Università degli studi di Ferrara
Dipartimento di Matematica
A.A. 2019/2020 – I semestre

STATISTICA MULTIVARIATA

SSD MAT/06

analisi dell'interdipendenza

LEZION 11 - PCA: utilizzi e applicazione in R

Docente: Valentina MINI

valentina.mini@unife.it

RICEVIMENTO: su appuntamento previa mail

DOMANDA CENTRALE

Quando le p variabili sono **numerose** è molto difficile riuscire a cogliere le **strutture** esistenti nei dati.

Si pone quindi il **PROBLEMA**:

è possibile sostituire le p variabili originarie con un numero minore di variabili "artificiali" (k<<p) (che impareremo a chiamare COMPONENTI PRINCIPALI) che garantiscono la SINTESI con la MINOR PERDITA DI INFORMAZIONE POSSIBILE?

Ossia, in termini geometrici, è possibile rappresentare le osservazioni, anziché nello spazio originario R^p, in uno spazio di dimensioni ridotte (R, R^2, R^3,……), con una perdita limitata d'informazione?

Ri-esprimere dati multivariati:

ridimensionamento VS

informazione

Origini dell' ACP

- Pearson (1901)

- Hotelling (1933)

-Ogni componente è **non correlata** con le altre (eliminiamo multicollinearità)

- Alla base: non un vero modello → ogni componente = **combinazione lineare** delle variabili originarie (es. somma pesata, regressione lineare ecc.)

Quando è utile l'ACP?

- Analisi di regressione in caso di collinearità;
- problemi di classificazione per gruppi ben separati;
- **riduzione delle dimensioni;**
- identificazione degli outliers.

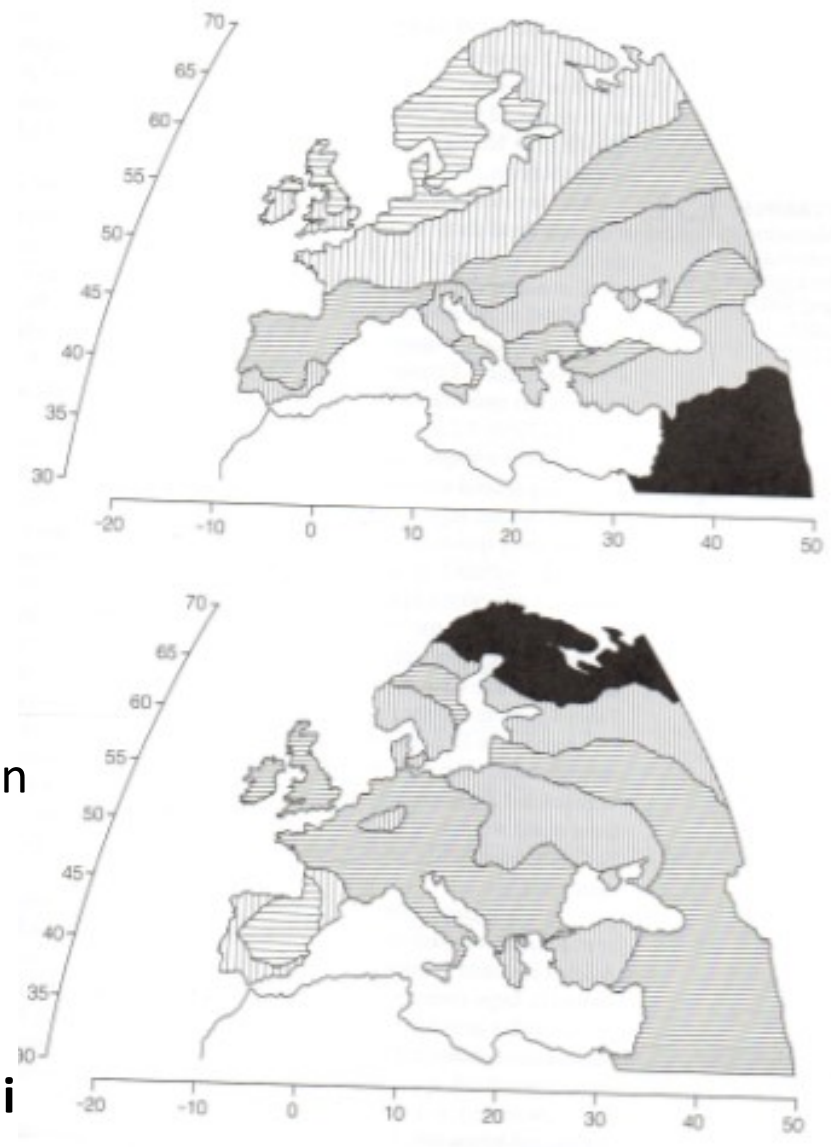
Primi utilizzi

Mappatura di eventi
Es. Mappa genetica
Mappa della distribuzione di 95 geni in Europa e Medio Oriente (Cavalli-Sforza, 1994)

Misurazione di situazioni non direttamente misurabili – variabili latenti - (es. applicazioni in studi psico-sociali)

Ricerche di Marketing per definire un profilo di acquirente (riducendo a poche componenti a totalità di variabili)

Analisi **delle associazioni tra variabili**



Es. analisi in campo psico-sociale (Cacioppo, petty and Kao, 1984)

