Analisi Fattoriale - Applicazioni

Analisi Esplorativa

Aldo Solari

1. Dati Esami

2. Stock-Price Data

Outline

Dati Esami

- Voto agli esami
- $n = 202$ studenti maschi
- $p = 6$

Variabili:

- Gaelic (non-math)
- English (non-math)
- History (non-math)
- Arithmetic (math)
- Algebra (math)
- Geometry (math)

Dati Esami: Correlazione

$$
\begin{array}{cc}
 & \begin{array}{cccccc} \text{Gaelic} & \text{English} & \text{History} & \text{Arithmetic} & \text{Algebra} & \text{Geometry} \end{array} \\
\mathbf{R} = & \begin{bmatrix} 1.0 & .439 & .410 & .288 & .329 & .248 \\ & 1.0 & .351 & .354 & .320 & .329 \\ & & 1.0 & .164 & .190 & .181 \\ & & & 1.0 & .595 & .470 \\ & & & & 1.0 & .464 \\ & & & & & 1.0 \end{bmatrix}
\end{array}
$$

Assunzioni

Se assumiamo che f e u hanno distribuzione congiunta Gaussiana con

$$\underset{k\times 1}{f} \sim \mathcal{N}_k(\underset{k\times 1}{0}, \underset{k\times k}{I}), \quad \underset{p\times 1}{u} \sim \mathcal{N}_p(\underset{p\times 1}{0}, \underset{p\times p}{\Psi})$$

allora x segue una distribuzione Gaussiana

$$\underset{p\times 1}{x} = \mu + \Lambda f + u \sim N_p(\underset{p\times 1}{\mu}, \underset{p\times p}{\Lambda\Lambda'} + \underset{p\times p}{\Psi})$$

Si noti che abbiamo rilassato l'assunzione $\mathbb{E}(x) = 0$, che equivale a $\mu = 0$

Funzione di log-verosimiglianza

$$
\begin{aligned}
\ell(X;\mu,\Sigma) &= -\frac{1}{2}n\log|2\pi\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)\Sigma^{-1}(x_i-\mu)' \\
&= -\frac{1}{2}n\log|2\pi\Sigma| - \frac{1}{2}n\mathrm{tr}(\Sigma^{-1}S) - \frac{1}{2}n(\bar{x}-\mu)\Sigma^{-1}(\bar{x}-\mu)'
\end{aligned}
$$

Sostituendo $\hat{\mu}=\bar{x}$

$$\ell(X;\hat{\mu},\Sigma) = -\frac{n}{2}\Big\{\log|2\pi\Sigma| + \mathrm{tr}(\Sigma^{-1}S)\Big\}$$

e per $\Sigma = \Lambda\Lambda' + \Psi$ otteniamo

$$\ell(X;\hat{\mu},\Lambda,\Psi) = -\frac{n}{2}\Big\{\log|2\pi(\Lambda\Lambda'+\Psi)| + \mathrm{tr}[(\Lambda\Lambda'+\Psi)^{-1}S]\Big\}$$

Stima di massima verosimiglianza

Massimizzare

$$\ell(X;\hat{\mu},\Lambda,\Psi) = -\frac{n}{2}\Big\{\log|2\pi(\Lambda\Lambda' + \Psi)| + \mathrm{tr}[(\Lambda\Lambda' + \Psi)^{-1}S]\Big\}$$

rispetto a Ψ e Λ

Stima iterativa

1. Per Ψ fissato, massimizza numericamente per Λ
2. Per Λ fissato, massimizza numericamente per Ψ

Commenti:

- Implementata nella funzione R `factanal()`
- Si possono ottenere casi di Heywood

Dati Esami: FA con $k = 2$ e stima di MV

Table 9.5			
Variable	Estimated factor loadings F_1	F_2	Communalities $\hat{h}_i^2$
1. Gaelic	.553	.429	.490
2. English	.568	.288	.406
3. History	.392	.450	.356
4. Arithmetic	.740	−.273	.623
5. Algebra	.724	−.211	.569
6. Geometry	.595	−.132	.372

- Stima di MV: $\hat{h}_1^2 = \hat{\lambda}_{11}^2 + \hat{\lambda}_{12}^2 = (0.553)^2 + (0.429)^2 \approx 0.490$
- Primo fattore: *intelligenza generale*
- Secondo fattore: *abilità matematica vs abilità verbale*

