

FACULTAD DE CIENCIAS EXACTAS, INGENIERÍA Y AGRIMENSURA

TÓPICOS DE MINERÍA DE DATOS

Trabajo Práctico 3: Clustering

Alumno: Jeremías Rodríguez

Profesor: Pablo Granitto

14 de enero de 2018

1. Ejercicio 1

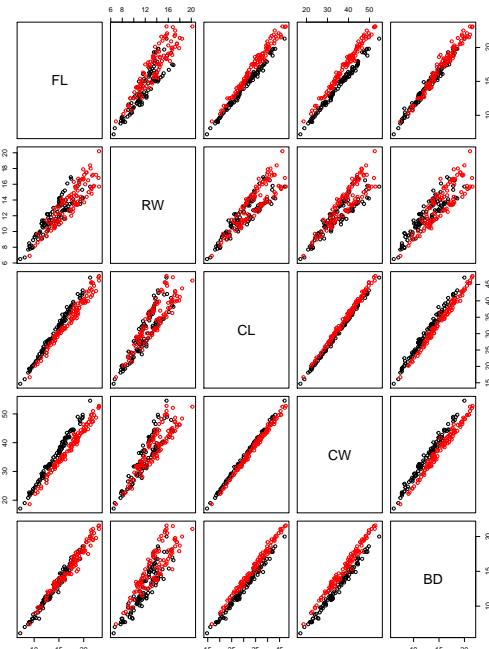
1.1. Apartado A: Dataset Crabs

Analizaré el dataset crabs, que describe cangrejos de la especie *Leptograpsus Variegatus*, y fue recolectado en Fremantle, W. Australia. Consta de 200 filas con 5 features numéricas y 2 clases (sexo y especie):

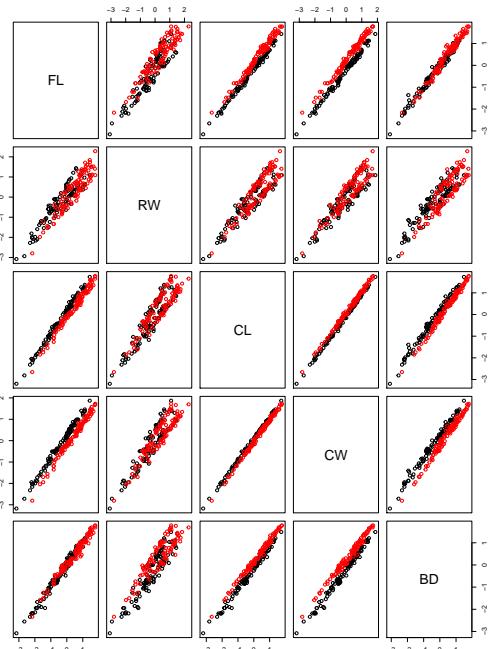
```
> summary(crabs)
   sp      sex       FL          RW          CL          CW      (..)
B:100  F:100  Min.   : 7.20  Min.   : 6.50  Min.   :14.70  Min.   :17.10
O:100  M:100  1st Qu.:12.90  1st Qu.:11.00  1st Qu.:27.27  1st Qu.:31.50
                  Median :15.55  Median :12.80  Median :32.10  Median :36.80
                  Mean   :15.58  Mean   :12.74  Mean   :32.11  Mean   :36.41
                  3rd Qu.:18.05  3rd Qu.:14.30  3rd Qu.:37.23  3rd Qu.:42.00
                  Max.   :23.10   Max.   :20.20   Max.   :47.60   Max.   :54.60
```

El objetivo es ver si, utilizando distintos pre-procesamientos en el dataset (escalado, PCA, logaritmos), los métodos de clustering estudiados (K-means y Hclust) logran agrupar por sexo o genero.

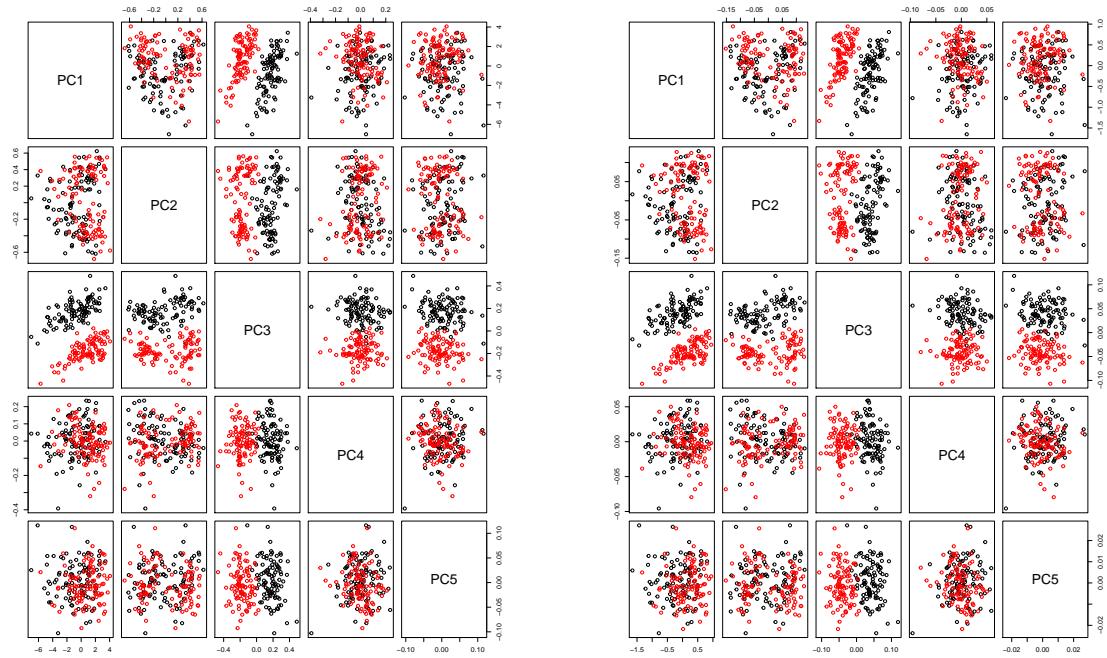
A continuación se muestran las 5 variantes del dataset crabs usadas. Los colores corresponden a la clase especie.



