Analisi in Componenti Principali (ACP)

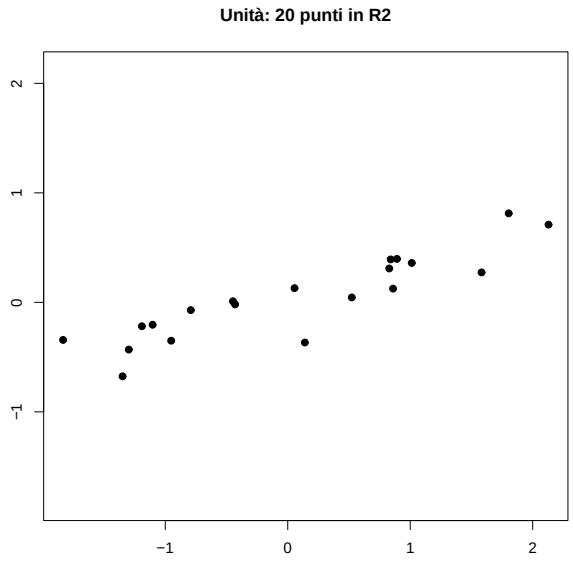
Ho una matrice di dati X (n osservazioni, k variabili quantitative $V_1, \cdots, V_k$).

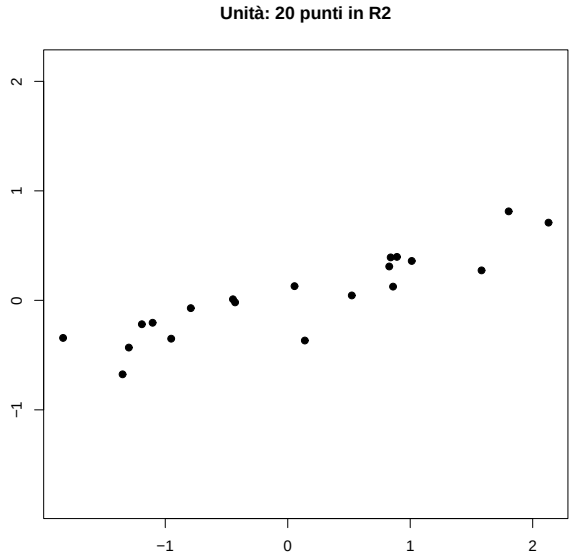
Ogni osservazione può essere rappresentata come un punto in $\mathbb{R}^k$, quindi ho n punti in $\mathbb{R}^k$.

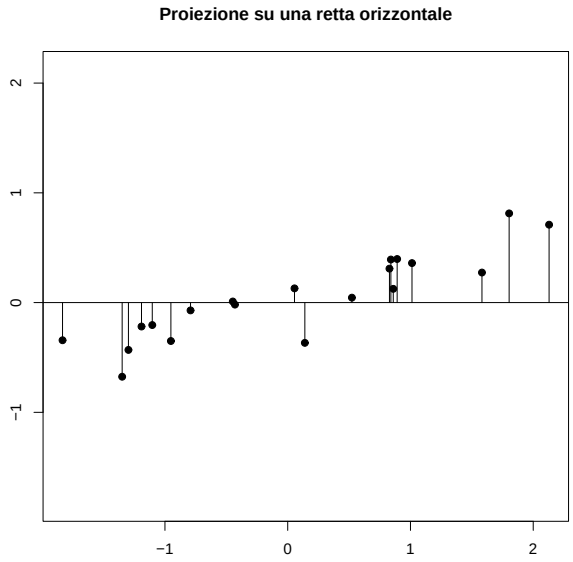
Voglio proiettare i dati
su uno spazio di dimensione inferiore ($\mathbb{R}^p, p < k$)
cercando di perdere meno informazione possibile.

