Sistemas de recomendación a partir de técnicas de clústering basadas en la estimación tipo núcleo de la densidad

Lucía López López[1] y Paula Saavedra-Nieves[2]

[1]Universidad de Santiago de Compostela

[2]CITMAga, Universidad de Santiago de Compostela

X Jornada de Usuarios de R en Galicia

Índice

Introducción

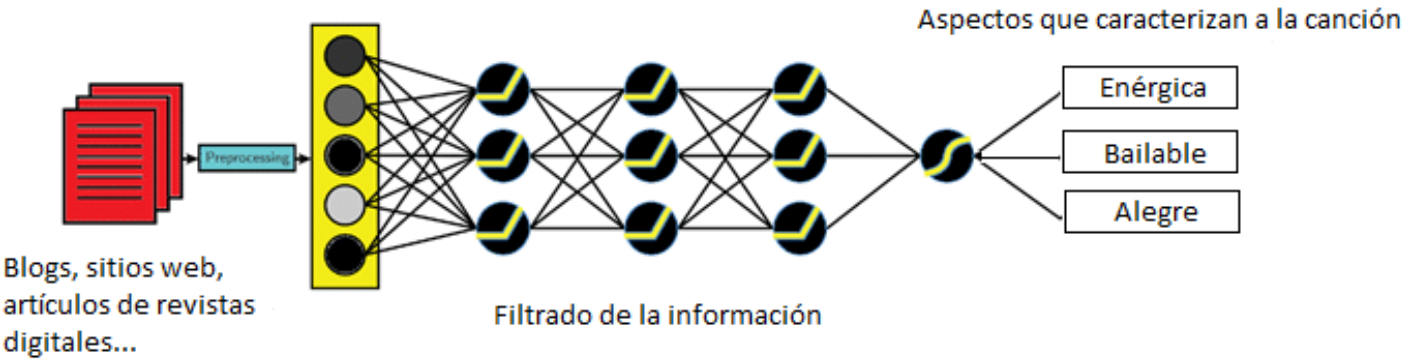
cuenta con un sistema de recomendación propio: BaRT (Bandits for Recommendations as Treatments) que combina tres modelos de recomendación diferentes:

- Modelo de filtrado colaborativo
- Modelo de procesamiento de lenguaje natural
- Modelo de audio sin procesar

Introducción

- Modelo de filtrado colaborativo

- Modelo de procesamiento de lenguaje natural



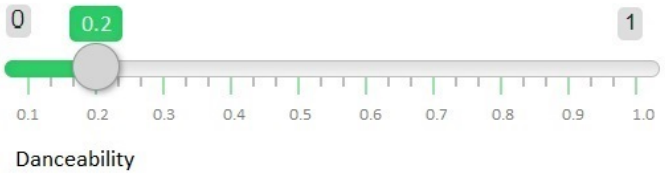
- Modelo de audio sin procesar

title	artist	album	release date	duration_ms	popularity	danceability	energy	key	loudness	mode	speechiness
Arni (Home World)	Yasunori Mitsuda	Chrono Cross Original Soundtrack	1999	203133	30	0.592	0.366	8	-11.695	1	0.0296
Fireflies	Owl City	Ocean Eyes	2009-01-01	228347	79	0.512	0.662	3	-6.797	1	0.0439
Zenzenzense - movie ver.	RADWIMPS	Your Name.	2016-08-24	285880	66	0.321	0.906	11	-3.967	1	0.1410
Saving Grace	Kodaline	One Day at a Time	2020-06-12	230899	55	0.450	0.718	3	-5.950	1	0.0295
New Page	INTERSECTION	New Page	2020-01-08	270904	54	0.604	0.733	2	-5.788	1	0.0348

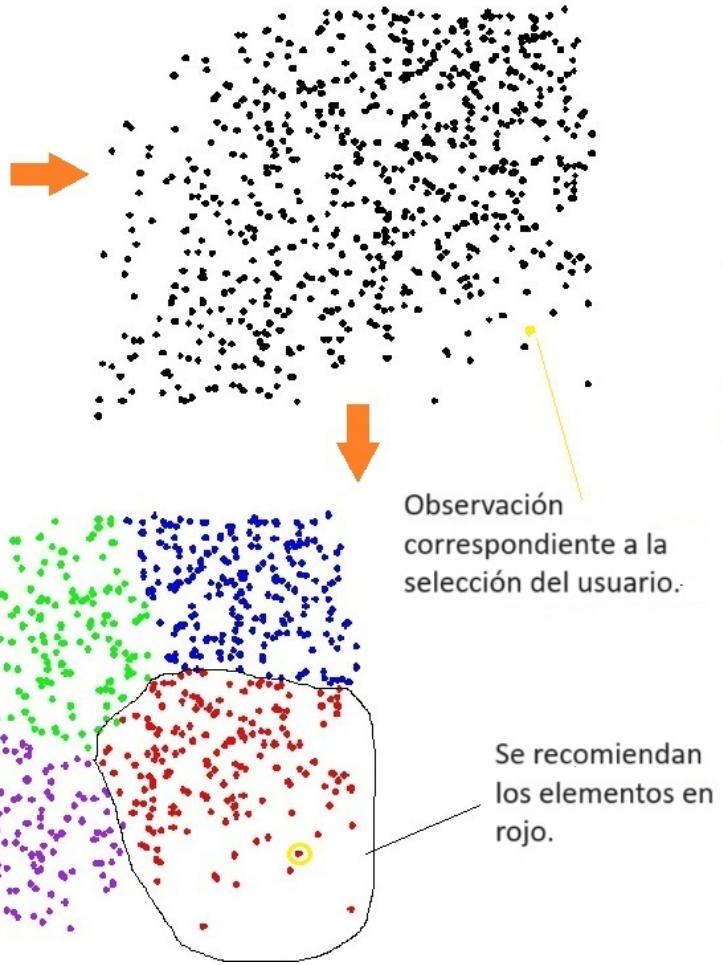
¿Inconvenientes del algoritmo de Spotify?

✗ Las recomendaciones basadas en nuestro historial son poco flexibles. Si nuestros gustos variasen, las sugerencias no se adaptarían a las preferencias actuales.
✗ Podemos obtener resultados poco satisfactorios si buscamos canciones en función de un tema.

Introducción



Ejemplo de barras deslizantes para la introducción de valores concretos de las variables *danceability* y *valence* deseados por el usuario.



Base de datos: Ay, Y. E. (2021). Spotify dataset 1921-2020, 600k+ tracks.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Borges, B. (2023). **shiny**: Web application framework for r [Manual de software informático].

