ANALISIS DE COORDENADAS PRINCIPALES

José Luis Vicente Villardón
Departamento de Estadística
Universidad de Salamanca

LOS METODOS DE REPRESENTACION DE DATOS

Introducción

El objeto del análisis de datos es representar un conjunto de individuos, objetos o subpoblaciones $\omega_1, \omega_2 \ldots, \omega_n$ pertenecientes a una población Ω, respecto a unas variables $X_1, X_2, \ldots, X_n$ que pueden ser cuantitativas, cualitativas o una combinación de ambas. La representación de los objetos se realiza en un espacio de dimensión reducida, normalmente 2 ó 3, o mediante diagramas de dispersión, árboles, dendogramas, etc ...
El propósito general es el de la reducción de la dimensión de los datos con el fin de interpretar las similitudes y las disimilitudes entre los individuos de manera simple, frente al del análisis factorial en el que se pretende explicar las relaciones entre las variables a partir de un número menor de factores comunes

Tipos de datos:

Los datos iniciales para el análisis pueden ser de varios tipos:

Datos brutos: Se dispone de la medida de *p* variables tomada en *n* individuos. Los datos se organizan en una matriz $\mathbf{X}_{nxp} = (x_{ij})$.

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

Distancias o disimilaridades entre pares de objetos: Se dispone de una matriz simétrica que contiene una medida de la disimilitud

entre los pares de objetos. $\Delta_{nxn} = (\delta_{ij})$. La diagonal principal contiene sólo ceros.

$$\Delta = \begin{pmatrix} 0 & \cdots & \delta_{1i} & \cdots & \delta_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \delta_{i1} & \cdots & 0 & \cdots & \delta_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \delta_{n1} & \cdots & x_{nj} & \cdots & 0 \end{pmatrix}$$

En general, una matriz nxn, $\Delta_{nxn} = (\delta_{ij})$, se dice que es una matriz de distancias si es simétrica y

$$\delta_{ii} = 0, \qquad \delta_{ij} \geq 0, \quad i \neq j$$

Similaridades entre pares de objetos: Se dispone de una medida de la similitud entre pares de objetos. Las medidas se organizan en una matriz simétrica, $\mathbf{S}_{nxn} = (s_{ij})$.

$$\mathbf{S} = \begin{pmatrix} s_{11} & \cdots & s_{1i} & \cdots & s_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ s_{i1} & \cdots & s_{ii} & \cdots & s_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ s_{n1} & \cdots & x_{nj} & \cdots & s_{nn} \end{pmatrix}$$

Generalmente las medidas de la similaridad están acotadas entre 0 y 1 de forma que la diagonal principal está formada por unos.
En general, una matriz nxn, $\mathbf{S}_{nxn} = (s_{ij})$, se dice que es una matriz de similaridades si es simétrica y

$$s_{ij} \leq s_{ii} \quad \forall i, j$$

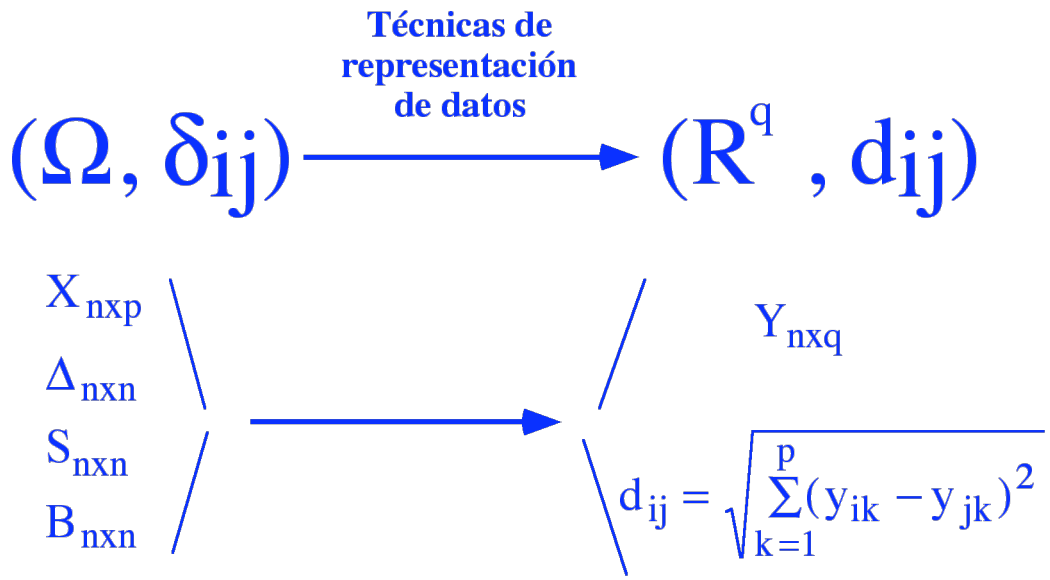
Productos escalares entre pares de objetos: $\mathbf{B}_{nxn} = (b_{ij})$

$$\mathbf{B} = \begin{pmatrix} b_{11} & \cdots & b_{1i} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ b_{i1} & \cdots & b_{ii} & \cdots & b_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{nj} & \cdots & b_{nn} \end{pmatrix}$$

Por ejemplo, la matriz de covarianzas contiene los productos escalares de las columnas de una matriz de datos. La matriz **XX'** contiene los productos escalaes entre las filas de una matriz de datos.

Propósito general de las técnicas de Análisis de datos

Partimos de un espacio cualquiera en el que las disimilaridades entre individuos se miden mediante una distancia cualquiera que no tiene que tener una interpretación física concreta, el resultado final son las coordenadas de los individuos en un espacio euclídeo en el que la distancia es la euclídea usual con la interpretación física usual. Una distancia cualquiera se dice que es euclidizable, o simplemente euclídea, si existe una configuración de puntos en un espacio euclídeo que reproduce las distancias iniciales.