(a) Dataset sin modificaciones

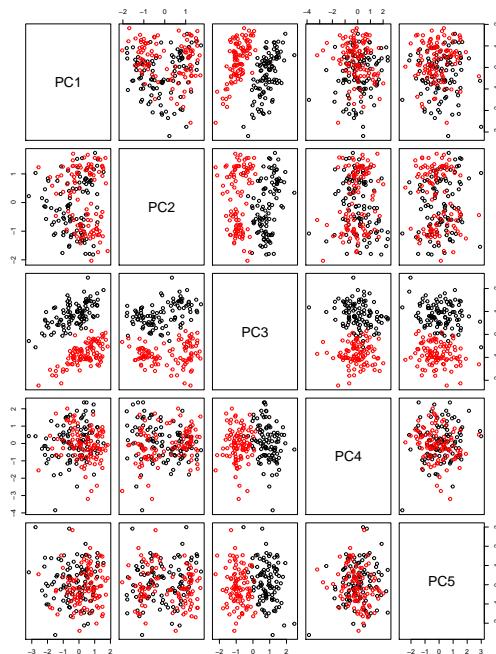


(b) Transformación logarítmica + escalado



(a) Transformación logarítmica + escalado + PCA

(b) Transformación logarítmica + PCA



(c) Transformación logarítmica + PCA + escalado

A continuación se encuentran los resultados, los porcentajes representan los casos macheados en pares.

	K-means	Hclust single	Hclust complete	Hclust average
Raw	57.5	50.5	59.5	61.5
Log + scale	60.5	50.5	60.5	50.5
Log + scale + PCA	60.5	50.5	60.5	50.5
Log + PCA	60.5	50.5	60.5	57
Log + PCA + Scale	100	50.5	53.5	50.5

Cuadro 1: Feature especie: Porcentaje de casos macheados en pares, método vs dataset

	K-means	Hclust single	Hclust complete	Hclust average
Raw	50.5	50.5	50.5	55.5
Log + scale	51.5	50.5	51	50.5
Log + scale + PCA	51.5	50.5	51	50.5
Log + PCA	51.5	50.5	51	51
Log + PCA + Scale	50	50.5	56.5	50.5

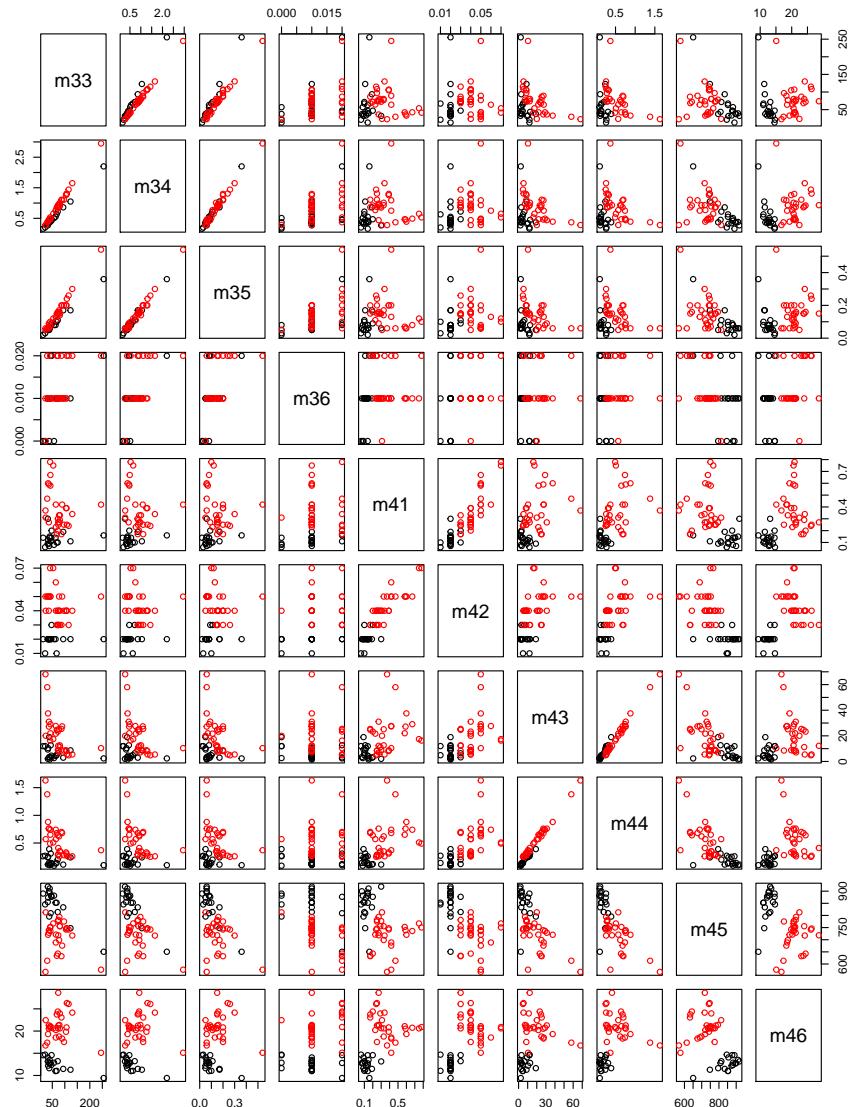
Cuadro 2: Feature sexo: Porcentaje de casos macheados en pares, método vs dataset

Como vemos, casi todos los métodos fallan en encontrar alguna clusterización coherente con nuestras clases. En particular, ningún método permitió recuperar la división por sexo (y los resultados fueron todos muy malos). En muchos casos se observó que en los clusters resultantes, o bien un cluster tenía casi todos los puntos del dataset; o bien ambos clusters tenían casi la misma proporción de puntos de cada clase. Incluso en algunos métodos se ven mejores resultados sin preprocessar el dataset.