El paquete `spotifyr` nos permite descargar datos de Spotify.

Función	Descripción
`get_album_tracks()`	Permite acceder a toda la información que recoge Spotify de las canciones de un determinado álbum.
`get_playlist_audio_features()`	Proporciona todas las características técnicas de las canciones de una playlist.
`get_my_playlists()`	Devuelve una lista con todas las playlists del usuario.
`get_my_saved_tracks()`	Devuelve una lista con las canciones que ha guardado el usuario en su biblioteca.

Thompson, C., Antal, D., Parry, J., Phipps, D., y Wolff, T. (2022). **spotifyr**: R wrapper for the 'spotify' web api [Manual de software informático].

Clústering basado en la estimación de regiones de elevada densidad

Hartigan (1975) estableció que "los clusters pueden considerarse como regiones de alta densidad separadas de otras tales regiones por áreas de baja densidad".

Bajo esta perspectiva, se identifican los clusters con las componentes conexas de los conjuntos de nivel t (con $t > 0$). Dada una muestra $\mathcal{X}_n = \{X_1, X_2, ..., X_n\}$ de un vector aleatorio X $d-$dimensional con función de densidad f, el conjunto de nivel t se define como

$$G(t) = \{x \in \mathbb{R}^d : f(x) \geq t\}.$$

Hartigan, J. (1975). Clustering Algorithms. J.Wiley & Sons.

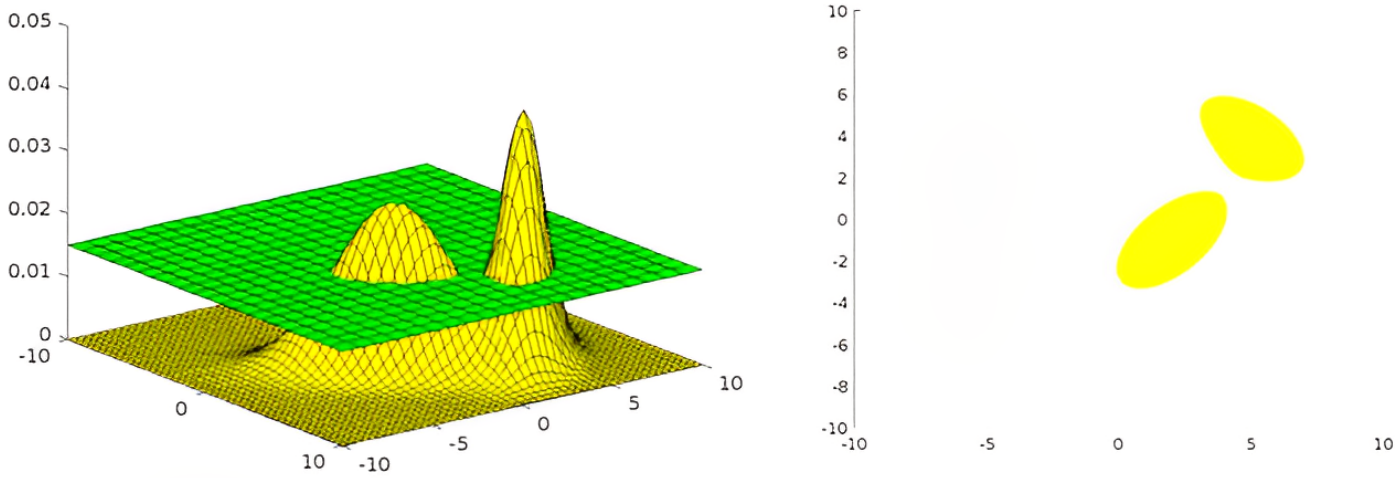


Figura 1: Función de densidad mixtura de dos normales bivariantes intersecada con el plano asociado al nivel t=0.015 (izquierda). Contornos del conjunto de nivel $G(t)$ con t=0.015 (derecha).

En la práctica, f suele ser desconocida, por ello, es más operativo establecer el contenido en probabilidad $1-\tau$, con $\tau \in (0,1)$, del conjunto de nivel que seleccionar t.

Clústering basado en la estimación de regiones de elevada densidad

Fijado τ, se define la región de elevada densidad (*highest density region*, HDR) como el conjunto

$$L(\tau) = \{x \in \mathbb{R}^d : f(x) \geq f_\tau\}$$

donde f_τ representa la mayor constante positiva tal que

$$\mathbb{P}(X \in L(\tau)) \geq 1 - \tau$$

con $\mathbb{P}$ denotando la distribución de probabilidad inducida por f.

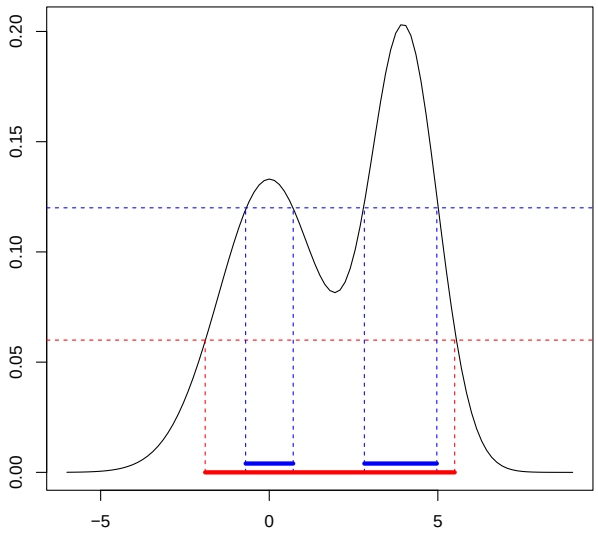


Figura 2: Función de densidad mixtura de dos normales univariantes junto con $L(\tau)$ para τ=0.1 (rojo) y τ=0.5 (azul).

Función de modas y árbol de clusters

Azzalini y Torelli (2007) introdujeron la función de modas m, una función escalonada que asigna el número de componentes conexas de $L(\tau)$ al contenido en probabilidad $1-\tau$, con τ variando entre 0 y 1.