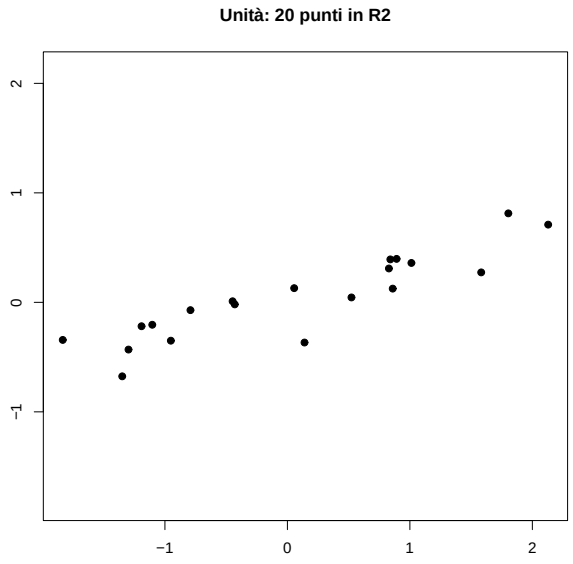
informazione=variabilità (qualcuno la chiama inerzia)

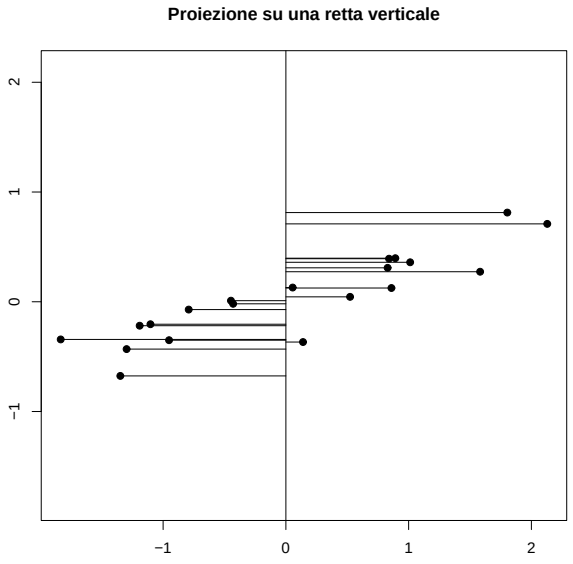
$$I_{tot} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij} - \bar{x}_j)^2 = \sum_{j=1}^{k}\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 = \sum_{j=1}^{k} Var(V_j).$$

