

Analisi in Componenti Principali

October 20, 2010

Outline

- 1 Definizione del metodo
- 2 Trasformazioni sulla matrice dei dati
- 3 Formalizzazione del metodo: analisi in R^p
- 4 Scelta del numero di dimensioni
- 5 Contributi agli assi e qualità della rappresentazione
- 6 Formalizzazione del metodo: analisi in R^n
- 7 Rappresentazione congiunta unità variabili

Analisi in Componenti Principali

Ottica analitica

- L'Analisi in Componenti Principali ha per obiettivo l'individuazione della struttura di fondo delle relazioni osservate.
- Questo avviene attraverso l'identificazione di un numero limitato di nuove variabili, risultanti dalla **combinazione lineare delle variabili di partenza**, che consentano un'adeguata descrizione del sistema osservato, eliminando ridondanze nelle informazioni contenute nei dati.

Ottica geometrica

- L'insieme dei vettori riga **unità** della matrice dei dati di partenza (di dimensioni $n \times p$) è interpretabile come una configurazione di punti nello spazio di dimensione R^p . Analogamente, i p vettori colonna **variabili** della matrice sono interpretabili come una nube di punti nello spazio R^n .
- L'ACP ha per obiettivo l'individuazione di un sottospazio di R^p (R^n) tale che **la somma dei quadrati delle distanze tra le proiezioni dei punti/unità (punti/variabile) sia massimizzata**.
- Ottenere una proiezione della configurazione dei punti di partenza che approssimi al meglio la forma originale

Matrice dei dati. Esempio: la dieta di 16 paesi europei

Si consideri di aver rilevato le quantità di sostanze nutritive ingerite in sedici paesi europei. Ciascuna delle $p = 10$ variabili **quantitative** rappresenta una dimensione dello spazio R^{10} in cui sono rappresentati gli $n = 16$ paesi europei. In modo duale, a ciascuna delle variabili corrisponde un punto nello spazio R^{16} .

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

Trasformazione sulla matrice dei dati

L'ACP si propone di individuare un sottospazio di dimensioni ridotte tale da massimizzare la somma delle distanze tra le proiezioni delle possibili coppie di punti. La somma di tali distanze è una misura della variabilità che caratterizza i dati. Dunque l'ACP mira a massimizzare la **variabilità spiegata** dalle proiezioni dei punti nel sottospazio.

- Per individuare le direzioni degli assi componenti del sottospazio di proiezione ottimale si fa riferimento al **teorema generale di Huyghens**

Teorema generale di Huyghens

- Assicura che le direzioni di massima variabilità passino per il **baricentro** della nube di punti e siano ortogonali tra loro
- il baricentro della nube corrisponde al vettore delle medie delle p variabili
- massimizzare la somma delle distanze quadratiche delle proiezioni equivale a **massimizzare la somma delle distanze delle proiezioni dal baricentro**

Centratura della matrice

- **analiticamente**: centrare la matrice dei dati consiste nel calcolare gli n **scarti dalla media** per ciascuna delle p variabili
- **geometricamente**: centrare la matrice dei dati consiste nel **traslare** la configurazione di punti in modo che il baricentro della nube si sposti nell'origine degli assi. Il problema diventa quindi massimizzare la somma delle distanze al quadrato dei punti dall'origine degli assi del sottospazio di proiezione, ovvero massimizzare la somma dei quadrati delle coordinate dei punti sugli assi

Rendere i dati omogenei: standardizzazione

Per trattare simultaneamente le variabili è necessario che siano **omogenee**: si procede pertanto alla standardizzazione, consistente nel dividere ciascuno degli elementi della matrice centrata $\mathbf{X}$ per lo scarto quadratico medio della variabile corrispondente. Formalmente,

$$x_{ij} = \frac{(y_{ij} - \mu_j)}{\sigma_j \sqrt{n}}$$

interpretazione geometrica della standardizzazione

- attraverso la standardizzazione i vettori colonna (variabili) vengono riscaldati, si collocano nella iper-sfera di raggio unitario e dimensione n .

Formalizzazione del problema

Il problema

La proiezione ortogonale della matrice $\mathbf{X}$ su un asse di versore $\mathbf{u}$ è $(OH) = \mathbf{X}\mathbf{u}$. Il problema della ricerca del sottospazio di proiezione che riproduca al meglio la variabilità originaria della nube viene formalizzato come segue

- funzione obbiettivo:

$$\max! \sum_{i=1}^n (OH_i)^2 = (\mathbf{X}\mathbf{u})'(\mathbf{X}\mathbf{u}) = \mathbf{u}'\mathbf{X}'\mathbf{X}\mathbf{u}$$