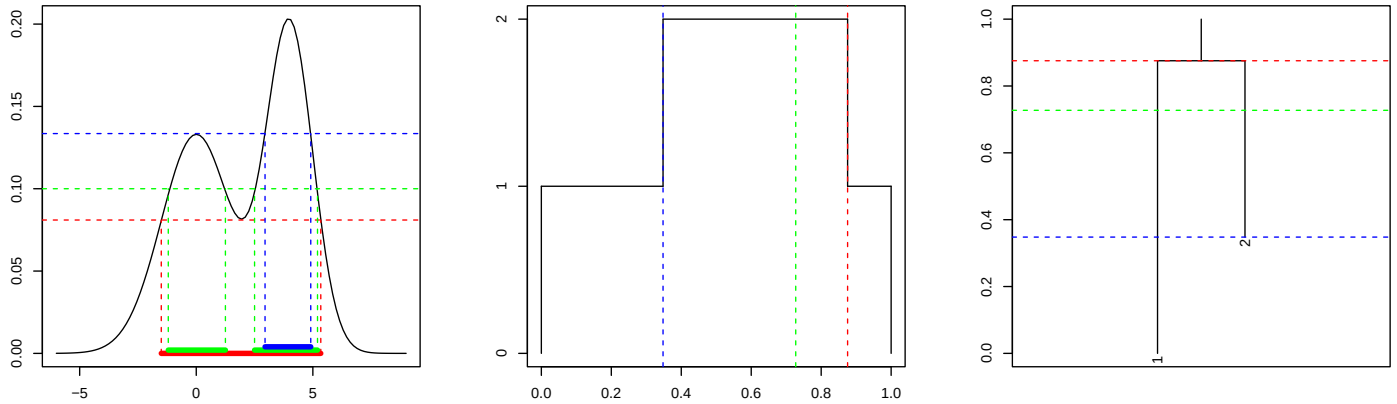


Figura 3: Función de densidad mixtura de dos normales univariantes junto con $L(\tau)$ para τ=0.65 (azul), τ=0.27 (verde) y τ=0.125 (rojo) (izquierda), función de modas (centro) y árbol de clusters asociados (derecha).

Azzalini, A., y Torelli, N. (2007). Clustering via nonparametric density estimation. Statistics and Computing, 17 (1), 71-80.

Estimación de regiones de elevada densidad

En la práctica, es habitual disponer de una muestra de un vector aleatorio con función de densidad f desconocida. Por lo tanto, será necesario estimar las HDR. En la literatura existen tres alternativas para este fin:

- La única información que tenemos son los puntos de la muestra:
  - Métodos *plug – in*
- Se conoce a priori alguna característica geométrica de las HDR:
  - Métodos de *exceso de masa*
  - Métodos *híbridos*

Estimación de regiones de elevada densidad

Los **métodos *plug-in*** proponen

$$\hat{L}(\tau) = \{x \in \mathbb{R}^d : f_n(x) \geq \hat{f}_\tau\}$$

como estimador de $L(\tau)$, donde

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - X_i), \tag{1}$$

siendo K una función de densidad simétrica, $K_H(z) = |H|^{-\frac{1}{2}} K(H^{-\frac{1}{2}} z)$, H la matriz de ventanas y $\hat{f}_\tau$ un estimador de f_τ.
Hyndmann (1996) propone tomar como $\hat{f}_\tau$ el cuantil τ de la distribución empírica de $f_n(X_1),\ldots,f_n(X_n)$.


Hyndman, R. J. (1996). Computing and graphing highest density regions. The American Statistician, 50 , 120-126.

Estimación de regiones de elevada densidad

- En general, la selección de H se realizaba minimizando criterios de error que miden discrepancias entre f y f_n.
- Sin embargo, si el objetivo es estimar las HDR y no la densidad, la ventana debería seleccionarse de forma diferente, pues la estimación de la densidad debe priorizarse en una cierta región en lugar de en todo el dominio.


Samworth, R. J., y Wand, M. P. (2010). Asymptotics and optimal bandwidth selection for highest density region estimation. The Annals of Statistics, 38 (3), 1767 – 1792.


Hyndman, R. J., Einbeck, J., y Wand, M. P. (2021). **hdrcde**: Highest density conditional density estimation [Manual de software informático].


Doss, C. R., y Weng, G. (2018). Bandwidth selection for kernel density estimators of multivariate level sets and highest density regions. Electronic Journal of Statistics, 12 (2), 4313 – 4376.


Weng, G. (2018). **lsbs**: Bandwidth selection for level sets and hdr estimation [Manual de software informático].

Estimación de regiones de elevada densidad

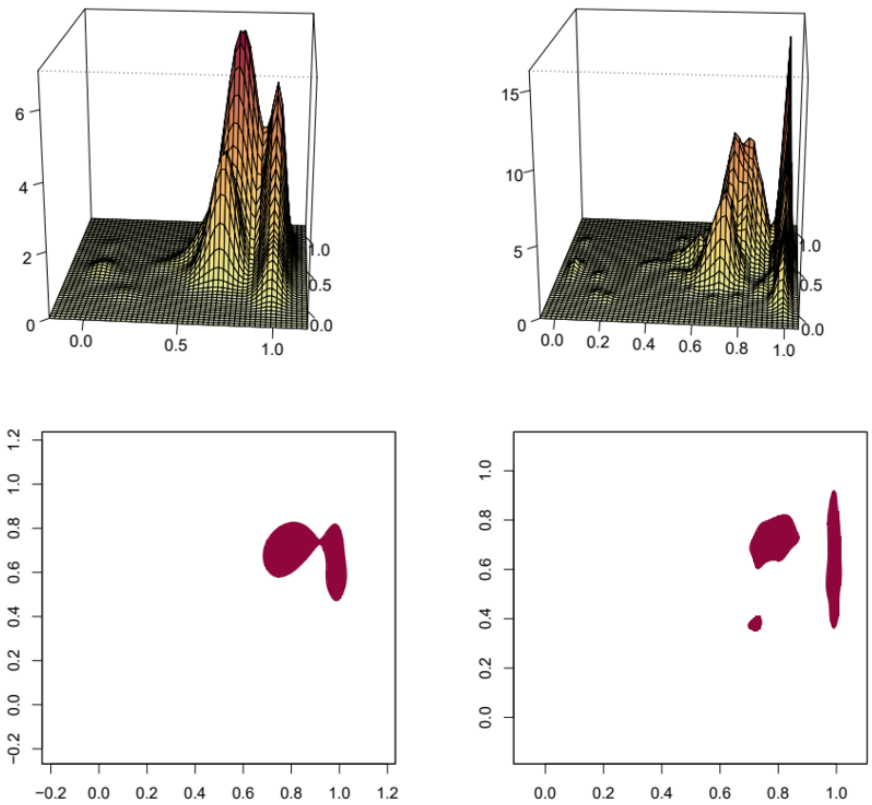


Figura 4: Estimación tipo núcleo de la densidad del vector aleatorio formado por las variables *acousticness* y *valence* junto con $L(\tau)$ para τ=0.5 utilizando la regla del pulgar de Silverman (1986) (izquierda) y el selector propuesto en Doss y Weng (2018) (derecha).