Supongamos que **D** es una matriz simétrica con elementos $-\frac{1}{2}\delta_{ij}^2$, **1** es un vector de n unos y **t** es un vector con n componentes tal que **1't**=1 y **Dt**≠0, entonces la distancia δ_{ij} es euclidizable si y solo si la matriz $(\mathbf{I}-\mathbf{1t}')\mathbf{D}(\mathbf{I}-\mathbf{1t}')$ es semidefinida positiva.
De aquí se deduce que puede encontrarse una representación euclídea con distancias $\delta_{ij}^{1/2}$ si **S** (la matriz de similaridades) es semidefinida positiva.

Dos son los tipos generales de técnicas del Análisis de Datos que presentan los resultados en forma de diagramas de dispersión.
- **Componentes Principales** (y métodos relacionados): Se parte de la matriz de datos completa **X** (en R^p) y se busca el subespacio de mejor ajuste en dimensión R^q. Se representan las coordenadas en el subespacio **Y** para interpretar las posiciones de los puntos con pérdida de información mínima. En este caso la medida de la disimilitud en el espacio original es la distancia euclídea.
- **Coordenadas principales** (y métodos relacionados): Se parte de una matriz de distancias **Δ** y se busca una configuración **Y**, en un

espacio euclídeo (R^q) en el que las distancias entre los puntos sean las contenidas en Δ.

Este tipo de técnicas permiten la representación euclídea en dimensión reducida de espacios abstractos en los que la medida de la similitud o de la disimiltud no tiene una interpretación física concreta.

MEDIDAS DE LA SIMILITUD, DISIMILITUD Y DISTANCIA CALCULADAS A PARTIR DE LA MATRIZ DE DATOS BRUTOS

Medidas de distancia para datos cuantitativos

Diferencias medias

Las medidas de disimilaridad/distancia para datos cuantitativos son las denominadas diferencias medias.

$$d(i,k) = \frac{1}{m}\sum_{j=1}^{p}(x_{ij} - x_{kj})$$

$$d(i,k) = \frac{1}{m}\sum_{j=1}^{m}\left|x_{ij} - x_{kj}\right|$$

Las diferencias medias suponen implícitamente que tratamos con variables de escalas comparables y generalmente son demasiado simples para ser utilizadas en la práctica, además el valor absoluto suele ser difícil de tratar.

Distancia euclídea (pitagórica)

Se trata de la distancia física usual que mide la distancia en línea recta entre dos puntos en el espacio multidimensional

$$\delta_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$$

Presenta la ventaja fundamental de que se trata de la distancia natural que estamos acostumbrados a interpretar. Es la distancia que utilizamos en las representaciones finales en dimension reducida.

Supone implícitamente que las variables tienen escalas comparables por lo que es muy sensible a las diferentes escalas de medida de las variables por lo que a veces se estandarizan las variables.

La distancia aumenta con el número de variables por lo que a veces se corrige obteniendo lo que se denomina distancia media..

$$d_{ik} = \sqrt{\frac{\delta_{ik}^2}{p}}$$

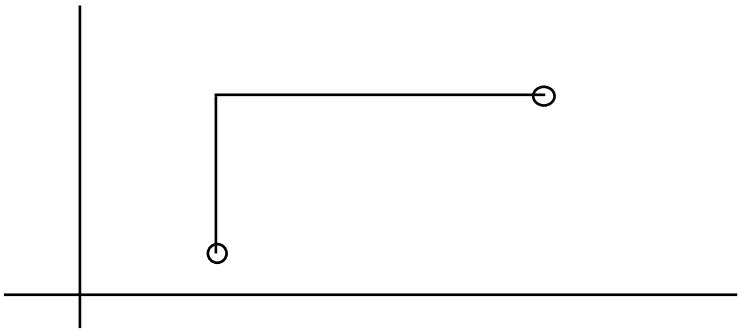
Distancia de Minkowsky

Se utiliza particularmente en estudios no métricos

$$d_r(i,j) = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^r \right)^{1/r}$$

Cuando r=1 se denomina Distancia "Ciudad" o Distancia media, ya que la distancia se mide como si recorriéramos las calles de una ciudad como se muestra en la figura siguiente.

$$d_1(i,j) = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right| \right)$$



Cuando r=2, obtenemos la distancia euclídea usual.

$$d_2(i,j) = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^2 \right)^{1/2}$$

Métrica de Canberra

Se trata de una distancia para datos positivos en la que se estandarizan las diferencias dividiendo por la suma de los valores. Se utiliza en estudios de taxonomía en Biología.

$$d_{CAMB}(i,j) = \sum_{k=1}^{p} \frac{\left| x_{ik} - x_{jk} \right|}{x_{ik} + x_{jk}}$$

Coeficiente de divergencia

$$D(i,j) = \left[\frac{1}{p} \sum_{k=1}^{p} \left(\frac{\left| x_{ik} - x_{jk} \right|}{x_{ik} + x_{jk}} \right)^2 \right]^{1/2}$$

Coeficiente de concordancia racial de Pearson

Se trata de una distancia entre poblaciones representadas por una muestra. Trata de medir la diferencia genética entre las mismas.

$$CCR(i,k) = \left[\frac{1}{p} \sum_{k=1}^{p} \left(\frac{(\bar{x}_{ik} - \bar{x}_{jk})^2}{(s_{ik}^2 / n_i) + (s_{jk}^2 / n_j)} \right) \right]^{1/2} - \frac{2}{p}$$

COEFICIENTES DE ASOCIACION (SIMILARIDAD) PARA DATOS BINARIOS (presencia/ausencia)

Disponemos ahora de una matriz de datos brutos en la que las variables son binarias y generalmente la presencia o ausencia de un carácter cualitativo. Generalmente la presencia se codifica con un 0 y la ausencia con un 1. El propósito es medir la similitud entre cada par de individuos a partir de la información que proporcionan los caracteres medidos sobre los mismos.