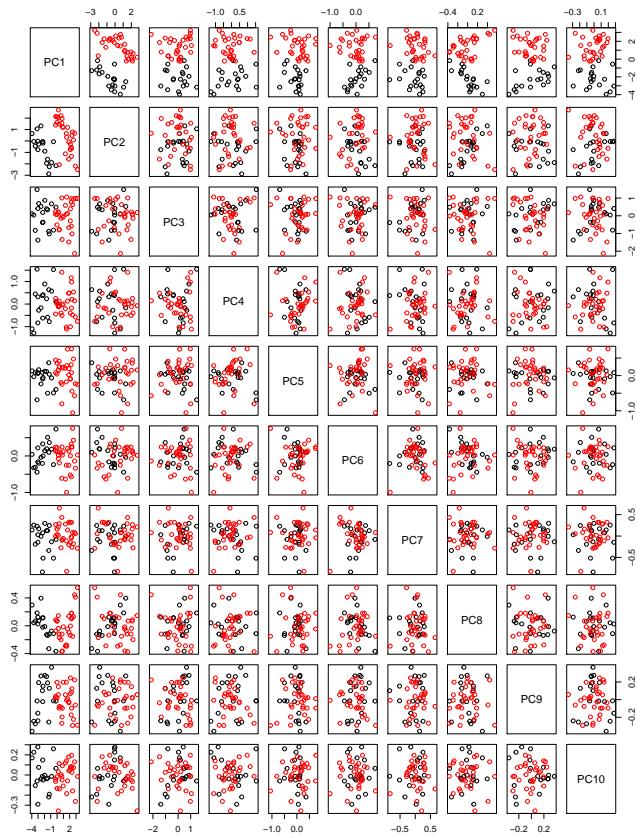
Por otro lado, respecto a recuperar la especie de los cangrejos, los porcentajes son en general mayores aunque la gran mayoría parecieran casi aleatorios. Sin embargo, vemos que k-means con el preprocessamiento adecuado logra clusterizar perfectamente las clases. Esto resalta la importancia de hacer el preprocessamiento adecuado antes de clusterizar, que es la diferencia entre obtener basura y obtener algo muy útil. Un motivo por el que k-means puede haber funcionado es que no hay jerarquías aparentes de datos, sino que estamos buscando una partición de ellos. En varios de los plots se ven dos cúmulos de puntos bastante diferenciados, por lo que la tendencia de k-means de buscar clusters fuertemente unidos también es conveniente al problema. Esto puede explicar porqué hclust con complete linkage arroja resultados ligeramente mejores que average y single, puesto que complete linkage también apunta a clusters compactos.

1.2. Apartado B: Dataset Lampone

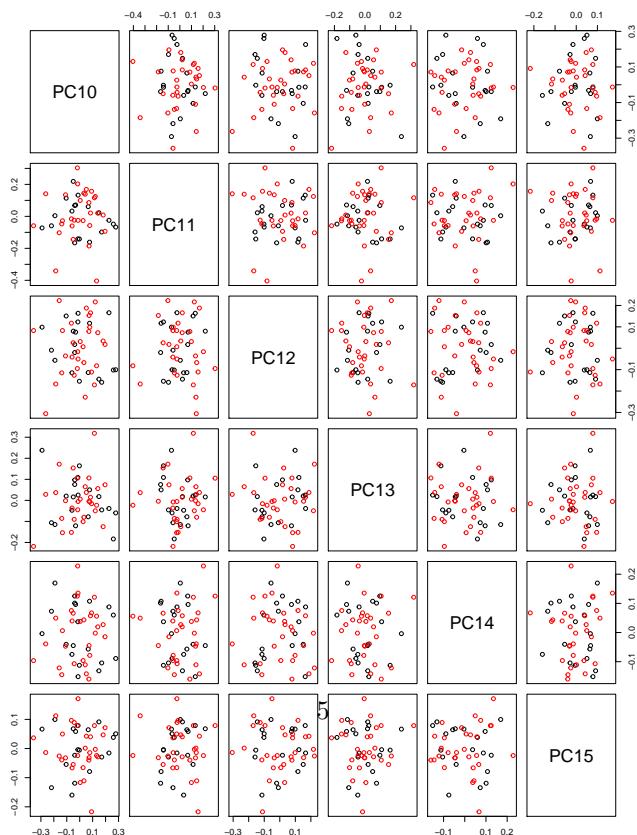
El dataset Lampone consta de 49 filas y 144 features referentes a arándanos, en el cuál se intentarán recuperar las clases Año de Medición y Especie de Arándano usando las mismas técnicas de clustering. Antes de poder aplicarle los métodos, se descartaron columnas no numéricas, y se usarán los mismos preprocesamientos que en el apartado A. Dada la cantidad de features, a continuación se muestran sólo algunos ploteos ilustrativos:



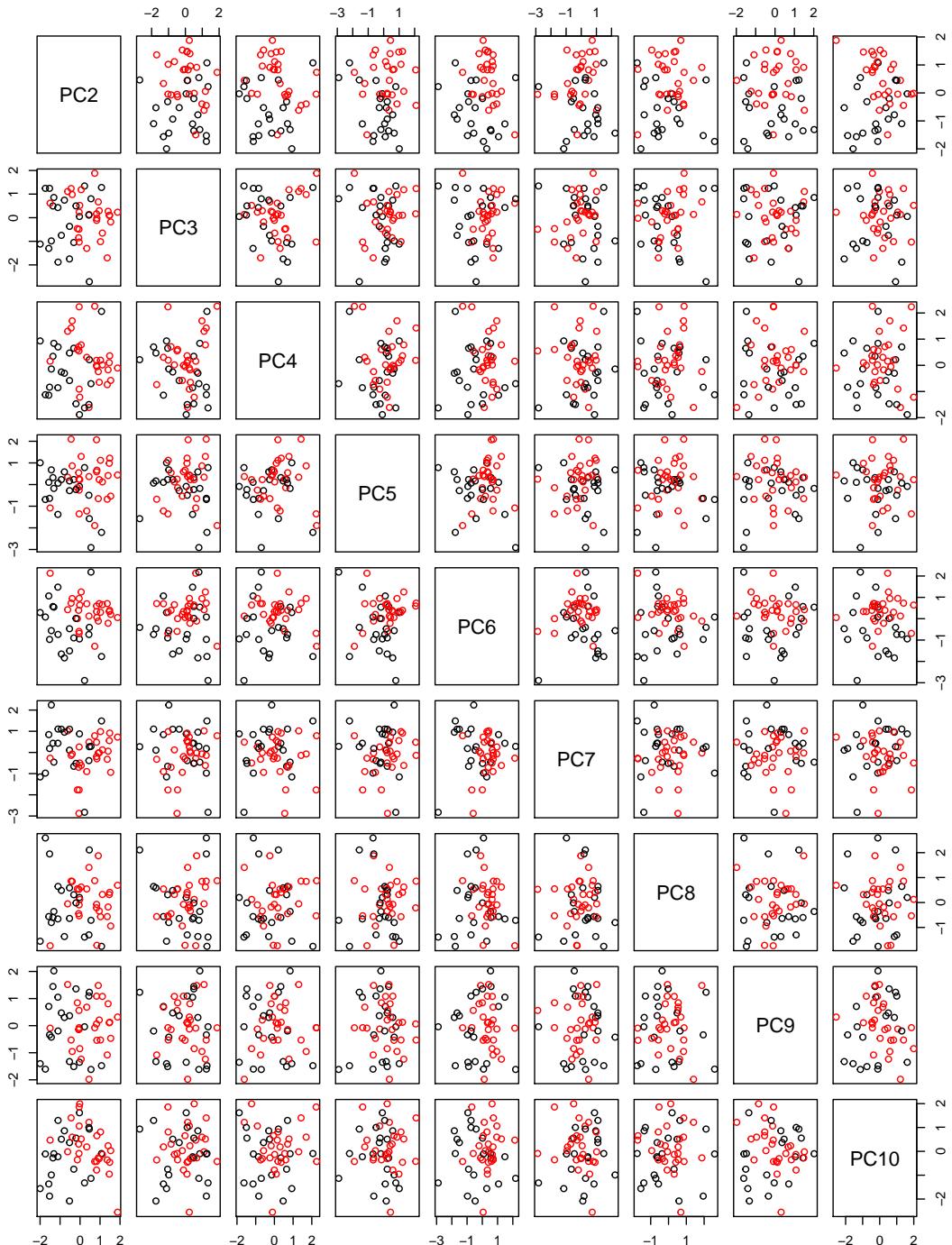
(a) Raw dataset. Primeras 10 componentes. Año de medición.



(a) log + PCA. Diez componentes más importantes. Año de medición



(b) log + PCA. Componentes 10-15. Año de medición



(a) PCA. Clases más importantes. Clase sp.

Los resultados obtenidos fueron:

	K-means	Hclust single	Hclust complete	Hclust average
Raw	69.39	57.14	57.14	57.14
Log + scale	97.96	59.18	83.67	59.18
Log + scale + PCA	97.96	59.18	83.67	59.18
Log + PCA	100	59.18	85.71	95.92
Log + PCA + Scale	59.18	61.22	61.22	62.22

Cuadro 3: Feature año de medición: Porcentaje de casos macheados en pares, método vs dataset

	K-means	Hclust single	Hclust complete	Hclust average
Raw	65.31	59.18	59.18	59.18
Log + scale	55.1	57.14	69.39	57.14
Log + scale + PCA	55.1	57.14	69.39	57.14
Log + PCA	53.06	57.14	53.06	57.14
Log + PCA + Scale	57.14	59.18	55.1	59.18

Cuadro 4: Feature especie de blueberry: Porcentaje de casos macheados en pares, método vs dataset

En primer lugar, ningún método logró dar un resultado aceptable recuperando la especie de arándano. Tal vez son necesarios más puntos (el dataset es muy pequeño) u otro tipo de features. Vemos en uno de los ploteos de PCA que los puntos de las dos especies están muy mezclados y no se ven dos clusters fácilmente. Esto explicaría porqué todos los métodos fallan.

Por otro lado, el año de medición sí es recuperado exitosamente. K-means consigue separar perfectamente incluso, y hclust complete y average arrojan buenos resultados. Vemos en los ploteos anteriores que la primer componente principal parece aportar la mayor parte de información para la separación. Mirando esta componente contra todas las otras, se pueden apreciar dos clusters con centros separados y puntos concentrados a su alrededor (a diferencia de lo que sucedía con la especie); esto favorece a k-means y a hclust complete linkage. Sin embargo, al no ser completamente compactos y diferenciados los clusters sino un poco más dispersos, se explicaría porqué average linkage tuvo mejor resultado que complete linkage bajo ciertos preprocesamientos.

2. Ejercicio 2

Las implementaciones se encuentran entre los archivos adjuntos. Algunas aclaraciones:

Gap Statistic:

- Mi implementación de GapStatistic está parametrizada por el método de clustering. Este metodo parámetro debe tomar como parámetro un dataset y un K, y devolver la suma total de las distancias w_k de todos los clusters. En particular para k-means, es simplemente realizar el clustering y extraer el campo tot.withinss; mientras que en hclust debí calcular a mano las distancias entre los puntos de un mismo cluster y luego sumarlas.
- Para calcular el gap, consideré la PCA del dataset argumento y generé puntos uniformes dentro de ella; a los que luego les apliqué el método de clustering. El algoritmo no vuelve a girar los puntos a los ejes originales, por lo que puede no funcionar correctamente para métodos que utilicen algo más que las distancias entre los puntos.