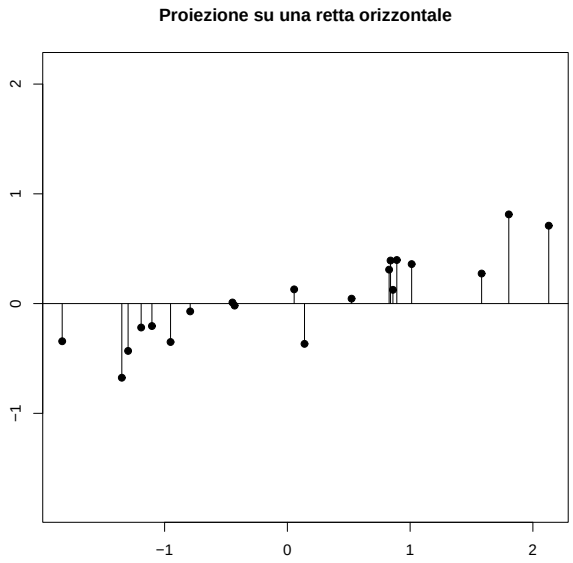

Unità: 20 punti in R2
2
1
0
−1
−1
0
1
2

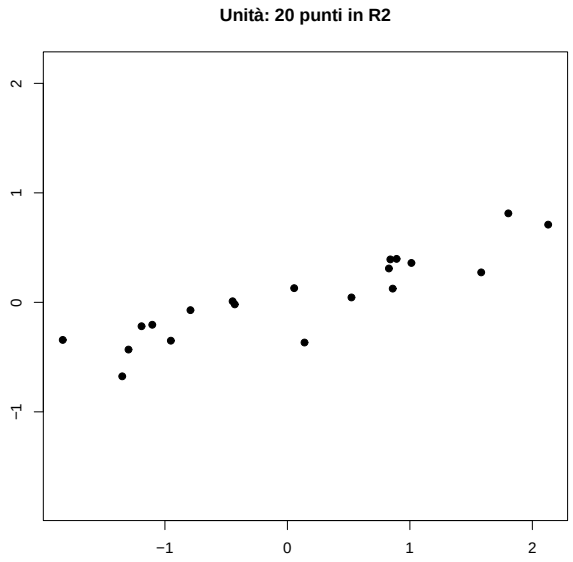

Unità: 20 punti in R2
2
1
0
−1
−1
0
1
2

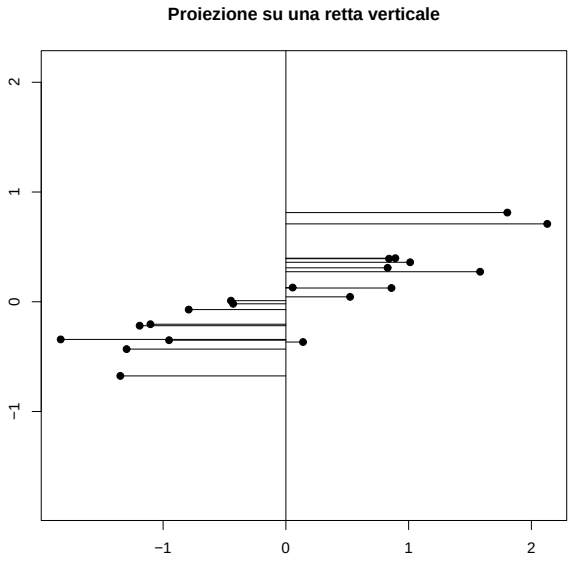

Proiezione su una retta orizzontale
2
1
0
−1
−1
0
1
2

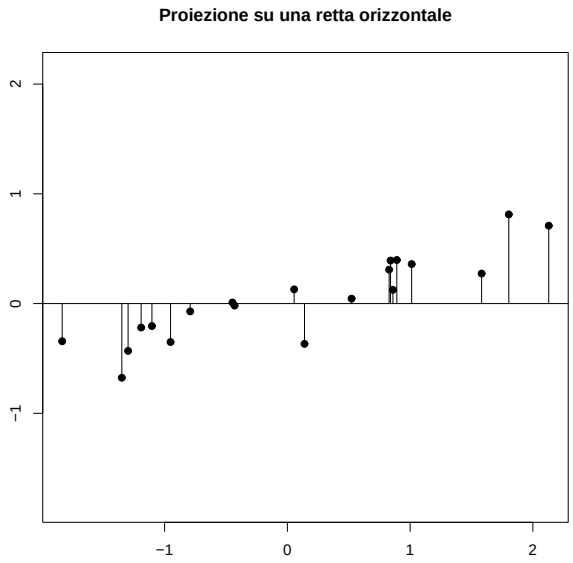

Unità: 20 punti in R2
2
1
0
−1
−1
0
1
2

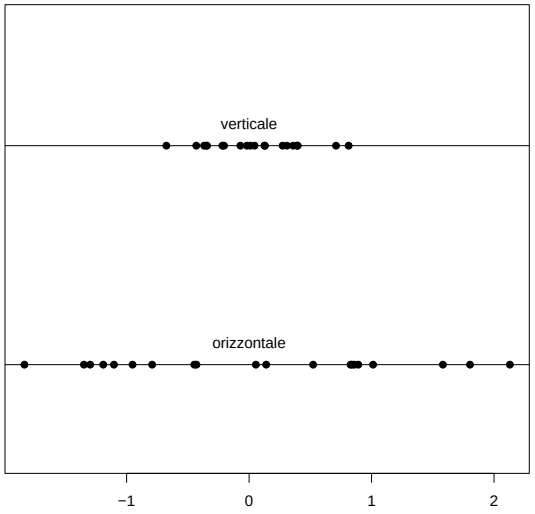

Proiezione su una retta verticale
2
1
0
−1
−1
0
1
2


Proiezione su una retta orizzontale
2
1
0
−1
−1
0
1
2


Unità: 20 punti in R2
2
1
0
−1
−1
0
1
2


Proiezione su una retta verticale
2
1
0
−1
−1
0
1
2


Proiezione su una retta orizzontale
2
1
0
−1
−1
0
1
2


Confronto tra le due proiezioni
verticale
orizzontale
−1
0
1
2

Cos'è l'ACP?

una tecnica di trasformazione di variabili:

partendo dalle variabili originarie $V_1, \cdots, V_k$

costruisco le nuove variabili $C_1, \cdots, C_k$ (poi non le uso tutte)

- le nuove variabili sono combinazioni lineari delle vecchie: $C_i = a_{i0} + \sum_{j=1}^{k} a_{ij} V_j$;
- sono ortogonali tra loro;
- sono costruite in modo sequenziale come segue:
  - la prima variabile C_1 spiega il massimo possibile di variabilità
  - la seconda C_2 deve essere ortogonale a C_1 e spiegare il massimo della variabilità che rimane
  - e così via

Che vuol dire spiega il massimo?