Para datos binarios es posible construir una tabla de contingencia para cada par de individuos donde se cuentan las presencias y ausencias comunes de cada uno de los caracteres estudiados.

		individuo **i**		
		Presente (1)	Ausente (0)	
individuo **k**	Presente (1)	***a***	***b***	***a+b***
	Ausente (0)	***c***	***d***	***c+d***
		a+c	***b+d***	***m=a+b+c+d***

Donde

a: numero de caracteres presentes en los dos individuos.

b: Número de caracteres presentes en **i** y ausentes en **k.**

c: Número de caracteres presentes en **k** y ausentes en **i.**

d: Número de caracteres ausentes en los dos.

A partir de la tabla de contingencia pueden construirse distintos coeficientes de similaridad. Algunos de estos coeficientes no consideran las dobles ausencias para no sobreestimar la similitud a partir de características que no están presentes en ninguno de los dos individuos, por ejemplo, la presencia de alas en dos mamíferos con características muy diferentes.

Coeficiente de Jaccard (Sneath)

$$S_J = \frac{a}{a+b+c}$$

Acotado entre cero y uno.

No considera las dobles ausencias

Coeficiente de Dice y Sorensen

$$S_D = \frac{2a}{2a+b+c}$$

Acotado entre cero y uno.

Da mayor importancia a las dobles presencias.

Coeficiente de Sokal y Michener (Coeficiente de concordancia simple)

$$S_{SM} = \frac{a+d}{a+b+c+d}$$

Acotado entre cero y uno.

Coeficiente de Rogers y Tanimoto

$$S_{RT} = \frac{a+d}{a+2b+2c+d}$$

Acotado entre cero y uno.

Coeficiente de Yule

$$S_Y = \frac{ad-bc}{ad+bc}$$

Acotado entre -1 y 1.

Coeficiente de Hamann

$$S_Y = \frac{a+d-b-c}{a+b+c+d}$$

Acotado entre -1 y 1.

Coeficiente General de Similaridad de Gower

Aplicable a todos los tipos de datos: binarios, multiestado (ordenados y cualitativos) y cuantitativos o auna combinación de varios tipos.

$$S_G = \frac{\sum_{k=1}^{p} w_{ijk}\ s_{ijk}}{\sum_{k=1}^{p} w_{ijk}}$$

donde se a asignado a cada par de individuos una puntuación $0 \le s_{ijk} \le 1$ y una ponderación w_{ijk} sobre el caracter (variable) k. La ponderación w_{ijk} es 1 cuando se considera que la commparación es válida para el caracter k y vale 0 cuando el valor del estado del caracter k es desconocido para uno o los dos individuos.

CARACTERES BINARIOS

$s_{ijk} = 1$ para coincidencias y $s_{ijk} = 0$ para divergencias.

$w_{ijk} = 0$ para dobles ausencias.

Para una matriz con solo caracteres binarios, el coeficiente Gower es igual al coeficiente de Jaccard.

CARACTERES MULTIESTADO

$s_{ijk} = 1$ para coincidencias y $s_{ijk} = 0$ para divergencias sin tener en cuenta el número de categorías.

Las ponderaciones son siempre 1 salvo para datos perdidos

CARACTERES CUANTITATIVOS

Para caracteres cuantitativos la similaridad se define como

$$s_{ijk} = 1 - \frac{\left| x_{ik} - x_{jk} \right|}{R_k}$$

donde R_k es el rango (diferencia entre el máximo y el mínimo) del caracter k-ésimo sobre toda la población conocida.

ANALISIS DE COORDENADAS PRINCIPALES

ALGUNOS RESULTADOS TEORICOS

Se dice que una matriz de distancias es *euclídea* si existe una configuración en algún espacio euclídeo cuyas distancias entre puntos estén dadas por $\Delta_{nxn} = (\delta_{ij})$; esto es si para algún p, existen puntos $\mathbf{x}_1,\ldots,\mathbf{x}_p \in R^p$ tales que

$$\delta_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)$$

Dada una matriz de disimilaridades/distancias es posible convertirla en una matriz de productos escalares tomando

$$\mathbf{B} = -1/2\ \mathbf{H}\,\Delta^{(2)}\ \mathbf{H'}$$

donde **H** (nxn) es la matriz de centrado :

$$\mathbf{H} = \mathbf{I} - (1/n)\ \mathbf{11'}$$

Resultado principal

Si $\Delta_{nxn} = (\delta_{ij})$ es una matriz de distancias y definimos **B** como antes. Entonces $\Delta_{nxn} = (\delta_{ij})$ es euclídea si y solo si **B** es semidefinida positiva. En particular se verifica lo siguiente:

a) Si $\Delta_{nxn} = (\delta_{ij})$ es una matriz de distancias euclídeas entre un conjunto de puntos para una configuración $\mathbf{Z} = (\mathbf{z}_1,\ldots,\mathbf{z}_n)'$ entonces

$$b_{ij} = (\mathbf{z}_i - \bar{\mathbf{z}})'(\mathbf{z}_j - \bar{\mathbf{z}}), \qquad i,j = 1,\ldots,n$$

En forma matricial es $\mathbf{B} = (\mathbf{HZ})(\mathbf{HZ})'$ de forma que $\mathbf{B} \geq 0$. Notese que la matriz **B** es la matriz de productos escalares para la configuración **Z**.