Estabilidad:

- El algoritmo recibe como parámetros un dataset, un entero K que indica el máximo k a analizar, y un entero B que indica cuántos datasets perturbados se generarán. Una vez generados los B datasets, se procede a comparar todas las parejas posibles y calcular el score correspondiente. Adicionalmente un parámetro proporción puede especificarse, pues el método de perturbación es subsampleo y podría desearse tomar distintas proporciones del dataset original. Finalmente, el método de clustering también es un parámetro.
- El algoritmo retorna una matriz de tamaño $K \times \binom{B}{2}$, con los scores correspondientes a cada pareja de los B conjuntos perturbados. El algoritmo no realiza análisis alguno sobre qué k elegir.

3. Ejercicio 3

Se corrieron los dos algoritmos del Ejercicio 2 sobre los datasets Iris, Lampone y 4-Gaussianas.

3.1. Gap Statistic

A continuación la moda¹ de 10 ejecuciones con cada método:

	Kmeans	HclS	HclA	HclC
Iris	4	2	3	3
Iris + log + scale	2	2	2	3
Lampone	1	1	1	1
Lampone + log + scale	3	1	1	2
4 Gaussianas	4	1	2(4)	2(4)

Cuadro 5: Número óptimo de clusters para cada problema según el metodo GapStatistic.

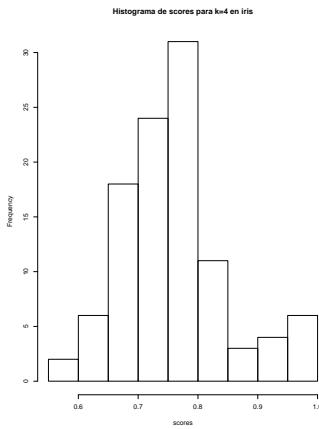
Se ve que k-means con los preprocesamientos adecuados a los conjuntos, indica el k óptimo correctamente en todos los casos. Se notó una gran fluctuación en los resultados para todos los métodos.

HCLust también retorna valores acertados, excepto en el problema de las 4 gaussianas (donde los resultados variaban mucho). En ambos casos el método funciona bastante rápido, y su costo es mucho menor al de estabilidad.

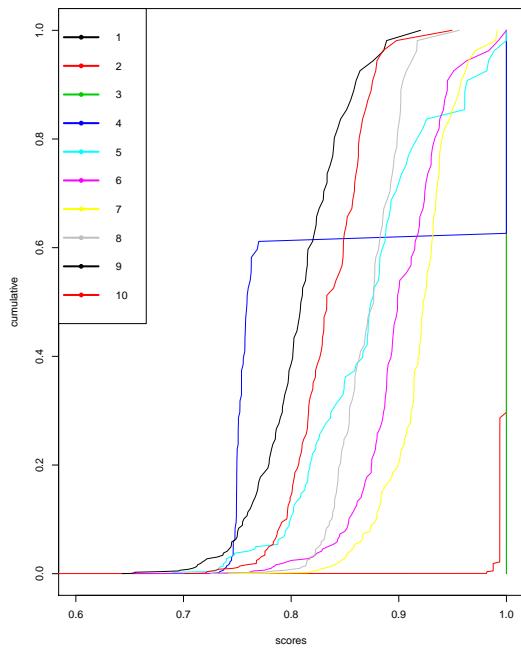
3.2. Estabilidad

Consideremos el dataset Iris. El algoritmo de estabilidad retorna una matriz con B columnas y K filas de scores. A continuación se muestra el histograma de la fila k=4:

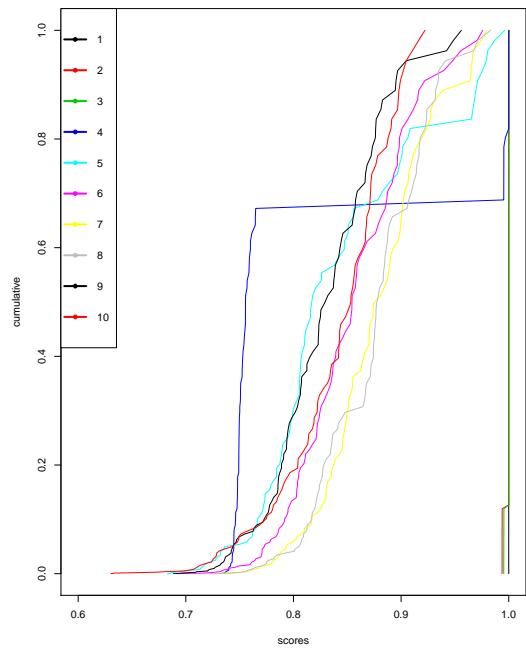
¹En algunos casos, el segundo valor más frecuente es indicadopués la relación de aparición es de 6 a 4



Considerando para cada k la acumulada del histograma, se generaron las siguientes gráficas para cada dataset:

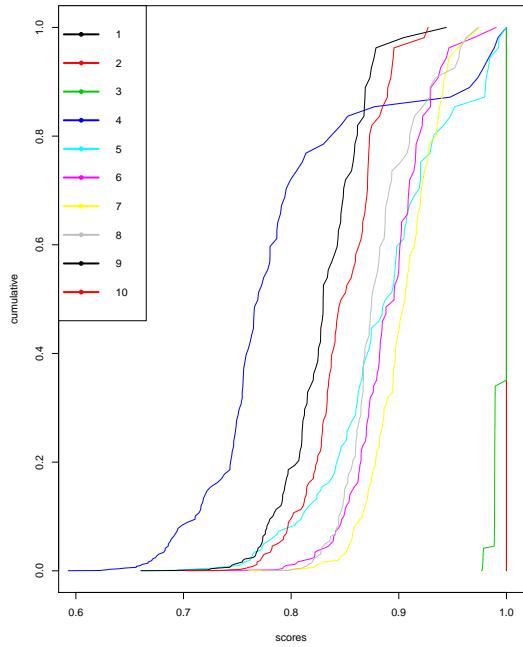


(a) Dataset 4-Gaus., estabilidad con K-means

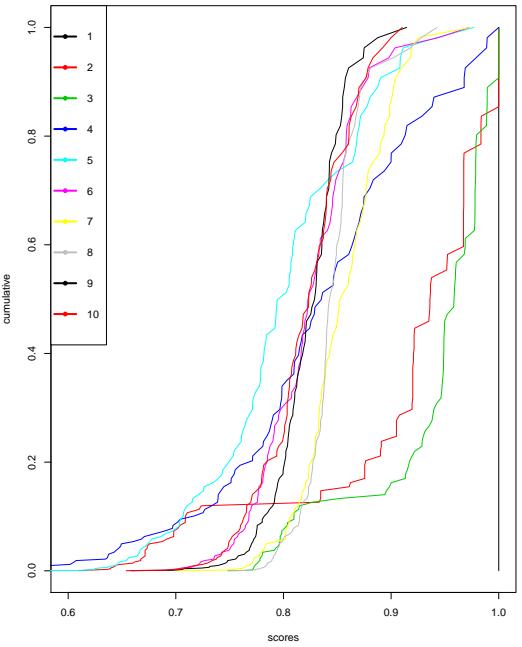


(b) Dataset 4-Gaus., estabilidad con Hclust-CL

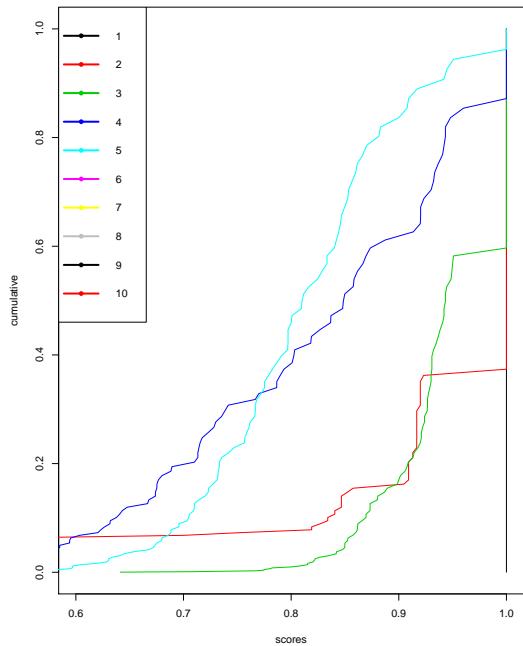
- En el dataset 4-Gaussianas, las soluciones estables son 1, 2 y 3. El mayor valor estable, 3, sería entonces razonable de elegir.
- En el dataset Iris vemos que las soluciones estables son 3, 2 y 1 con k-means; por lo que un valor de k razonable a elegir sería el mayor -3-. Sin embargo para h-clust el único valor estable es 1.
- En el dataset lampone no está muy claro cuáles valores son estables o no, excepto el 1.



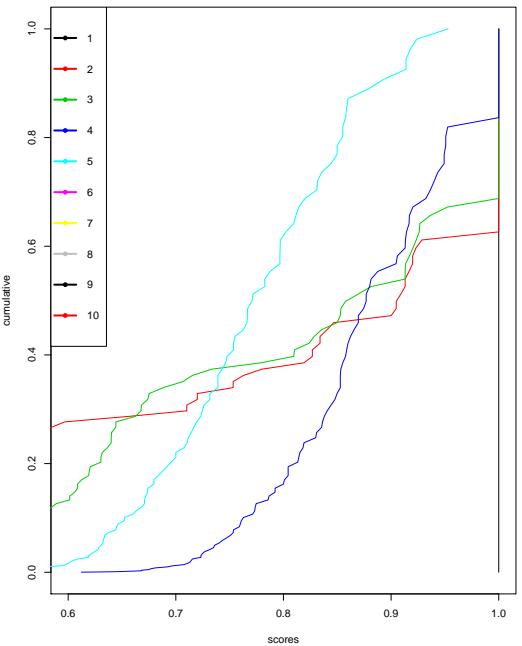
(a) Dataset Iris, estabilidad con K-means



