27 Congreso Nacional de Estadística e Investigación Operativa
Lleida, 8-11 de abril de 2003

GINKGO, UN PROGRAMA DE ANÁLISIS MULTIVARIANTE ORIENTADO A LA CLASIFICACIÓN BASADA EN DISTANCIAS

M. De Cáceres[1], F. Oliva[2], Xavier Font[1].

[1]Departamento de Biologia Vegetal
Universidad de Barcelona, 08028 Barcelona, España
E-mail: mcaceres@bio.ub.es
[2]Departamento de Estadística
Universidad de Barcelona, 08028 Barcelona, España
E-mail: francesc@bio.ub.es

RESUMEN

GINKGO es una aplicación orientada a la representación y clasificación de individuos a partir de datos multivariantes. Las principales técnicas que contempla son:

- Cluster analysis: algoritmos jerárquicos, partitivos (K-means y fuzzy C-means) y de clusters independientes (Possibilistic C-means). Presenta además la particularidad de poder utilizar todas las técnicas anteriores a partir de una matriz de proximidades.
- Análisis discriminante: lineal, cuadrático y basado en distancias.
- Reducción de la dimensión y representación: PCA, MDS, NMDS y CA

GINKGO ha sido desarrollado en lenguaje Java y tiene una distribución libre, basada en la tecnología Java Web Start, que permite actualizaciones automáticas.

Palabras y frases clave: Software multivariante, Clasificación, Cluster Analysis, Análisis Discriminante, Representación de dados, Reducción de la dimensión.
Clasificación AMS: 62H30 62-07.

1. Introducción

A pesar de la existencia de numerosos programas en estadística multivariante, la mayoria de ellos presentan siempre los mismos métodos clásicos, con lo que los usuarios no expertos desconocen técnicas de análisis más especializadas que pueden resultar más adecuadas para sus necesidades. Además, dichos programas suelen ser difíciles de emplear para usuarios no especializados.

GINKGO intenta acercar varias herramientas de análisis multivariante hacia usuarios no expertos en estadística, a través de una interfaz gráfica de usuario sencilla. La interfaz de usuario presenta 3 ventanas internas: el editor de datos, el gestor de análisis y el gestor de gráficos. Las operaciones que generan información añaden nuevos elementos a la ventana correspondiente. El programa permite guardar tanto matrices de datos como resultados de análisis en un solo fichero de proyecto, con lo que un proceso largo de análisis de datos puede realizarse en varias sesiones.

La interfaz de usuario descrita proporciona un marco integrado de trabajo que permite la exploración paso a paso de datos multivariantes. En primer lugar, permite elegir entre distintos coeficientes de similaridad y disimilaridad adecuados a los datos de los que se dispone. A continuación, la estructura de éstos puede ir siendo dilucidada empleando las distintas técnicas de reducción de la dimensionalidad y clasificación incluidas en el programa. Finalmente, los resultados obtenidos con el empleo de distintos espacios y/o técnicas de análisis pueden ser luego comparados en el mismo programa. Cabe resaltar, que es sencillo traspasar de nuevo al editor de datos las matrices generadas en un análisis, con lo que se pueden realizar procesos de análisis moderadamente complejos.

2. Operaciones básicas

2.1 Edición de datos multivariantes

El editor de datos de GINGKO permite una edición directa de matrices de datos. Se diferencian dos tipos de matrices, rectangulares (objeto-descriptor) y simétricas (objeto-objeto o descriptor-descriptor). Ambos tipos de matrices de datos se pueden crear dentro del programa aunque también se permite la importación de datos en formato de texto ASCII. Además de los estadísticos univariantes habituales, el programa ofrece otras operaciones comunes, como son la estandarización de variables, la transposición de matrices o el cálculo de matrices de covariancia/correlación.

