Elisa Merigo email: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola: 869436 - 29 aprile 2024

Censimento pacchetti R

1. index.Gap{clusterSim} → https://search.r-project.org/CRAN/refmans/clusterSim/html/index.GAP.html

   Questo pacchetto ha una documentazione non troppo fornita. Spiegata bene la parte della distribuzione uniforme. Gli argomenti non poco dettagliati. Nella parte “Details” viene condiviso il link del file pdf inerente alla documentazione del pacchetto, ma non si riesce ad aprire. Come “value” vengono proposti:
   - Gap → riferendosi a Tibshirani
   - diffu → valore necessario per scegliere il corretto numero di cluster → mette anche la formula per trovare k

   Presenta tre esempi:

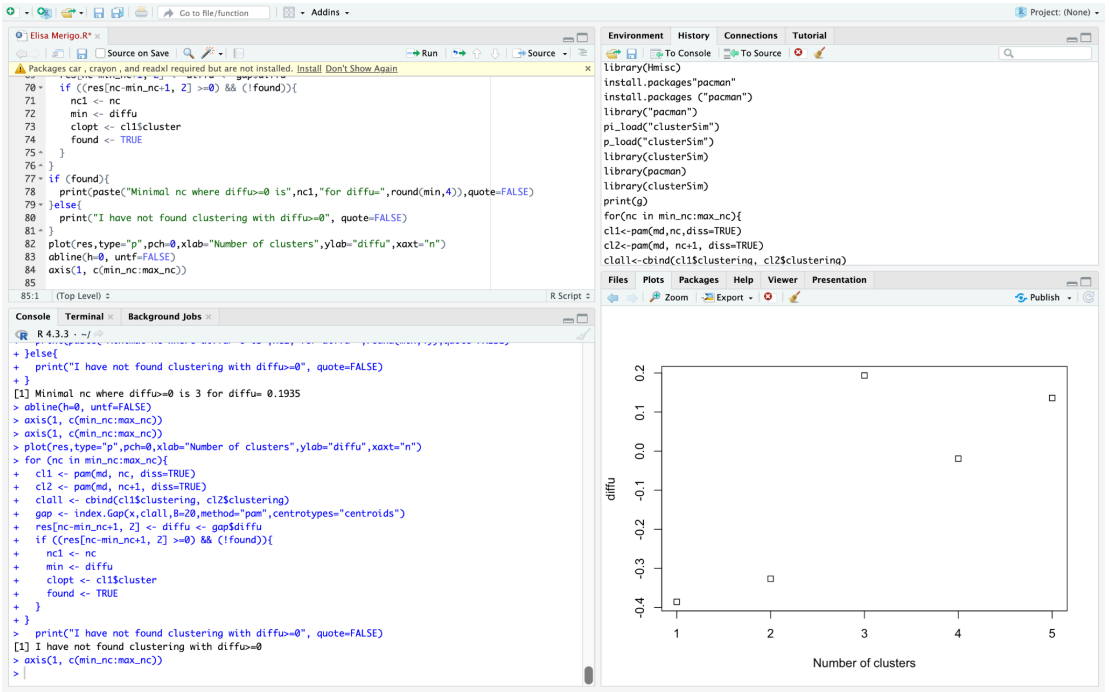
   Esempio 1:

```
# Example 1
library(clusterSim)
data(data_ratio)
cl1<-pam(data_ratio,4)
cl2<-pam(data_ratio,5)
clall<-cbind(cl1$clustering,cl2$clustering)
g<-index.Gap(data_ratio, clall, reference.distribution="unif", B=10,
   method="pam")
print(g)
```

```
> library(clusterSim)
> print(g)
[1] 1 5 7
>
```