Silverman, B. (1986). Density estimation for statistics and data analysis. Chapman & Hall.

Duong, T. (2022). **ks**: Kernel smoothing [Manual de software informático].

Cálculo del número de componentes conexas

El enfoque $plug - in$ nos permitirá construir la versión empírica de la función de modas, $\hat{m}$, que a cada $1 - \tau$ (con $\tau \in (0, 1)$) le asignará el número de componentes conexas de $\hat{L}(\tau)$.

? Pero, ¿cómo calculamos el número de componentes conexas de $\hat{L}(\tau)$?

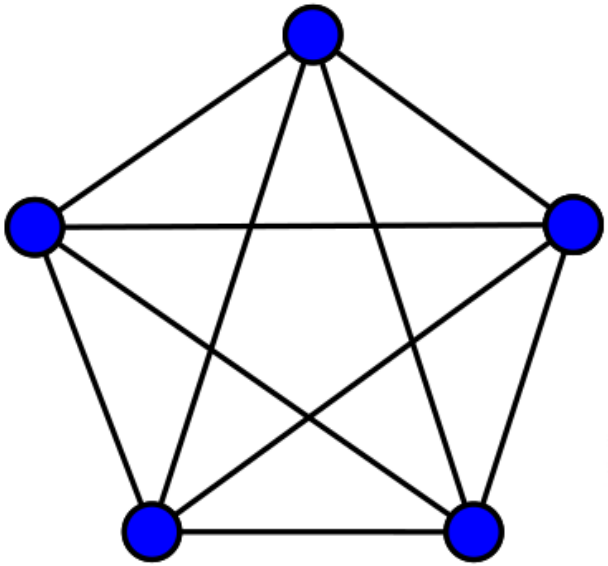
- Azzalini y Torelli (2007) proponen un método basado en la construcción de la triangulación de Delaunay de la muestra.

  La complejidad computacional de este procedimiento crece exponencialmente con la dimensión de los datos.

Azzalini, A., y Torelli, N. (2007). Clustering via nonparametric density estimation. Statistics and Computing, 17 (1), 71-80.

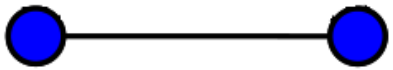
Cálculo del número de componentes conexas

- Stuetzle y Nugent (2010) proponen construir un grafo completo ponderado tomando como vértices los puntos de la muestra.



Peso arista $X_i - X_j$:

$$\min_{T \in [0,1]} f_n((1-T)X_i + TX_j)$$



Peso vértice X_i: $f_n(X_i)$

Peso vértice X_j: $f_n(X_j)$

Luego, se extrae el subgrafo formado por los arcos y nodos con peso mayor o igual que $\hat{f}_\tau$ y se calcula el número de componentes conexas.


Stuetzle, W., y Nugent, R. (2010). A generalized single linkage method for estimating the cluster tree of a density. Journal of Computational and Graphical Statistics, 19(2), 397-418.


Even, S. (2011). Graph algorithms 2nd ed. Cambridge University Press

Cálculo del número de componentes conexas

- Menardi y Azzalini (2014) establecen una perspectiva similar, pero teniendo en cuenta la variabilidad de f_n.
  Dadas dos observaciones X_i, X_j, la evaluación de la densidad estimada a lo largo del segmento que las une se define como

$$\varphi(T) = f_n(TX_i + (1-T)X_j), \quad T \in [0,1].$$

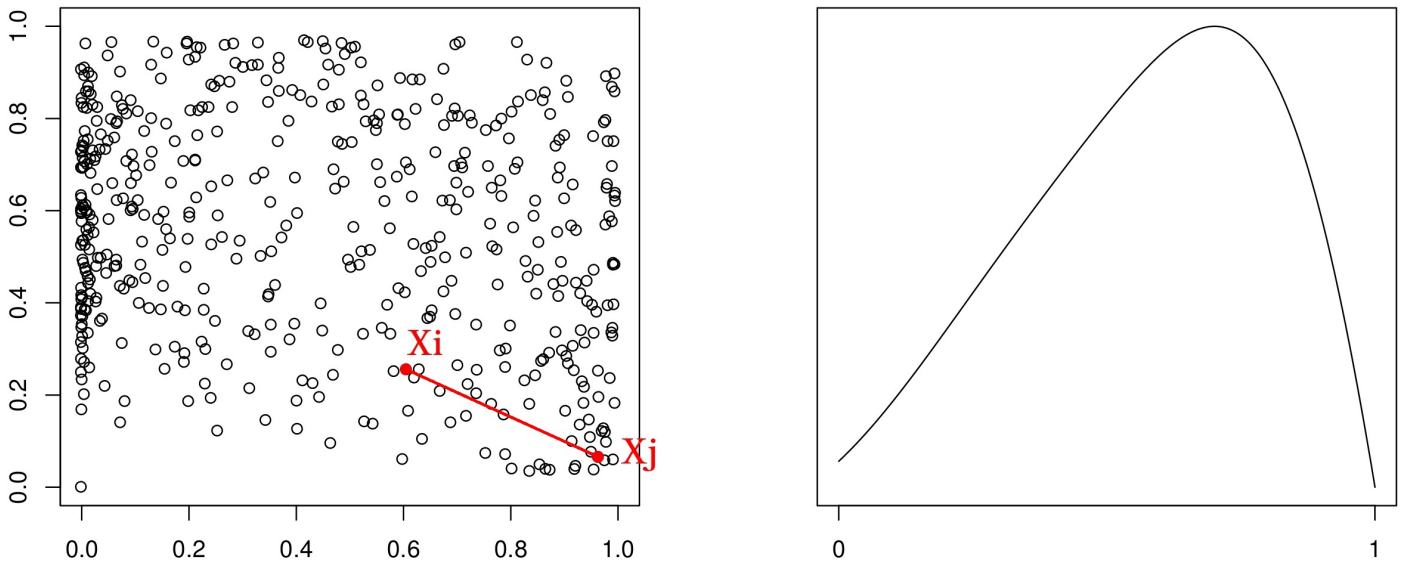


Figura 5: Diagrama de dispersión de *acousticness* frente a *valence* (izquierda) y $\varphi(T)$ para el par de puntos marcados en rojo (derecha).

