4R\_Esercitazione2

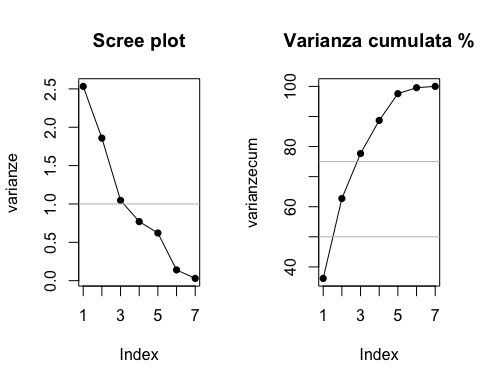
Francesca Pecorari

2024-01-19

# Report analisi di “Dataset alimenti.txt”

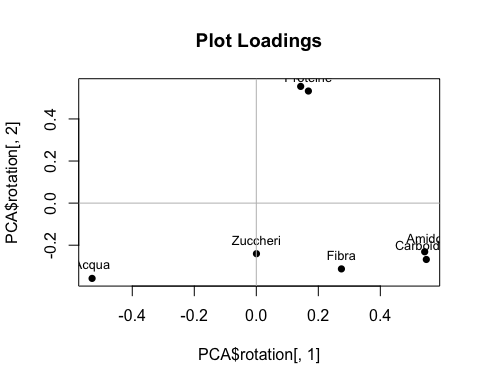
## PCA

### Scree plot e grafico della varianza cumulata



Dai grafici deduciamo che sono sufficienti 3 componenti per descrivere la varianza del dataset. Osservando lo Scree plot non è possibile notare un gomito ben definito, ma si vede che la terza componente è l’ ultima a mostrare varianza > 1. Dal plot della varianza cumulata notiamo che tra le 2 e le 3 componenti sono sufficienti a spiegare tra il 50 % e l’ 80 % della varianza del dataset.

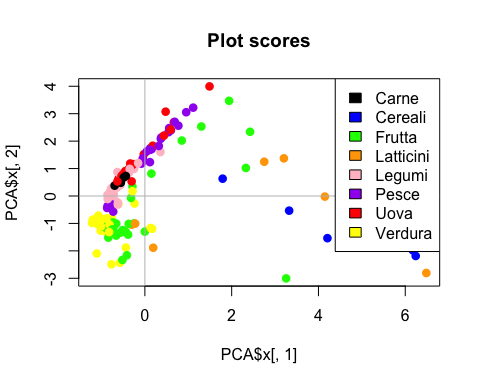
### Plot loadings



Interpretando il grafico, si potrebbe affermare che l’unica variabile a contribuire poco alla direzione di una delle nostre componenti principali, in particolare la PC2, sia la variabile Zuccheri, in quanto è proprio sullo zero.

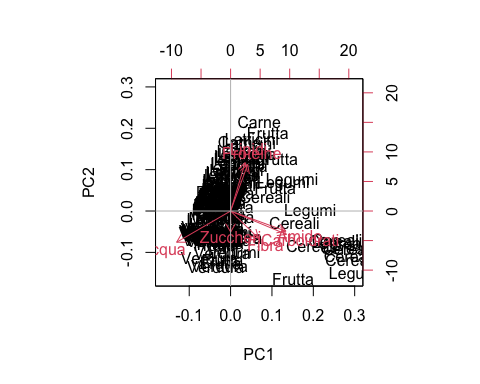
Si nota che le variabili Amido, Carboidrati e Proteine e Lipidi sembrano contribuire in modo simile alla direzione delle componenti; per il resto le variabili non sembrano avere valori di loading simili e questo indica, in generale, che non sono correlate e che avranno effetti diversi sulle componenti.

### Plot scores



Dal grafico si nota che ci sono delle osservazioni delle categorie Verdura e Frutta vicine all’ origine, questo indica che i loro valori sono prossimi alla media per la maggior parte delle variabili, inoltre sono posizionate vicine, il che significa che sono osservazioni simili. Per le categorie Latticini, Cereali, Frutta e Uova notiamo dei valori molto lontani dall’origine, che potrebbero essere dati estremi o outliers.

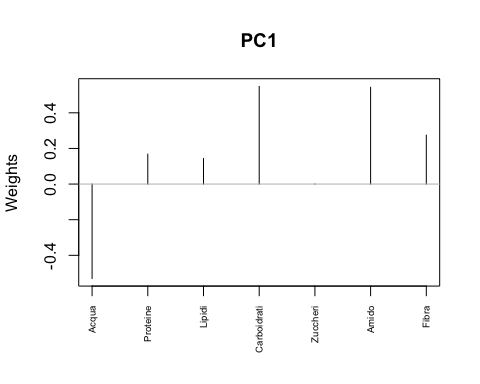
### Biplot



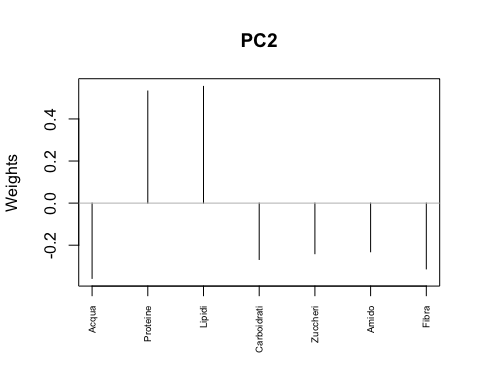
Non ci sono vettori paralleli agli assi di nessuna delle due PC, questo significa che nessuna variabile originale contribuisce in modo esclusivo a nessuna delle componenti di questo grafico. La variabile originale che sembra avere la maggiore variabilità spiegata dalle PC utilizzate per costruire questo grafico è Acqua, mentre la variable Zucchero, quella con il vettore più corto, sarebbe rappresentata meglio da altre PC.