- La variabile C_1 è quella di varianza massima tra tutte le possibili combinazioni lineari $C = a_0 + \sum_{j=1}^{k} a_j V_j$
- La variabile C_2 è quella di varianza massima tra tutte le possibili combinazioni lineari $C = a_0 + \sum_{j=1}^{k} a_j V_j$ perpendicolari a C_1
- E così via.

Si può dimostrare che

trovare variabili $C_1, \cdots C_k$ con queste caratteristiche

- È possibile!!!
- È facile!!!
- le nuove variabili conservano tutta l'informazione
  $I_{tot} = \sum_{j=1}^{k} Var(V_j) = \sum_{j=1}^{k} Var(C_j)$.

Come si fa?

- Si prende la matrice di varianze e covarianze
  $\Sigma = \frac{1}{n}(X - \bar{X})'(X - \bar{X})$
- si trovano autovalori di Σ e si ordinano $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$.
- si trova l'autovettore[a] $(a_{11}, \cdots, a_{1j}, \cdots, a_{1k})$ corrispondente a λ_1, quindi si costruisce $C_1 = \sum_{j=1}^{k} a_{1j}(V_j - \bar{V}_j)$
- le altre componenti principali $C_2, \cdots, C_k$ si costruiscono usando gli autovettori (normalizzati) corrispondenti a $\lambda_2, \cdots, \lambda_k$.

Si può vedere che $Var(C_j) = \lambda_j$,
questo aiuta a decidere *quante* componenti principali considerare.

[a] attenzione, l'autovettore deve essere normalizzato, nel senso che $\sum_{j=1}^{k} a_{1j}^2 = 1$

Perché funziona?

Indichiamo con Λ la matrice che ha gli autovalori sulla diagonale e con Q la matrice che ha in ogni colonna un autovettore normalizzato

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_k \end{pmatrix} \qquad Q = \begin{pmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_k \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

Si può dimostrare che $\Sigma = Q\Lambda Q'$ e che $Q' = Q^{-1}$.

La matrice con le componenti principali è $Y = (X - \bar{X})Q$, per cui la matrice di covarianze di Y è

$$\frac{1}{n}Y'Y = \frac{1}{n}Q'(X - \bar{X})'(X - \bar{X})Q = Q'\Sigma Q = Q'Q\Lambda Q'Q = \Lambda$$

Esempio 1:

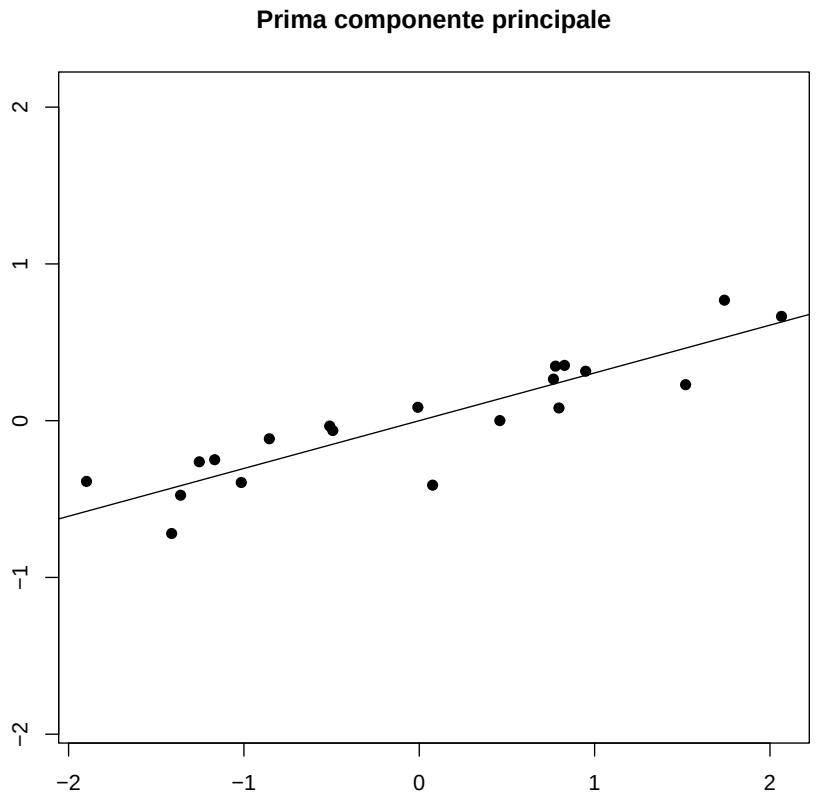
scegliamo la retta migliore su cui proiettare la nuvola di punti del grafico precedente