C_1	I prefer complex to simple problems.
C_2	I like to have the responsibility of handling a situation that requires a lot of thinking.
C_3	Thinking is not my idea of fun. (R)
C_4	I would rather do something requiring little thought than something that is sure to challenge my thinking abilities. (R)
C_5	I try to anticipate and avoid situations where there is a likely chance that I will have to think in depth about something. (R)
C_6	I find satisfaction in deliberating hard for long hours.
C_7	I only think as hard as I have to. (R)
C_8	I prefer to think about small daily projects to long-term ones. (R)
C_9	I like tasks that require little thought once I've learned them. (R)
C_{10}	The idea of relying on thought to make my way to the top appeals to me.
C_{11}	I really enjoy a task that involves coming up with new solutions to problems.
C_{12}	Learning new ways to think doesn't excite me much. (R)
C_{13}	I prefer my life to be filled with puzzles that I must solve.
C_{14}	The notion of thinking abstractly is appealing to me.
C_{15}	I prefer tasks that are intellectual, difficult, and important to ones that do not require much thought.
C_{16}	I feel relief rather than satisfaction after completing a task that required a lot of mental effort. (R)
C_{17}	It's enough for me that something gets the job done; I don't care how or why it works. (R)
	I usually end up deliberating about issues even when they do not affect me personally.

Capacità di un individuo in termini di problem solving

→Soluzione 1) indice somma dei valori

→Soluzione 2) PCA (primo pc combinazione lineare delle 18 variabili e cattura ampia parte di variabilità nel campione)

Es. analisi in campo psico-sociale (Cacioppo, petty and Kao, 1984)

Primo passo dell'analisi: guardare alla struttura dei dati originari in termini di matrice di correlazione= identificare pattern di associazione tra le variabili

Correlation matrix for 18 items measuring need for cognition (n = 201)

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	C_{17}
C_2	0.445																
C_3	−0.239	−0.454															
C_4	−0.270	−0.375	0.365														
C_5	−0.326	−0.487	0.421	0.558													
C_6	0.248	0.274	−0.187	−0.187	−0.250												
C_7	−0.268	−0.338	0.319	0.345	0.343	−0.235											
C_8	−0.270	−0.355	0.228	0.306	0.340	−0.165	0.314										
C_9	−0.320	−0.328	0.389	0.415	0.310	−0.174	0.221	0.312									
C_{10}	0.289	0.375	−0.411	−0.338	−0.336	0.279	−0.268	−0.299	−0.380								
C_{11}	0.364	0.516	−0.325	−0.405	−0.384	0.297	−0.262	−0.148	−0.338	0.520							
C_{12}	−0.366	−0.415	0.258	0.373	0.555	−0.275	0.310	0.132	0.245	−0.261	−0.392						
C_{13}	0.393	0.382	−0.245	−0.288	−0.322	0.220	−0.172	−0.245	−0.250	0.321	0.418	−0.350					
C_{14}	0.341	0.376	−0.257	−0.250	−0.295	0.291	−0.231	−0.331	−0.145	0.338	0.310	−0.397	0.454				
C_{15}	0.268	0.354	−0.228	−0.185	−0.165	0.186	−0.066	−0.181	−0.177	0.223	0.236	−0.199	0.274	0.230			
C_{16}	−0.280	−0.192	0.166	0.320	0.242	−0.155	0.212	0.150	0.214	−0.147	−0.242	0.192	−0.058	−0.132	−0.071		
C_{17}	−0.273	−0.425	0.282	0.412	0.332	−0.232	0.251	0.364	0.373	−0.226	−0.361	0.284	−0.114	−0.090	−0.174	0.331	
C_{18}	0.126	0.166	−0.140	−0.069	−0.105	0.162	−0.131	−0.042	−0.007	−0.021	0.100	−0.119	0.029	0.209	0.084	−0.194	−0.087

Es. analisi in campo psico-sociale (Cacioppo, petty and Kao, 1984)

Risultato della prima componente estratta dalla ricerca

Results from principal components analysis of need for cognition data in Table 4.2

	$\mathbf{u}_1$		$\mathbf{u}_1$
C_1	0.251	C_{10}	0.259
C_2	0.309	C_{11}	0.282
C_3	−0.253	C_{12}	−0.259
C_4	−0.275	C_{13}	0.232
C_5	−0.289	C_{14}	0.230
C_6	0.183	C_{15}	0.169
C_7	−0.227	C_{16}	−0.164
C_8	−0.206	C_{17}	−0.229
C_9	−0.234	C_{18}	0.087

Eigenvalue $\lambda_1 = 5.7794$

Proportion of variance accounted for 32.1 percent.