Rotazione dei pesi fattoriali

- Per la rotazione dei pesi fattoriali $\underset{p\times k}{\Lambda}$, dobbiamo cercare una matrice ortogonale $\underset{k\times k}{A}$ $(A'A = AA' = I)$ tale per cui i pesi fattoriali ruotati $\underset{p\times k}{\Lambda}^{*} = \underset{p\times k}{\Lambda}\underset{k\times k}{A}$ sono più facilmente interpretabili
- $\underset{2\times 2}{A} = \left[\begin{array}{cc} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{array}\right]$ rotazione oraria per $k = 2$
- Questo non cambia la soluzione del modello, solo la sua descrizione
- Situazione desiderata per i fini interpretativi:
  - i pesi fattoriali sono tutti grandi e positivi o prossimi a 0 (con pochi valori intermedi)
  - ogni variabile osservabile è legata in modo pesante al più ad un solo fattore
- Per $k > 2$ il metodo *varimax* identifica la rotazione massimizzando un'opportuna funzione dei pesi fattoriali ruotati che misura la variabilità dei pesi

Dati Esami: rotazione dei pesi fattoriali

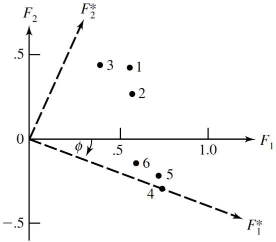

Figure 9.1 Factor rotation for test scores.

Table 9.6

Variable	Estimated rotated factor loadings F_1^*	F_2^*	Communalities $\hat{h}_i^{*2} = \hat{h}_i^2$
1. Gaelic	.369	$\boxed{.594}$	.490
2. English	.433	$\boxed{.467}$	.406
3. History	.211	$\boxed{.558}$	.356
4. Arithmetic	$\boxed{.789}$	.001	.623
5. Algebra	$\boxed{.752}$	.054	.568
6. Geometry	$\boxed{.604}$	.083	.372

- Primo fattore: *abilità matematica*
- Secondo fattore: *abilità verbale*

Verifica d'ipotesi sul numero di fattori

- Un vantaggio della stima di massima verosimiglianza e che permette un test di ipotesi sul numero di fattori
- Ipotesi nulla H_0: il modello fattoriale con k fattori è corretto

$$\Sigma = \underset{p\times k}{\Lambda}\underset{k\times p}{\Lambda}' + \Psi$$

- Ipotesi alternativa H_1: Σ è una matrice definitiva positiva diversa da quella specificata sotto l'ipotesi nulla
- Rifiuto l'ipotesi nulla con un p-value $\leq 5\%$
- Test sequenziali: parto da $k = 1$, se rifiuto proseguo con $k = 2, 3, \ldots$ fino a quando fallisco di rifiutare l'ipotesi

Test rapporto di verosimiglianza

- Siano $\hat{\Lambda}$ e $\hat{\Psi}$ le stime di massima verosimiglianza per il k specificato dall'ipotesi nulla
- La statistica test rapporto di verosimiglianza è data da

$$T = -2\log\Big(\frac{\text{MV sotto } H_0}{\text{MV}}\Big) = n\log\Big(\frac{|\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}|}{|S|}\Big)$$

e sotto H_0 segue asintoticamente una distribuzione

$$\chi^2_{\frac{1}{2}[(p-k)^2-p-k]}$$

- Il p-value del test si calcola come $\Pr(\chi^2_{\frac{1}{2}[(p-k)^2-p-k]} > t)$ dove t è il valore osservato della statistica test
- L'approssimazione χ^2 può essere migliorata utilizzando la statistica test con la correzione di Bartlett:

$$T_{Bartlett} = [(n-1) - (2p+4k+5)/6]\log\Big(\frac{|\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}|}{|S|}\Big)$$

Distribuzione di $f|x$

Se assumiamo che f e u hanno distribuzione congiunta Gaussiana con

$$\underset{k\times 1}{f} \sim \mathcal{N}_k(\underset{k\times 1}{0}, \underset{k\times k}{I}), \quad \underset{p\times 1}{u} \sim \mathcal{N}_p(\underset{p\times 1}{0}, \underset{p\times p}{\Psi})$$

allora la loro combinazione lineare x ha distribuzione gaussiana

$$\underset{p\times 1}{x} = \Lambda f + u \sim N_p(\underset{p\times 1}{0}, \underset{p\times p}{\Lambda\Lambda'} + \underset{p\times p}{\Psi})$$

e la distribuzione di f condizionata ad x è anch'essa Gaussiana:

$$\underset{k\times 1}{f} | \underset{p\times 1}{x} \sim \mathcal{N}_k(\Lambda'(\Lambda\Lambda' + \Psi)^{-1}x, I - \Lambda'(\Lambda\Lambda' + \Psi)^{-1}\Lambda)$$