b) Si **B** es semidefinida positiva de rango p entonces una configuración correspondiente a **B** puede construirse a partir de los valores y vectores propios de **B** como

$$\mathbf{Z} = \mathbf{U}\,\mathbf{D}_{\boldsymbol{\lambda}}^{1/2}$$

donde

$$\mathbf{B} = \mathbf{U}\,\mathbf{D}_{\boldsymbol{\lambda}}\,\mathbf{U}' \quad (\text{donde } \mathbf{U}'\,\mathbf{U} = \mathbf{I})$$

es la descomposición espectral de la matriz **B**. **U** contiene los vectores propios en columnas y $\mathbf{D}_{\boldsymbol{\lambda}}$ es una matriz diagonal que contiene los correspondientes valores propios ordenados en orden decreciente.

$$\mathbf{D}_{\lambda} = diag(\lambda_1,\ldots,\lambda_p), \qquad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$$

Entonces los puntos en R^p con coordenadas $\mathbf{z}_i = (z_{i1},\ldots,z_{ip})'$ (es decir la i-ésima fila de **Z**) tienen interdistancias dadas en $\Delta_{nxn} = (\delta_{ij})$. Además la configuración tiene centro de gravedad $\bar{\mathbf{z}} = \mathbf{0}$, y **B** representa la matriz de productos escalares para la configuración.
La demostración de los resultados puede encontrarse en Mardia, Kent and Bibby (1979).

ALGORITMO PRACTICO

Supongamos que tenemos una matriz de distancias observadas $\Delta_{nxn} = (\delta_{ij})$ y queremos representarla mediante una configuración de puntos en un espacio euclídea de dimensión reducida (2 ó 3). Normalmente aunque la distancia sea *eucílidea* el número de dimensiones necesarias para representarla suele ser demasiado

elevado. La solución consiste en seleccionar los primeros vectores propios correspondientes a los valores propios más grandes. Si los primeros valores propios son grandes en comparación con el resto, cabe esperar que tengamos una representación bastante aproximada para $\Delta_{nxn} = (\delta_{ij})$. A la configuración así obtenida es a la que denominamos *Coordenadas Principales* o solución clásica del problema de escalado multidimensional.

Un algoritmo práctico de cálculo sería el siguiente:

1.- A partir de $\Delta_{nxn} = (\delta_{ij})$ construir $\mathbf{A} = \left(-\frac{1}{2}\delta_{ij}^2\right)$

2.- Obtener la matriz **B** cuyos elementos son $b_{ij} = a_{ij} - \bar{a}_{i\bullet} - \bar{a}_{\bullet j} + \bar{a}_{\bullet\bullet}$, es decir restando a cada elemento de **A** la media de su fila y la de su columna y sumando la media de todos los elementos.

3.- Obtener la descomposición espectral de **B** (**B = U** $\mathbf{D}_{\lambda}$ **U′**) y seleccionar los vectores propios correspondientes a los mayores valores propios siempre que éstos sean positivos.

4.- Las coordenadas buscadas están en las primeras columnas de **Z = U** $\mathbf{D}_{\lambda}^{1/2}$

5.- Si todos los valores propios son positivos, la bondad del ajuste de la representación se calcula como

$$\frac{\sum_{j=1}^{r} \lambda_j}{\sum_{i=1}^{p} \lambda_i}$$

donde *r* es la dimensión de la representación final y *p* es el rango de **B**.

COORDENADAS PRINCIPALES A PARTIR DE UNA MATRIZ DE SIMILARIDADES

Para utilizar la técnica que se muestra en los apartados anteriores con una matriz de similaridades es necesarios primero convertirla en una matriz de distancias.

La transformación estándar que convierte una matriz de similaridades **S** en una matriz de distancias $\Delta_{nxn} = (\delta_{ij})$ es la siguiente

$$\delta_{ij} = \sqrt{s_{ii} - 2s_{ij} + s_{jj}}$$

El resultado siguiente permite la utilización de las coornadas principales sobre la matriz de similaridades.

- Si **S** es semidefinida positiva, entonces la matriz de distancias definida por la transformación estándar anterior es euclídea con matriz de productos escalares centrados **B=HSH** y **H** la matriz de centrado **H** = **I –** (1/n) **11′.**

RELACION ENTRE EL ANALISIS DE COMPONENTES Y EL DE COORDENADAS PRINCIPALES

De todas las posibles elecciones de la matriz de distancias $\Delta_{nxn} = (\delta_{ij})$ cuando esta se calcula a partir de la matriz datos brutos **X**, la más sencilla es la distancia eucídea usual en el espacio p-dimensional

$$\delta_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

En este caso hay una estrecha relación entre el análisis de coordenadas principales y el de componentes principales.
Supongamos que **X** está centrada por columnas de forma que la matriz de productos escalares centrados (en el espacio completo) se obtiene como **B=XX′**, además **B** es la matriz de productos escalares que se obtendría transformando la matriz de distancias como en el apartado anterior.
Si $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ son los valores propios de **X′X** (a partir de los cuales se obtienen las componentes principales), entonces también son los valores propios no nulos de **B=XX′.**

Además las coordenadas principales para la matriz de distancias euclídeas coinciden con las coordenadas de los individuos sobre las componentes principales.

EJEMPLOS

Ejemplo 1: Ordenación de varias especies de arañas sobre un gradiente ambiental hipotético.
Supongamos que tenemos la matriz de presencia ausencia de un grupos de 13 especies de arañas en 28 lugares. Los datos han sido tomados de Ter Braak (1986). El propósito del estudio es clasificar las especies de arañas teniendo en cuenta la similitud entre las mismas de acuerdo con el hábitat en que se desarrollan. Se entiende que los lugares de muestreo corresponden a los posibles hábitats de las especies.

```
Arct lute 0 0 1 1 1 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Pard lugu 0 1 1 1 1 0 1 1 0 0 0 0 1 1 1 1 1 1 1 1 1 0 1 0 1 0 0 0
Zora spin 1 1 1 1 1 1 1 1 0 0 0 0 1 1 1 1 1 0 1 1 1 0 0 0 1 0 0 0
```