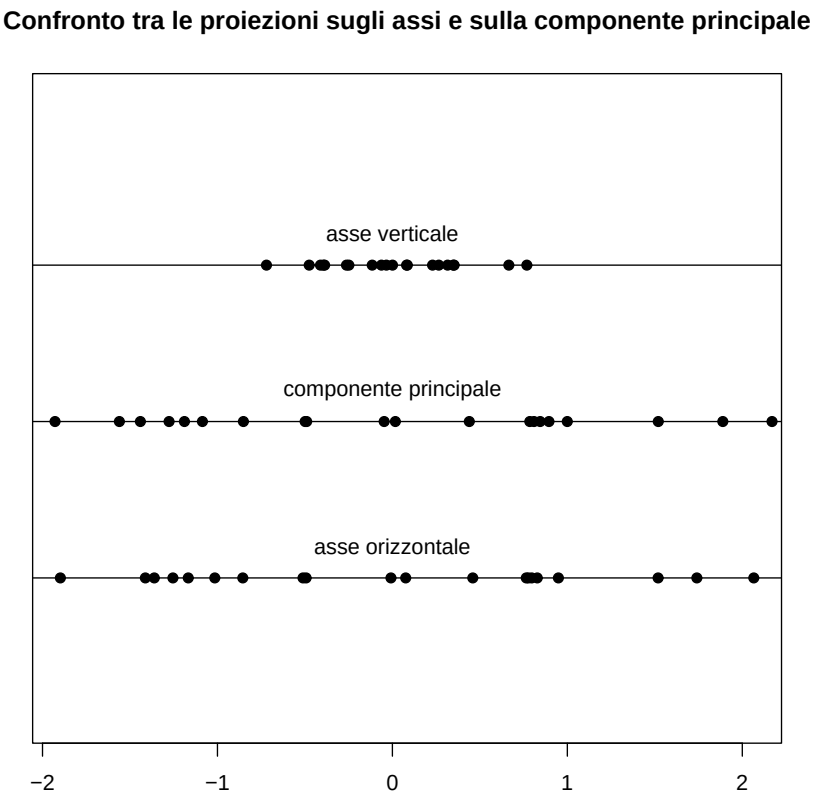
```
X <- read.table(file="es1.dati.txt")

n <- dim(X)[1]
k <- dim(X)[2]

# "centro" la matrice di dati X
X <- transform(X, V1=V1-mean(V1), V2=V2-mean(V2))

# calcolo la matrice di covarianza
# ATTENZIONE: R calcola la varianza e covarianza corrette
SIGMA <- cov(X)
SIGMA <- SIGMA*(n-1)/n
```

```
# trovo autovalori e autovettori di SIGMA
autotutto <- eigen(SIGMA)

# mi servono l'autovalore più grande e
# l'autovettore corrispondente
# ATTENZIONE: l'autovettore deve essere normalizzato
lambda <- autotutto$values
avett1 <- autotutto$vectors[,1]
avett1 <- avett1/sum(avett1^2)

# troviamo le coordinate rispetto alla nuova variabile
compprinc <- X$V1 * avett1[1] + X$V2 * avett1[2]
```

```
# la retta su cui proietto passa per l'origine e per avett1
intercetta <- 0
coeffang <- avett1[2]/avett1[1]

plot(X, xlim=range(X), ylim=range(X), xlab="", ylab="", pch=19)
abline(a=intercetta, b=coeffang)
title("Prima componente principale")
```

```
# confronto con le proiezioni fatte prima
win.graph()
plot(X[,1], X[,1]*0-1, xlim=range(X), ylim=range(X),
        xlab="", ylab="", yaxt="n", pch=19)
abline(h=-1)
text(x=0, y=-0.8, labels="asse orizzontale")
title("Confronto tra le proiezioni sugli assi
        e sulla componente principale")
points(X[,2], X[,2]*0+1, pch=19)
abline(h=1)
text(x=0, y=1.2, labels="asse verticale")

points(compprinc, compprinc*0, pch=19)
abline(h=0)
text(x=0, y=0.2, labels="componente principale")
```


