1. **ESEMPIO DI PCA**

A <- read.table ('clipboard', header=TRUE)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | cereali | tuberi | ortaggi | frutta | agrumi | vite | olivo |
| Abru | 3590855 | 1751220 | 5449426 | 708934 | 560 | 3365847 | 1294269 |
| Basi | 4363879 | 62500 | 3708553 | 1641121 | 1559906 | 390346 | 364225 |
| Cala | 1739823 | 662162 | 5519088 | 1452920 | 14683931 | 741594 | 10077833 |
| Camp | 3924660 | 3003977 | 10458024 | 7590865 | 704345 | 2496557 | 2405586 |
| EmRo | 28225968 | 2270305 | 23975740 | 15426392 | 0 | 8595036 | 69929 |
| FrVG | 9025287 | 205466 | 168715 | 394817 | 0 | 1548909 | 2274 |
| Lazi | 4623582 | 733404 | 6841368 | 2694107 | 71986 | 2642408 | 1716086 |
| Ligu | 18120 | 94820 | 230370 | 41257 | 4880 | 127165 | 180400 |
| Lomb | 32935561 | 290328 | 6959450 | 912570 | 0 | 1876437 | 45364 |
| Marc | 6804551 | 137124 | 1947801 | 256951 | 0 | 0 | 231802 |
| Moli | 2080439 | 136160 | 919430 | 215050 | 0 | 457307 | 449975 |
| Piem | 20721193 | 475128 | 2508111 | 4382869 | 0 | 3876660 | 824 |
| Pugl | 10020854 | 797250 | 28828207 | 2052407 | 2782495 | 18336000 | 12139400 |
| Sard | 1113011 | 469177 | 3242480 | 543111 | 853664 | 1329089 | 498888 |
| Sici | 8802313 | 2236681 | 11053669 | 3467058 | 18483078 | 10624445 | 3361878 |
| Tosc | 5384898 | 1167367 | 3288339 | 828962 | 955 | 4023110 | 1174819 |
| TrAA | 20535 | 219900 | 113826 | 15099290 | 0 | 1631673 | 10500 |
| Umbr | 7091903 | 77400 | 1725350 | 91299 | 0 | 1296170 | 409179 |
| ValA | 1665 | 22000 | 0 | 39450 | 0 | 45000 | 0 |
| Vene | 31121694 | 1233404 | 5441741 | 4449690 | 0 | 11464375 | 77865 |

**cereali tuberi ortaggi frutta agrumi vite olivo**

**Abru 3590855 1751220 5449426 708934 560 3365847 1294269**

**Basi 4363879 62500 3708553 1641121 1559906 390346 364225**

**Cala 1739823 662162 5519088 1452920 14683931 741594 10077833**

**Camp 3924660 3003977 10458024 7590865 704345 2496557 2405586**

**EmRo 28225968 2270305 23975740 15426392 0 8595036 69929**

**FrVG 9025287 205466 168715 394817 0 1548909 2274**

**Lazi 4623582 733404 6841368 2694107 71986 2642408 1716086**

**Ligu 18120 94820 230370 41257 4880 127165 180400**

**Lomb 32935561 290328 6959450 912570 0 1876437 45364**

**Marc 6804551 137124 1947801 256951 0 0 231802**

**Moli 2080439 136160 919430 215050 0 457307 449975**

**Piem 20721193 475128 2508111 4382869 0 3876660 824**

**Pugl 10020854 797250 28828207 2052407 2782495 18336000 12139400**

**Sard 1113011 469177 3242480 543111 853664 1329089 498888**

**Sici 8802313 2236681 11053669 3467058 18483078 10624445 3361878**

**Tosc 5384898 1167367 3288339 828962 955 4023110 1174819**

**TrAA 20535 219900 113826 15099290 0 1631673 10500**

**Umbr 7091903 77400 1725350 91299 0 1296170 409179**

**ValA 1665 22000 0 39450 0 45000 0**

**Vene 31121694 1233404 5441741 4449690 0 11464375 77865**

PCA = princomp(A)

biplot(PCA)



Qualche informazione è deducibile anche da questo disegno ma le frecce di lunghezza diversa segnalano che involontariamente abbiamo operato una scelta, quella di non standardizzare le grandezze, che di solito con PCA non ripaga. Non è obbligatorio standardizzare: ma se il nostro scopo è migliorare la visibilità, questa è un po’ compromessa dalle diverse unità di misura.

Standardizziamo.

