

L'Analisi in Componenti Principali

(Caso semplice: dati centrati o standardizzati)

Analisi dei Dati¹

¹ Corso di Laurea in Scienze Statistiche e Attuariali
Dipartimento di Diritto, Economia, Management e Metodi Quantitativi (DEMM)
Università degli Studi del Sannio

Prof. Pietro Amenta

Fonte: *Pietro Amenta. Appunti di Analisi dei Dati Multidimensionali*

- Sia \mathbf{Y} la matrice dei dati da analizzare
- I pesi sono raccolti nella matrice diagonale \mathbf{D}

$$\mathbf{D} = \begin{pmatrix} p_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & p_n \end{pmatrix}$$

- I pesi hanno quindi la caratteristica che $\sum_{i=1}^n p_i = 1$.
- Abbiamo già visto che quando le unità sono costituite da osservazioni casuali o quando non vi sia motivo di assegnare pesi differenti alle diverse unità statistiche, i pesi risultano tutti uguali e pari a $1/n$, cioè:

$$p_i = \frac{1}{n}, \forall i \in \{1, \dots, n\}$$

e la matrice \mathbf{D} può essere scritta come $\mathbf{D} = \frac{1}{n} \times \mathbf{I}_n$.

Metriche più utilizzate

- La metrica $\mathbf{M} = \mathbf{I}_p$ non comporta alcuna ponderazione e viene utilizzata su dati omogenei.
- La metrica $\mathbf{M} = \mathbf{D}_{1/\sigma^2}$ equivale a dividere ciascun elemento per lo scarto quadratico medio della variabile corrispondente e, quindi, considerando che i dati sono centrati, a standardizzare le variabili iniziali. E' una operazione molto utile quando si analizzano variabili eterogenee o espresse in unità di misura molto differenti.
- Le due situazioni possono essere equivalenti. L'analisi può essere effettuata o sulla matrice dei dati iniziali utilizzando la metrica $\mathbf{M} = \mathbf{D}_{1/\sigma^2}$ oppure sulla matrice dei dati trasformati $\mathbf{X} = \mathbf{Y}\mathbf{D}_{1/\sigma}$ utilizzando la metrica euclidea $\mathbf{M} = \mathbf{I}_p$.

- Si dimostra che per ogni matrice simmetrica definita positiva \mathbf{M} esiste una matrice \mathbf{T} tale che $\mathbf{M} = \mathbf{T}^T \mathbf{T}$. Indicati con \mathbf{e}_1 e \mathbf{e}_2 i vettori contenenti i valori delle variabili osservate su due unità, il prodotto scalare è pari a

$$\begin{aligned} \langle \mathbf{e}_1, \mathbf{e}_2 \rangle_{\mathbf{M}} &= \mathbf{e}_1^T \mathbf{M} \mathbf{e}_2 = \mathbf{e}_1^T \mathbf{T}^T \mathbf{T} \mathbf{e}_2 = \\ &= (\mathbf{T} \mathbf{e}_1)^T (\mathbf{T} \mathbf{e}_2) = \langle \mathbf{T} \mathbf{e}_1, \mathbf{T} \mathbf{e}_2 \rangle_{\mathbf{I}_p} \end{aligned}$$

- Effettuare il prodotto scalare tra due vettori \mathbf{e}_1 e \mathbf{e}_2 utilizzando la metrica \mathbf{M} equivale quindi ad effettuare il prodotto scalare ordinario, con metrica $\mathbf{M} = \mathbf{I}$, dopo aver sostituito la matrice iniziale \mathbf{Y} con la matrice trasformata $\mathbf{X} = \mathbf{Y} \mathbf{T}^T$.

Il caso più comune è quello di considerare le metriche $\mathbf{D} = \frac{1}{n} \times \mathbf{I}_n$ e $\mathbf{M} = \mathbf{I}$ oppure $\mathbf{M} = \mathbf{D}_{1/\sigma^2}$. Tali metriche possono essere incorporate con un pretrattamento iniziale nei dati \mathbf{Y} .

Dati centrati. $\mathbf{D} = \frac{1}{n} \times \mathbf{I}_n$ e $\mathbf{M} = \mathbf{I}_p$.