Cálculo del número de componentes conexas

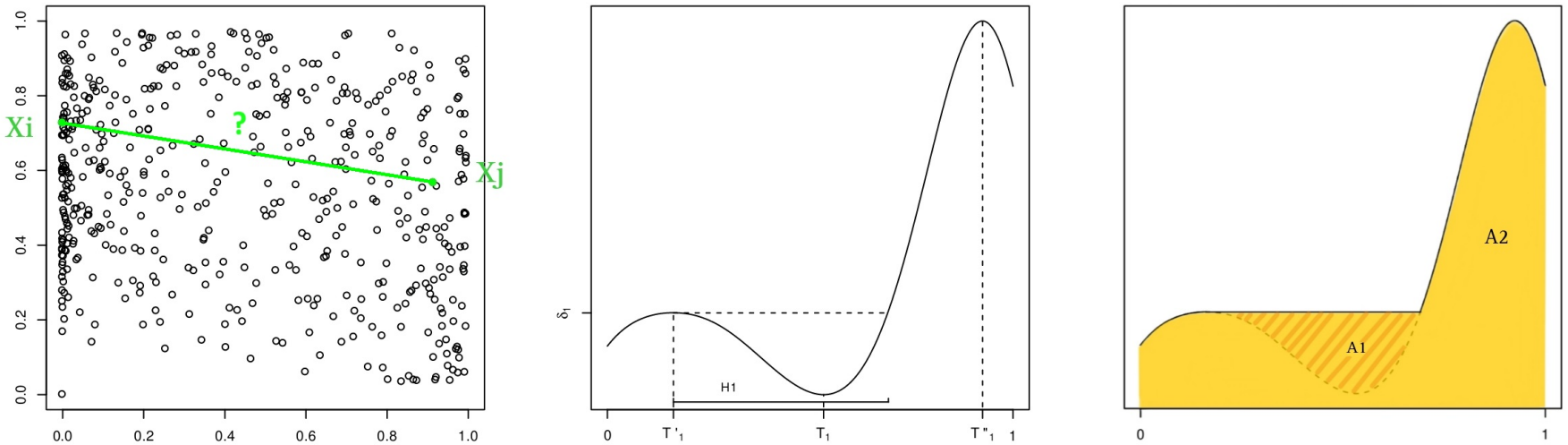


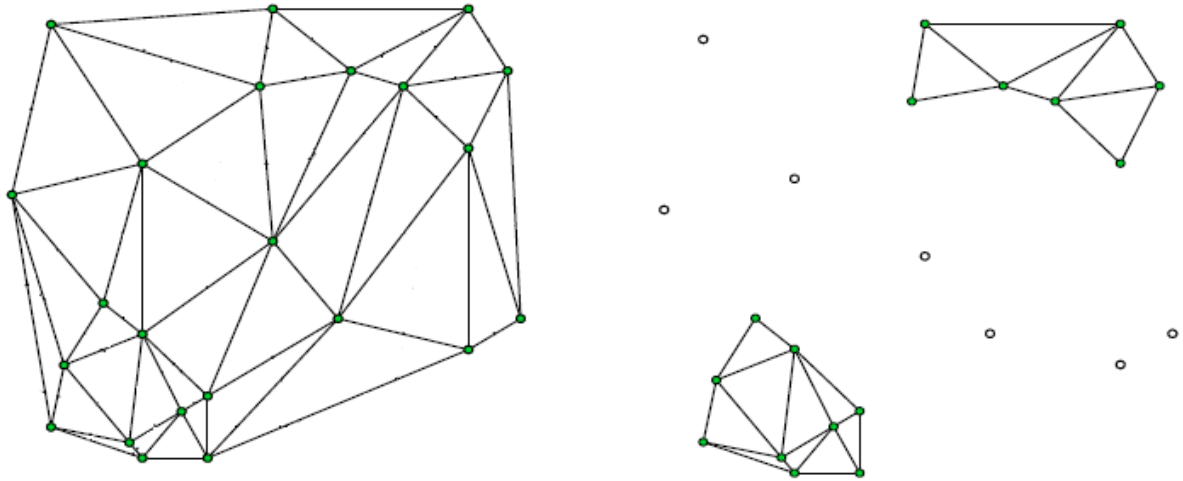
Figura 6: Diagrama de dispersión de *acousticness* frente a *valence* (izquierda), evaluación de f_n a lo largo del segmento que une los puntos marcados en verde ($\varphi(T)$) (centro), modificación de $\varphi(T)$ para calcular la profundidad del valle.

No hay un método establecido para seleccionar el umbral. Los autores apuntan, que según su experiencia numérica, λ =0.1 suele ser una elección adecuada.

Menardi, G., y Azzalini, A. (2014). An advancement in clustering via nonparametric density estimation. Statistics and Computing, 24 , 753-767.

Cálculo del número de componentes conexas

Una vez construido el grafo, dado $\tau \in (0,1)$ se extrae del grafo inicial el subgrafo formado por los vértices con densidad igual o superior a $\hat{f}_\tau$. Por último, se identifican las componentes conexas de dicho subgrafo.




Menardi, G., y Azzalini, A. (2014). An advancement in clustering via nonparametric density estimation. Statistics and Computing, 24 , 753-767.


Even, S. (2011). Graph algorithms 2nd ed. Cambridge University Press

Función de modas y árbol de clusters empíricos

Para construir la versión empírica de la función de modas, bastará con considerar una rejilla de valores equiespaciados de $1 - \tau$ suficientemente fina y aplicar, por ejemplo, el algoritmo de Menardi y Azzalini (2014).