AA =A

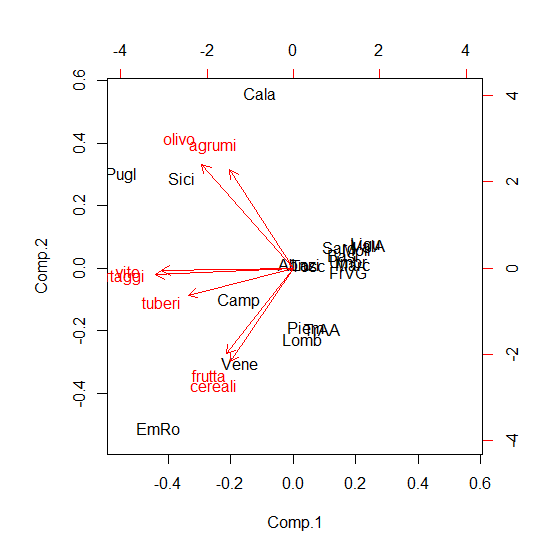
for ( i in 1:7) {

AA[,i] = ( A[,i] - mean(A[,i]) ) / ( sd(A[,i]) )

}

PCAA = princomp(AA)

biplot(PCAA)



Ora la leggibilità delle frecce è molto aumentata. Come in ogni esempio, si può ora fare un’analisi non matematica ma qualitativa delle associazioni tra regioni, tra frecce e tra regioni e frecce.

Possiamo poi chiederci se per caso le due componenti principali hanno un significato. A questo scopo visualizziamo anche la matrice dei loadings:

In realtà è più facile farsi delle idee col biplot, ad es. che la prima componente principale è quella della produttività in generale, la seconda differenzia tra colture più fredde o più calde.

**Esercizio per casa**. I) normalizzare la tabella al numero di abitanti. Si può usare questo vettore che offre una misura approssimata della popolazione regionale:

pop<-c(1.3,0.6,1.9,6,4.4,1.2,5.6,1.6,9.9,1.5,0.3,4.4,4.1,1.6,5,3.7,1,0.9,0.1,4.9)

ii) ripetere l’analisi PCA evidenziando alcune differenze nei risultati.

1. **AUTOVALORI, DEV ST, VARIANZE SPIEGATE**

plot(PCAA)



> PCAA$sdev

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7

1.6931828 1.2520744 0.9412341 0.7724487 0.6585050 0.4465577 0.3159122

> PCAA$sdev^2

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7

2.86686806 1.56769023 0.88592165 0.59667696 0.43362880 0.19941375 0.09980054

Sono i numeri visualizzati da plot.

summary(PCAA)

> summary(PCAA)

Importance of components:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7

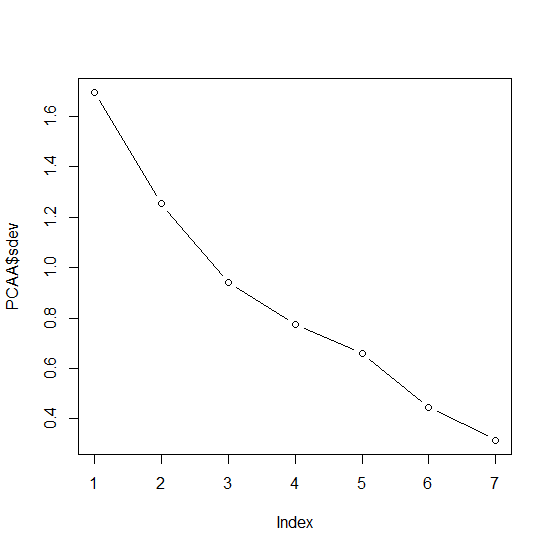
St dev 1.693183 1.2520744 0.9412341 0.77244868 0.65850498 0.44655767 0.3159122

Prop Var 0.431108 0.2357429 0.1332213 0.08972586 0.06520734 0.02998703 0.0150076

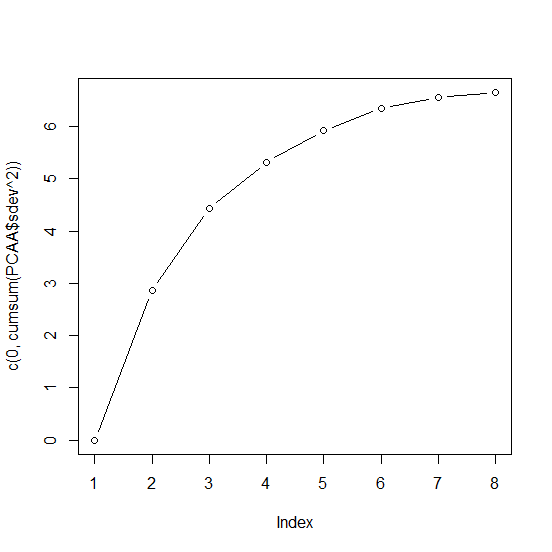
Cum Pro 0.431108 0.6668509 0.8000722 0.88979803 0.95500537 0.98499240 1.0000000

1. **ARRICCHIMENTO GRAFICO (diagnostica)**

plot(PCAA$sdev, type="b")



plot(c(0,cumsum(PCAA$sdev^2)), type="b")



varianza.cumulativa=c(0, 0.431108, 0.6668509, 0.8000722, 0.88979803, 0.95500537, 0.98499240, 1.0000000)

plot(varianza.cumulativa,type="b")



Questi disegni quantificano anche la bontà della riduzione.