- vincolo: $\mathbf{u}'\mathbf{u} = 1$

Soluzione del problema: ricerca del miglior asse

Metodo di risoluzione del problema

La ricerca del **miglior asse** di versore $\mathbf{u}_1$ è dunque un problema di **massimizzazione vincolata**. Per risolverlo si fa ricorso al **metodo dei moltiplicatori di Lagrange**. Tale metodo consiste nel ridefinire la funzione obbiettivo in modo da includere il vincolo (in questo caso $\mathbf{u}_1' \mathbf{u}_1 = 1$). La funzione da massimizzare diventa

$$\max \mathcal{L} = \mathbf{u}_1' \mathbf{X}' \mathbf{X} \mathbf{u}_1 - \lambda (\mathbf{u}_1' \mathbf{u}_1 - 1)$$

Soluzione del problema

La soluzione del problema

Per massimizzare $\mathcal{L}$ si deve porre a zero la derivata prima rispetto ad $\mathbf{u}$

$$\max \mathcal{L} = \mathbf{u}_1' \mathbf{X}' \mathbf{X} \mathbf{u}_1 - \lambda_1 (\mathbf{u}_1' \mathbf{u}_1 - 1)$$

$$\partial_{\mathbf{u}_1} \mathcal{L} = 2\mathbf{X}' \mathbf{X} \mathbf{u}_1 - 2\lambda_1 \mathbf{u}_1 = 0$$

$$\mathbf{X}' \mathbf{X} \mathbf{u}_1 = \lambda_1 \mathbf{u}_1$$

Soluzione del problema

Risoluzione dell'equazione caratteristica

$$\mathbf{X}'\mathbf{X}\mathbf{u}_1 = \lambda_1\mathbf{u}_1$$

$$\mathbf{X}'\mathbf{X}\mathbf{u}_1 - \lambda_1\mathbf{u}_1 = 0$$

$$(\mathbf{X}'\mathbf{X} - \lambda_1)\mathbf{u}_1 = 0$$

quest'ultima rappresenta l'**equazione caratteristica**: di fatto la soluzione del problema di massimizzazione vincolata si ottiene mediante la **ricerca di autovalori ed autovettori della matrice $\mathbf{X}'\mathbf{X}$** .

Soluzione del problema

Metodo di risoluzione del problema

La ricerca del **secondo miglior asse** di versore $\mathbf{u}_2$ è un problema di **massimizzazione vincolata** (Lagrange). I vincoli da includere nel lagrangiano sono in questo caso $\mathbf{u}_2' \mathbf{u}_2 = 1$ (norma unitaria) e $\mathbf{u}_1' \mathbf{u}_2 = 0$ (vincolo di ortogonalità) (per il teorema di Huyghens). La funzione da massimizzare diventa

$$\max \mathcal{L} = \mathbf{u}_2' \mathbf{X}' \mathbf{X} \mathbf{u}_2 - \lambda (\mathbf{u}_2' \mathbf{u}_2 - 1) - \mu (\mathbf{u}_1' \mathbf{u}_2)$$

Soluzione del problema

Per massimizzare $\mathcal{L}$ si deve porre a zero la derivata prima rispetto ad $\mathbf{u}$

$$\max \mathcal{L} = \mathbf{u}_2' \mathbf{X}' \mathbf{X} \mathbf{u}_2 - \lambda_2 (\mathbf{u}_2' \mathbf{u}_2 - 1) - \mu (\mathbf{u}_1' \mathbf{u}_2)$$

$$\partial_{\mathbf{u}_2} \mathcal{L} = 2\mathbf{X}' \mathbf{X} \mathbf{u}_2 - 2\lambda_2 \mathbf{u}_2 - \mu \mathbf{u}_1 = 0$$

Premoltiplicando $\mathbf{u}_1'$

$$\partial_{\mathbf{u}_2} \mathcal{L} = 2\mathbf{u}_1' \mathbf{X}' \mathbf{X} \mathbf{u}_2 - 2\lambda_2 \mathbf{u}_1' \mathbf{u}_2 - \mu \mathbf{u}_1' \mathbf{u}_1 = 0$$

i primi due termini sono nulli per l'ortogonalità tra $\mathbf{u}_1$ e $\mathbf{u}_2$;
perchè il termine $\mu \mathbf{u}_1' \mathbf{u}_1$ si annulli deve essere $\mu = 0$

$$\mathbf{X}' \mathbf{X} \mathbf{u}_2 = \lambda_2 \mathbf{u}_2$$

Soluzione del problema

Risoluzione dell'equazione caratteristica

$$\mathbf{X}'\mathbf{X}\mathbf{u}_2 = \lambda_2\mathbf{u}_2$$

$$(\mathbf{X}'\mathbf{X} - \lambda_2)\mathbf{u}_2 = 0$$

la soluzione del problema di massimizzazione vincolata si ottiene mediante la **ricerca dell'autovettore associato al secondo autovalore più grande della matrice $\mathbf{X}'\mathbf{X}$.**

gli **assi successivi al secondo** avranno per versori gli autovettori associati agli autovalori successivi, presi in ordine decrescente.