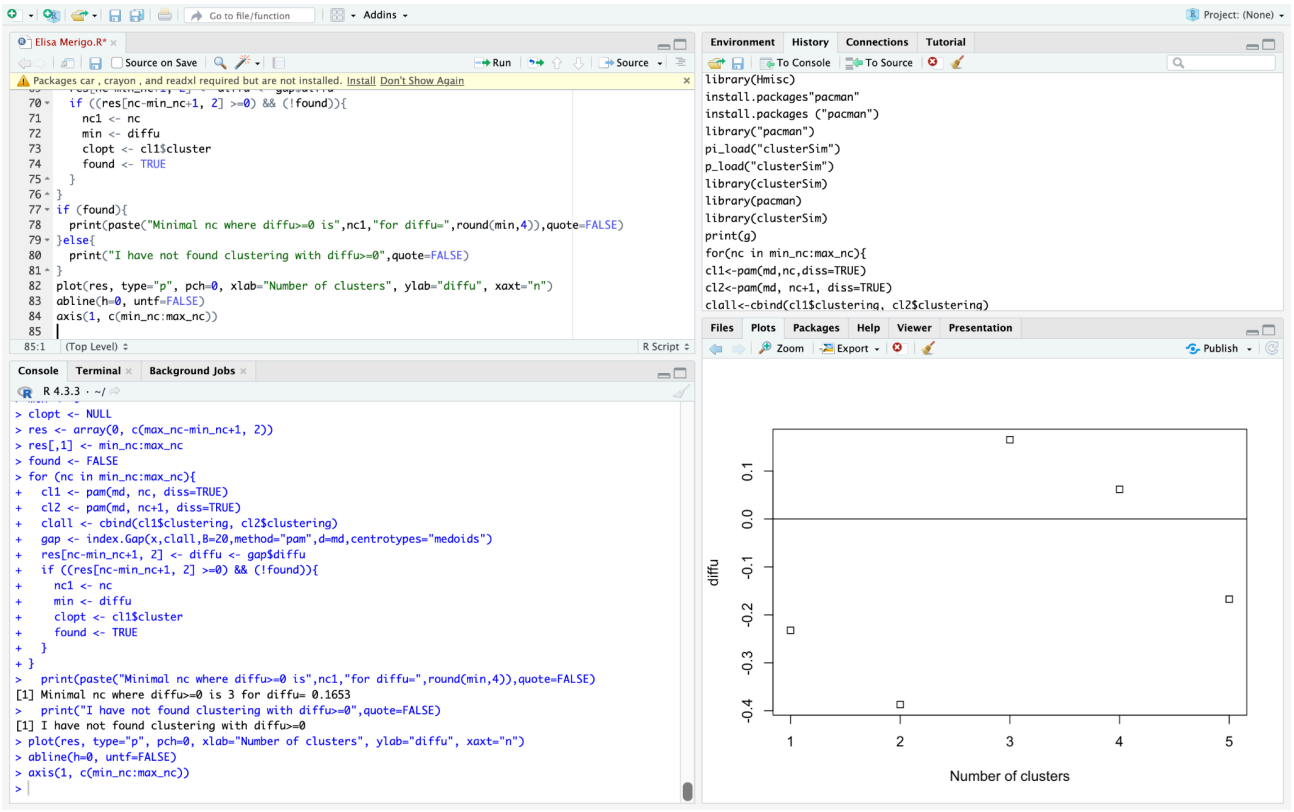
   Al posto di “pam” ho provato a mettere “k-means” e ho ottenuto lo stesso risultato.