1. **PUNTEGGI E CLASSIFICHE**

I punteggi di ogni regione rispetto a ciascuna componente principale si ottengono con:

predict(PCAA)

round(predict(PCAA),3)

Vediamo solo i punteggi rispetto alla prima componente principale:

punteggi = -predict(PCAA)[,1]

round(punteggi,3)

Vediamo infine la classifica

classifica =sort(round(punteggi,3))

> sort(round(class,3))

ValA Ligu Moli Marc FrVG Umbr Basi Sard TrAA Tosc Piem Lazi Lomb Abru

-1.820 -1.742 -1.556 -1.445 -1.339 -1.336 -1.208 -1.111 -0.709 -0.362 -0.318 -0.315 -0.216 -0.051

Cala Vene Camp Sici EmRo Pugl

0.802 1.275 1.320 2.692 3.269 4.170

1. **NOMI SU UN GRAFICO**

Svolgiamo un esercizio preparatorio che risolve anche una vecchia questione. Supponiamo di voler esaminare graficamente il legame tra frutta e vite nelle varie regioni italiane:

plot(A$frutta, A$vite)

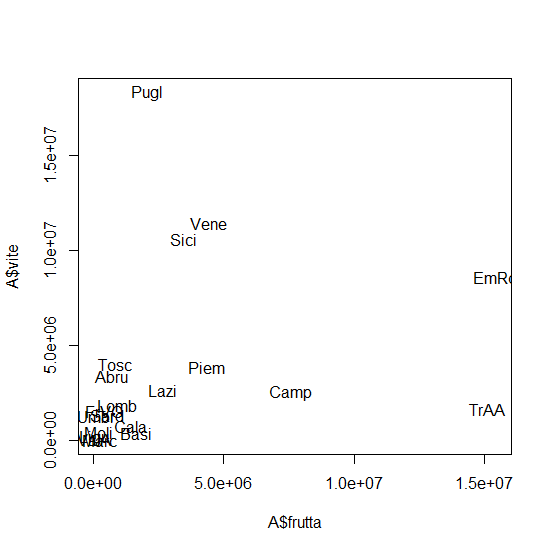
Possiamo arricchire il disegno dei nomi delle regioni?

text(A$frutta, A$vite, labels=as.character(row.names(A)))

E’ quasi meglio evitare il pallino, a questo punto:

plot(A$frutta, A$vite, type="n")

text(A$frutta, A$vite, labels=as.character(row.names(A)))



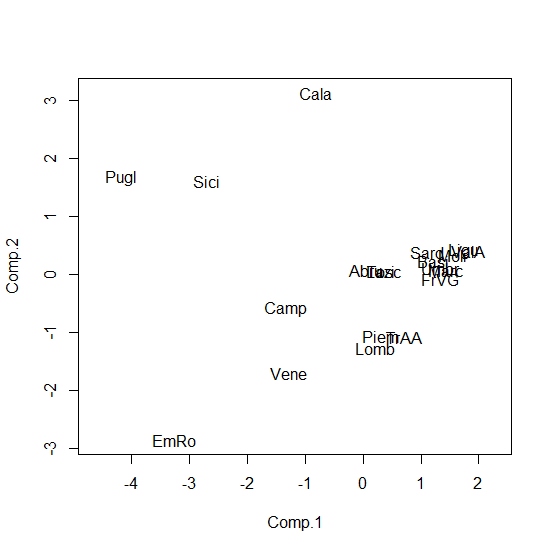
1. **ESPLORAZIONE DI TUTTI I PIANI PRINCIPALI**

C<-predict(PCAA)

i=1; j=2

plot(C[,c(i,j)],type="n",asp=1)

text(C[,c(i,j)],labels=as.character(row.names(AA)))



Questo caso ora è in scala. Provare a ragionare piano per piano.

Mettiamo le esplorazioni insieme al biplot

par(mfrow=c(1,2))

i=2; j=3

biplot(PCAA)

plot(C[,c(i,j)],type="n",asp=1)

text(C[,c(i,j)],labels=as.character(row.names(AA)))

