ANÁLISIS DE COORDENADAS PRINCIPALES

3. La matriz simétrica n x n

$$S = XX' = \begin{bmatrix} \sum x_{1j}^2 & \sum x_{1j}x_{2j} & \cdots & \sum x_{1j}x_{nj} \\ \sum x_{2j}x_{1j} & \sum x_{2j}^2 & \cdots & \sum x_{2j}x_{nj} \\ . & . & & . \\ . & . & & . \\ \sum x_{nj}x_{1j} & \sum x_{nj}x_{2j} & \cdots & \sum x_{nj}^2 \end{bmatrix} \qquad [3]$$

donde se puede pensar que las sumatorias para j de 1 a p contienen **medidas de las similitudes entre los n objetos** considerados. Esto no es evidente inmediatamente, pero se justifica al considerar la distancia euclidiana al cuadrado del objeto i al objeto k, que es:

ANÁLISIS DE COORDENADAS PRINCIPALES

$$d_{ik}^2 = \sum_{j=1}^{p} (x_{ij} - x_{kj})^2$$

Desarrollando el lado derecho de esta ecuación se tiene que

$$d_{ik}^2 = s_{ii} + s_{kk} - 2s_{ik} \qquad [4]$$

donde s_{ik} es el elemento en la i-ésima fila y la k-ésima columna de **XX'**. Se deduce que s_{ik} es una **medida de la similitud entre los objetos i y k** puesto que aumentar s_{ik} significa que la distancia d_{ik} entre los objetos disminuye. Además, se ve que s_{ik} toma el valor máximo de $(s_{ii} + s_{kk})/2$ cuando $d_{ik} = 0$, que ocurre cuando los objetos **i** y **k** tienen valores idénticos para las variables X_1 a X_p.

ANÁLISIS DE COORDENADAS PRINCIPALES

4. Si la matriz $\mathbf{Z} = \begin{bmatrix} z_{i1} & z_{i2} & \cdots & z_{ip} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ . & . & & . \\ . & . & & . \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix}$

contiene los valores de los p C.P. para los n objetos considerados, luego puede reescribirse en términos de la matriz de datos **X** como

$$\mathbf{Z = X A'} \qquad [5]$$

donde la fila i de **A** es $\mathbf{a'}_i$, el i-ésimo vector propio de la matriz de covarianza C de la muestra. Es una propiedad de **A** que **A'A = I**; es decir, la transpuesta de **A** es la inversa de **A**. Por lo tanto, posmultiplicando ambos lados de la ecuación precedente por **A** se tiene

$$\mathbf{X = Z A} \qquad [6]$$

ANÁLISIS DE COORDENADAS PRINCIPALES

Esta enumeración de resultados ha sido larga, pero necesaria para explicar el análisis de coordenadas principales (ACoP) en relación con el análisis de componentes principales (ACP). Para identificar esta relación, tengan en cuenta que a partir de las ecuaciones [1] y [2]

$$\boldsymbol{X'X}\frac{\boldsymbol{a}_i}{n-1} = \lambda_i \boldsymbol{a}_i$$

Luego premultiplicando ambos términos de la ecuación por X y utilizando la ecuación [3] se tiene

$$\mathbf{S}(\mathbf{X}\boldsymbol{a}_i) = ((n-1)\lambda_i \boldsymbol{X}\boldsymbol{a}_i)$$

o

$$\mathbf{S}\boldsymbol{z}_i = (n-1)\lambda_i \boldsymbol{z}_i \qquad [7]$$

ANÁLISIS DE COORDENADAS PRINCIPALES

$z_i = Xa_i$ es un vector de n elementos, que contiene los valores de Z_i para los n objetos considerados. Por lo tanto, el i-ésimo valor propio de la matriz de similitud **S = X'X** es (n-1) λ_i, y el correspondiente vector propio da los valores del el i-ésimo C.P. para los n objetos.

*El **análisis de coordenadas principales (ACoP)** consiste en aplicar la ecuación [7] a una matriz **S** (n x n) de similitudes entre n objetos, que se calcula utilizando cualquiera de los índices de similitud disponibles.* De esta forma, es posible encontrar los C.P. correspondientes a **S** sin necesariamente medir ninguna variable en los objetos de interés. Los componentes tendrán las propiedades de los C.P. y, en particular, no estarán correlacionados para n objetos.