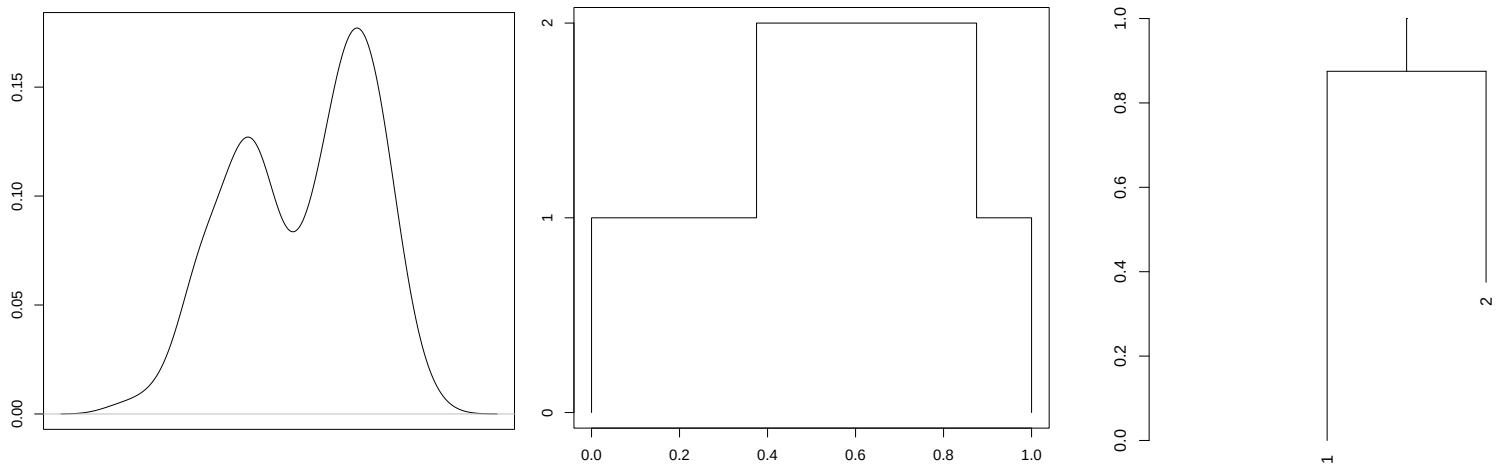


Figura 7: Estimación tipo núcleo de la función de densidad (izquierda), función de modas empírica (centro) y árbol de clusters empírico (derecha) asociados a una muestra de tamaño 450 de una mixtura de dos normales univariantes.

Azzalini, A., y Menardi, G. (2014). Clustering via nonparametric density estimation: The R package **pdfCluster**. Journal of Statistical Software, 57 (11), 1–26.

Función de modas y árbol de clusters empíricos

- Para $1 - \tau$ =0.875:

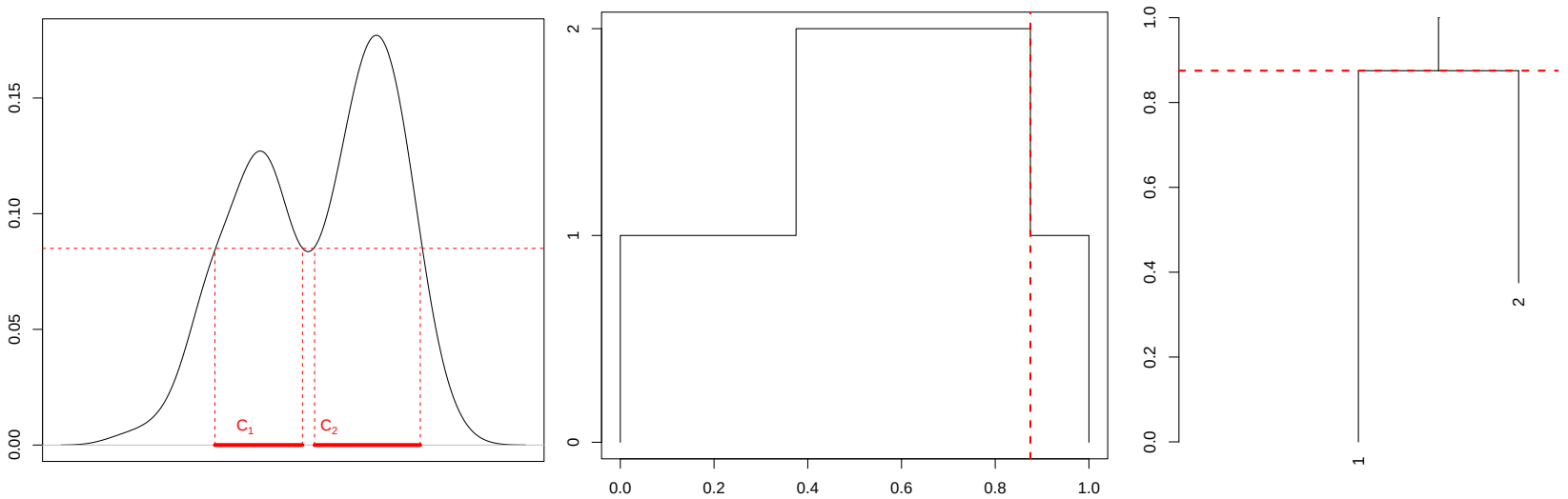


Figura 8: Estimación tipo núcleo de la función de densidad (izquierda), función de modas empírica (centro) y árbol de clusters empírico (derecha) asociados a una muestra de tamaño 450 de una mixtura de dos normales univariantes.

Azzalini, A., y Menardi, G. (2014). Clustering via nonparametric density estimation: The R package **pdfCluster**. Journal of Statistical Software, 57 (11), 1–26.

Función de modas y árbol de clusters empíricos

Los puntos de la muestra que pertenecen a las mayores componentes conexas que contienen una única moda, son denominadas por Azzalini y Torelli (2007) como *cluster cores*. Denotaremos por M el número de *cluster cores*.