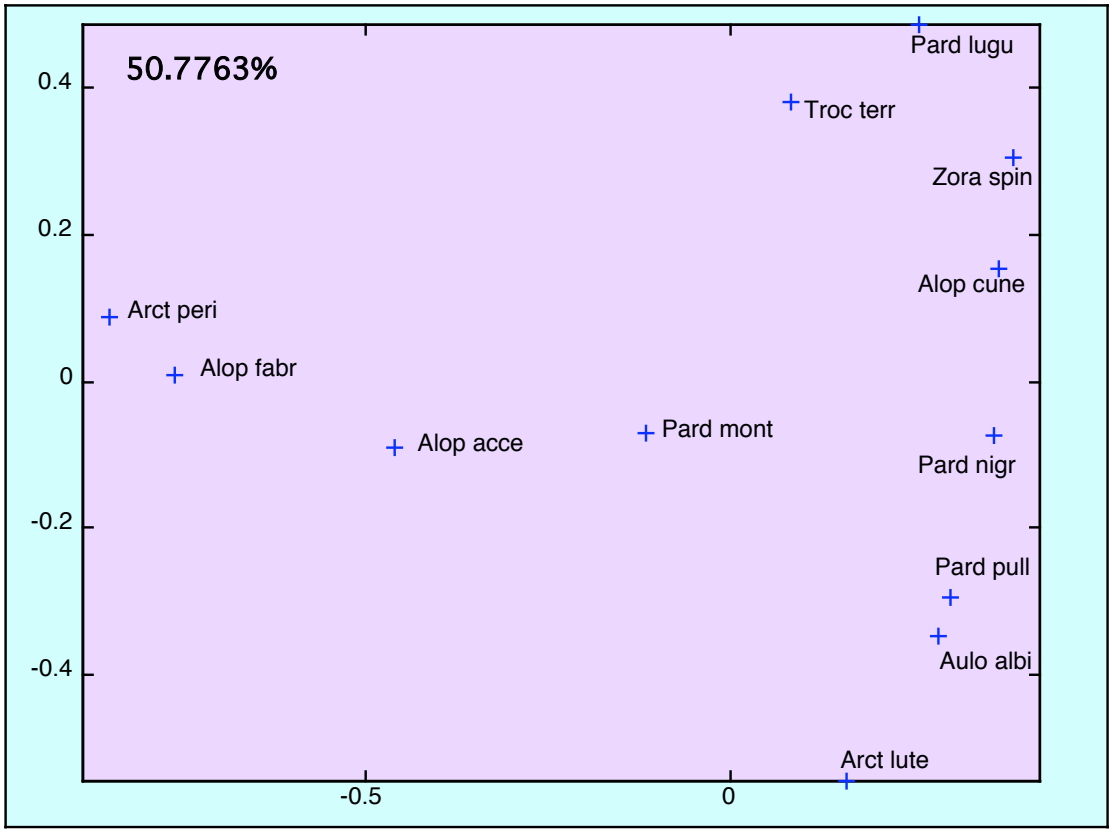
```
Pard nigr 1 1 1 1 1 1 1 1 1 0 0 1 1 1 0 1 0 0 1 0 0 0 0 0 1 0 0 0
Pard pull 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0
Aulo albi 1 1 1 1 1 1 1 1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0
Troc terr 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 1
Alop cune 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 0 1 1 1 1 0 0 0 1 0 0 0
Pard mont 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 1 1 1 1 1 0 1 1
Alop acce 1 0 1 1 1 0 1 0 1 1 1 1 1 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1
Alop fabr 0 0 1 0 0 0 0 0 0 1 1 0 1 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1
Arct peri 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 1 1 1
```

Convertimos los datos en una matriz de similitudes entre especies a partir del coeficiente de Jaccard.