- Il vettore del baricentro \mathbf{g} della nube dei punti in \mathbb{R}^p è dato da $\mathbf{g} = \mathbf{Y}^T \mathbf{D} \mathbf{1} = \mathbf{Y}^T \mathbf{1} / n = [\bar{y}_1, \dots, \bar{y}_p]$ di dimensione $(1 \times p)$
- La nube dei punti viene quindi centrata (traslata) nel baricentro

$$\mathbf{X} = (\mathbf{Y} - \mathbf{1g}) \times \frac{1}{\sqrt{n}}$$

- L'elemento generico della matrice \mathbf{X} risulta allora essere

$$x_{ij} = \frac{y_{ij} - \bar{y}_j}{\sqrt{n}}$$

con $i = 1, \dots, n$ e $j = 1, \dots, p$.

Dati centrati e standardizzati. $\mathbf{D} = \frac{1}{n} \times \mathbf{I}_n$ e $\mathbf{M} = \mathbf{D}_{1/\sigma_j^2}$.

- Il vettore del baricentro \mathbf{g} della nube dei punti in \mathbb{R}^p è dato da $\mathbf{g} = \mathbf{Y}^T \mathbf{D} \mathbf{1} = \mathbf{Y}^T \mathbf{1}/n = [\bar{y}_1, \dots, \bar{y}_p]$ di dimensione $(1 \times p)$
- La nube dei punti viene quindi centrata (traslata) nel baricentro

$$\mathbf{X} = \frac{1}{\sqrt{n}} \times (\mathbf{Y} - \mathbf{1g}) \mathbf{D}_{1/\sigma_j}$$

- L'elemento generico della matrice \mathbf{X} risulta allora essere

$$x_{ij} = \frac{y_{ij} - \bar{y}_j}{\sigma_j \sqrt{n}}$$

con $i = 1, \dots, n$ e $j = 1, \dots, p$, dove $\sigma_j = \sqrt{\frac{1}{n} \sum_j (y_{ij} - \bar{y}_j)^2}$.

La matrice $\mathbf{X}^T \mathbf{X}$ assume un preciso significato statistico

Dati centrati. Matrice di Varianza e Covarianza $\mathbf{V} = \mathbf{X}^T \mathbf{X}$

$$\begin{aligned}
 \mathbf{X}^T \mathbf{X} &= \frac{1}{n} \times (\mathbf{Y} - \mathbf{1g})^T (\mathbf{Y} - \mathbf{1g}) \\
 &= \frac{1}{n} \times (\mathbf{Y}^T \mathbf{Y} - \mathbf{Y}^T \mathbf{1g} - \mathbf{g}^T \mathbf{1}^T \mathbf{Y} + \mathbf{g}^T \mathbf{1}^T \mathbf{1g}) \\
 &= \frac{1}{n} \times \mathbf{Y}^T \mathbf{Y} - \mathbf{g}^T \mathbf{g} - \mathbf{g}^T \mathbf{g} + \mathbf{g}^T \mathbf{g} \\
 &= \frac{1}{n} \times \mathbf{Y}^T \mathbf{Y} - \mathbf{g}^T \mathbf{g} = \mathbf{V}
 \end{aligned}$$

$$\bullet \quad \mathbf{V} = \mathbf{X}^T \mathbf{X} = \begin{pmatrix} \text{Var}(\mathbf{y}_1) & \dots & \text{Cov}(\mathbf{y}_1, \mathbf{y}_p) \\ \vdots & \ddots & \vdots \\ \text{Cov}(\mathbf{y}_p, \mathbf{y}_1) & \dots & \text{Var}(\mathbf{y}_p) \end{pmatrix}$$

$$\bullet \quad V_{jj'} = \text{Cov}(\mathbf{y}_j, \mathbf{y}_{j'})$$

$$\bullet \quad \text{Inerzia totale } \mathcal{I} = \text{tr}(\mathbf{V}) = \text{tr}(\mathbf{X}^T \mathbf{X}) = \sum_j \text{Var}(j) = \sum_j \sigma_j^2$$