Prima componente principale
2
1
0
−1
−2
−2
−1
0
1
2


Confronto tra le proiezioni sugli assi e sulla componente principale
asse verticale
componente principale
asse orizzontale
−2
−1
0
1
2

Attenzione:

- Orientamento: la direzione degli assi delle componenti principali è arbitraria (cambiando a_{ij} con $-a_{ij}$ non cambierebbe nulla in termini di informazione né di interpretazione dei risultati).
- Standardizzazione: Il risultato dell'ACP varia se lavoro con le variabili originarie (matrice Σ) oppure standardizzate (matrice di correlazione). Questa scelta è importante e dipende dallo scopo dell'analisi.
- Se ci sono correlazioni vicine a ± 1: le variabili di partenza V_j non sono linearmente indipendenti e gli autovalori positivi sono $r < k$. In questo caso i dati originari sono su uno spazio di dimensione r (minore di k).
- Se tutte le correlazioni sono vicine a 0: l'ACP è inutile perché ripropone le stesse variabili in ordine decrescente di varianza.

Esempio un po' più complesso

Dati: Percentuale di occupati nei diversi settori industriali nei paesi europei – anno 1979.

Agr	Agricoltura
Min	Ind. Mineraria
Man	Ind. Manufatturiera
PS	Settore Energetico
Con	Costruzioni
SI	Terziario
Fin	Settore Finanziario
SPS	Servizi Sociali
TC	Trasporti e Comunicazioni

Obiettivo: Verificare se la struttura occupazionale è legata ai diversi sistemi economico-politici presenti in Europa in quel periodo.

Le variabili sono espresse tutte nella stessa unità di misura (percentuale di occupati), ma hanno deviazioni standard molto diverse.

- se lavoriamo con la matrice di covarianza teniamo conto della diversa variabilità
- se lavoriamo sulla matrice di correlazione (variabili standardizzate) perdiamo quest'informazione e diamo pari importanza a tutti i settori

Siccome ci sono settori (ad esempio quello minerario) che pesano poco in termini di occupazione ma sono fortemente caratterizzanti il tipo di economia, decidiamo di lavorare con la matrice di covarianza

(spesso si fanno entrambe le analisi e poi si sceglie quella che sembra migliore)

ACP in R

```
# leggi i dati.
# ATTENZIONE1: la prima riga da' i nomi alle variabili
# ATTENZIONE2: la prima colonna dà i nomi alle unità
X <- read.table(file="europeanstat.data", header=TRUE, row.names=1)

# prima di cominciare diamo un'occhiata alle correlazioni
cor(X)

# facciamo l'ACP
# cor=TRUE perché lavoriamo con la matrice di correlazione
output <- princomp(X, cor=TRUE)

# vediamo quanto spiegano le componenti principali
summary(output)
```

```
# cerchiamo di interpretare le componenti principali
# guardando le correlazioni con le variabili originarie
# ci fermiamo al terzo decimale
correlazioni <- cor(X,output$scores)
correlazioni <- round(correlazioni, 3)

# rappresentiamo le unità sul piano delle prime
# due componenti principali
biplot(output)

# è un po' disordinato, quindi rifacciamo il grafico
# riducendo i nomi degli stati a 3 lettere

win.graph() # serve per aprire un'altra finestra grafica
biplot(output, xlabs=substr(row.names(X), 1,3))
```

L'output di R

`sdev:` deviazione standard delle componenti principali (radice degli autovalori)

`loadings:` matrice Q (le cui colonne sono gli autovettori)

`center:` medie delle variabili originarie

`scale:` deviazioni standard delle variabili originarie

`n.obs:` numero di unità

`scores:` coordinate delle unità nelle nuove variabili (solo se `scores = TRUE`)

`call:` il comando richiesto

Interpretiamo i risultati

Prima di iniziare: La somma per righe è sempre 100 (più o meno!!!). Quindi c'è una relazione lineare tra le variabili. Di fatto siamo già in un sottospazio di dimensione 8.