Autovalori λ , autovettori U

Calcolando gli autovalori e gli autovettori di norma 1 ad essi associati, si è individuata la **base ortogonale del sottospazio di proiezione ottimale**.

```
> lambdas
[1] 3.89203172 2.29661854 1.21257231 0.96721852 0.55945203 0.40642560 0.30504962 0.21476359 0.10914340 0.03672468
> |
```

```
> print("autovettori U")
[1] "autovettori U"
>
> print(round(eigen(uvecu,4))
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,] -0.2988 0.2366 -0.1753 0.6107 -0.2711 0.1289 -0.4712 0.1658 -0.1059 0.3170
[2,] -0.3503 0.1794 0.1279 -0.6197 0.0896 0.1738 -0.0575 0.3712 -0.0412 0.5135
[3,] -0.1078 0.0464 0.8587 0.1000 -0.1530 0.1652 -0.0234 -0.1551 -0.3818 -0.1333
[4,] 0.4414 -0.0072 -0.1467 0.1783 -0.0473 0.4094 0.5069 0.2488 -0.4380 0.2668
[5,] -0.4498 -0.0315 -0.1073 0.1300 0.3685 0.3625 0.1276 0.3934 -0.0428 -0.5729
[6,] -0.3703 -0.2337 -0.2692 -0.1746 -0.4578 -0.3971 0.1812 0.0199 -0.5347 -0.1373
[7,] -0.1226 -0.5364 0.2896 0.2560 -0.1976 -0.1665 0.2828 0.4053 0.4410 0.2056
[8,] 0.4020 0.2573 0.1415 -0.0225 0.0248 -0.4512 -0.2446 0.6432 -0.1399 -0.2333
[9,] 0.2530 -0.4225 -0.0783 -0.2831 -0.4152 0.4685 -0.4621 0.1186 0.0534 -0.2212
[10,] 0.0378 -0.5672 0.0051 0.0875 0.5792 -0.1357 -0.3387 -0.0393 -0.3805 0.2331
```

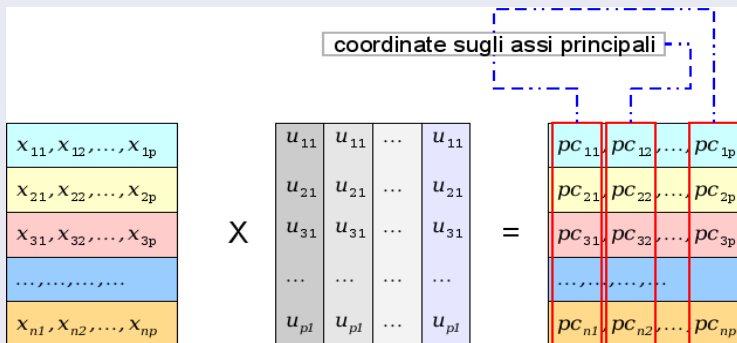
autovettori degli assi di proiezione (1,2 e 3).

Proiezione ortogonale sugli assi

Le colonne della matrice degli autovettori $\mathbf{U}$ rappresentano il sistema di pesi da applicare alle righe della matrice $\mathbf{X}$ per ottenere le coordinate delle unità nel sottospazio ottimale ottenuto. Ricordando la proiezione ortogonale, le coordinate principali degli individui.