Dati centrati e standardizzati. Matrice di Correlazione R

- $$\mathbf{X}^T \mathbf{X} = \frac{1}{n} \times \mathbf{D}_{1/\sigma_j} (\mathbf{Y} - \mathbf{1g})^T (\mathbf{Y} - \mathbf{1g}) \mathbf{D}_{1/\sigma_j}$$

$$= \frac{1}{n} \times \mathbf{D}_{1/\sigma_j} \mathbf{Y}^T \mathbf{Y} \mathbf{D}_{1/\sigma_j} - \mathbf{D}_{1/\sigma_j} \mathbf{g}^T \mathbf{g} \mathbf{D}_{1/\sigma_j} = \mathbf{R}$$
- $$\mathbf{R} = \mathbf{X}^T \mathbf{X} = \begin{pmatrix} 1 & \dots & r(\mathbf{y}_1, \mathbf{y}_p) \\ \vdots & \ddots & \vdots \\ r(\mathbf{y}_p, \mathbf{y}_1) & \dots & 1 \end{pmatrix}$$
- $$R_{jj'} = \frac{\text{Cov}(\mathbf{y}_j, \mathbf{y}_{j'})}{\sigma_j \sigma_{j'}} = r(\mathbf{y}_j, \mathbf{y}_{j'})$$
- Inerzia totale $\mathcal{I} = \text{tr}(\mathbf{R}) = \text{tr}(\mathbf{X}^T \mathbf{X}) = p$

L'Analisi in Componenti Principali

- Verrà sempre ipotizzato, nel prosieguo, che le variabili \mathbf{y}_j siano state preventivamente centrate rispetto al baricentro (origine) e, eventualmente, anche standardizzate.
- Abbiamo visto che il primo asse ricercato deve essere tale da massimizzare la varianza dei punti proiettati su di esso (asse principale dell'elissoide)
- Il vettore \mathbf{c}_1 delle proiezioni degli n individui sull'asse \mathbf{u}_1 sarà dato da

$$\mathbf{c}_1 = \mathbf{X}\mathbf{u}_1$$

- Il vettore \mathbf{u}_1 sarà di norma unitaria: $\mathbf{u}_1^T \mathbf{u}_1 = 1$

- Sappiamo che la quantità da massimizzare sarà data dalla somma dei quadrati delle proiezioni delle n unità

$$\max \sum_{i=1}^n c_i^2 = \max \mathbf{u}_1^T \mathbf{X}^T \mathbf{X} \mathbf{u}_1$$

con il vincolo $\mathbf{u}_1^T \mathbf{u}_1 = 1$

- Il Lagrangiano associato sarà

$$L = \mathbf{u}_1^T \mathbf{X}^T \mathbf{X} \mathbf{u}_1 - \lambda_1 (\mathbf{u}_1^T \mathbf{u}_1 - 1)$$

- Derivando rispetto a \mathbf{u}_1 ed annullando le derivate parziali, si ottiene

$$\frac{\partial L}{\partial \mathbf{u}_1} = 2\mathbf{X}^T \mathbf{X} \mathbf{u}_1 - 2\lambda_1 \mathbf{u}_1 = 0$$

da cui

$$\mathbf{X}^T \mathbf{X} \mathbf{u}_1 = \lambda_1 \mathbf{u}_1$$

- Premoltiplicando i due membri per \mathbf{u}_1^T , si ottiene

$$\lambda_1 = \mathbf{u}_1^T \mathbf{X}^T \mathbf{X} \mathbf{u}_1 = \mathbf{c}_1^T \mathbf{c}_1 = \|\mathbf{c}_1\|^2$$

La varianza dei punti proiettati sul primo asse è quindi pari a λ_1 e coincide con la norma quadratica del vettore \mathbf{c}_1

- La relazione $\mathbf{X}^T \mathbf{X} \mathbf{u}_1 = \lambda_1 \mathbf{u}_1$ implica che la soluzione per il primo asse è data dall'autovettore \mathbf{u}_1 associato all'autovalore λ_1 più grande derivante alla diagonalizzazione della matrice simmetrica e semidefinita positiva $\mathbf{X}^T \mathbf{X}$.
- La prima componente principale sarà fornita da una combinazione lineare delle variabili \mathbf{x}_j con coefficienti pari agli elementi u_{j1} dell'autovettore \mathbf{u}_1 : $\mathbf{c}_1 = \mathbf{X} \mathbf{u}_1 = \sum_{j=1}^p \mathbf{x}_j u_{j1}$
- La nube degli individui è centrata sul primo asse: $\mathbf{1}^T \mathbf{c}_1 = 0$
- La seconda componente principale è definita dalla combinazione lineare di varianza massima $\mathbf{c}_2 = \mathbf{X} \mathbf{u}_2$ soggetta ai vincoli di normalità ($\mathbf{u}_2^T \mathbf{u}_2 = 1$) e di ortogonalità ($\mathbf{u}_1^T \mathbf{u}_2 = 0$).