2.2 Matrices de semejanza

Es un problema habitual en usuarios de aplicaciones de estadística multivariante el estar limitado al uso de la Distancia Euclídea en los paquetes estadísticos convencionales. Esto es así debido a la carencia de otras medidas de distancia en dichos paquetes y al empleo implícito o explícito de la Distancia Euclídea dentro de las mismas técnicas multivariantes que los programas comunes ofrecen. Los usuarios avezados a este problema acostumbran a transformar previamente los datos antes de empezar el análisis o emplear técnicas de *scaling*. No obstante, muchos usuarios inexpertos se conforman con lo que el paquete estadístico les ofrece. GINKGO intenta suplir estas carencias de varios modos. En primer lugar, incorpora medidas de similaridad y distancia poco comunes en los paquetes estadísticos estándar y que son frecuentemente utilizadas en ámbitos de aplicación, como en ecología. Las medidas de similaridad y disimilaridad disponibles se encuentran listadas en la tabla 1 (para una referencia más detallada ver p.e. Legendre & Legendre 1998). En segundo lugar, GINKGO ofrece la posibilidad de transformar matrices de similaridad en disimilaridad (p.e. con la transformación de Gower) y viceversa. Finalmente, proporciona técnicas de ordenación y clasificación aplicables a matrices de disimilaridades, conservando las propiedades de la métrica usada íntegramente en los análisis posteriores. GINKGO permite además la comparación de matrices de semejanza mediante el cálculo de correlaciones y los diagramas de Shepard.

3. Reducción de la dimensionalidad

GINKGO permite el uso de varias técnicas clásicas de reducción de la dimensionalidad:

- Análisis de Componentes Principales (*PCA*)
- Análisis de Coordenadas Principales (o *Metric scaling*, Gower 1966)
- *Multidimensional scaling* No Métrico (*NMDS*, Kruskal 1964a, 1964b).
- Análisis Factorial de Correspondencias (*CA*, Hill 1973).

Una caracteristica que GINKGO ofrece y es poco común en otros programas, es permitir la utilización de forma inmediata de los resultados de *clustering* para distinguir los grupos en las representaciones gráficas de dimensionalidad reducida. De este modo, la interpretación de ambas técnicas de análisis se pueden complementar y contrastar con facilidad.