Punteggi fattoriali

- I **punteggi fattoriali** $\underset{k\times 1}{\hat{f}} = (\hat{f}_1, \ldots, \hat{f}_k)'$ sono le "stime" delle variabili non osservabili $\underset{k\times 1}{f} = (f_1, \ldots, f_k)'$
- Nel **metodo di Thompson** i punteggi fattoriali corrispondono alla (stima della) media di $f|x$:

$$\widehat{\mathbb{E}(f|x)} = \hat{\Lambda}'(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi})^{-1}x$$

  In pratica si preferisce utilizzare S^{-1} (o R^{-1}) al posto di $(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi})^{-1}$
- Per l'i-sima osservazione x_i (centrata):

$$\hat{f}_i = \hat{\Lambda}' S^{-1} x_i$$

- Per l'i-sima osservazione standardizzata z_i

$$\hat{f}_i = \hat{\Lambda}' R^{-1} z_i$$

Punteggi fattoriali

- Nel **metodo di Bartlett** i punteggi fattoriali sono calcolati come soluzione al problema di minimizzazione

$$\hat{f} = \arg\min_{f} \, (x - \hat{\Lambda} f)' \hat{\Psi}^{-1} (x - \hat{\Lambda} f)$$

  dove $\hat{\Lambda}$ e $\hat{\Psi}$ sono le stime di MV (che soddisfano il Vincolo 1)

- Per l'i-sima osservazione x_i (centrata):

$$\hat{f}_i = (\hat{\Lambda}' \hat{\Psi}^{-1} \hat{\Lambda})^{-1} \hat{\Lambda}' \hat{\Psi}^{-1} x_i$$

- Per l'i-sima osservazione standardizzata z_i

$$\hat{f}_i = (\hat{\Lambda}' \hat{\Psi}^{-1} \hat{\Lambda})^{-1} \hat{\Lambda}' \hat{\Psi}^{-1} z_i$$

Outline

Table 8.4 Stock-Price Data (Weekly Rate Of Return)

Week	J P Morgan	Citibank	Wells Fargo	Royal Dutch Shell	Exxon Mobil
1	0.01303	−0.00784	−0.00319	−0.04477	0.00522
2	0.00849	0.01669	−0.00621	0.01196	0.01349
3	−0.01792	−0.00864	0.01004	0	−0.00614
4	0.02156	−0.00349	0.01744	−0.02859	−0.00695
5	0.01082	0.00372	−0.01013	0.02919	0.04098
6	0.01017	−0.01220	−0.00838	0.01371	0.00299
7	0.01113	0.02800	0.00807	0.03054	0.00323
8	0.04848	−0.00515	0.01825	0.00633	0.00768
9	−0.03449	−0.01380	−0.00805	−0.02990	−0.01081
10	−0.00466	0.02099	−0.00608	−0.02039	−0.01267
⋮	⋮	⋮	⋮	⋮	⋮
94	0.03732	0.03593	0.02528	0.05819	0.01697
95	0.02380	0.00311	−0.00688	0.01225	0.02817
96	0.02568	0.05253	0.04070	−0.03166	−0.01885
97	−0.00606	0.00863	0.00584	0.04456	0.03059
98	0.02174	0.02296	0.02920	0.00844	0.03193
99	0.00337	−0.01531	−0.02382	−0.00167	−0.01723
100	0.00336	0.00290	−0.00305	−0.00122	−0.00970
101	0.01701	0.00951	0.01820	−0.01618	−0.00756
102	0.01039	−0.00266	0.00443	−0.00248	−0.01645
103	−0.01279	−0.01437	−0.01874	−0.00498	−0.01637

Stock-Price Data

- Rendimento (settimanale) di cinque titoli
- Gen 04 - Dic 05
- $n = 103$
- $p = 5$

Variabili:

- JP Morgan (bank)
- Citibank (bank)
- Wells Fargo (bank)
- Royal Dutch Shell (oil)
- Exxon-Mobil (oil)

Stock-Price Data: correlazione

$$\bar{\mathbf{x}}' = [.0011, .0007, .0016, .0040, .0040]$$

$$\mathbf{R} = \begin{bmatrix} 1.000 & .632 & .511 & .115 & .155 \\ .632 & 1.000 & .574 & .322 & .213 \\ .511 & .574 & 1.000 & .183 & .146 \\ .115 & .322 & .183 & 1.000 & .683 \\ .155 & .213 & .146 & .683 & 1.000 \end{bmatrix}$$