Il Lagrangiano associato sarà

$$L_2 = \mathbf{u}_2^T \mathbf{X}^T \mathbf{X} \mathbf{u}_2 - \lambda_2 (\mathbf{u}_2^T \mathbf{u}_2 - 1) - \mu (\mathbf{u}_1^T \mathbf{u}_2)$$

Derivando L_2 rispetto a \mathbf{u}_2 ed annullando le derivate parziali, si ottiene

$$\frac{\partial L_2}{\partial \mathbf{u}_2} = 2\mathbf{X}^T \mathbf{X} \mathbf{u}_2 - 2\lambda_2 \mathbf{u}_2 - \mu \mathbf{u}_1 = 0$$

da cui

$$\mathbf{X}^T \mathbf{X} \mathbf{u}_2 = \lambda_2 \mathbf{u}_2 + \frac{\mu}{2} \mathbf{u}_1$$

Premoltiplicando i due membri per \mathbf{u}_2^T , si ottiene

$$\lambda_2 = \mathbf{u}_2^T \mathbf{X}^T \mathbf{X} \mathbf{u}_2$$

Si noti che la condizione $\mathbf{u}_2^T \mathbf{u}_1 = 0$ equivale a quella di ortogonalità tra le componenti \mathbf{c}_2 e \mathbf{c}_1 :

$$\langle \mathbf{c}_2, \mathbf{c}_1 \rangle = \mathbf{c}_2^T \mathbf{c}_1 = \mathbf{u}_2^T \mathbf{X}^T \mathbf{X} \mathbf{u}_1 = \mathbf{u}_2^T \lambda_1 \mathbf{u}_1 = \lambda_1 \mathbf{u}_2^T \mathbf{u}_1 = 0$$

quindi premoltiplicando per \mathbf{u}_1^T i due membri dell'identità

$$\mathbf{X}^T \mathbf{X} \mathbf{u}_2 = \lambda_2 \mathbf{u}_2 + \frac{\mu}{2} \mathbf{u}_1$$

abbiamo che

$$\boxed{\underbrace{\mathbf{u}_1^T \mathbf{X}^T \mathbf{X} \mathbf{u}_2}_{=0} = \lambda_2 \underbrace{\mathbf{u}_1^T \mathbf{u}_2}_{=0} + \frac{\mu}{2} \underbrace{\mathbf{u}_1^T \mathbf{u}_1}_{=1}} \implies \mu = 0$$

quindi si perviene all'identità $\mathbf{X}^T \mathbf{X} \mathbf{u}_2 = \lambda_2 \mathbf{u}_2$ da cui deriva che la soluzione per il secondo asse è data dall'autovettore \mathbf{u}_2 associato all'autovalore λ_2 più grande derivante alla diagonalizzazione della matrice $\mathbf{X}^T \mathbf{D} \mathbf{X}$.

La seconda componente principale sarà fornita da una combinazione lineare delle variabili \mathbf{x}_j con coefficienti pari agli elementi u_{j2} dell'autovettore \mathbf{u}_2 : $\mathbf{c}_2 = \mathbf{X} \mathbf{u}_2 = \sum_{j=1}^p \mathbf{x}_j u_{j2}$, ed è centrata sul secondo asse: $\mathbf{1}^T \mathbf{c}_2 = 0$

Come visto nel caso generale, le soluzioni per gli assi successivi corrispondono agli autovettori associati agli autovalori della matrice $\mathbf{X}^T \mathbf{D} \mathbf{X}$ ordinati in modo decrescente.

La matrice delle componenti principali è allora data da $\mathbf{C} = \mathbf{X} \mathbf{U}$ dove $\mathbf{C} = (\mathbf{c}_1, \dots, \mathbf{c}_p)$ e $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_p)$.