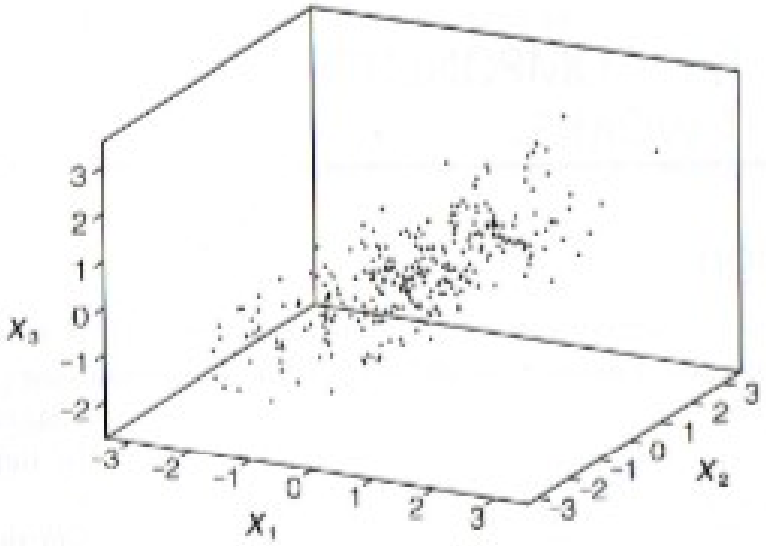
PCA: intuizione grafica

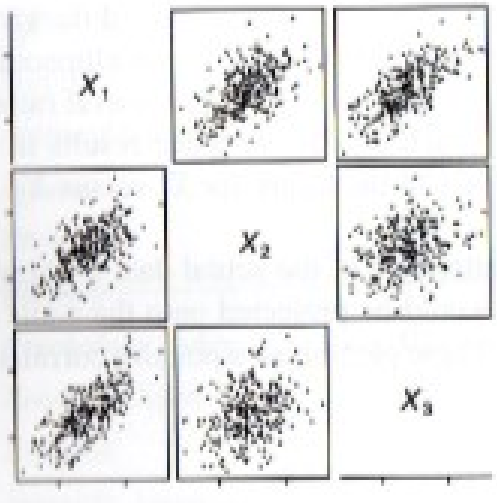
- Matrice di correlazione per tre variabili che sono state standardizzate

Correlation matrix for X_1, X_2, and X_3

	X_1	X_2	X_3
X_1	1.000	0.562	0.704
X_2	0.562	1.000	0.304
X_3	0.704	0.304	1.000
	$\text{var}(X_1) = 1.00$	$\text{var}(X_2) = 1.00$	$\text{var}(X_3) = 1.00$

PCA: intuizione grafica





• per analizzare i dati graficamente è possibile riportarli su grafico

1. tridimensionale nel nostro caso in cui abbiamo tre variabili

2. scatterplot per coppie di variabili = per ovviare alla difficoltà n-dimensioni

PCA: struttura

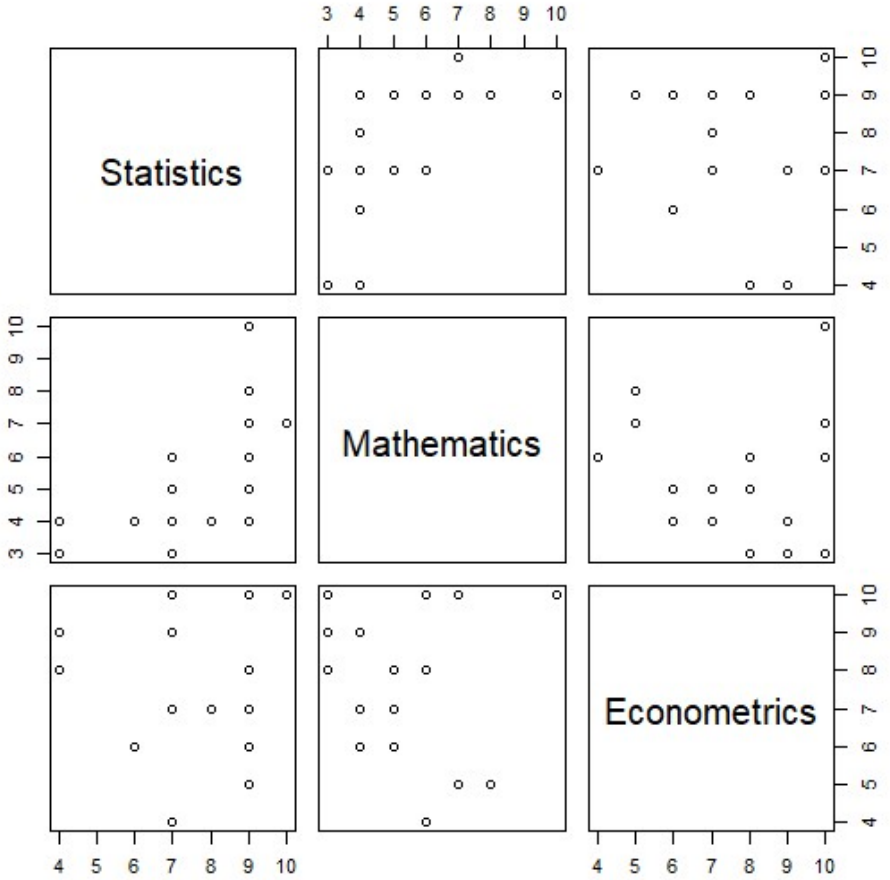
esercitazione in R per la comprensione dei risultati e la definizione della struttura