   Esempio 2:



Elisa Merigo email: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola: 869436 - 29 aprile 2024

Esempio 3:

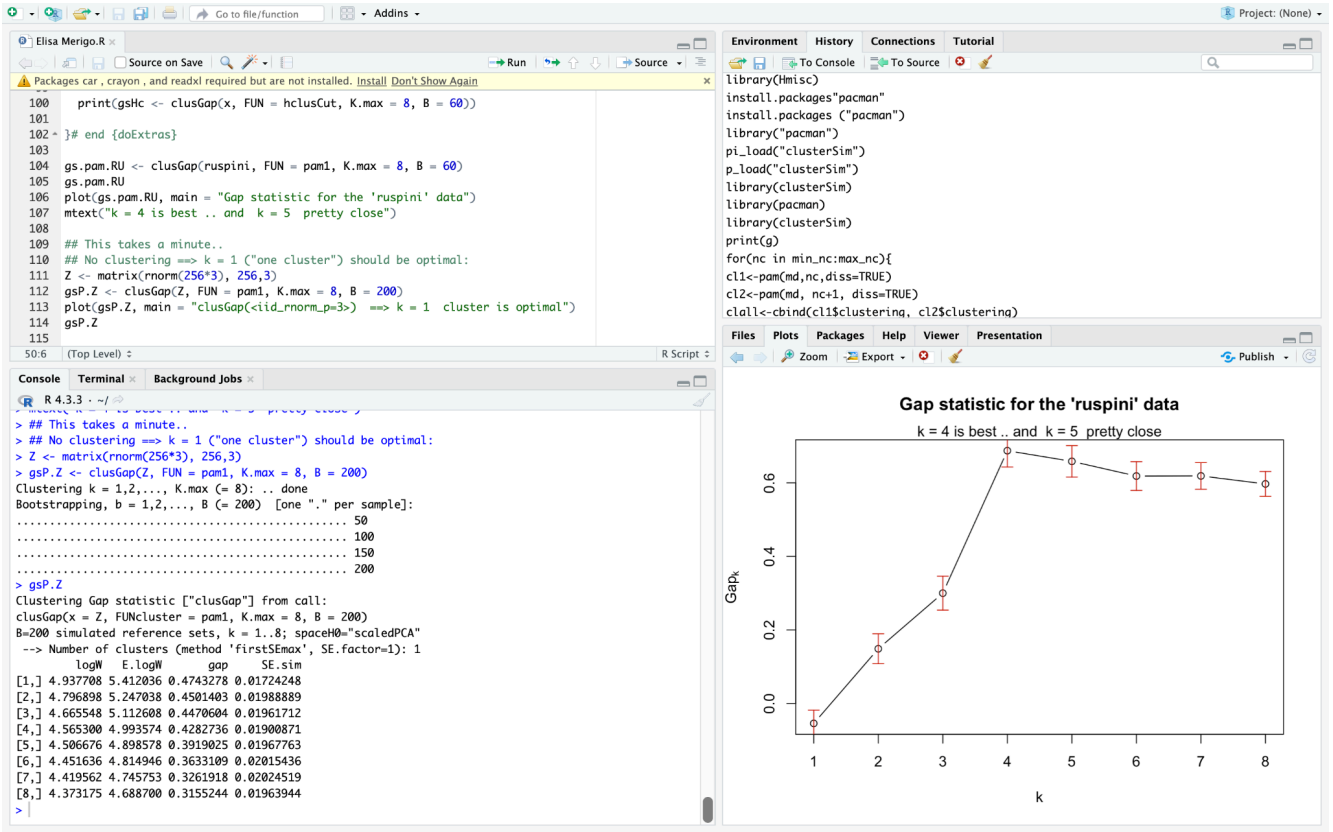


2. clusGap{cluster} → https://stat.ethz.ch/R-manual/R-devel/library/cluster/html/clusGap.html

Questo pacchetto ha una documentazione ben fornita: viene spiegato in modo esaustivo cosa fa la gap statistic, quali sono le variabili utilizzate, la distribuzione di riferimento, 1- standard error e che tipo di k deve prendere. Gli argomenti sono molto dettagliati e ben spiegati. Nella parte “details” viene spiegato l’utilizzo della variabile result. L’analisi dei valori è più esaustiva rispetto al pacchetto precedente.