$$C_U = \mathbf{X}\mathbf{u}$$

Calcolo delle coordinate principali



Coordinate principali dei paesi europei

Le coordinate principali delle unità sono date da

```
> print("Coordinate principali individui")
[1] "Coordinate principali individui"
>
> print(round(PCUnit,4))
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
Belgio	0.2255	-1.9606	0.6208	-0.0976	-0.1228	1.3489	-0.0127	-0.1090	-0.0715	0.2536
Danimarca	1.5216	-1.4778	-0.4322	1.0050	0.2818	-1.0462	0.3980	-0.0970	0.4846	0.1477
Germania	0.8518	-2.1236	-0.3028	0.6383	0.1161	0.2729	-0.2413	-0.8474	0.2578	-0.3614
Grecia	-3.4524	1.4321	-0.3170	1.1366	0.7786	1.1307	-0.2771	0.1331	0.0331	-0.0555
Spagna	-2.9206	-0.9516	1.2268	-0.3920	1.7174	-0.7729	0.3752	0.3294	-0.4114	-0.0661
Francia	-0.6756	-3.0773	-0.3176	-0.7658	-1.1035	-0.1774	-0.4717	0.3034	-0.4920	-0.0772
Irlanda	0.5796	0.6237	2.9337	1.7822	-0.8096	-0.3458	-0.6111	0.3262	0.1167	-0.0064
Italia	-3.4193	0.3122	-2.2674	0.7491	-0.5714	-0.2659	-0.2491	0.2337	0.3348	0.0686
Olanda	0.7899	0.0444	0.6719	-1.3590	0.0502	0.6486	0.9380	0.3805	0.5554	-0.2931
Portogallo	-2.3981	1.9556	0.4194	-0.8334	-1.5534	-0.4104	0.6998	-0.5398	-0.2088	-0.0897
Regno Unito	0.7043	1.1704	0.5403	-0.2514	0.0097	0.3894	0.4008	-0.4357	0.0067	0.3963
Austria	0.1814	-0.8868	-0.4847	-0.4181	0.2144	-0.4739	0.3867	0.0127	0.0804	0.2640
Finlandia	1.4830	1.4328	-0.2587	-1.6333	0.0501	-0.2001	-0.9241	0.7826	0.3237	0.0122
Islanda	3.5754	0.9519	-1.3509	1.3594	-0.1435	0.2961	0.7119	0.6552	-0.5862	-0.1039
Norvegia	1.4648	2.0160	-0.2914	0.0691	0.6252	-0.5476	-0.3199	-0.6458	-0.1608	-0.1370
Svezia	1.4885	0.5386	-0.3904	-0.9893	0.4608	0.1536	-0.8035	-0.4821	-0.2625	0.0479

Scelta delle dimensioni

Per scegliere il numero di dimensioni è necessario trovare un compromesso tra la variabilità spiegata dal sottospazio e il potere di sintesi della soluzione adottata: più alto è il numero q di dimensioni considerate minore sarà la sintesi dell'informazione. In particolare, se la dimensione del sottospazio è $q > 3$ non sarà possibile rappresentare la proiezione della nube di punti ottenuta.

Per le proprietà degli autovalori

$$tr(\mathbf{X}'\mathbf{X}) = \sum_{\alpha=1}^p \lambda_{\alpha}$$

la traccia della matrice $tr(\mathbf{X}'\mathbf{X})$ corrisponde al valore della variabilità complessiva della nube di punti corrispondente alla matrice dei dati standardizzati: ciascun autovalore esprime una parte della variabilità complessiva.

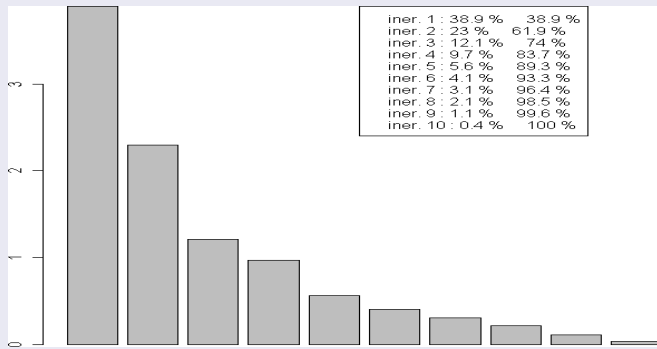
inertia

La variabilità spiegata dall'asse associato all'autovalore λ_{α} si definisce **inertia** ed è data da

$$inertia_{\alpha} = \frac{\lambda_{\alpha}}{\sum_{\alpha=1}^p \lambda_{\alpha}} \times 100$$

Scelta delle dimensioni

racpresentazione inertia di ciascuna delle dimensioni del sottospazio di R^p



criteri di scelta delle dimensioni

La scelta del numero di dimensioni non può essere univocamente determinata: sta all'analista valutare quale sia la scelta migliore sulla base dei seguenti **criteri**:

- **criterio della variabilità spiegata**: si sceglie il numero delle dimensioni sulla base della percentuale cumulata di variabilità spiegata dagli assi (**inertia cumulata**)
- **criterio dell' eigenvalue-one**: poichè per effetto della standardizzazione le variabili di partenza hanno tutte varianza pari ad 1 , si scelgono le variabili di sintesi la cui varianza sia maggiore di 1. La variabilità sugli assi è data dall'autovalore, quindi si sceglieranno le dimensioni il cui autovalore è maggiore di 1
- **criterio dello scree-test**: si considera il diagramma di Pareto relativo agli autovalori, quando il salto da una barra alla successiva si regolarizza, l'apporto alla variabilità di ogni dimensione aggiuntiva è irrilevante. Si sceglieranno le dimensioni corrispondenti alle barre che precedono la regolarizzazione del decremento della distribuzione

Interpretazione dei fattori e qualità della rappresentazione

La proiezione nel sottospazio ridotto determina delle distorsioni della configurazione di punti originaria; inoltre, gli assi di proiezione vengono determinati sulla base dei valori delle variabili di partenza osservati sulle diverse unità. È dunque necessario valutare, per ciascuna unità,

- contributo alla determinazione di ciascun asse di proiezione
- qualità della rappresentazione sugli assi

contributo assoluto

Il contributo di una unità alla variabilità dell'asse α si definisce **contributo assoluto** e si misura come segue:

$$CA_{i\alpha} = \frac{c_{i\alpha}^2}{n\lambda_\alpha}$$

il fattore $\frac{1}{n}$ rappresenta il peso dell'unità, che nell'ACP è uguale per le n unità.

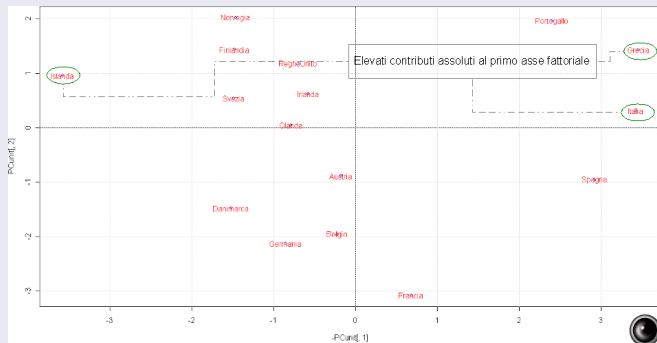
qualità della rappresentazione: il $\cos^2$

La qualità della rappresentazione di ciascun vettore unità dipende dal rapporto tra la lunghezza (norma) del vettore nello spazio originario e la lunghezza della sua proiezione ortogonale su ciascun asse. Il rapporto tra tali norme quadratiche equivale al coseno dell'angolo formato tra il vettore nello spazio originario e la sua proiezione sull'asse.

$$\cos_{i\alpha}^2 = \frac{\|\hat{x}_{i\alpha}\|^2}{\|\hat{x}_i\|^2}$$

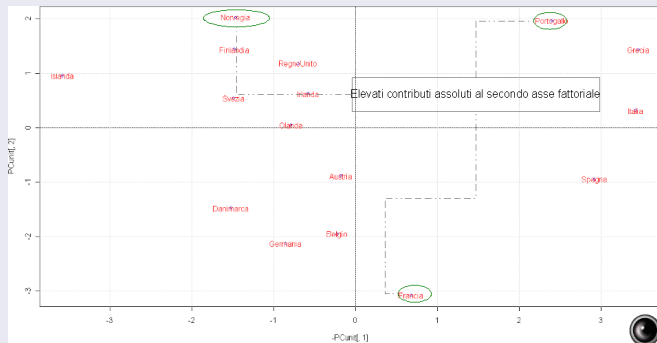
Rappresentazione degli individui

Contributi assoluti sul primo asse



Rappresentazione degli individui

Contributi assoluti sul secondo asse



Contributi assoluti degli individui

Contributi assoluti degli individui