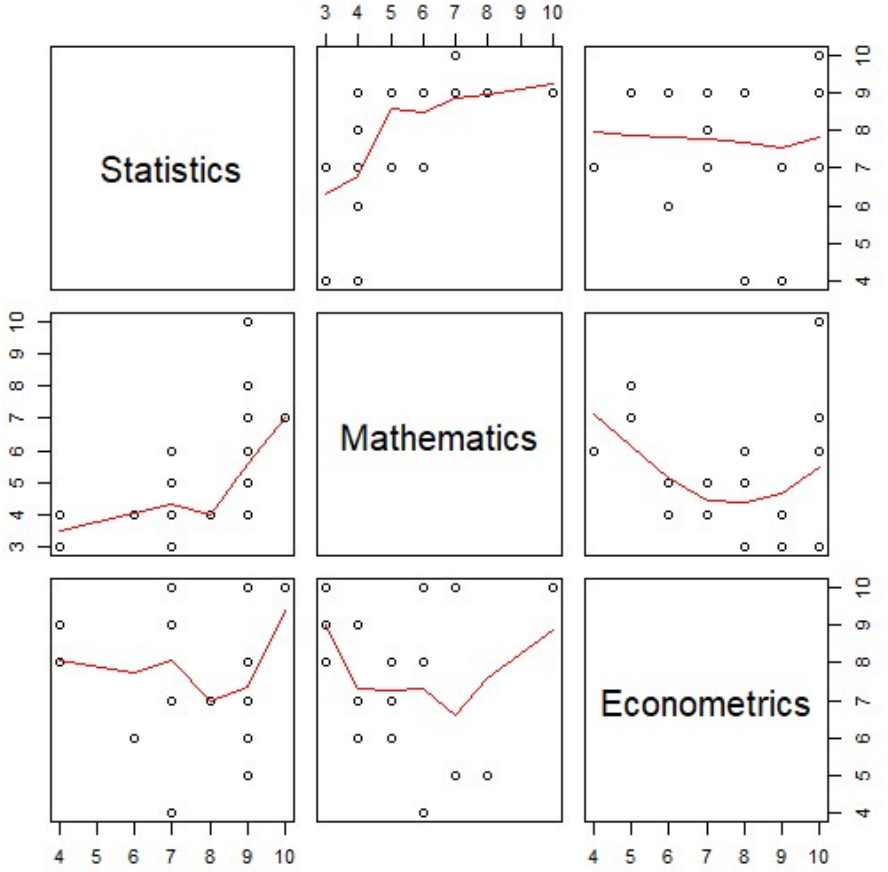
1. STRUTTURA DEL DATABASE

```
> voti
   Student Statistics Mathematics Econometrics
1        1          9           5            8
2        2          4           3            8
3        3          9           6            8
4        4          9           5            8
5        5          9          10           10
6        6          7           3           10
7        7          7           5            7
8        8          7           6            4
9        9          4           4            9
10      10          7           4            9
11      11          6           4            6
12      12          7           6           10
13      13          9           7            5
14      14          8           4            7
15      15         10           7           10
16      16          9           8            5
17      17          9           4            7
18      18          6           4            6
19      19          9           5            6
20      20          7           3            9
> str(voti)
'data.frame':   20 obs. of  4 variables:
 $ Student     : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Statistics  : int  9 4 9 9 9 7 7 7 4 7 ...
 $ Mathematics : int  5 3 6 5 10 3 5 6 4 4 ...
 $ Econometrics: int  8 8 8 8 10 10 7 4 9 9 ...
```


Eventuale riduzione del database di interesse

2. VISUALIZZAZIONE GRAFICA





Commento al grafico: quali relazioni possiamo notare? Quali variabili coinvolgono?

3. ANALISI DELLE CORRELAZIONI TRA VARABILI

```
> round(res,2)
             Statistics Mathematics Econometrics
Statistics         1.00        0.58        -0.05
Mathematics        0.58        1.00        -0.06
Econometrics      -0.05       -0.06         1.00
```

COMMENTO ALLA MATRICE DELLE CORRELAZIONI

Ha senso procedere con una analisi per componenti principali?

4. La PCA

Opzione importante da impostare: variabili standardizzate
pca=prcomp(dataset, scale=TRUE)

```
Standard deviations (1, .., p=3):
[1] 1.2629141 0.9945984 0.6448426

Rotation (n x k) = (3 x 3):
                    PC1         PC2           PC3
Statistics   -0.7004326 0.09961408  0.706732738
Mathematics  -0.7011109 0.08926511 -0.707442795
Econometrics  0.1335578 0.99101401 -0.007316321
```

Deviazione standard delle componenti

Pesi fattoriali (correlazione tra un fattore e quella variabile)

$Y_i = b_1X_1 + b_2X_2 + \ldots + b_nX_n + E_i$, *dove bi è il peso fattoriale*

Determinazione delle componenti principali

L'ACP è una metodologia statistica multivariata che, partendo da una matrice dei dati nxp con variabili quantitative, consente di sostituire alle p variabili (tra loro correlate) un nuovo insieme di variabili artificiali dette **COMPONENTI PRINCIPALI** (**CP**) che:

1. sono tra loro **INCORRELATE** (**ORTOGONALI**);
2. sono **elencate in ordine decrescente rispetto alla loro varianza.**

- La **prima CP Y_1** è la **COMBINAZIONE LINEARE** delle p variabili di partenza avente **MAX VARIANZA**.

- La **seconda CP Y_2** è la **COMBINAZIONE LINEARE** delle p variabili di partenza con **VARIANZA IMMEDIATAMENTE INFERIORE**, soggetta al vincolo di essere **ORTOGONALE** alla CP precedente.

- La **terza CP Y_3**.....etc....

Se le p variabili sono **FORTEMENTE CORRELATE**, un numero **k<<p** di CP tiene conto di una **ELEVATA QUOTA DI VARIANZA TOTALE**.
Quindi, possiamo considerare solo tali k CP, **trascurando le restanti p-k**, ottenendo una **SENSIBILE PARSIMONIA** nella descrizione dei dati.

5. QUANTE COMPONENTI ESTRAIAMO? Non vi è metodo univoco

1

Metodo della varianza comune cumulata spiegata dai componenti >70%

2

Metodo degli autovalori (o metodo di Kaiser 1960) = si tengono tutti i componenti con autovalori >=1

3

Criterio grafico dello screeplot: teniamo tutti i componenti fino al punto di "gomito" del grafico