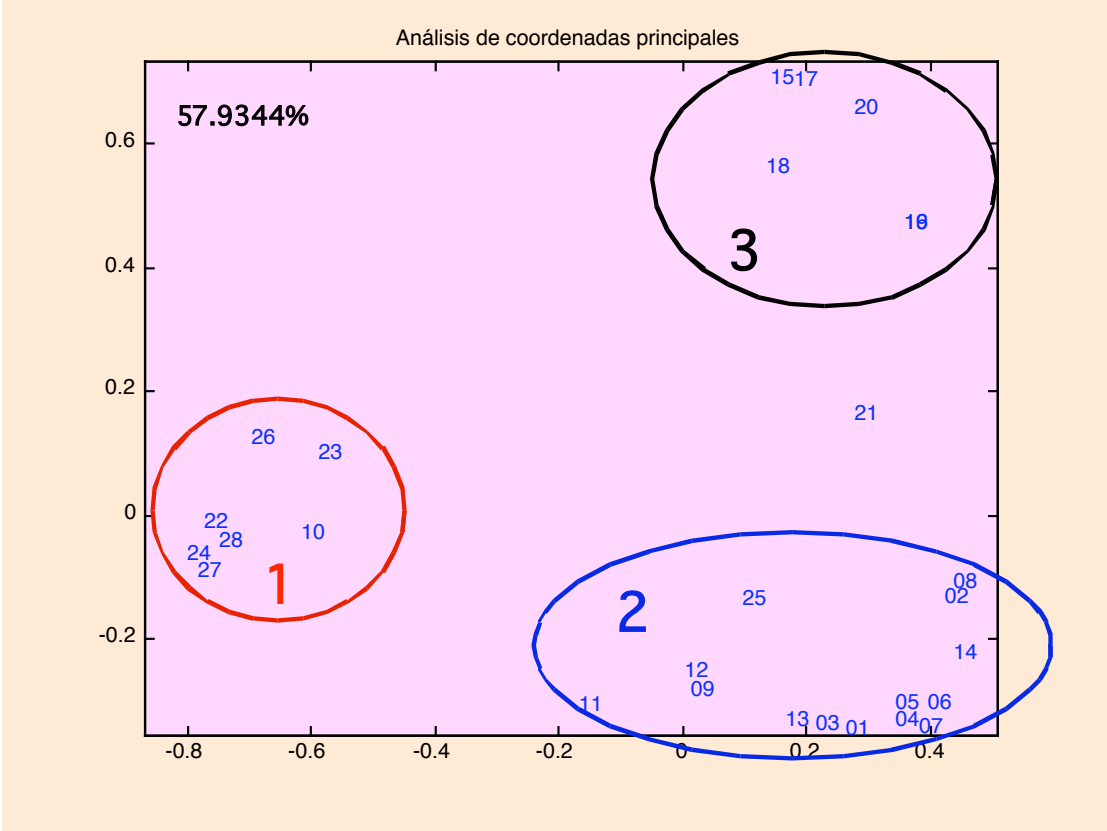
Arct lute	1	0.333	0.412	0.467	0.5	0.583	0.269	0.368	0.333	0.263	0.125	0
Pard lugu	0.333	1	0.789	0.524	0.409	0.45	0.654	0.636	0.407	0.259	0.167	0.045
Zora spin	0.412	0.789	1	0.684	0.55	0.611	0.654	0.714	0.462	0.259	0.12	0
Pard nigr	0.467	0.524	0.684	1	0.706	0.688	0.577	0.789	0.565	0.391	0.13	0
Pard pull	0.5	0.409	0.55	0.706	1	0.733	0.538	0.737	0.667	0.409	0.136	0
Aulo albi	0.583	0.45	0.611	0.688	0.733	1	0.462	0.632	0.571	0.381	0.211	0
Troc terr	0.269	0.654	0.654	0.577	0.538	0.462	1	0.731	0.679	0.536	0.321	0.143
Alop cune	0.368	0.636	0.714	0.789	0.737	0.632	0.731	1	0.6	0.385	0.154	0
Pard mont	0.333	0.407	0.462	0.565	0.667	0.571	0.679	0.6	1	0.727	0.455	0.227
Alop acce	0.263	0.259	0.259	0.391	0.409	0.381	0.536	0.385	0.727	1	0.647	0.353
Alop fabr	0.125	0.167	0.12	0.13	0.136	0.211	0.321	0.154	0.455	0.647	1	0.545
Arct peri	0	0.045	0	0	0	0	0.143	0	0.227	0.353	0.545	1

En lugar de estudiar las similitudes directamente, trataremos de buscar una configuración euclídea con puntos que representan a cada una de las especies y de forma que las distancias entre los puntos aproximen las similitudes en el sentido de que dos puntos próximos deben ser similares.

La representación recoge el 50.7763 % de la variabilidad.


50.7763%
0.4
0.2
0
-0.2
-0.4
-0.5
0
Pard lugu
Troc terr
Zora spin
Alop cune
Arct peri
Alop fabr
Alop acce
Pard mont
Pard nigr
Pard pull
Aulo albi
Arct lute

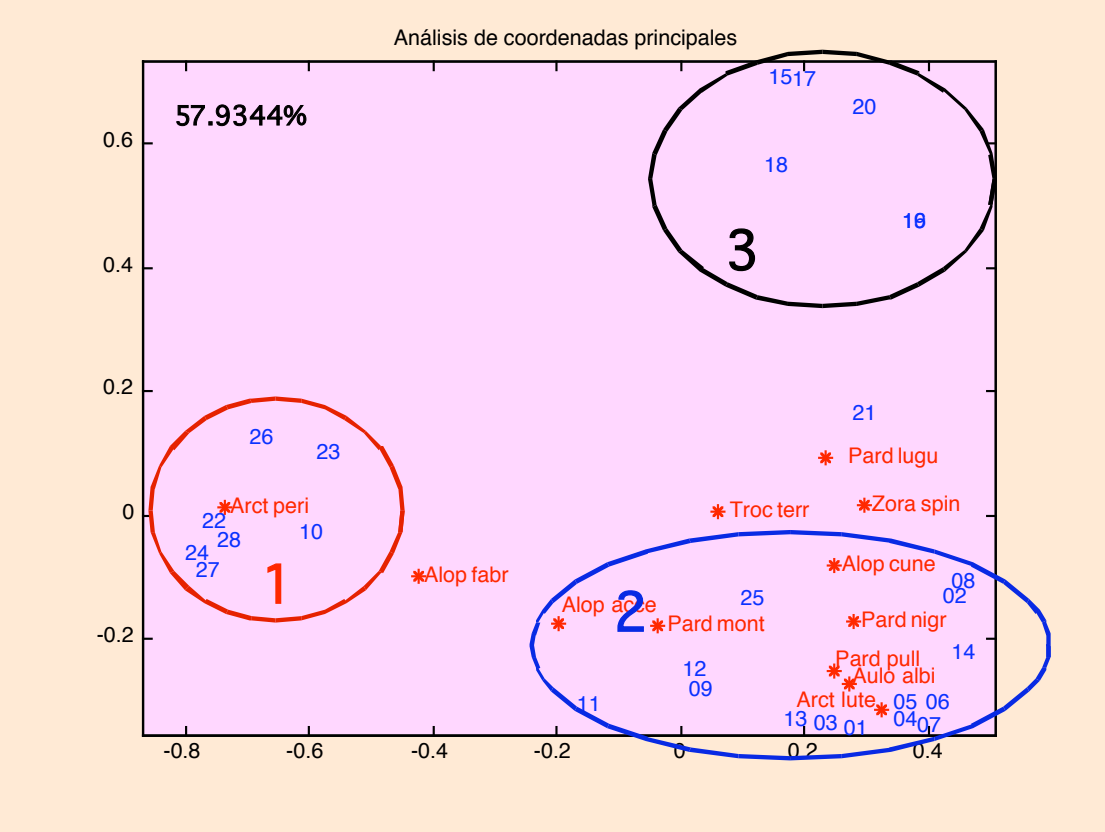
De la misma forma podemos hacer una representación para los lugares calculando similitudes entre los mismos.