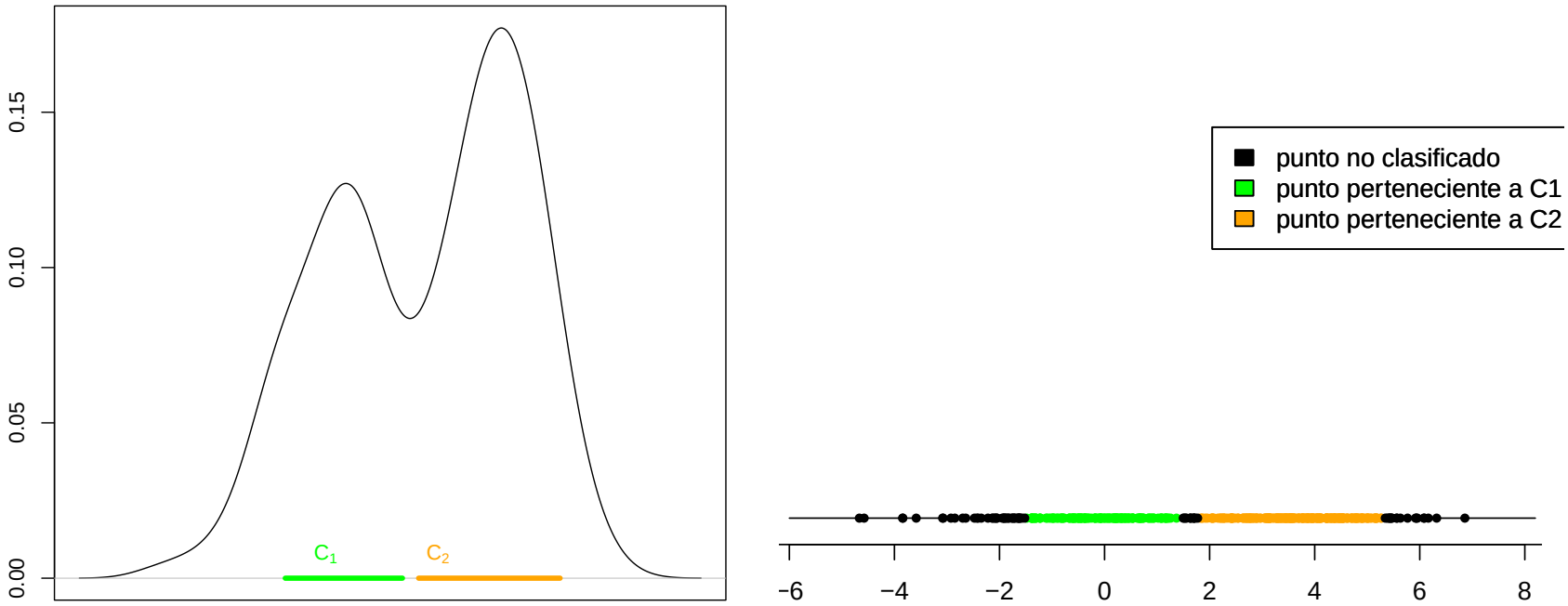


Figura 9: *Cluster cores* ($M = 2$) identificados para la muestra de una mixtura de dos normales univariantes.

Clasificación propuesta por Azzalini y Torelli (2007)

Dado un elemento no etiquetado $x_0 \in \mathcal{X}_n$:

1. Determinar la estimación tipo núcleo de la densidad $f_{n,m}(x_0)$ basándose en las observaciones de la muestra que han sido asignadas al grupo m para cada $m = 1, ..., M$.
2. Para cada $m = 1, ..., M$, calcular

$$r_m(x_0) = \frac{f_{n,m}(x_0)}{\max\limits_{i \neq m} f_{n,i}(x_0)}.$$

   Si $f_{n,m}(x_0) = 0 \ \forall m \in \{1, 2, ..., M\}$, entonces x_0 se considerará un dato aislado y no se tendrá en cuenta en la agrupación.

3. Asignar x_0 al grupo $m_0 \in \{1, ..., M\}$ que verifique

$$r_{m_0}(x_0) = \max\{r_m(x_0), m = 1, ..., M\}.$$

Paquete pdfCluster de

Función	Descripción
`pdfCluster()`	Realiza clústering basado en la estimación de regiones de elevada densidad empleando el método de Azzalini y Torelli (2007) o el de Menardi y Azzalini (2014).
`groups()`	Extrae los grupos detectados de los objetos de la clase `pdfCluster`. Si igualamos el argumento `stage` a 0 obtendremos los *clúster cores*.
`plot()`	Aplicando esta función a un objeto de clase `pdfCluster` obtendremos la gráfica de la función de modas y el árbol de clústers.
`dbs()`	Calcula la información de *Silhouette* basada en densidad.

Azzalini, A., y Menardi, G. (2014). Clustering via nonparametric density estimation: The R package **pdfCluster**. Journal of Statistical Software, 57 (11), 1–26.

Validación para métodos de clústering basados en densidad

- Información de *Silhouette* (s_i)
- Información de *Silhouette* basada en densidad (dbs_i)

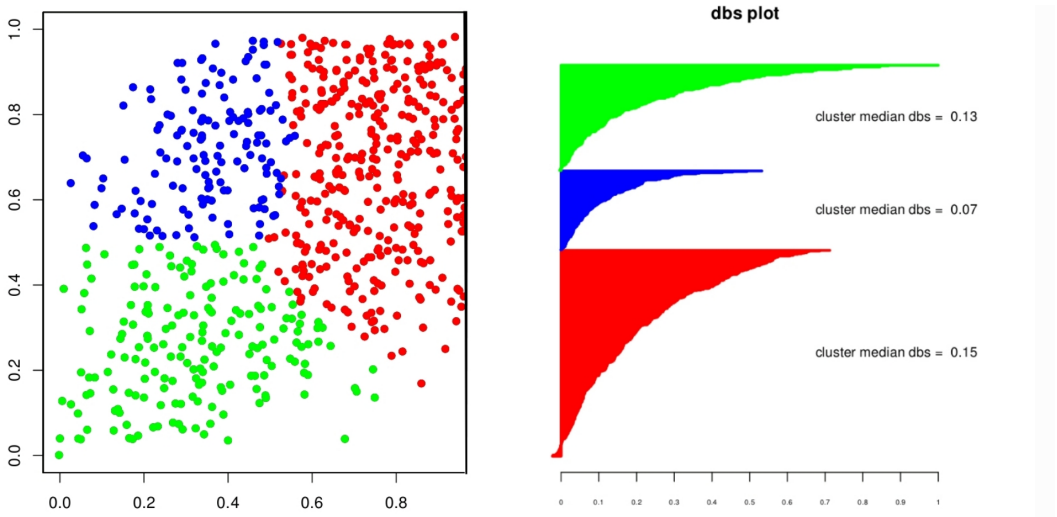


Figura 10: Clasificación de una muestra de tamaño 700 de las variables *energy* y *valence* empleando el método de Menardi y Azzalini (2014) (izquierda). Diagrama de silueta basada en densidad (derecha).

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20 , 53-65.

Menardi, G. (2011). Density-based silhouette diagnostics for clustering methods. Statistics and Computing, 21 , 295–308.

duration_ms	*danceability*	*energy*	*loudness*	*speechiness*
400000	0.7	0.7	-9	0
acousticness	*instrumentalness*	*liveness*	*valence*	*tempo*
0	0.2	0	0.9	100

Cuadro 1: Selección del usuario.