Stock-Price Data: stima di MV

Table 9.8

Variable	Maximum likelihood estimates of factor loadings F_1	F_2	Rotated estimated factor loadings F_1^*	F_2^*	Specific variances $\hat{\psi}_i^2 = 1 - \hat{h}_i^2$
J P Morgan	.115	.755	.763	.024	.42
Citibank	.322	.788	.821	.227	.27
Wells Fargo	.182	.652	.669	.104	.54
Royal Dutch Shell	1.000	−.000	.118	.993	.00
ExxonMobil	.683	.032	.113	.675	.53
Cumulative proportion of total sample variance explained	.323	.647	.346	.647	

- Primo fattore (ruotato): *bank*
- Secondo fattore (ruotato): *oil*
- Proporzione di varianza spiegata dal j-mo fattore $= \frac{\sum_{i=1}^{p} \hat{\lambda}_{ij}^2}{p}$

Stock-Price Data: residui

Massima Verosimiglianza

$$\mathbf{R} - \hat{\mathbf{L}}\hat{\mathbf{L}}' - \hat{\mathbf{\Psi}} = \begin{bmatrix} 0 & .001 & -.002 & .000 & .052 \\ .001 & 0 & .002 & .000 & -.033 \\ -.002 & .002 & 0 & .000 & .001 \\ .000 & .000 & .000 & 0 & .000 \\ .052 & -.033 & .001 & .000 & 0 \end{bmatrix}$$

Stock-Price Data: test di $k = 2$

Ipotesi nulla

$$H_0 : \Sigma = \Lambda\Lambda' + \Psi \quad (k = 2)$$

Statistica test

$$n \ln \frac{|\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}_z|}{|R|} = n \ln \frac{0.17898}{0.17519} = n \ln 1.0216$$

p**-value**

$$\mathbb{P}(\chi^2_1 > n \ln(1.0216)) \approx 0.138 > 5\%$$

(p-value $= 0.15$ utilizzando la correzione di Bartlett)

Stock-Price Data: punteggi fattoriali

Decomposizione di R

$$\hat{\mathbf{L}}_{\mathbf{z}}^{*} = \begin{bmatrix} .763 & .024 \\ .821 & .227 \\ .669 & .104 \\ .118 & .993 \\ .113 & .675 \end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{\Psi}}_{\mathbf{z}} = \begin{bmatrix} .42 & 0 & 0 & 0 & 0 \\ 0 & .27 & 0 & 0 & 0 \\ 0 & 0 & .54 & 0 & 0 \\ 0 & 0 & 0 & .00 & 0 \\ 0 & 0 & 0 & 0 & .53 \end{bmatrix}$$

Una osservazione

$$\mathbf{z}' = [.50, -1.40, -.20, -.70, 1.40]$$

Metodo di Bartlett

$$\hat{\mathbf{f}} = (\hat{\mathbf{L}}_{\mathbf{z}}^{*\prime}\hat{\boldsymbol{\Psi}}_{\mathbf{z}}^{-1}\hat{\mathbf{L}}_{\mathbf{z}}^{*})^{-1}\hat{\mathbf{L}}_{\mathbf{z}}^{*\prime}\hat{\boldsymbol{\Psi}}_{\mathbf{z}}^{-1}\mathbf{z} = \begin{bmatrix} -.61 \\ -.61 \end{bmatrix}$$

Metodo di Thompson

$$\hat{\mathbf{f}} = \hat{\mathbf{L}}_{\mathbf{z}}^{*\prime}\mathbf{R}^{-1}\mathbf{z} = \begin{bmatrix} .331 & .526 & .221 & -.137 & .011 \\ -.040 & -.063 & -.026 & 1.023 & -.001 \end{bmatrix} \begin{bmatrix} .50 \\ -1.40 \\ -.20 \\ -.70 \\ 1.40 \end{bmatrix} = \begin{bmatrix} -.50 \\ -.64 \end{bmatrix}$$

Stock-Price Data: punteggi fattoriali

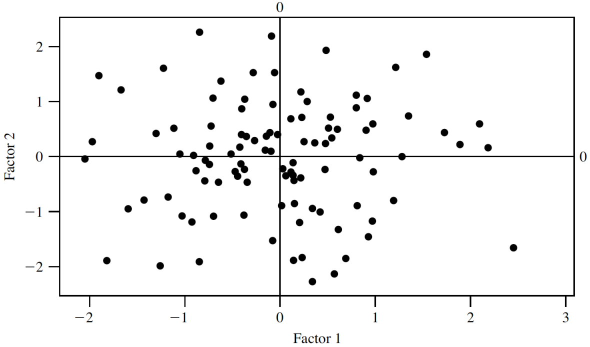

Figure 9.4 Factor scores using (9-58) for factors 1 and 2 of the stock-price data (maximum likelihood estimates of the factor loadings).