```
> print("Contributi assoluti delle unità")
[1] "Contributi assoluti delle unità"
> print(round(CoAsunit,4))
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
Belgio	0.0008	0.1046	0.0199	0.0006	0.0017	0.2798	0.0000	0.0035	0.0029	0.1094
Danimarca	0.0372	0.0594	0.0096	0.0653	0.0089	0.1683	0.0325	0.0027	0.1345	0.0371
Germania	0.0117	0.1227	0.0047	0.0263	0.0015	0.0115	0.0119	0.2090	0.0380	0.2223
Grecia	0.1914	0.0558	0.0052	0.0835	0.0677	0.1966	0.0157	0.0052	0.0006	0.0052
Spagna	0.1370	0.0246	0.0776	0.0099	0.3295	0.0919	0.0288	0.0316	0.0969	0.0074
Francia	0.0073	0.2577	0.0052	0.0379	0.1360	0.0048	0.0456	0.0268	0.1386	0.0102
Irlanda	0.0054	0.0106	0.4436	0.2053	0.0732	0.0184	0.0765	0.0310	0.0078	0.0001
Italia	0.1878	0.0027	0.2650	0.0363	0.0365	0.0109	0.0127	0.0159	0.0642	0.0080
Olanda	0.0100	0.0001	0.0233	0.1194	0.0003	0.0647	0.1803	0.0421	0.1766	0.1462
Portogallo	0.0923	0.1041	0.0091	0.0449	0.2696	0.0259	0.1003	0.0848	0.0250	0.0137
Regno Unito	0.0080	0.0373	0.0150	0.0041	0.0000	0.0233	0.0329	0.0552	0.0000	0.2673
Austria	0.0005	0.0214	0.0121	0.0113	0.0051	0.0345	0.0306	0.0000	0.0037	0.1186
Finlandia	0.0353	0.0559	0.0035	0.1724	0.0003	0.0062	0.1750	0.1783	0.0600	0.0003
Islanda	0.2053	0.0247	0.0941	0.1194	0.0023	0.0135	0.1038	0.1249	0.1968	0.0184
Norvegia	0.0345	0.1106	0.0044	0.0003	0.0437	0.0461	0.0210	0.1214	0.0148	0.0319
Svezia	0.0356	0.0079	0.0079	0.0632	0.0237	0.0036	0.1323	0.0676	0.0395	0.0039

Coseni al quadrato degli individui

Coseni al quadrato degli individui

```
> print("Coseni al quadrato")
[1] "Coseni al quadrato"
> print(round(Cos2unit,4))
```