Medidas de Disimilaridad	Medidas de Similaridad
• Euclidean Distance (ED): $d_{EC}(\mathbf{x}_1,\mathbf{x}_2)=\sqrt{\sum_{j=1}^{P}\left(x_{1j}-x_{2j}\right)^2}$ • Squared ED: $d_{SQEC}(\mathbf{x}_1,\mathbf{x}_2)=\sum_{j=1}^{P}\left(x_{1j}-x_{2j}\right)^2$ • Binary ED: $d_{BINEC}(\mathbf{x}_1,\mathbf{x}_2)=\sqrt{\sum_{j=1}^{P}\left(I(x_{1j}>0)-I(x_{2j}>0)\right)^2}$ • Squared Binary ED: $d_{SQBINEC}(\mathbf{x}_1,\mathbf{x}_2)=\sum_{j=1}^{P}\left(I(x_{1j}>0)-I(x_{2j}>0)\right)^2$ • Absolute Value Distance: $d_{ABS}(\mathbf{x}_1,\mathbf{x}_2)=\sqrt{\sum_{j=1}^{P}\left\|x_{1j}-x_{2j}\right\|}$ • Manhattan metric: $d_{Man}(\mathbf{x}_1,\mathbf{x}_2)=\sum_{j=1}^{P}\left\|x_{1j}-x_{2j}\right\|$ • Bray-Curtis Distance: $d_{BC}(\mathbf{x}_1,\mathbf{x}_2)=\frac{\sum_{j=1}^{P}\left\|x_{1j}-x_{2j}\right\|}{\sum_{j=1}^{P}(x_{1j}+x_{2j})}$ • Chord distance: $d_{chord}(\mathbf{x}_1,\mathbf{x}_2)=\sqrt{2\left(1-\frac{\sum_{j=1}^{P}x_{1j}x_{2j}}{\sqrt{\sum_{j=1}^{P}x^2{}_{1j}\sum_{j=1}^{P}x^2{}_{2j}}}\right)}$ • Hellinger Distance: $d_{Hell}(\mathbf{x}_1,\mathbf{x}_2)=\sqrt{\sum_{j=1}^{P}\left[\sqrt{\frac{x_{1j}}{x_{1+}}}-\sqrt{\frac{x_{2j}}{x_{2+}}}\right]^2}$ • χ^2 Distance: $d_{\chi^2}(\mathbf{x}_1,\mathbf{x}_2)=\sqrt{\sum_{j=1}^{P}\frac{1}{x_{+j}/x_{++}}\left(\frac{x_{1j}}{x_{1+}}-\frac{x_{2j}}{x_{2+}}\right)^2}$ • χ^2 Metric: $d_{m\chi^2}(\mathbf{x}_1,\mathbf{x}_2)=\sqrt{\sum_{j=1}^{P}\frac{1}{x_{+j}}\left(\frac{x_{1j}}{x_{1+}}-\frac{x_{2j}}{x_{2+}}\right)^2}$ • Canberra Metric: $d_{Camb}(\mathbf{x}_1,\mathbf{x}_2)=\sum_{j=1}^{P}\left[\frac{\left\|x_{1j}-x_{2j}\right\|}{(x_{1j}+x_{2j})}\right]$ • Mahalanobis Distance: $d_{Mah}(\mathbf{x}_1,\mathbf{x}_2)=\mathbf{d_{12}S^{-1}d'_{12}}$ donde $x_{i+}=\sum_{j=1}^{P}x_{ij}$, $x_{+j}=\sum_{i=1}^{N}x_{ij}$, $x_{++}=\sum_{j=1}^{P}\sum_{i=1}^{N}x_{ij}$, $\mathbf{d_{12}}=\left(x_{11}-x_{21},...,x_{1P}-x_{2P}\right)'$, i S^{-1} matriz var-cov.	• *Simple Matching Coefficient*: $s_1(\mathbf{x}_1,\mathbf{x}_2)=\frac{a+d}{a+b+c+d}$ • Jaccard index: $s_2(\mathbf{x}_1,\mathbf{x}_2)=\frac{a}{a+b+c}$ • Ellenberg index) $s_3(\mathbf{x}_1,\mathbf{x}_2)=\frac{\frac{1}{2}Ma}{\frac{1}{2}Ma+Mb+Mc}$ • Gleason index: $s_4(\mathbf{x}_1,\mathbf{x}_2)=\frac{Ma}{Ma+Mb+Mc}$ • Sørensen index: $s_5(\mathbf{x}_1,\mathbf{x}_2)=\frac{2a}{2a+b+c}=\frac{a}{(2a+b+c)/2}$ • Motyka index: $s_6(\mathbf{x}_1,\mathbf{x}_2)=\frac{\sum_{j=1}^{P}min(x_{1j},x_{2j})}{\sum_{j=1}^{P}x_{1j}+\sum_{j=1}^{P}x_{2j}}$ • Spatz index: $s_7(\mathbf{x}_1,\mathbf{x}_2)=s_4(\mathbf{x}_1,\mathbf{x}_2)\cdot\frac{1}{P}\sum_{ji=1}^{P}\frac{max(x_{1j},x_{2j})}{min(x_{1j},x_{2j})}$ • Kulczynski index: $s_8(\mathbf{x}_1,\mathbf{x}_2)=\frac{1}{2}\cdot\left(\frac{\sum_{j=1}^{P}min(x_{1j},x_{2j})}{\sum_{j=1}^{P}x_{1j}}+\frac{\sum_{j=1}^{P}min(x_{1j},x_{2j})}{\sum_{j=1}^{P}x_{2j}}\right)$ • χ^2 *similarity*: $s_9(\mathbf{x}_1,\mathbf{x}_2)=1-\sqrt{\sum_{j=1}^{P}\frac{1}{x_{+j}}\left(\frac{x_{1j}}{x_{1+}}-\frac{x_{2j}}{x_{2+}}\right)^2}$ • Pearson linear correlation index. donde a son dobles presencias, d son dobles ausencias, b y c son valores distintos de 0 solo en un objeto o el otro, Mc = Σ(valores no nulos en ambos), Ma = Σ (valores no nulos solo en O_1), Mb = Σ (valores no nulos solo en O_2)

Tabla 1: Medidas de disimilaridad y similaridad disponibles en GINKGO

4. Clasificación

4.1 Técnicas de *clustering*

En cuanto a técnicas de *clustering* se refiere, GINKGO permite aplicar 3 modelos distintos de clasificación:

a) *Clustering* Jerárquico Aglomerativo. Se permite la selección entre los algoritmos *Single Linkage*, *Complete Linkage*, *UPGMA*, *WPGMA*, *UPGMC*, *WPGMC*, método de Ward, *Flexible clustering*. La salida de estos algoritmos es una matriz ultramétrica, usada para la representación gráfica de un dendrograma. Cabe resaltar que el programa permite "cortar" un dendrograma por el nivel de similaridad deseado para producir particiones que pueden ser comparadas con el uso de otras técnicas de clasificación.

b) Algoritmos partitivos: Sea **X** una matriz de *N* observaciones *P*-dimensional donde queremos establecer una partición en *K* grupos o *clusters*:

1) *K-means* (MacQueen, 1967). Este algoritmo partitivo se basa en la optimización del siguiente funcional:

$$TESS_K = \sum_{k=1}^{K} E_{(k)}^2 = \sum_{k=1}^{K} \sum_{i=1}^{N} I[O_i \subset C_k]\, e_{i(k)}^2$$

donde el error de cada objeto equivale a la desviación respecto al centroide del grupo al que ha sido asignado:

$$e_{i(k)}^2 = \sum_{j=1}^{P} (x_{ij} - \bar{x}_{(k)j})^2 = (\mathbf{x}_i - \bar{\mathbf{x}}_{(k)})'(\mathbf{x}_i - \bar{\mathbf{x}}_{(k)})$$

2) *Fuzzy C-means* (*FCM,* Bezdek, 1981). Esta generalización de K-means al enfoque basado en la lógica difusa optimiza el siguiente funcional:

$$FTESS_{K,m} = \sum_{k=1}^{K} J_{(k),m}^2 = \sum_{k=1}^{K} \sum_{i=1}^{N} u_{i(k)}^m\, e_{i(k)}^2$$

y las membresias (*memberships*) se calculan usando la siguiente equación:

$$u_{i(k)} = \frac{1}{\sum_{l=1}^{K} \left[\frac{e_{i(k)}}{e_{i(l)}} \right]^{2/(m-1)}}$$

c) *Clustering* no partitivo: *Possibilistic C-means* (*PCM*, Krishnapuran & Keller, 1993, 1996) surge de la relajación del concepto de partición. Con este modelo de clasificación cada *cluster* se determina independientemente comparando la distancia del objeto al centroide con una distancia de referencia, que es un parámetro del algoritmo:

$$t_{i(k)} = \frac{1}{1+\left(\frac{e_{i(k)}}{\eta_k}\right)^{2/(m-1)}}$$

El funcional que minimiza *PCM* es el siguiente:

$$PCM_{K,m,\eta} = \sum_{k=1}^{K}\sum_{i=1}^{N} t_{i(k)}^{m}\, e_{i(k)}^{2} + \sum_{k=1}^{K}\eta_k^2 \sum_{i=1}^{N}(1-t_{i(k)})^{m}$$

4.2 Clasificaciones basadas en matrices de distancias

Sea **X** un vector aleatorio *P*-dimensional definido sobre un espacio de probabilitat (Π, $\mathcal{A}$, ***P***) que toma valores $S \subset \Re^P$ con una función de densitat de probabilidad f respecto a una medida adequada λ. Consideremos $d(\cdot,\cdot)$ una función de dissimilaridad definida sobre pares de elementos de Π, tal que su cuadrado sea integrable en *S*. La *variabilidad geométrica* de X respecto $d(\cdot,\cdot)$ (Cuadras & Fortiana, 1995) se define como:

$$V_d(\mathbf{X}) = \frac{1}{2}E\left[d^2(\mathbf{X}_1,\mathbf{X}_2)\,\middle|\,\mathbf{X}_1,\mathbf{X}_2 \in S\right] = \frac{1}{2}\int_{S\times S} d^2(\mathbf{x}_1,\mathbf{x}_2) f(\mathbf{x}_1) f(\mathbf{x}_2)\,\lambda(d\,\mathbf{x}_1)\,\lambda(d\,\mathbf{x}_2)$$

Dado $\mathbf{x}_o \in \Re^P$, definimos la proximidad de $\mathbf{x}_o$ a la población Π respecto $d(\cdot,\cdot)$ como (Cuadras *et al*., 1997):

$$\phi_d^2(\mathbf{x}_o,\Pi) = \int_S d^2(\mathbf{x}_o,\mathbf{x}) f(\mathbf{x})\,\lambda(d\,\mathbf{x}) - V_d(\mathbf{X})$$

Aplicando lo anterior al cálculo de las distancias a los centroides en K-means se obtiene:

$$e_{i(k)}^2 = \frac{1}{n_k}\sum_{h=1}^{N} I[O_h \subset C_k]\, d_{ih}^2 - \frac{1}{2n_k^2}\sum_{h,l=1}^{N} I[O_h \subset C_k]\, I[O_l \subset C_k]\, d_{hl}^2 \qquad (1)$$

Por otro lado, en el caso difuso, para FCM/PCM, la ecuación equivalente es:

$$e_{i(k)}^2 = \frac{1}{\sum_{h=1}^{N} u_{h(k)}^m}\sum_{h=1}^{N} u_{h(k)}^m d_{ih}^2 - \frac{1}{2\left(\sum_{h=1}^{N} u_{h(k)}^m\right)^2}\sum_{h,l=1}^{N} u_{h(k)}^m u_{l(k)}^m d_{hl}^2 \qquad (2)$$

Una aportación interesante de GINKGO respecto a las técnicas disponibles en otros programas es la posibilidad de ejecutar los algoritmos de *clustering* como K-means y FCM a partir de una matriz de disimilaridades cualquiera, opción raramente disponible en programas comerciales.