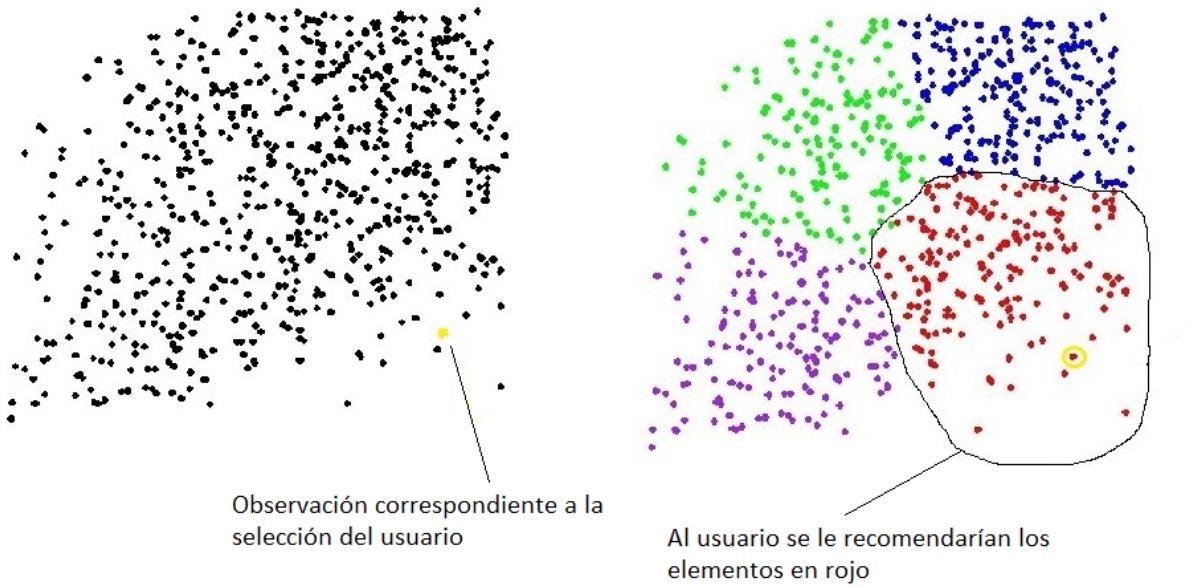


Figura 11: Ilustración del funcionamiento del sistema.

Cluster core	Centroide (vector de medias)	Miembros
C1	(212245.81, 0.63, 0.61, -8.28, 0.08, 0.34, 0.01, 0.15, 0.56, 117.69)	677
C2	(111245.23, 0.70, 0.21, -22.77, 0.96, 0.68, 0.00, 0.25, 0.59, 100.59)	52
C3	(169297.88, 0.42, 0.13, -23.17, 0.07, 0.99, 0.88, 0.15, 0.45, 90.30)	26
C4	(180200.00, 0.76, 0.36, -12.66, 0.09, 0.98, 0.84, 0.17, 0.90, 110.00)	5
C5	(104298.00, 0.64, 0.64, -10.34, 0.91, 0.24, 0.00, 0.34, 0.51, 101.97)	6

Cuadro 2: Resultados obtenidos con el método de Menardi y Azzalini (2014).

Clúster	Centroide (vector de medias)	Miembros
C'1	(206846.60, 0.51, 0.29, -13.10, 0.06, 0.80, 0.02, 0.19, 0.41, 106.32)	248
C'2	(277944.86, 0.49, 0.72, -7.72, 0.08, 0.17, 0.10, 0.27, 0.45, 131.43)	318
C'3	(113049.11, 0.68, 0.34, -19.00, 0.92, 0.57, 0.00, 0.30, 0.58, 98.10)	84
C'4	(207930.11, 0.72, 0.71, -7.20, 0.11, 0.23, 0.03, 0.16, 0.70, 118.46)	467
C'5	(184338.12, 0.49, 0.24, -18.49, 0.08, 0.85, 0.80, 0.16, 0.49, 106.36)	84

Cuadro 3: Resultados obtenidos con el método $k-$medias para $k = 5$, implementado en el paquete `stats` de R.

Información de *Silhouette* basada en densidad

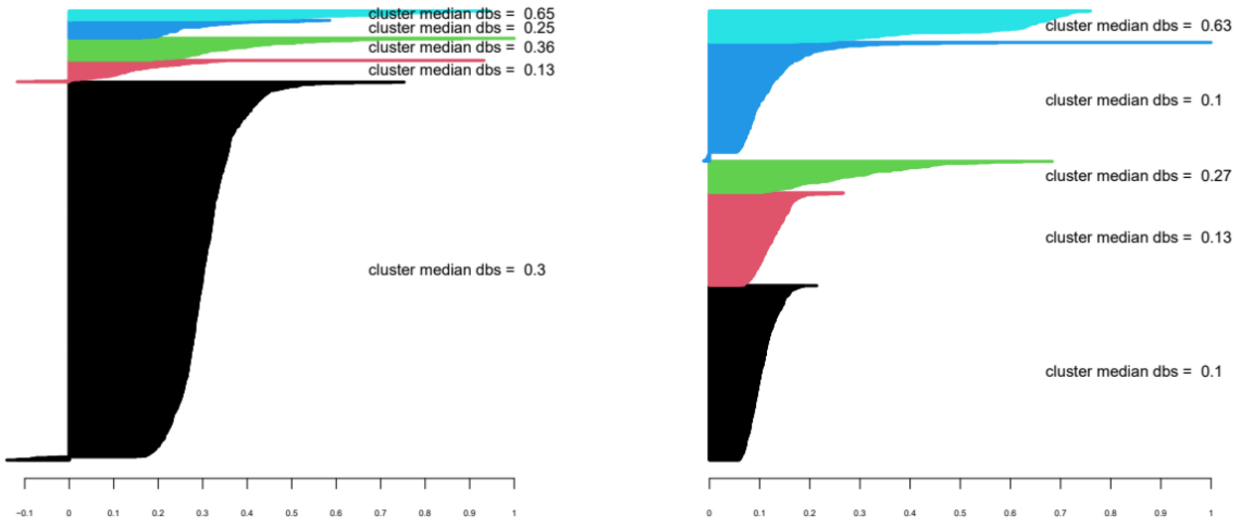


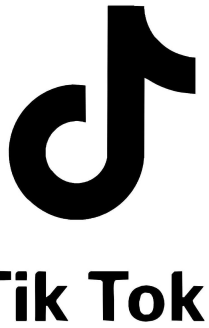
Figura 12: Gráficos de *Silhouette* basados en densidad para las particiones alcanzadas con el método de Menardi y Azzalini (2014) (izquierda) y con el método de $k-$medias para $k = 5$ (derecha).

- El valor para la selección del usuario es de 0.28 con el método basado en densidad y de 0.11 con $k-$medias.
- La información de *Silhouette* estándar para la selección del usuario es de 0.17 en ambos casos.

Conclusiones

Posibles extensiones de este trabajo:

- Introducción de variables continuas de usuario como la edad.
- Introducción de variables categóricas (género, nacionalidad, tipo de suscripción...). Método descrito en Azzalini y Menardi (2016).
- Aplicación a otros contextos (chequeando previamente el comportamiento práctico). Por ejemplo a plataformas de *streaming* de series y películas o a redes sociales.



Azzalini, A., y Menardi, G. (2016). Density-based clustering with non-continuous data. Journal of the American Statistical Association(31), 771–798.

Gracias.