5. QUANTI COMPONENTI TRATTENIAMO?

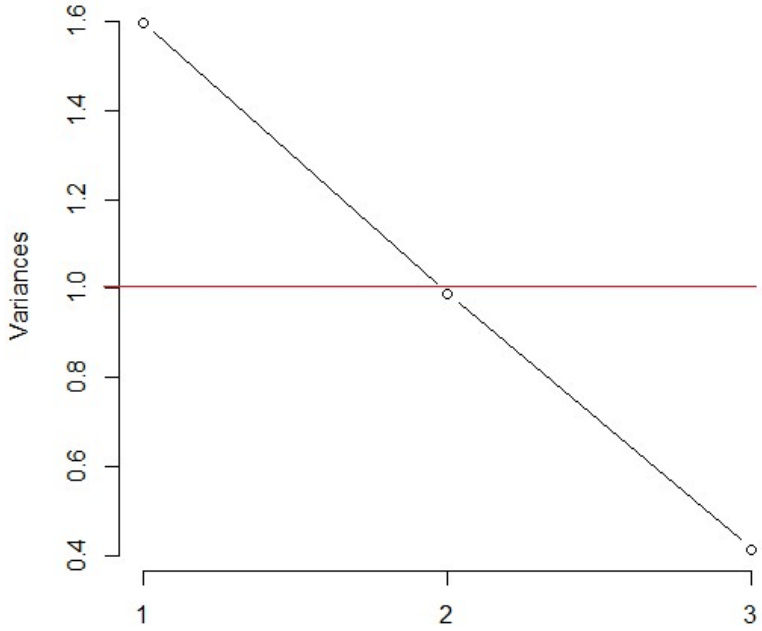
1

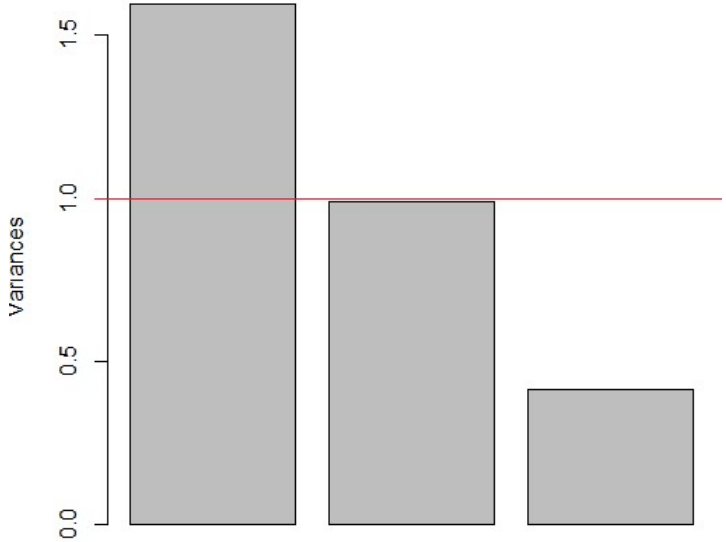
```
> summary(pcal)
Importance of components%s:
                          PC1    PC2    PC3
Standard deviation     1.2629 0.9946 0.6448
Proportion of Variance 0.5316 0.3297 0.1386
Cumulative Proportion  0.5316 0.8614 1.0000
```

2

```
> e_values=pcal$sdev^2
> e_values
[1] 1.5949521 0.9892259 0.4158220
```

3





3

Osservazione 1

Non sempre l'andamento del grafico fornisce una risposta univoca, poiché la diminuzione degli autovalori può essere graduale, senza salti evidenti.

Osservazione 2

Alcuni autori suggeriscono di **escludere** tra le CP scelte quelle sul gomito (Harman, 1976). Altri suggeriscono di **includere** tra le CP scelte quelle sul gomito (Cattell, 1966).

7. CORRELAZIONE TRA LE VARIABILI OROGINARIE E LE COMPONENTI

```
> cor.pcal=cor(voti,pcal$scores)
> cor.pcal
             Statistics Mathematics Econometrics
Statistics    1.00000000  0.58414703  -0.05369974
Mathematics   0.58414703  1.00000000  -0.05968737
Econometrics -0.05369974 -0.05968737   1.00000000
```

COSA RACCONTA LA MATRICE DI CORRELAZIONE?

8. IDENTIFICARE I VALORI FINALI (SCORES DELLE COMPONENTI) = LE NUOVE VARIABILI CREATE

```
> yscores
              PC1        PC2        PC3
 [1,] -0.4905158  0.2893869  0.6394995
 [2,]  2.3444694 -0.1022851 -0.6614091
 [3,] -0.8769271  0.3385846  0.2495984
 [4,] -0.4905158  0.2893869  0.6394995
 [5,] -2.2779323  1.6086209 -1.3179294
 [6,]  1.2518119  1.1469264  0.5790940
 [7,]  0.2620292 -0.3645465 -0.1888231
 [8,] -0.3413423 -1.9252171 -0.5668391
 [9,]  2.0303781  0.4835354 -1.0552719
[10,]  0.7930806  0.6595013  0.1931546
[11,]  0.9885530 -1.0090223 -0.2111025
[12,]  0.0925780  1.2945195 -0.5906093
[13,] -1.4802986 -1.2220860 -0.1284176
[14,]  0.2360080 -0.3550889  0.6172202
[15,] -1.5311309  1.5196831  0.2679161
[16,] -1.8667099 -1.1728883 -0.5183187
[17,] -0.1764245 -0.2964336  1.0333623
[18,]  0.9885530 -1.0090223 -0.2111025
[19,] -0.6351559 -0.7838587  0.6474229
[20,]  1.1794919  0.6103036  0.5830557
```

```
> finaly=yscores[,1:2]
> finaly
              PC1        PC2
 [1,] -0.4905158  0.2893869
 [2,]  2.3444694 -0.1022851
 [3,] -0.8769271  0.3385846
 [4,] -0.4905158  0.2893869
 [5,] -2.2779323  1.6086209
 [6,]  1.2518119  1.1469264
 [7,]  0.2620292 -0.3645465
 [8,] -0.3413423 -1.9252171
 [9,]  2.0303781  0.4835354
[10,]  0.7930806  0.6595013
[11,]  0.9885530 -1.0090223
[12,]  0.0925780  1.2945195
[13,] -1.4802986 -1.2220860
[14,]  0.2360080 -0.3550889
[15,] -1.5311309  1.5196831
[16,] -1.8667099 -1.1728883
[17,] -0.1764245 -0.2964336
[18,]  0.9885530 -1.0090223
[19,] -0.6351559 -0.7838587
[20,]  1.1794919  0.6103036
```