La ordenación de los lugares sobre los ejes de la representación puede entenderse como el gradiente ambiental hipotético que proporciona una mayor diferenciación entre las características ambientales de los mismos.

En la representación pueden distinguirse claramente tres grupos de puntos con características diferentes. El eje 1 separa los grupos 1 del 2 y 3, mientras que el eje 2 separa el 3 del 2, ocupando el 1 posiciones intermedias.

La representación para los lugares puede complementarse con una representación superpuesta para las especies calculando, sobre los ejes de ordenación, la media de los valores en los que la especie está presente.



El grupo 1 se caracteriza por una mayor presencia de las especies *Alop Peri* y *Alop fabra,* el grupo 2 por la mayor presencia del resto y el grupo 3 por la presencia, aunque parece que en menor medida de *Pard lugu, Troc terr y Zora Spin.*

La interpretación de los gradientes ambientales depende del conocimiento a priori del investigador o de la medida de algunas variables ambientales relacionadas con el problema.

Supongamos que disponemos de un conjunto de variables ambientales

1.-Contenido de agua en suelo.
2.-% de cubierta de arena.
3.-% cubierta de musgo.
4.-Reflejo de la superficie del suelo en un día sin nubes.
5.-Cubierta de hojas caidas y ramas.
6.-Cubierta de hierba

1.-Agua	2.-Arena	3.-Musgo	4.-Reflejo	5.-Hojas	6.-Hierbas
5	0	7	8	0	9
8	0	2	3	3	9
6	0	5	8	0	9
6	0	5	6	0	9
8	0	0	5	0	9
9	5	5	1	7	6
8	0	1	5	0	9
6	0	2	1	9	6
5	0	9	7	0	6
4	8	7	8	0	5
4	0	9	8	0	7
5	0	8	8	0	8
9	3	1	7	3	9
8	0	4	2	0	9
9	0	1	1	9	5
8	0	1	0	9	0
9	0	1	2	9	5
8	0	0	2	9	5
7	0	3	0	9	2
8	0	1	0	9	0
7	0	1	0	9	2
1	7	9	8	0	0
0	6	9	9	0	6
2	7	9	9	0	5
3	7	2	5	0	8
0	9	4	9	0	2
0	5	8	8	0	6
0	7	8	8	0	6

Podemos incluir la información ambiental sobre el gráfico mediante regresiones, obteniendo lo que en el contexto del Análisis de Proximidades se conoce como "modelo vectorial".

<u>Modelos vectoriales:</u> Ordenación de las especies con información ambiental añadida

La ordenación, mediante las técnicas clásicas de análisis de proximidades, puede entenderse como la búsqueda de gradientes ambientales hipotéticos. Una vez que los gradientes han sido

encontrados trataremos de buscar su relación con las variables ambientales observadas.

Sea **X** la matriz de coordenadas para los lugares obtenida a partir de cualquier técnica de ordenación. Supongamos que queremos colocar q vectores $\mathbf{b}_k$ (k=1, ... , q), sobre el diagrama de ordenación, de forma que el producto escalar de una fila de **X**, $\mathbf{x}_i$ (i=1, ... , n) , por cada uno de esos vectores, $\mathbf{x}_i' \mathbf{b}_k$, aproxime los elementos de **Z**, (z_{ik}) tan bien como sea posible. Si tomamos los vectores $\mathbf{b}_k$ como filas de una matriz **B**, el problema es encontrar una matriz **B** que haga mínimo

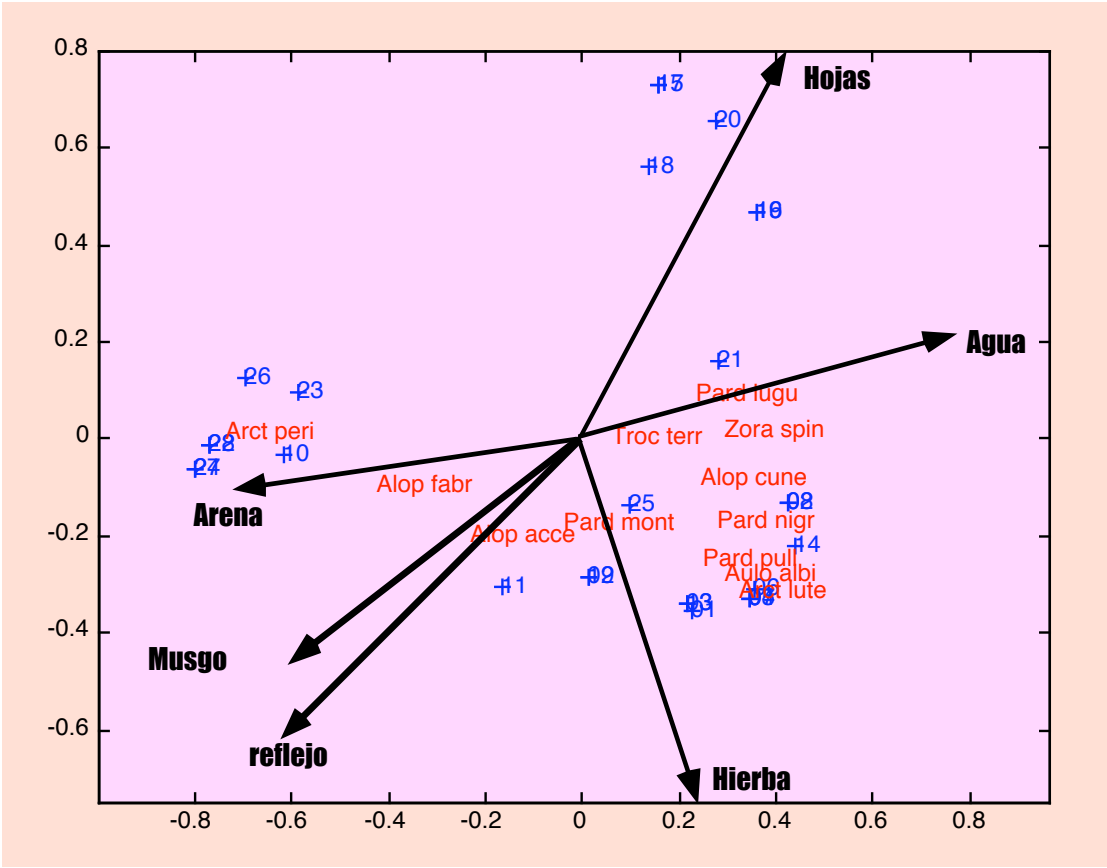
$$L = \|\mathbf{Z} - \mathbf{X}\mathbf{B}'\| = tr[(\mathbf{Z} - \mathbf{X}\mathbf{B}')'(\mathbf{Z} - \mathbf{X}\mathbf{B}')] =$$
$$= tr(\mathbf{Z}'\mathbf{Z}) - tr(\mathbf{Z}'\mathbf{X}\mathbf{B}') - tr(\mathbf{B}\mathbf{X}'\mathbf{Z}) + tr(\mathbf{B}\mathbf{X}'\mathbf{X}\mathbf{B}')$$