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	0.0082	0.6194	0.0621	0.0015	0.0024
[2,]	0.3174	0.2994	0.0256	0.1385	0.0109
[3,]	0.1068	0.6636	0.0135	0.0600	0.0020
[4,]	0.6871	0.1182	0.0058	0.0745	0.0349
[5,]	0.5662	0.0601	0.0999	0.0102	0.1958
[6,]	0.0367	0.7621	0.0081	0.0472	0.0980
[7,]	0.0244	0.0282	0.6247	0.2306	0.0476
[8,]	0.6452	0.0054	0.2837	0.0310	0.0180
[9,]	0.1309	0.0004	0.0947	0.3875	0.0005
[10,]	0.4149	0.2759	0.0127	0.0501	0.1741
[11,]	0.1722	0.4756	0.1014	0.0219	0.0000
[12,]	0.0191	0.4558	0.1361	0.1013	0.0266
[13,]	0.2557	0.2387	0.0078	0.3102	0.0003
[14,]	0.6814	0.0483	0.0973	0.0985	0.0011
[15,]	0.2840	0.5380	0.0112	0.0006	0.0517
[16,]	0.4595	0.0602	0.0316	0.2030	0.0440

Ricerca del sottospazio ottimale di rappresentazione delle variabili: formalizzazione del problema

Metodo dei minimi quadrati

Il metodo dei minimi quadrati ha per obiettivo la ricerca dell'asse U che minimizzi la somma delle distanze al quadrato dei punti dall'asse. L'identificazione dell'asse U corrisponde all'identificazione del suo versore $\mathbf{v}$: la ricerca del vettore $\mathbf{v}$ che massimizzi la quantità $\sum_{i=1}^n (OH_i)^2$ è vincolata al fatto che il vettore abbia lunghezza 1 (norma unitaria) tale condizione sussiste se $\mathbf{v}'\mathbf{v} = 1$.

In questo caso, volendo proiettare i punti variabile, dobbiamo considerare i vettori colonna della matrice $\mathbf{X}$, la proiezione ortogonale sull'asse di versore $\mathbf{v}$ è data da $\mathbf{X}'\mathbf{v}$.

Il problema

Ricordando che $\mathbf{X}'\mathbf{v}$:

- funzione obiettivo:

$$\max! \sum_{i=1}^n (OH_i)^2 = (\mathbf{X}'\mathbf{v})'(\mathbf{X}'\mathbf{v}) = \mathbf{v}'\mathbf{X}\mathbf{X}'\mathbf{v}$$

- vincolo: $\mathbf{v}'\mathbf{v} = 1$

Ricerca del sottospazio ottimale di rappresentazione delle variabili: ricerca del miglior asse

Metodo di risoluzione del problema

La ricerca del miglior asse di versore $\mathbf{v}_1$ è un problema di **massimizzazione vincolata**. Per risolverlo si fa ricorso al **metodo dei moltiplicatori di Lagrange**. Tale metodo consiste nel ridefinire la funzione obbiettivo in modo da includere il vincolo (in questo caso $\mathbf{v}_1' \mathbf{v}_1 = 1$). La funzione da massimizzare diventa

$$\max \mathcal{L} = \mathbf{v}_1' \mathbf{X} \mathbf{X}' \mathbf{v}_1 - \mu (\mathbf{v}_1' \mathbf{v}_1 - 1)$$

Soluzione del problema

La soluzione del problema

Per massimizzare $\mathcal{L}$ si deve porre a zero la derivata prima rispetto ad $\mathbf{u}$

$$\max \mathcal{L} = \mathbf{v}_1' \mathbf{X} \mathbf{X}' \mathbf{v}_1 - \mu (\mathbf{v}_1' \mathbf{v}_1 - 1)$$

$$\partial_{\mathbf{v}_1} \mathcal{L} = 2\mathbf{X} \mathbf{X}' \mathbf{v}_1 - 2\mu_1 \mathbf{v}_1 = 0$$

$$\mathbf{X} \mathbf{X}' \mathbf{v}_1 = \mu_1 \mathbf{v}_1$$

Soluzione del problema

Risoluzione dell'equazione caratteristica

$$\mathbf{X}\mathbf{X}'\mathbf{v}_1 = \mu_1\mathbf{v}_1$$

$$\mathbf{X}\mathbf{X}'\mathbf{v}_1 - \mu_1\mathbf{v}_1 = 0$$

$$(\mathbf{X}\mathbf{X}' - \mu_1)\mathbf{v}_1 = 0$$

quest'ultima rappresenta l'**equazione caratteristica**: di fatto la soluzione del problema di massimizzazione vincolata si ottiene mediante la **ricerca di autovalori ed autovettori della matrice $\mathbf{X}\mathbf{X}'$** .

Coordinate dei punti variabile sul sottospazio di proiezione

Analogamente a quanto descritto nell'analisi dello spazio di rappresentazione degli individui R^p , gli autovalori μ_α catturano una quota della variabilità complessiva. La **coordinata** della variabile j sull'asse individuato dall'autovalore μ_α ($\alpha = 1, \dots, n$) è data da