Questi valori sono quelli che andranno a costituire le nuove variabili inserite nel nostro database. In questo esempio: riduzione da 3 variabili originarie a 2 variabili (componenti) estratte

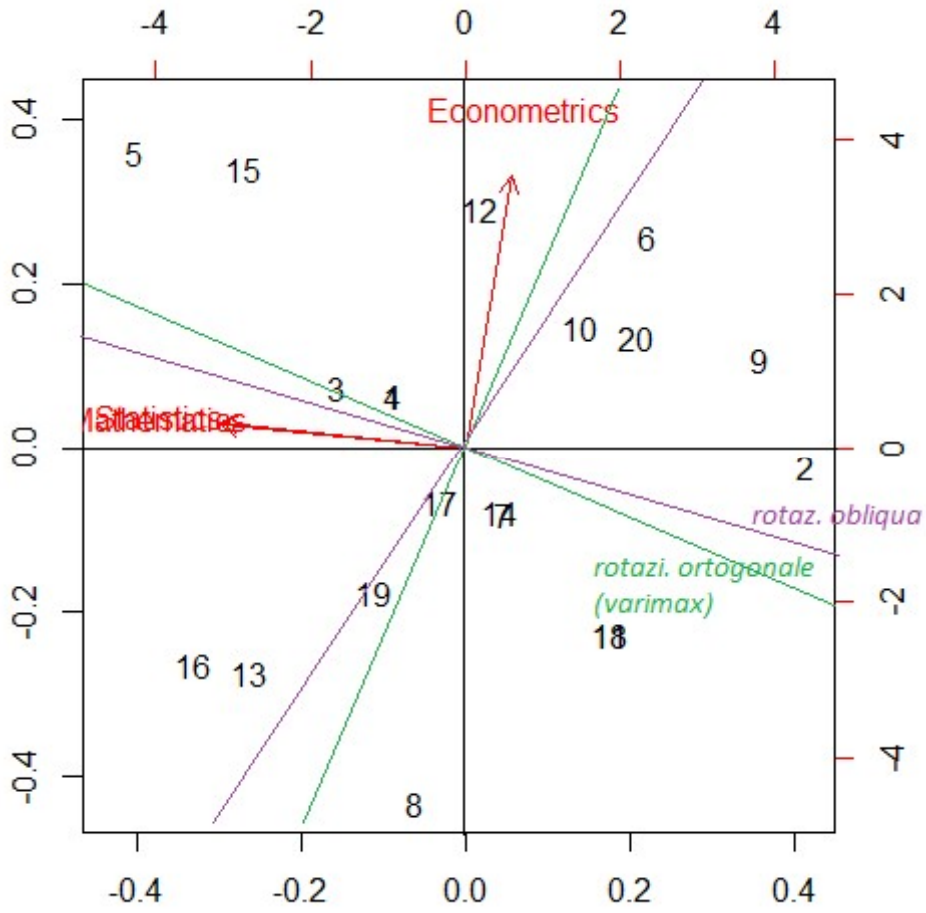
9. MIGLIORARE L'INTERPRETAZIONE CON LA ROTAZIONE DEGLI ASSI

Rotazione ortogonale

```
Loadings:
      PC1    PC2    PC3
 [1,]         0.748  0.417
 [2,]  1.456 -1.909 -0.424
 [3,] -0.354  0.905
 [4,]         0.748  0.417
 [5,] -0.994  2.115 -2.013
 [6,]  1.791
 [7,]        -0.478
 [8,] -1.619 -1.188  0.332
 [9,]  1.491 -1.462 -1.053
[10,]  1.045
[11,]        -1.394  0.312
[12,]  0.725  0.548 -1.099
[13,] -1.881  0.195  0.352
[14,]  0.115 -0.168  0.722
[15,]         2.110 -0.516
[16,] -2.237  0.353
[17,]         0.301  1.047
[18,]        -1.394  0.312
[19,] -0.780  0.169  0.895
[20,]  1.400 -0.215  0.311
```

Rotazione obliqua

```
Loadings:
      PC1    PC2    PC3
 [1,]  0.172  0.615  0.363
 [2,]  1.144 -2.122  0.226
 [3,] -0.216  1.051 -0.246
 [4,]  0.172  0.615  0.363
 [5,] -0.921  3.588 -3.219
 [6,]  1.869 -0.210  0.500
 [7,] -0.171 -0.527
 [8,] -1.825 -1.280  0.289
 [9,]  1.178 -1.289 -0.651
[10,]  1.060 -0.170  0.161
[11,] -0.167 -1.732  0.697
[12,]  0.707  1.098 -1.326
[13,] -1.869  0.298
[14,]  0.176 -0.590  0.970
[15,]  0.201  2.653 -1.122
[16,] -2.256  0.734 -0.656
[17,]  0.137 -0.217  1.240
[18,] -0.167 -1.732  0.697
[19,] -0.672 -0.177  0.901
[20,]  1.447 -0.606  0.769
```



Questi valori sono quelli che andranno a costituire le nuove variabili inserite nel nostro database. In questo esempio: riduzione da 3 variabili originarie a 2 variabili (componenti) estratte