4.3 Análisis discriminante

Las técnicas de análisis discriminante actualmente disponibles en GINKGO son:

a) Análisis Discriminante Lineal (Canónico). Entre las opciones disponibles se encuentran elegir el número de ejes canónicos y la normalización de los vectores propios.
b) Análisis Discriminante Cuadrático.
c) Análisis Discriminante Basado en Distancias (Cuadras *et al*., 1997). Empleando ecuación (1) podemos emplear cualquier medida de disimilaridad para un análisis discriminante. Así, la matriz de entrada para este análisis es una matriz simétrica de disimilaridades entre objetos. Análogamente, empleando la ecuación (2), se puede definir un análisis discriminante *fuzzy* basado en distancias.

5. Pantallas de ejemplo

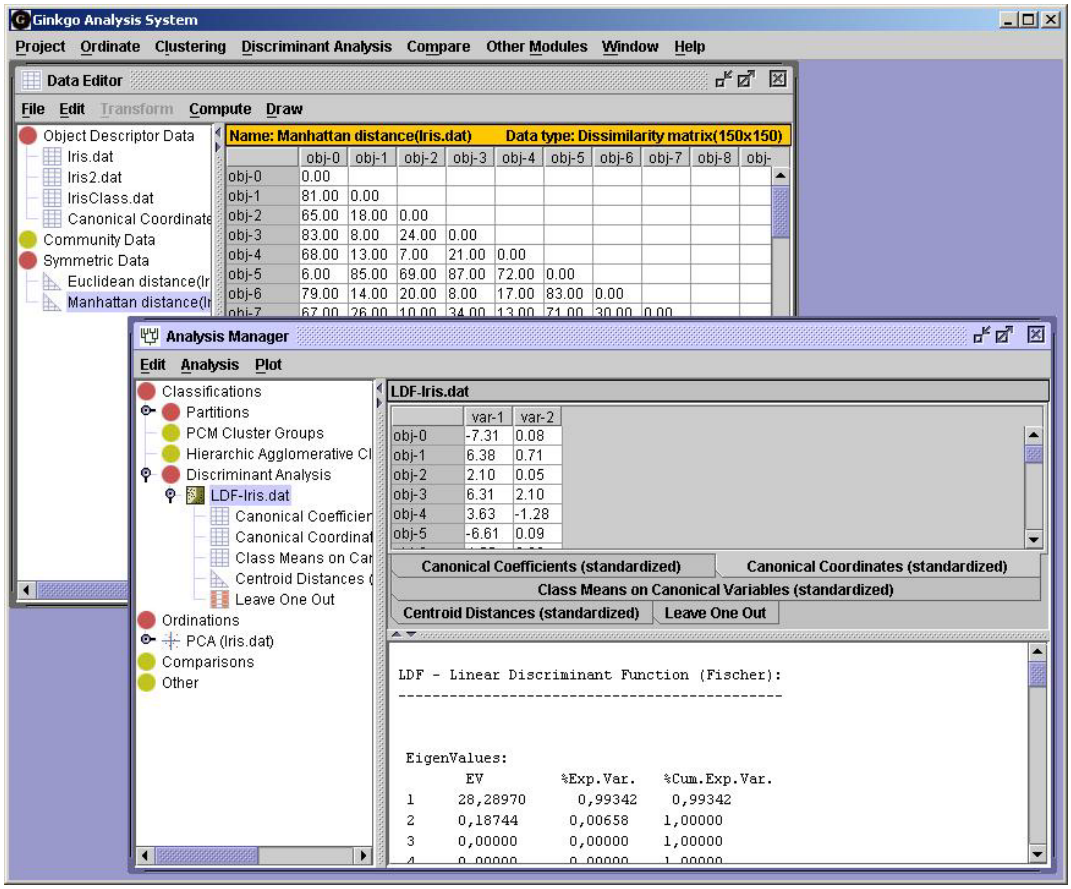

Figura 1: Aspecto general de la interfaz del programa

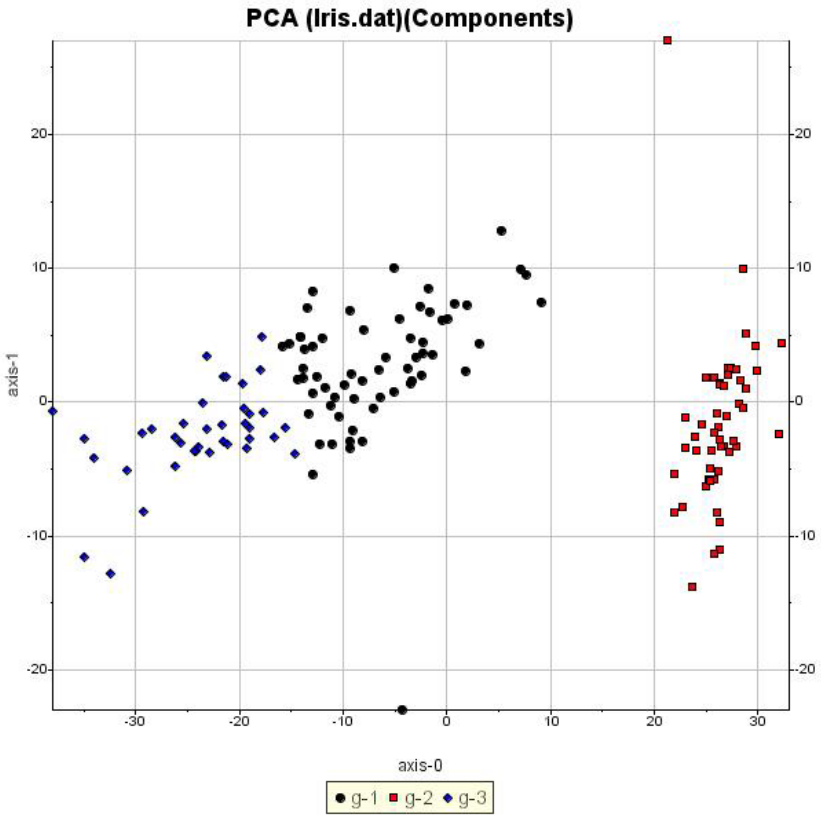

Figura 2: Elemplo de gráfico generado usando los 2 primeros ejes de un PCA y la clasificación generada en K-means.

6. Disponibilidad

El software presentado ha sido enteramente desarrollado en Java. Su distribución es gratuita y las actualizaciones del programa se realizan automáticamente gracias a la tecnología Java Web Start. GINKGO forma parte de un entorno de trabajo, llamado VEGANA. Su página principal se encuentra en: http:\\biodiver.bio.ub.es\vegana.

7. Agradecimientos

El presente trabajo se ha realizado con el soporte del "Comissionat per a Universitats i Recerca" (1999SGR00059), del "Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya" (2001 FI 00269) y de la "Secretaria de Estado de Educación, Universidades, Investigación y Desarrollo" (BFM 2000-0801).