$$C_\alpha(j) = \mathbf{x}'_j\mathbf{v}_\alpha$$

autovalori λ , autovettori U

autovalori e autovettori

Gli autovalori risultanti dalla diagonalizzazione della matrice XX' sono n di cui $n - p$ nulli. Gli autovalori non nulli coincidono con quelli della matrice $X'X$.

```
> lambdas
[1] 3.89203172 2.29661854 1.21257231 0.96721852 0.55945203 0.40642560 0.30504962 0.21476359 0.10914340 0.03672468
> |
```

```
> print("Autovettori V")
[1] "Autovettori V"
>
> print(round(autvecv,4))
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]      [,9]      [,10]      [,11]      [,12]      [,13]      [,14]      [,15]      [,16]
[1,] -0.0286  0.3234  0.1409 -0.0248 -0.0410 -0.5290 -0.0057 -0.0588  0.0541 -0.3308 -0.0760 -0.5941 -0.1992  0.2137  0.1389  0.1156
[2,] -0.1928  0.2438 -0.0981  0.2555  0.0942  0.4103  0.1802 -0.0523 -0.3667 -0.1927  0.4817 -0.3770  0.1268 -0.1623 -0.1598  0.0478
[3,] -0.1079  0.3503 -0.0687  0.1623  0.0388 -0.1070 -0.1092 -0.4572 -0.1951  0.4714 -0.0010  0.1353 -0.4285  0.1686 -0.2707 -0.1918
[4,]  0.4375 -0.2362 -0.0720  0.2889  0.2602 -0.4434 -0.1254  0.0718 -0.0251  0.0724  0.2838 -0.1368  0.1835 -0.1815 -0.1788 -0.4186
[5,]  0.3701  0.1570  0.2785 -0.0997  0.5740  0.3031  0.1698  0.1777  0.3113  0.0862  0.0842 -0.0583 -0.2286  0.2675 -0.1348  0.1107
[6,]  0.0856  0.5076 -0.0721 -0.1947 -0.3688  0.0696 -0.2135  0.1637  0.3723  0.1008  0.0787 -0.0469  0.4085  0.0664 -0.3501 -0.1663
[7,] -0.0734 -0.1029  0.6660  0.4530 -0.2706  0.1356 -0.2766  0.1760 -0.0883  0.0083  0.0847  0.0439  0.0206  0.3122  0.0718 -0.1266
[8,]  0.4333 -0.0515 -0.5148  0.1904 -0.1910  0.1043 -0.1127  0.1261 -0.2534 -0.0895 -0.0319  0.0221 -0.0308  0.5499  0.0335  0.2328
[9,] -0.1001 -0.0073  0.1525  0.3455  0.0168 -0.2543  0.4246  0.2053 -0.4203  0.3823  0.1294  0.0008  0.3274  0.3336  0.0397 -0.0167
[10,]  0.3039 -0.3226  0.0952 -0.2119 -0.5192  0.1610  0.3168 -0.2912  0.1580  0.1170  0.2574 -0.2711 -0.2701 -0.0284 -0.0089 -0.1295
[11,] -0.0892 -0.1931  0.1227 -0.0639  0.0032 -0.1527  0.1814 -0.2350 -0.0051 -0.5171  0.0239  0.2960  0.0642  0.2352 -0.6378  0.0486
[12,] -0.0230  0.1463 -0.1100 -0.1063  0.0717  0.1859  0.1750  0.0068 -0.0609 -0.3444 -0.1406  0.0965 -0.0571  0.2036  0.2807 -0.7807
[13,] -0.1879 -0.2364 -0.0587 -0.4152  0.0167  0.0785 -0.4183  0.4222 -0.2449 -0.0159  0.0211 -0.2219 -0.3828 -0.0255 -0.3111 -0.1395
[14,] -0.4531 -0.1570 -0.3067  0.3456 -0.0480 -0.1161  0.3223  0.3535  0.4436  0.1356  0.1336 -0.0368 -0.1846  0.1835 -0.0752 -0.0653
[15,] -0.1856 -0.3326 -0.0662  0.0176  0.2090  0.2147 -0.1448 -0.3484  0.1217  0.1787 -0.3618 -0.4634  0.3670  0.2741 -0.0939 -0.0847
[16,] -0.1886 -0.0888 -0.0886 -0.2515  0.1540 -0.0602 -0.3637 -0.2600  0.1987 -0.0625  0.6385  0.1553  0.0607  0.2690  0.3232  0.0498
```