CONCETTI RIASSUNTIVI

APC: riassunto dei concetti principali

- Scopo primario di PCA: riduzione del numero di variabili nel database (rappresentanti altrettante caratteristiche del fenomeno analizzato) in alcune variabili latenti (**feature reduction**).
- Questo avviene tramite trasformazione lineare = proietta le variabili originarie in un nuovo sistema cartesiano
  - Nel sistema cartesiano: la nuova variabile con maggiore varianza viene proiettata sul primo asse; la seconda (per dimensione della varianza) sul secondo asse – e via via.
- Riduzione ← selezione delle principali tra le nuove variabili dette componenti
  - Si definiscono principali in termini di varianza
- Sono gli stessi dati che determinano i vettori di trasformazione

APC: riassunto dei concetti principali

Identificazione dei componenti: presentata tecnica che parte dalla matrice dei coefficienti di correlazione (R)

Ci sono altre tecniche per calcolare le componenti:

1) Assumendo che a ciascuna delle variabili originarie venga sottratta la loro media e pertanto la nuova variabile (X_i) abbia media nulla,

$$\mathbf{w}_1 = \arg\max_{\|\mathbf{w}\|=1} E\left\{\left(\mathbf{w}^T\mathbf{x}\right)^2\right\}$$

(Dove arg max indica l'insieme degli argomenti *w* in cui è raggiunto il massimo.) Con i primi (k-1) componenti, il k-esimo componente può essere trovato sottraendo i primi (k-1) componenti principali a *X*

$$\hat{\mathbf{x}}_k = \mathbf{x} - \sum_{i=1}^{k-1} \mathbf{w}_i \mathbf{w}_i^T \mathbf{x}$$

e sostituendo questo

$$\mathbf{w}_k = \arg\max_{\|\mathbf{w}\|=1} E\left\{\left(\mathbf{w}^T\hat{\mathbf{x}}_{k-1}\right)^2\right\}$$

2) Utilizzando la matrice delle covarianze di x.

APC: riassunto dei concetti principali

Come si procede teoricamente con **l'individuazione delle componenti principali**:

- si identificano gli **autovalori della matrice di correlazione** (R)
- si ottengono tanti autovalori quante sono le variabili esplicative (xi)
- utilizzando la matrice R, si identifica **l'autovalore relativo alla prima componente** principale (ovvero con varianza max)
- **l'autovalore con il maggiore valore corrisponde alla varianza della componente pri**ncipale.
- in ordine decrescente il secondo autovalore è la varianza della seconda componente principale (e via via per n autovalori).
- per ciascun autovalore viene calcolato l'autovettore corrispondente: vettore riga dei coefficienti che moltiplicano le variabili originarie (K) x_i nella combinazione lineare per ottenere le nuove variabili (o componenti, talvolta indicate con W_i). Tali coefficienti sono definiti pesi fattoriali (*factor loadings*)
- la **matrice di autovettori viene definita matrice di rotazione V**.
- eseguendo l'operazione V=W*X si possono trovare le **coordinate di ciascun punto** nel nuovo spazio vettoriale
- le coordinate per ciascun punto relative alle componenti principali permettono di ottenere il grafico: esso permette di vedere quali dati sono simili tra loro, ovvero quali si stanno muovendo nella stessa direzione

APC: riassunto dei concetti principali

Come si procede con l'individuazione delle componenti principali (continua):

- gli elementi di autovettore colonna corrispondente ad un autovalore esprimono quindi il legame tra le variabili di partenza e la componente considerata attraverso dei pesi. **Il numero di variabili latenti da considerare come componenti principali si fonda sulla grandezza relativa di un autovalore rispetto agli altri**

- **Matrice dei fattori**: la costruiamo elencando per riga le variabili originarie e per colonna le variabili latenti; con valori da 0 a 1 ogni valore ci dice quanto le variabili incidano sui fattori.

- **La matrice del punteggio** (valori) fattoriale ha la stessa struttura, ma ci dice quanto le variabili originarie hanno pesato sulla determinazione della grandezza di quelle latenti

APC: riassunto dei concetti principali

ESEMPIO RIASSUNTIVO:

- Facciamo una simulazione. Poniamo di disporre di un'indagine che ci riporta per 10 soggetti: voto medio (da 0 a 33), intelligenza (da 0 a 10), media ore studiate in un giorno e zona d'origine (che varia da 1 a 3). Standardizziamo i valori con la formula:
- $z=(Xi-E(X))/SD$
- (con "E(x)" che è X medio).
- **Dopo di che calcoliamo la matrice dei coefficienti di correlazione che sarà:**

		Zscore (VotoMedio)	Zscore (Intelligenza)	Zscore (Provenienza)	Zscore (OreMed Studio)
Correlation	Zscore(VotoMedio)	1,000	,600	-,838	,788
	Zscore(Intelligenza)	,600	1,000	-,222	,022
	Zscore(Provenienza)	-,838	-,222	1,000	-,918
	Zscore(OreMedStudio)	,788	,022	-,918	1,000

- Chiaramente la diagonale principale è composta da valori uguali ad 1 (il coefficiente di correlazione di una variabile con se stessa deve dare necessariamente questo valore).
- È pure una matrice simmetrica (il coefficiente di correlazione tra la variabile "x" e la variabile "y" sarà uguale a quello tra "y" e "x"): vediamo come ci sia un forte legame tra voto, media ore studio e intelligenza.

APC: riassunto dei concetti principali

ESEMPIO RIASSUNTIVO:

Studiamo allora gli autovalori (eigenvalues) e quanto spiegano:

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,828	70,708	70,708	2,828	70,708	70,708
2	1,070	26,755	97,463	1,070	26,755	97,463
3	,084	2,088	99,551			
4	,018	,449	100,000			

Abbiamo posto gli autovalori più alti per primi, e, come detto, il loro rapporto con la somma degli autovalori ci dà la varianza che spiegano. Abbiamo selezionato (arbitrariamente) solo quelli che hanno valore maggiore di 1: i più significativi, che ci spiegano il 70,708% e il 26,755% rispettivamente.