### Grafico del peso delle variabili sperimentali su una PC

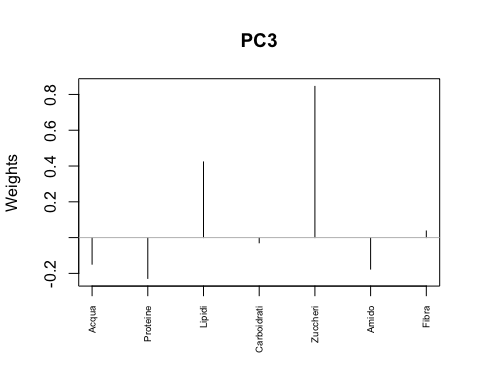
Le variabili sperimentali che hanno peso maggiore per la PC1 sono: Acqua, Carboidrati e Amido



Le variabili sperimentali che hanno peso maggiore per la PC2 sono: Proteine e Lipidi

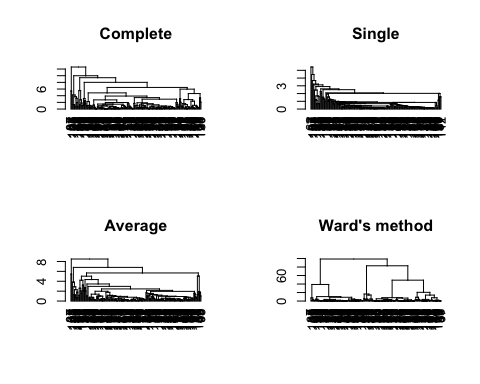


Le variabili sperimentali che hanno peso maggiore per la PC3 sono: Zuccheri e Lipidi

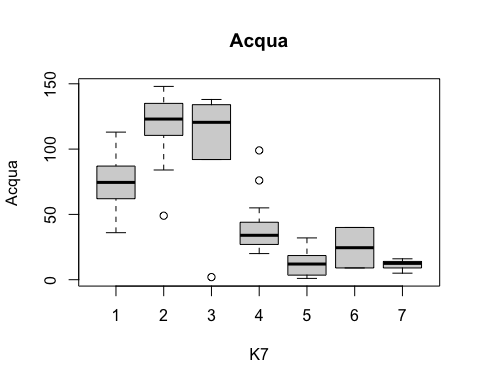


## CLUSTERING GERARCHICO

### Dendrogrammi

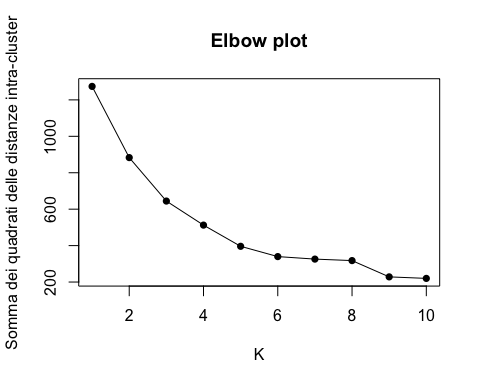


Ci sono vari metodi per dividere gerarchicamente i dati in cluster e dipendono dal criterio linkage, nei grafici riportati si può notare, concentrandosi sulle altezze a cui avvengono le varie divisioni, come questi metodi lavorino facendo scelte differenti.



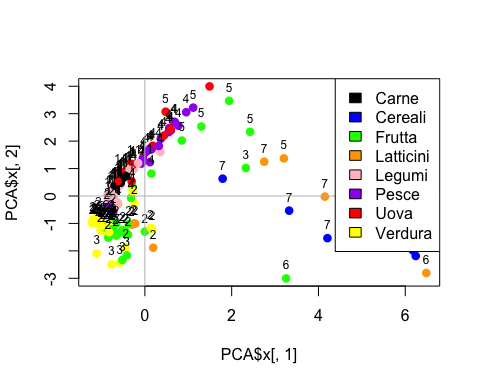
Dal boxplot si nota che in ogni cluster si ha una distribuzione diversa della variabile Acqua, probabilmente se i cluster corrispondessero effettivamente a alimenti di categorie diverse, questo si potrebbe spiegare dicendo che hanno tutti valori di acqua diversi e per questo si distinguono.

## K-MEANS CLUSTERING



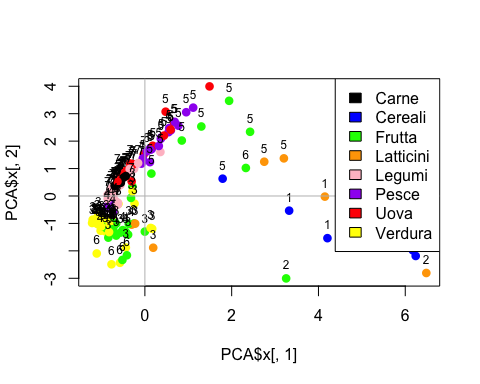
Nell’ interpretazione dell’elbow plot si cerca un “gomito” in cui la discesa della somma dei quadrati degli errori inizia a rallentare significativamente; idealmente si sceglie il numero di cluster appena prima del punto di gomito. Nel nostro grafico, si potrebbe scegliere il valore 8 e avere quindi 7 cluster.

## Risultato HCA proiettato sul grafico degli score della PCA



Notiamo che osservazioni vicine, quindi simili, sono spesso assegnate allo stesso cluster.

## Risultato KM proiettato sul grafico degli score della PCA



Anche in questo caso, osservazioni vicine tendono ad essere assegnate allo stesso cluster.