Coordinate principali delle variabili

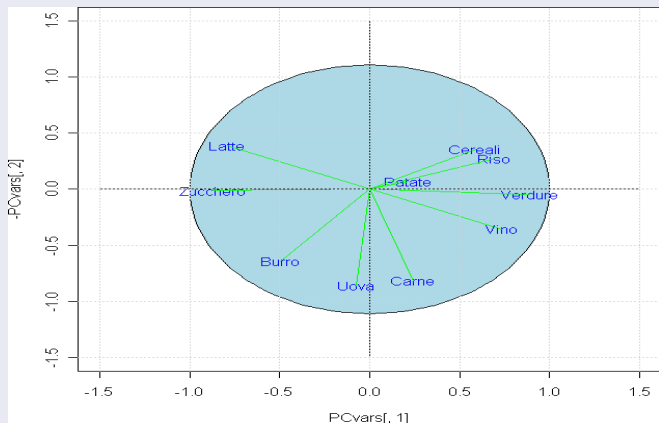
Coordinate principali delle variabili

```
> print("Coordinate principali variabili")
[1] "Coordinate principali variabili"
>
> print(round(PCvars,4))
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]
[1,]	0.5894	-0.3580	-0.1930	0.6006	-0.2028	-0.0821	-0.2602	0.0768	0.0350	-0.0607	0	0	0	0	0	0
[2,]	0.6910	-0.2719	0.1408	-0.6095	0.0671	-0.1108	-0.0318	0.1720	0.0136	-0.0984	0	0	0	0	0	0
[3,]	0.2126	-0.0703	0.9456	0.0983	-0.1145	-0.1053	-0.0129	-0.0719	0.1261	0.0255	0	0	0	0	0	0
[4,]	-0.8709	0.0108	-0.1616	0.1754	-0.0354	-0.2610	0.2800	0.1153	0.1447	-0.0511	0	0	0	0	0	0
[5,]	0.8875	0.0477	-0.1181	0.1279	0.2756	-0.2311	0.0705	0.1823	0.0141	0.1098	0	0	0	0	0	0
[6,]	0.7306	0.3542	-0.2965	-0.1717	-0.3424	0.2532	0.1001	0.0092	0.1767	0.0263	0	0	0	0	0	0
[7,]	0.2418	0.8129	0.3189	0.2518	-0.1478	0.1061	0.1562	0.1878	-0.1457	-0.0394	0	0	0	0	0	0
[8,]	-0.7930	-0.3900	0.1558	-0.0221	0.0186	0.2877	-0.1351	0.2981	0.0462	0.0447	0	0	0	0	0	0
[9,]	-0.4992	0.6403	-0.0862	-0.2784	-0.3105	-0.2986	-0.2552	0.0550	-0.0177	0.0424	0	0	0	0	0	0
[10,]	-0.0745	0.8595	0.0057	0.0861	0.4332	0.0865	-0.1871	-0.0182	0.1257	-0.0447	0	0	0	0	0	0

Rappresentazione delle variabili: il cerchio delle correlazioni

Sottospazio di approssimazione di R^n

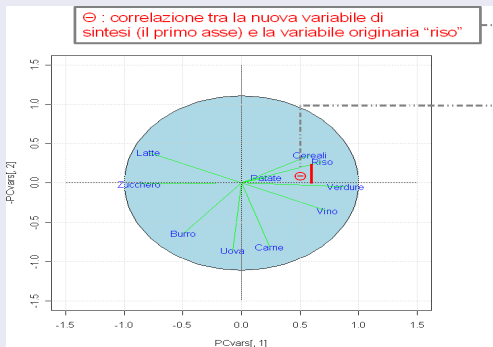


Rappresentazione delle variabili: guida all'interpretazione

Sottospazio di approssimazione di R^n

Mentre per i punti individuo, la distanza euclidea che li separa è indice di dissimilarità, nel caso delle variabili si interpreta l'angolo formato tra i diversi vettori/variabile.

- minore è l'angolo θ tra un vettore e ciascun asse fattoriale, maggiore sarà la **correlazione** tra la variabili originaria corrispondente e le variabili di sintesi
- minore è l'angolo θ tra due vettori, maggiore sarà la **correlazione** tra le variabili corrispondenti



Possibile rappresentazione congiunta unità/variabili

Da un punto di vista interpretativo può essere di aiuto avere una rappresentazione congiunta tra unità e variabili. Per fare questo, si proiettano le variabili iniziali nel sottospazio di proiezione delle unità.

Possibilità di una rappresentazione congiunta

- Si definiscono p vettori variabile di lunghezza 1 che rappresentano le **direzioni**. Poichè ad ogni dimensione dello spazio R^p corrisponde una variabile, i vettori unitari associati a ciascuna variabile hanno tutti gli elementi uguali a zero, fatta eccezione per la dimensione corrispondente alla variabile che rappresentano.
- Gli operatori di proiezione ortogonale sono, come per le unità, gli autovettori (colonne della matrice U).

direzione delle p variabili nel sottospazio delle unità

$$\begin{bmatrix} 1,0,0,\dots,0 \\ 0,1,0,\dots,0 \\ 0,0,1,\dots,0 \\ \dots,\dots,\dots,\dots \\ 0,0,0,\dots,1 \end{bmatrix} \times \begin{bmatrix} u_{11} & u_{11} & \dots & u_{11} \\ u_{21} & u_{21} & \dots & u_{21} \\ u_{31} & u_{31} & \dots & u_{31} \\ \dots & \dots & \dots & \dots \\ u_{p1} & u_{p1} & \dots & u_{p1} \end{bmatrix} = \begin{bmatrix} pv_{11} & pv_{12} & \dots & pv_{1p} \\ pv_{21} & pv_{22} & \dots & pv_{2p} \\ pv_{31} & pv_{32} & \dots & pv_{3p} \\ \dots & \dots & \dots & \dots \\ pv_{p1} & pv_{p2} & \dots & pv_{pp} \end{bmatrix}$$

Rappresentazione congiunta unità/variabili