la solución viene dada por

$$\mathbf{B}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}$$

es decir., los coeficientes de regresión de cada variable ambiental sobre los ejes de ordenación.

Esto puede interpretarse como un biplot en el que uno de los conjuntos de coordenadas es fijo. Los marcadores (coordenadas) para las variables pueden interpretarse como un conjunto de ejes de predicción GOWER (1996), con un conjunto de variables externas a la ordenación para predecir. (La proyección del lugar sobre el eje biplot predice el valor de la variable ambiental correspondiente).

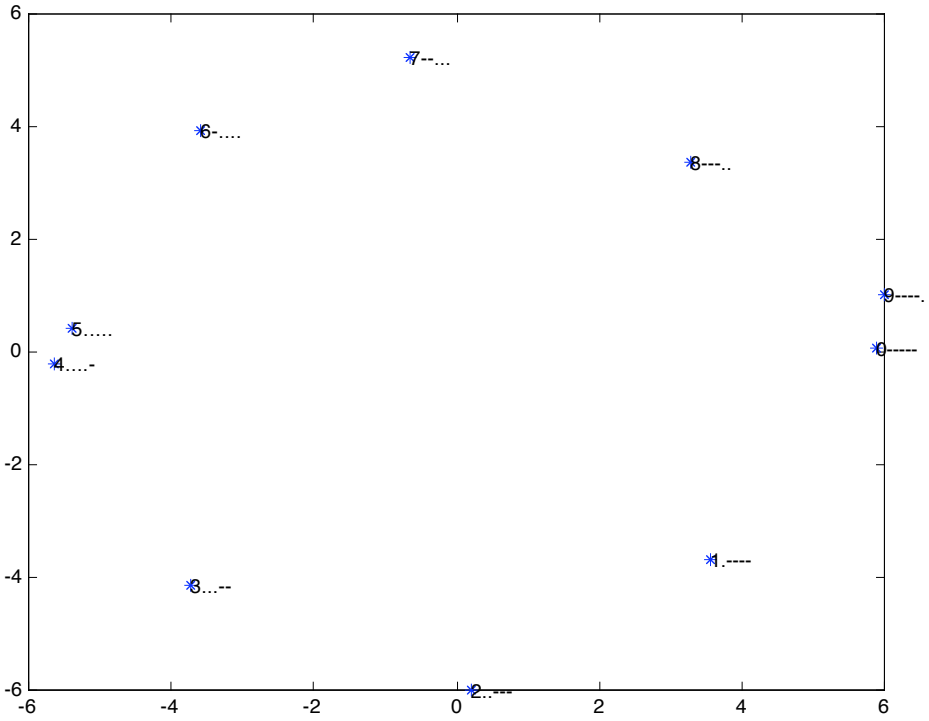

Hojas
Agua
Arena
Musgo
reflejo
Hierba
Arct peri
Alop fabr
Troc terr
Zora spin
Pard lugu
Alop cune
Pard mont
Alop acce
Pard nigr
Pard pull
Aulo albi
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8

Ejemplo 2: Confusión entre los códigos Morse correspondiente a los números.

En el ejemplo siguiente mostramos un caso en el que los datos obtenidos corresponden a una medida de la similaridad que no se ha calculado a partir de datos brutos sino que ha sido directamente observada.

La tabla siguiente contiene el porcentaje de personas que pensaron que las secuencias de códigos Morse correspondientes a cada pareja eran idénticas después de oirlas en una sucesión rápida.

	1	2	3	4	5	6	7	8	9	0
1	84									
2	62	89								
3	16	59	86							
4	6	23	38	89						
5	12	8	27	56	90					
6	12	14	33	34	30	86				
7	20	25	17	24	18	65	85			
8	37	25	16	13	10	22	65	88		
9	37	28	9	7	5	8	31	58	91	
0	52	18	9	7	5	18	15	39	79	94



Matriz de coordenadas

3.5740	-3.6755
0.2215	-5.9905
-3.7185	-4.1499
-5.6454	-0.1988
-5.3898	0.4145
-3.5772	3.9301
-0.6375	5.2245
3.2870	3.3661
5.9986	1.0095
5.8872	0.0701

Valores propios

182.2194
121.9238

Bondad del ajuste

32.3396
21.6386

Bondad del ajuste acumulada

32.3396
53.9782