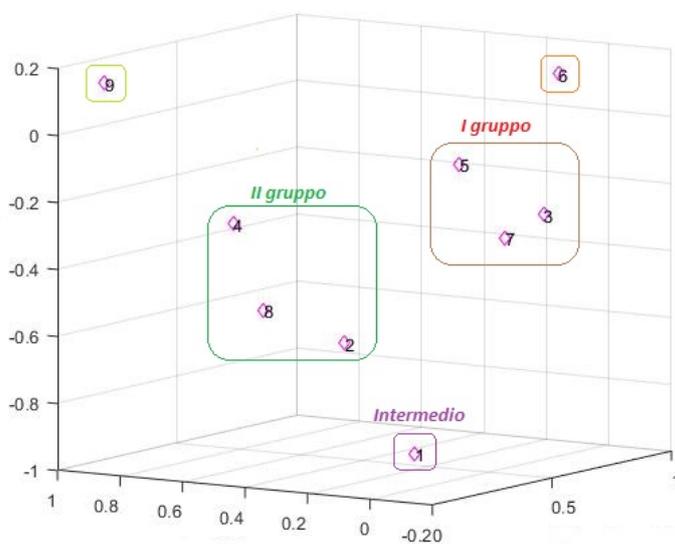


Vediamo ora cosa succede se applichiamo la rotazione dei fattori a questo secondo esempio, utilizzando la funzione `rotatefactors` al caso  $m=3$ :

Criteria	Carichi stimati			Carichi ruotati		
1.Clima	0.381	0.240	-0.737	0.012	-0.134	<b>-0.853</b>
2.Alloggi	0.658	0.276	-0.222	0.301	0.313	<b>-0.609</b>
3.Salute	0.850	-0.330	-0.008	<b>0.856</b>	0.099	-0.297
4.Criminalità	0.519	0.392	0.198	0.203	<b>0.594</b>	-0.262
5.Trasporti	0.648	-0.198	0.156	<b>0.651</b>	0.212	-0.123
6.Educazione	0.508	-0.533	0.245	<b>0.761</b>	-0.020	0.149
7.Arte	0.855	-0.215	-0.028	<b>0.787</b>	0.169	-0.361
8.Divertimento	0.605	0.424	-0.0543	0.203	<b>0.496</b>	<b>-0.512</b>
9.Economia	0.250	0.519	0.649	0.002	<b>0.851</b>	0.169

Quello che notiamo è che rispetto al primo fattore i criteri 3.,5.,7. e 6. hanno un carico alto; rispetto al secondo fattore sono i criteri 2.,4.,8. e 9. ad avere carico alto; nella terza colonna infine vediamo che i carichi si dispongono in maniera circa uniforme nell'intervallo tra -0.123 e -0.608 poi ci sono i picchi negativi del fattore 1. (-0.853) e positivi dei fattori 6. (0.149) e 9. (0.169).

Per trarre le conclusioni ci facciamo aiutare anche dalla rappresentazione grafica che sembra comunque confermare la divisione nei due gruppi che avevamo fatto precedentemente, si vedono anche le caratteristiche del criterio 6., simili rispetto al primo gruppo e del 9. simili rispetto al secondo; quello che è meno evidente dopo la rotazione è la posizione del criterio 1. che sembra avere caratteristiche più intermedie.



L'utilità della rotazione dei fattori la capiamo ancora di più se applichiamo l'analisi con  $m=4$ . Vediamo come sono i carichi stimati e quelli ruotati:

Criteri	Carichi stimati				Carichi ruotati			
1.Clima	0.381	0.240	-0.737	0.132	0.012	-0.134	<b>-0.853</b>	-0.215
2.Alloggi	0.658	0.276	-0.222	0.491	0.301	0.313	<b>-0.609</b>	<b>0.384</b>
3.Salute	0.850	-0.330	-0.008	0.014	<b>0.856</b>	0.099	-0.297	-0.013
4.Criminalità	0.519	0.392	0.198	-0.517	0.203	<b>0.594</b>	-0.262	<b>0.111</b>
5.Trasporti	0.648	-0.198	0.156	-0.291	<b>0.651</b>	0.212	-0.123	-0.51
6.Educazione	0.508	-0.533	0.245	0.322	<b>0.761</b>	-0.020	0.149	<b>0.168</b>
7.Arte	0.855	-0.215	-0.028	-0.097	<b>0.787</b>	0.169	-0.361	-0.032
8.Divertimento	0.605	0.424	-0.0543	-0.182	0.203	<b>0.496</b>	<b>-0.512</b>	<b>0.171</b>
9.Economia	0.250	0.519	0.649	0.405	0.002	<b>0.851</b>	0.169	<b>0.941</b>
<b>Porzione di varianza spiegata</b>	0.743							

Il quarto fattore non ruotato sembrerebbe smentire le osservazioni sui due gruppi trovati precedentemente, infatti i criteri 3,5,7. e 2,4,8. non hanno carichi simili tra loro. Dopo la rotazione però vediamo proprio come i criteri del primo gruppo sono quelli che hanno carico negativo, mentre quelli del secondo gruppo sono quelli che hanno carico positivo. L'economia (9.) ha ancora un picco rispetto agli altri, sempre nella direzione del secondo gruppo, mentre il clima (1.) e l'educazione (6.) hanno caratteristiche intermedie. Con questo ultimo esempio abbiamo raggiunto una precisione molto soddisfacente, il 74,3%, e possiamo quindi dire che le conclusioni tratte siano attendibili. In particolare abbiamo raggiunto lo scopo di trovare delle particolari relazioni di covarianza tra due diversi gruppi: il primo formato dai criteri della salute, dei trasporti, dell'arte e in parte dell'educazione, il secondo formato dai criteri degli alloggi, della criminalità, del divertimento e in parte dell'economia; ed un ultimo criterio intermedio: il clima.

# Bibliografia

- [1] RICHARD A. JOHNSON, DEAN W. WICHERN. *Applied Multivariate Statistical Analysis*. Prentice Hall, 2001.
- [2] DAVIDE PALITTA, VALERIA SIMONCINI. *Dispense del corso di Calcolo Numerico, Modulo di Algebra Lineare Numerica*. 2016.
- [3] RICHARD BOYER, DAVID SAVAGEAU copyrighted and published by Rand McNally. *Places Rated Almanac*.
- [4] VALERIANO COMINCIOLI. *Analisi Numerica, Metodi Modelli Applicazioni*. McGraw-Hill, 1995.