Presenta un esempio ma ben commentato:

Elisa Merigo email: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola: 869436 - 29 aprile 2024



ESEMPIO MATRICE CON “clusGap”
Installato pacchetto “ggpubr” e “ggplot2” per creare cluster plot.
Installato “FactoMineR” e “factoextra” per creare cluster plot.

https://www.datanovia.com/en/blog/k-means-clustering-visualization-in-r-step-by-step-guide/

Creando la matrice 8x4 con 4 righe di 0 e 4 righe di 1 e numero k = 2, ottengo una gap negativa nel cluster 1 e infinito nel cluster 2. Il cluster 1 è formato dagli uno e il cluster 2 dagli zeri. Con il grafico della gap si nota che il valore ottimale del numero di cluster è 2.
Modificando la settima riga (composta da 1) ottengo una gap negativa nel cluster 1 e positiva nel cluster 2. Il cluster 1 è formato dagli zeri e alcuni outliers, il cluster 2 è formato dagli uno. Con il grafico della gap si può notare che il valore ottimale del numero di cluster è 2.
Modificando l’ottava riga ottengo una gap negativa nel cluster 1 e positiva nel cluster 2. Il cluster 1 è formato dagli uno e valori vicini, il cluster 2 dagli zeri. Con il grafico della gap si può notare che il valore ottimale del numero di cluster è 2.
Modificando la sesta riga ottengo una gap negativa nel cluster 1 e positiva nel cluster 2. Il cluster 1 è formato dagli zeri, il cluster 1 dagli uno e dai valori vicini. Con il grafico della gap si può notare che il valore ottimale del numero di cluster è 2.
Modificando la quinta riga ottengo una gap negativa nel cluster 1 e positiva nel cluster 2. Il cluster 1 è formato da valori diversi da 0.00, il cluster 2 è formato dagli zeri. Con il grafico della gap si può notare che il valore ottimale del numero di cluster è 2.

Elisa Merigo email: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola: 869436 - 29 aprile 2024

Considerazioni: I grafici della gap più belli sono quelli in cui si modificano le righe. Ci aspettavamo una gap positiva nel primo caso, ma abbiamo ottenuto il valore migliore solo nel secondo cluster. Nel primo cluster abbiamo ottenuto un valore negativo poichè il valore atteso del lnW è minore del lnW.

Provo ad utilizzare index.Gap in confronto al clusGap per vedere se escono valori positivi della gap.

ESEMPIO MATRICE CON “index.Gap”

Installato pacchetto “clusterSim” per poter usare index.Gap.
Per il grafico finale della gap statistic non riesco a trovare una funzione che implementa index.Gap.

Dopo aver scritto la prima matrice, creo le funzioni cl1 e cl2 utilizzando il kmeans. In entrambe ho due cluster:

- in cl1: cluster 1 formato dagli zeri e cluster 2 dagli uno
- in cl2: cluster 1 formato dagli uno e cluster 2 dagli zeri

Successivamente creo clall che prende cl1$cluster e cl2$cluster e crea una matrice disponendo i risultati per colonne utilizzando cbind. In questo modo ho una matrice composta dai vettori dei cluster.
Ora calcolo la gap statistic utilizzando index.Gap e ottengo come valore **+Inf.**
Stampo il grafico utilizzando fviz_nbclust, numero ottimale = 2.
Faccio la stessa cosa con la matrice in cui cambio la settima riga.
Il valore della gap è pari a **+0.5803108**.
Stampo grafico, numero ottimale pari a 2.
Creo la matrice in cambio l’ottava riga. La prima volta che si calcola viene un valore negativo, bisogna far girare di nuovo cl5, cl6, clall3. Il valore della gap è **+0.0238918**.
Nel grafico il numero ottimale è pari a 2.
Nella matrice cambio la sesta riga.
Il valore della gap è **+0.01161746**.
Stampo il grafico, numero ottimale pari a 2.
Stampo la matrice con la quinta riga cambiata.
Il valore della gap è **+0.1997138**.
Nel grafico il valore ottimale è pari a 2.