L'Analisi in Componenti Principali

Esempio ACP: i consumi alimentari 16 paesi, 10 variabili continue

	Cere	Riso	Pata	Zucc	Verd	Vino	Carn	Latt	Burr	Uova
Belgio	72,20	4,20	98,80	40,40	103,2	20,90	102,0	80,00	7,70	14,20
Danimarca	70,50	2,20	57,00	39,50	50,00	22,00	105,8	145,2	4,10	14,30
Germania	71,30	2,30	74,10	37,10	83,10	22,80	97,20	90,70	6,90	14,80
Grecia	109,8	5,40	90,00	30,00	229,5	25,30	77,10	63,10	0,90	11,30
Spagna	71,40	5,80	107,8	26,80	191,7	43,00	102,1	98,40	0,60	15,30
Francia	73,00	4,30	78,20	34,10	95,00	64,50	110,5	98,90	8,90	15,00
Irlanda	93,40	3,20	151,5	34,80	55,00	3,90	105,0	185,9	3,40	11,40
Italia	110,2	4,80	38,60	27,90	181,9	61,60	88,00	65,00	2,40	11,10
Olanda	54,60	5,00	86,70	39,70	99,00	14,00	89,40	136,2	5,40	10,70
Portogallo	86,00	5,70	106,6	29,40	100,0	57,00	75,50	96,00	1,50	7,70
RegnoUnito	74,30	4,50	94,10	39,80	60,00	10,40	74,40	129,3	3,20	10,80
Austria	68,70	4,20	62,60	37,10	81,90	34,30	93,40	121,3	4,30	13,40
Finlandia	70,10	5,40	61,60	35,70	52,60	10,20	65,00	208,4	5,80	10,90
Islanda	79,70	1,90	50,20	54,90	50,00	6,20	71,70	205,6	4,60	11,30
Norvegia	76,90	3,50	73,20	37,30	48,30	6,60	54,90	176,5	2,10	11,30
Svezia	69,30	4,30	70,00	37,50	48,50	12,30	60,50	154,1	5,70	12,90

L'Analisi in Componenti Principali

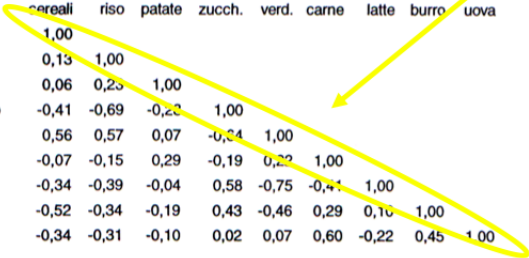
Matrice di correlazione:

	cereali	riso	patate	zucch.	verd.	carne	latte	burro	uova
cereali	1,00								
riso	0,13	1,00							
patate	0,06	0,23	1,00						
zucchero	-0,41	-0,69	-0,28	1,00					
verdure	0,56	0,57	0,07	-0,64	1,00				
carne	-0,07	-0,15	0,29	-0,19	0,22	1,00			
latte	-0,34	-0,39	-0,04	0,58	-0,75	-0,41	1,00		
burro	-0,52	-0,34	-0,19	0,43	-0,46	0,29	0,10	1,00	
uova	-0,34	-0,31	-0,10	0,02	0,07	0,60	-0,22	0,45	1,00

L'Analisi in Componenti Principali

$$\begin{aligned} \text{tr}(R) &= \sum_i r_{ii} \\ &= \sum_{\alpha} \lambda_{\alpha} \\ &(\alpha = 1, \dots, p) \end{aligned}$$

Matrice di correlazione:



	cereali	riso	patate	zucch.	verd.	carne	latte	burro	uova
cereali	1,00								
riso	0,13	1,00							
patate	0,06	0,23	1,00						
zucchero	-0,41	-0,69	-0,23	1,00					
verdure	0,56	0,57	0,07	-0,34	1,00				
carne	-0,07	-0,15	0,29	-0,19	0,22	1,00			
latte	-0,34	-0,39	-0,04	0,58	-0,75	-0,41	1,00		
burro	-0,52	-0,34	-0,19	0,43	-0,46	0,29	0,10	1,00	
uova	-0,34	-0,31	-0,10	0,02	0,07	0,60	-0,22	0,45	1,00

$$\begin{aligned} tr(R) &= \sum_i r_{ii} \\ &= \sum_{\alpha} \lambda_{\alpha} \\ &\quad (\alpha = 1, \dots, p) \end{aligned}$$

	1	2	3	4	5	6	7	8	9
1	3.4591	38.43	38.43						
2	2.1630	24.03	62.47						
3	1.1394	12.66	75.13						
4	0.9345	10.38	85.51						
5	0.4842	5.38	90.89						
6	0.3251	3.61	94.50						
7	0.2401	2.67	97.17						
8	0.2143	2.38	99.55						
9	0.0403	0.45	100.00						