Matrice di correlazione: L'agricoltura è correlata negativamente con tutte le altre (tranne il minerario). In ogni caso le variabili sono correlate, ma non sembrano esserci correlazioni fortissime.

Autovalori: L'ultimo è praticamente 0, come era lecito aspettarsi perché c'è multicollinearità. I primi 2 sono oltre il 62% del totale, i primi 3 il 75%.

I e II componente principale

La **prima** ha una fortissima correlazione negativa con l'agricoltura, positiva o nulla con tutte le altre. Discrimina paesi con economia agricola (Turchia, Jugoslavia, in parte Grecia e Romania) da quelli con economia industriale.
La **seconda** è correlata positivamente con i Settori Terziario, Finanziario e servizi Sociali, negative le correlazioni con gli altri settori (quasi nullo con l'agricoltura). In sostanza sembra discriminare i "servizi" dall'"industria" (intesi in senso un po' ampio).

Osservando la rappresentazione delle unità sul piano, l'analisi riesce a vedere chiaramente la distinzione tra i paesi occidentali e quelli del Patto di Varsavia, con la Turchia e Yugoslavia isolate (erano "paesi non allineati").

Qualità di rappresentazione delle unità

Per valutare se ciascuna unità x_i (i-esima riga della matrice originaria) è ben rappresentata nello spazio di dimensione ridotta $C_{1,\cdots,p}$ (con $p < k$) si calcola il quadrato del coseno dell'angolo formato dall'unità e il sottospazio

$$\cos^2(x_i, C_{1,\cdots,p}) = \frac{\sum_{j=1}^{p} c_{ij}^2}{\sum_{j=1}^{k} c_{ij}^2}.$$

Una ulteriore indicazione sulla qualità di rappresentazione di ogni unità è la distanza tra unità e sottospazio (rapportata alla distanza media)

$$\frac{\sqrt{\sum_{j=p+1}^{k} c_{ij}^2}}{\sqrt{\sum_{j=p+1}^{k} \lambda_j}}$$

R non li dà, dobbiamo calcolarli "a mano"

```
# qualità di rappresentazione:
# calcoliamo i cosen quadri
coord2 <- output$scores^2
sommarighe <- apply(coord2, 1, sum)
cosquadri <- apply(coord2[,1:2], 1, sum)/sommarighe
sort(cosquadri) # li ordiniamo per leggerli meglio

# la distanza rapportata alla distanza media
autovalori <- output$sdev^2
denom <- sqrt(sum(autovalori[3:9]))
rappdist <- sqrt(apply(coord2[,3:9], 1, sum))/denom
sort(rappdist) # li ordiniamo per leggerli meglio
```

Nell'esempio

Tra i paesi con cosen quadro molto basso (Irlanda, Spagna, Italia, URRS)

l'Irlanda in fondo non è rappresentata troppo male (la distanza è più o meno nella media)

mentre la Spagna è rappresentata molto molto male.

Italia e URSS sono un po' intermedie, con cosen quadro migliore di Spagna e Irlanda ma con distanza abbastanza alta.

Attenzione alla Yugoslavia: anche se il cosen quadro è abbastanza buono, la distanza è al di fuori della norma (il boxplot la individua come outlier).

Estensioni:

È possibile considerare una funzione di distanza sullo spazio delle unità diversa dalla distanza euclidea standard. Questo comporta alcune differenze nella quantificazione dell'informazione totale e nella determinazione delle componenti principali. In questo corso non ce ne occuperemo.

Può accadere che le unità non abbiano tutte la stessa rilevanza. Molti software (SAS, SPSS, ...) permettono di dare pesi diversi alle varie unità.
In `R` non è previsto, per farlo bisogna costruire una nuova matrice di dati ripetendo ogni osservazione (riga di X) un numero di volte proporzionale al peso che le si vuole assegnare.