(b) Dataset Iris, estabilidad con Hclust-CL



(c) Dataset Lampone, estabilidad con K-means



(d) Dataset Lampone, estabilidad con Hclust-CL

4. Ejercicio 4: Opcional

Analizaré el dataset wine², relacionado al uso de análisis químicos para identificar clases de vinos. En concreto, este dataset consiste en 13 features numéricas que representan distintas propiedades químicas de una muestra de vino, y una feature 'clase' que varía entre tres tipos de vinos. 178 muestras fueron tomadas.

```
> summary(wine)
  class      alcohol      malic acid      ash
  Min.   :1.000   Min.   :11.03   Min.   :0.740   Min.   :1.360
  1st Qu.:1.000   1st Qu.:12.36   1st Qu.:1.603   1st Qu.:2.210
  Median :2.000   Median :13.05   Median :1.865   Median :2.360
  Mean   :1.938   Mean   :13.00   Mean   :2.336   Mean   :2.367
  3rd Qu.:3.000   3rd Qu.:13.68   3rd Qu.:3.083   3rd Qu.:2.558
  Max.   :3.000   Max.   :14.83   Max.   :5.800   Max.   :3.230

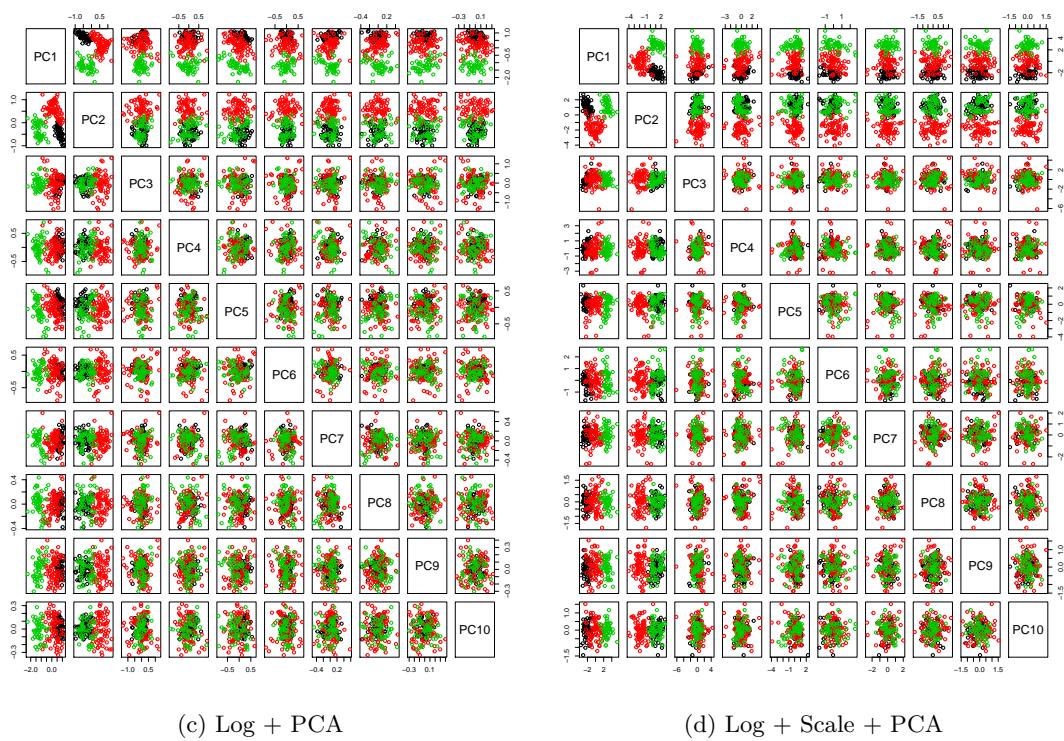
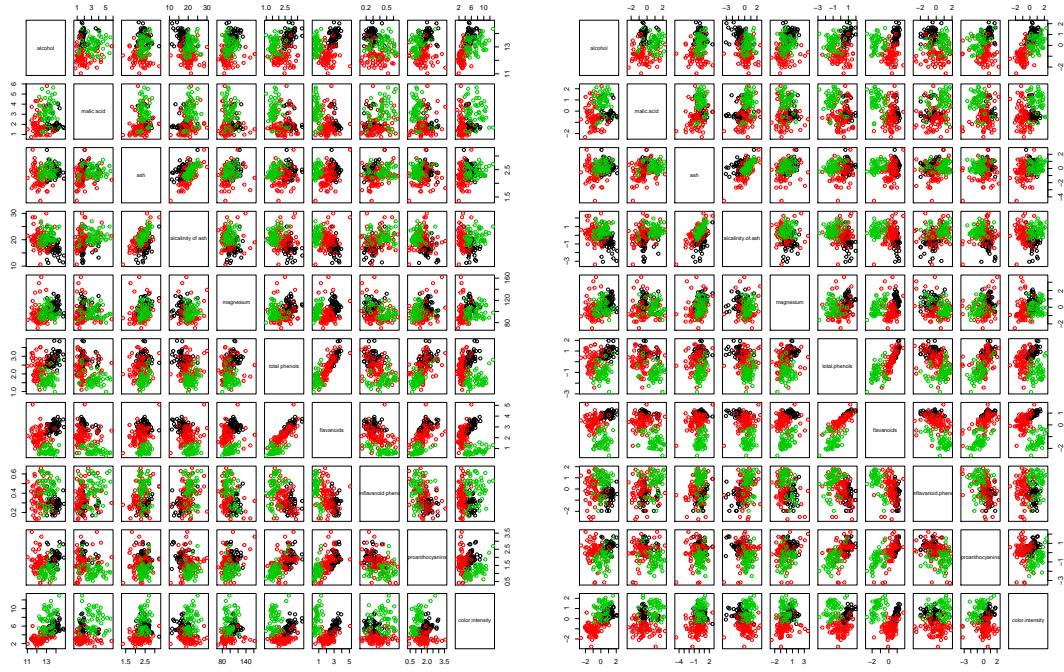
  alcalinity of ash    magnesium    total phenols    flavanoids
  Min.   :10.60       Min.   : 70.00   Min.   :0.980   Min.   :0.340
  1st Qu.:17.20       1st Qu.: 88.00   1st Qu.:1.742   1st Qu.:1.205
  Median :19.50       Median : 98.00   Median :2.355   Median :2.135
  Mean   :19.49       Mean   : 99.74   Mean   :2.295   Mean   :2.029
  3rd Qu.:21.50       3rd Qu.:107.00  3rd Qu.:2.800   3rd Qu.:2.875
  Max.   :30.00       Max.   :162.00   Max.   :3.880   Max.   :5.080

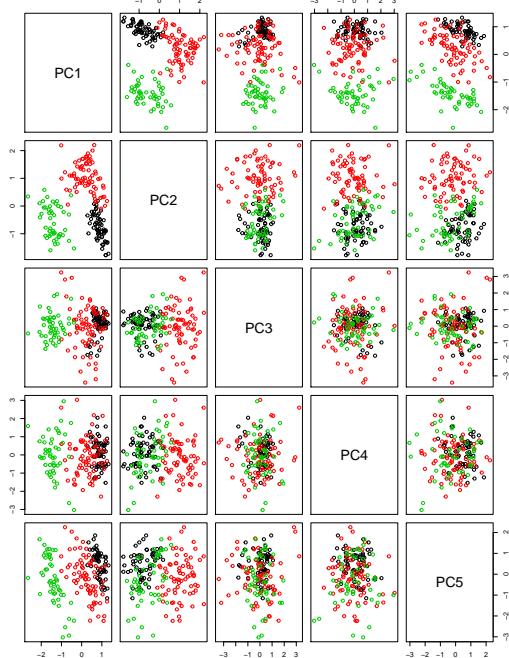
  nonflavanoid phenols proanthocyanins color intensity      hue
  Min.   :0.1300       Min.   :0.410   Min.   : 1.280   Min.   :0.4800
  1st Qu.:0.2700       1st Qu.:1.250   1st Qu.: 3.220   1st Qu.:0.7825
  Median :0.3400       Median :1.555   Median : 4.690   Median :0.9650
  Mean   :0.3619       Mean   :1.591   Mean   : 5.058   Mean   :0.9574
  3rd Qu.:0.4375       3rd Qu.:1.950   3rd Qu.: 6.200   3rd Qu.:1.1200
  Max.   :0.6600       Max.   :3.580   Max.   :13.000   Max.   :1.7100

  OD280 of diluted wines    proline
  Min.   :1.270          Min.   : 278.0
  1st Qu.:1.938          1st Qu.: 500.5
  Median :2.780          Median : 673.5
  Mean   :2.612          Mean   : 746.9
  3rd Qu.:3.170          3rd Qu.: 985.0
  Max.   :4.000          Max.   :1680.0
```