8. Referencias bibliográficas

Bezdek J. C. (1981). Pattern recognition with fuzzy objective functions. Plenum Press. New York.

Cuadras,C.M. & Fortiana,J. (1995): A continuous metric scaling solution for a random variable. *Journal of Multivariate Analysis* 52, 1-14.

Cuadras,C.M., Fortiana,J. & Oliva,F. (1997): The proximity of an individual to a population with applications in discriminant analysis. *Journal of Classification* 14, 117-136.

Gower, J.C. (1966): Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53, 325-338.

Hill M. O. (1973): Reciprocal averaging: An eigenvector method of ordination. *Journal of ecology* 61, 237-249.

Krishnapuram R. & Keller J. M. (1993): A possibilistic approach to clustering. *IEEE transactions on fuzzy systems* 1, 98-110.

Krishnapuram R. & Keller J. M. (1996): The possibilistic c-means algorithm: Insights and recommendations. *IEEE transactions on fuzzy systems* 4, 385-393.

Kruskal, J.B. (1964a): Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. 29(1), 1-27.

Kruskal, J.B. (1964b): Non-metric Multidimensional scaling: A numerical method. *Psychometrika* 29(2), 115-129.

Legendre P. & Legendre L. (1998). Numerical Ecology. Second english edition. Elsevier.

MacQueen J. (1967): Some methods for classification and analysis of multivariate observation. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, pp. 281-297.

Oliva, F., De Caceres, M., Font, X., and Cuadras, C. M. (2001): Contribuciones desde una perspectiva basada en distancias al fuzzy C-means clustering. XXV Congreso Nacional de Estadistica e Investigación Operativa. Úbeda, 2001.