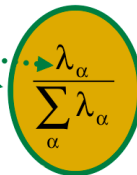
$$\frac{\lambda_{\alpha}}{\sum_{\alpha} \lambda_{\alpha}}$$

L'Analisi in Componenti Principali

$$\begin{aligned} \text{tr}(R) &= \sum_i r_{ii} \\ &= \sum_{\alpha} \lambda_{\alpha} \\ &(\alpha = 1, \dots, p) \end{aligned}$$

Autovalori e percentuale di variabilità spiegata

Num	Valore	%	% cum.	Istogramma
1	3.4591	38.43	38.43
2	2.1630	24.93	62.47
3	1.1394	12.66	75.13
4	0.9345	10.38	85.51
5	0.4842	5.38	90.89
6	0.3251	3.61	94.50
7	0.2401	2.67	97.17
8	0.2143	2.38	99.55
9	0.0403	0.45	100.00



$$\frac{\lambda_{\alpha}}{\sum_{\alpha} \lambda_{\alpha}}$$

Quante componenti scegliere?

Num	Valore	%	% cum.	Istogramma
1	3.4591	38.43	38.43
2	2.1630	24.03	62.47
3	1.1394	12.66	75.13
4	0.9345	10.38	85.51
5	0.4842	5.38	90.89
6	0.3251	3.61	94.50
7	0.2401	2.67	97.17	...
8	0.2143	2.38	99.55	..
9	0.0403	0.45	100.00	.

L'Analisi in Componenti Principali

Quante componenti scegliere ?

Autovalori e percentuale di variabilità spiegata

Num	Valore	%	% cum.	Istogramma
1	3.4591	38.43	38.43
2	2.1630	24.03	62.47
3	1.1394	12.66	75.13
4	0.9345	10.38	85.51
5	0.4842	5.38	90.89
6	0.3251	3.61	94.50
7	0.2401	2.67	97.17	...
8	0.2143	2.38	99.55	..
9	0.0403	0.45	100.00	•

L'Analisi in Componenti Principali

Quante componenti scegliere ?

Autovalori e percentuale di variabilità spiegata


Num	Valore	%	% cum.	Istogramma
1	3.4591	38.43	38.43
2	2.1630	24.03	62.47
3	1.1394	12.66	75.13
4	0.9345	10.38	85.51
5	0.4842	5.38	90.89
6	0.3251	3.61	94.50
7	0.2401	2.67	97.17	...
8	0.2143	2.38	99.55	..
9	0.0403	0.45	100.00	•

L'Analisi in Componenti Principali

Quante componenti scegliere ?

Autovalori e percentuale di variabilità spiegata

Num	Valore	%	% cum.	Istogramma
1	3.4591	38.43	38.43
2	2.1630	24.03	62.47
3	1.1394	12.66	75.13
4	0.9345	10.38	85.51
5	0.4842	5.38	90.89
6	0.3251	3.61	94.50
7	0.2401	2.67	97.17	...
8	0.2143	2.38	99.55	..
9	0.0403	0.45	100.00	•



L'Analisi in Componenti Principali

Coordinate degli individui:

	(1)	(2)	(3)	(4)	(5)
Belgio	-0,39	2,06	-0,35	-0,19	-0,94
Danimarca	-1,73	1,20	0,64	0,90	0,85
Germania	-1,13	2,01	0,62	0,41	-0,16
Grecia	3,84	-0,64	0,95	0,65	-0,15
Spagna	2,70	1,48	-0,93	-0,45	1,87
Francia	-0,42	2,64	-0,01	-0,59	-0,78
Irlanda	-0,02	-0,19	-2,47	2,39	-0,25
Italia	2,92	-0,25	2,29	0,36	-0,46
Olanda	-0,61	-0,03	-0,89	-1,25	-0,35
Portogallo	2,22	-1,93	-1,05	-0,26	-0,80
Regno Unito	-0,29	-1,08	-0,60	-0,14	-0,17
Austria	-0,48	0,69	0,36	-0,48	0,43
Finlandia	-1,12	-1,62	-0,23	-1,57	-0,04
Islanda	-3,29	-1,53	1,31	1,26	-0,03
Norvegia	-0,92	-2,09	0,18	0,05	0,84
Svezia	-1,27	-0,72	0,16	-1,08	0,15