Analizaré este dataset con las técnicas estudiadas para ver si es posible recuperar la clase de vino, y luego veré cuál es el número de clusters real según los métodos vistos. Los análisis se aplicarán sobre las siguientes 5 variantes del dataset:

²<https://archive.ics.uci.edu/ml/datasets/Wine>





(a) Log + PCA + Scale

Resultados:

	K-means	Hclust single	Hclust complete	Hclust average
Raw	70.22	42.7	67.42	61.24
Log + scale	91.57	38.2	64.04	64.04
Log + scale + PCA	95.51	38.76	61.8	38.76
Log + PCA	72.47	38.76	58.43	35.39
Log + PCA + Scale	96.63	38.76	61.8	62.22

Cuadro 6: Porcentajes de casos macheados en pares, dataset wine.

El problema se resuelve muy bien aplicando k-means, y la clasificación mejora mucho escalando los datos. Esto tiene sentido porque hay muchas variables numéricas y las escalas son hasta dos órdenes de magnitud distintas. H-clust complete es el mejor método jerárquico en cuanto a resultado, aunque no tiene mucho sentido aplicarlo porque no estamos buscando jerarquías.

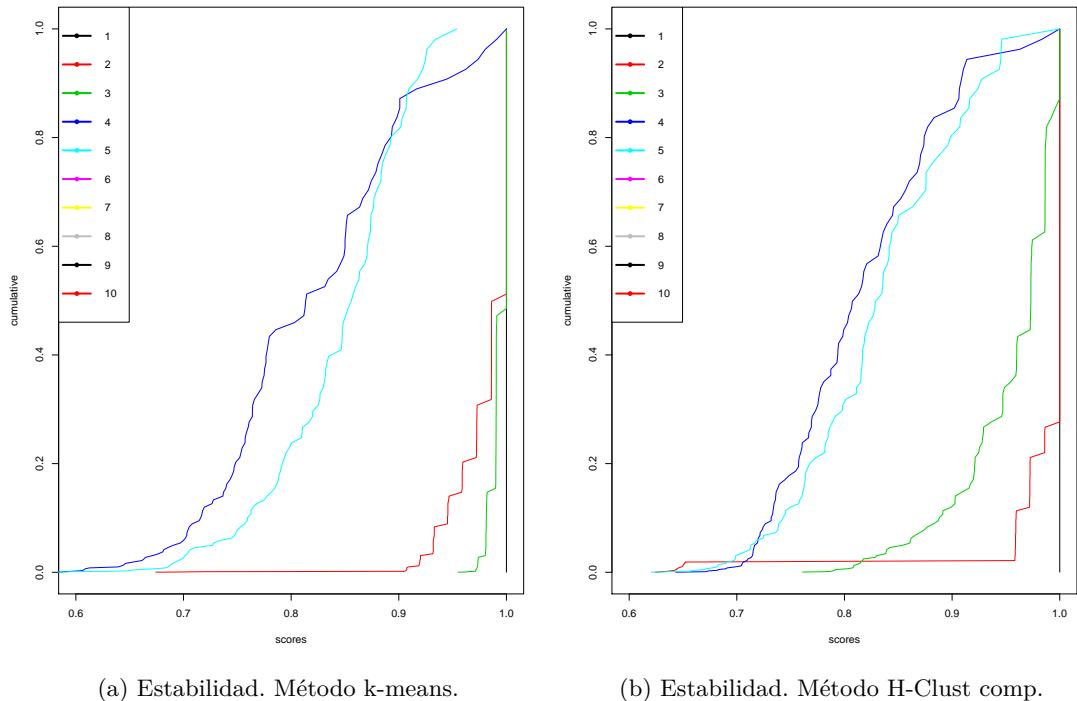
Respecto al número de clusters óptimo, gapStatistic arroja los siguientes números:

	Kmeans	HclA	HclC
Wine raw	1	1	1
Wine + log + pca	3	1	1
Wine + log + scale	3	1	1
Wine + log + scale + pca	1	1	1(2)
Wine + log + pca + scale	3	1	1

Cuadro 7: Número óptimo de clusters para cada problema según el metodo GapStatistic. Moda de 10 ejecuciones.

Nuevamente, k-means recupera el resultado correcto con gap-statistic. Observando los ploteos, vemos que los tres clusters son bastante compactos y diferenciados; por lo tanto tiene sentido que un método sencillo como k-means funcione bien.

Para el análisis de estabilidad, veamos la gráfica de las acumuladas:



Las soluciones más estables son 1, 2 y 3. Y eligiendo la mayor, el número de clusters indicado sería 3 (guiándonos por k-means, que el más indicado para este problema).