Guardiamo ora alla **matrice delle componenti principali**:

	Component	
	1	2
Zscore(VotoMedio)	,966	,204
Zscore(Intelligenza)	,442	,894
Zscore(Provenienza)	-,947	,228
Zscore(OreMedStudio)	,897	-,420

Il fattore 1 (che, facendo una congettura, si potrebbe chiamare bravura) pesa dunque fortemente sul voto medio. Sembrerebbe pure che pesi in maniera negativa sulla variabile della zona di origine (chiaramente questa affermazione non avrebbe senso perché invertiremmo il nesso di causalità, spetta infatti allo statistico saper dare una spiegazione e una lettura sensate).

ACP: summary

ACP – i concetti principali

- **Goals**:
  - Reduce the dimensionality of the dataset
  - Detect new informative variables which can replace the observed original variables
  - Use a graphical representation of data to get some preliminary information previous to a following analysis
  - Reduce the number of explanatory variables in a multiple regression model in the presence of multicollinearity

- Result:
  - The original variability of the observed response variables $X_1,\ldots,X_k$ (which usually are correlated between each other) can be described by new uncorrelated variables $Y_1,\ldots,Y_k$, which are linear combinations of the original observed variables $X_1,\ldots,X_k$
  - The variables $Y_1,\ldots,Y_k$ are sorted according to the degree of importance, i.e. Y_1 is the variable which «explains» the greatest variance
  - $Y_1,\ldots,Y_k$, are called PRINCIPAL COMPONENTS

- Assumptions
  - $X_1,\ldots,X_k$ follow a (multivariate) distributionnwith mean vector μ and covariance matrix Σ
  - The values in μ and Σ are finite;
  - The rank of Σ is $q < k$;
  - The dataset is given by the $n \times k$ matrix $[x_{ij}]$, $i=1,..,n;\ j=1,\ldots,q$

- S and R matrices:

- covariance matrix $S=[s_{ij}]$ which includes the necessary information for PCA
  - A suitable estimate of Σ is provided by the sampling
- As a matter of fact the information for PCA is usually provided by the matrix of sampling correlations $R=[r_{ij}]$, especially when the magnitudes, the units of measurement or the variabilities of the original variables are very much different

- Principal Components (PC) extraction from R is equivalent to PC extraction from S after standardization of the original variables

- First Principal Comporert:

1. $Y_1 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{k1}X_k$

2. Detect the values $a_{11}, \cdots, a_{k1}$ which maximize the variance of Y_1, formally:
   - fird $a_{11}^*, a_{21}^*, \cdots, a_{k1}^*$ such that
     1. $\max[Var(Y_1)] = Var(a_{11}^*X_1 + a_{21}^*X_2 + \cdots + a_{k1}^*X_k) = \sum_{j,r} a_{j1}^* a_{r1}^* s_{jr}$
     2. $\sum_j (a_{j1}^*)^2 = 1$
   - $\lambda_1 = \max[Var(Y_1)] = \sum_{j,r} a_{j1}^* a_{r1}^* s_{jr}$ max eigenvalue of S
   - $(a_{11}^*, \cdots, a_{k1}^*)'$ eigenvector of S which corresponds to λ_1

- Second Principal Component:

1. $Y_2 = a_{12}X_1 + a_{22}X_2 + \cdots + a_{k2}X_k$

2. Detect the values $a_{12}, \cdots, a_{k2}$ which maximize the variance of Y_2, formally:
   - find $a_{12}^*, a_{22}^*, \cdots, a_{k2}^*$ such that
     1. $\max[Var(Y_2)] = Var(a_{12}^*X_1 + a_{22}^*X_2 + \cdots + a_{k2}^*X_k) = \sum_{j,r} a_{j2}^* a_{r2}^* s_{jr}$
     2. $\sum_j (a_{j2}^*)^2 = 1$
     3. $\sum_j a_{j1}^* a_{j2}^* = 0$
   - $\lambda_2 = \max[Var(Y_2)] = \sum_{j,r} a_{j2}^* a_{r2}^* s_{jr}$ 2^{nd} max eigenvalue of S
   - $(a_{12}^*, \cdots, a_{k2}^*)'$ eigenvector of S which corresponds to λ_2

- Following Principal Components: same iterative procedure ..

- Main differences between FA and PCA:

1. In FA we distinguish between common factors and unique factors while in PCA we have only common factors

2. In FA the communality is unknown and must be estimated while in PCA it is equal to 1

3. In FA the number of common factors is less than the number of observed original variables ($q<k$) while in PCA the number of components is equal to the number of observed orginale variables ($q=k$)

4. In FA the estimation of the communality follows an iterative method while PCA does not include iterations

ESERCIZI IN R

3 Esercizi in R

Problem 1 – Passito

- Eseguire una PCA sulle17 variabili risposta che rappresentano il questionario sulle abitudini, il comportamento e le preferenze dei consumatori di vino (dalla variabile LIKE_WINE alla variabile PRICE) per identificare *q<17* nuove variabili che “spiegano” I dati

Problem 2 – centro commerciale

- Eseguire una PCA sulle 5 variabili risposta per individuare $q<5$ nuove variabili che “spiegano” i dati

Problem 3 – abitudini alimentari

- Eseguire una PCA sulle12 variabili risposta osservate (da *Alcoholic.Beverages a Milk*) per individuare *q<12* nuove variabili che "spiegano" i dati