ANÁLISIS DE COORDENADAS PRINCIPALES

Aplicando **ACoP** a la matriz **XX'** se obtiene esencialmente la misma ordenación que resulta de aplicar **ACP** a la matriz de datos **X**.

Diferencias en términos de escalado:

Método	Escalado/Varianza
ACoP	$(n-1)\lambda_i$
ACP	λ_i

Esta diferencia es inmaterial porque son solo importantes los valores relativos de los objetos en los ejes de ordenación.

ANÁLISIS DE COORDENADAS PRINCIPALES

Complicaciones:

1) La matriz de similitud no cumple todas las propiedades de una matriz calculada a partir de los datos según **S = XX '**

De la ecuación [3], las sumas de las filas y columnas de **XX'** son todas cero. Por ej., la suma de la primera fila es

$$\sum x_{1j}^2 + \sum x_{1j}x_{2j} + \ldots + \sum x_{1j}x_{nj} = \sum x_{ij}\left(x_{1j} + x_{2j} + \ldots + x_{nj}\right)$$

donde las sumatorias son para **j** de **1** a **p**. Esto es cero porque es **n** veces la media de $\mathbf{X_j}$, y se supone que todas las variables **X** tienen media cero. Por lo tanto, ***se requiere que la matriz de similitud S tenga sumas cero para filas y columnas.***

ANÁLISIS DE COORDENADAS PRINCIPALES

Complicaciones:

Si este no es el caso, entonces la matriz inicial puede ser doblemente centrada reemplazando el elemento $\mathbf{s_{ik}}$ en la **fila i** y la **columna k** por $\mathbf{s_{ik}-s_i-s_k + s}$, donde $\mathbf{s_i}$ es la media de la i-ésima fila de **S**, $\mathbf{s_k}$ es la media de la k-ésima columna de **S**, y **s** es la media de todos los elementos en **S**. La matriz de similitud doble-centrada tendrá cero media en filas y columnas y, por lo tanto, es más adecuada para el análisis.

ANÁLISIS DE COORDENADAS PRINCIPALES

Complicaciones:

2) Algunos de los **valores propios** de la matriz de similitud pueden ser **negativos**, con lo que los correspondientes C.P. parecen tener **varianzas negativas**.

Sin embargo, la ordenación solo utiliza los componentes asociados con los valores propios mayores, por lo que unos pocos valores propios negativos pequeños carecen de importancia.

En caso que los valores propios mayores sean negativos, esto sugiere que la matriz de similitud no es adecuada para la ordenación.

ANÁLISIS DE COORDENADAS PRINCIPALES

Ej. de uso del ACoP - Especies de plantas en la Reserva Natural Steneryd:

Los datos sobre abundancia de especies en parcelas se volvieron a analizar usando ***distancias de Manhattan*** entre las parcelas. Esto es, la distancia entre las parcelas i y k se midieron por $d_{ik} = \sum|x_{ij} - x_{kj}|$, donde la sumatoria aplica a j sobre las 25 especies y x_{ij} representa la abundancia de las especies j en la parcela i. Las similitudes se calcularon como $s_{ik}=-d_{ik}^2/2$ y luego se centraron doblemente antes de calcular los valores y vectores propios.

ANÁLISIS DE COORDENADAS PRINCIPALES

Ej. de uso del ACoP - Especies de plantas en la Reserva Natural Steneryd:

XLSTAT 2015.2.01.17502 - Matrices de similitud/disimilitud (correlación...) - el 08/05/2018 a las 18:24:05
Datos: Libro = Ejemplo 1.12.xlsm / Hoja = Datos / Rango = Datos!$B$1:$R$26 / 25 filas y 17 columnas
Etiquetas de las filas: Libro = Ejemplo 1.12.xlsm / Hoja = Datos / Rango = Datos!$A$1:$A$26 / 25 filas y 1 columna
Disimilitud: Distancia de Manhattan
Calcular proximidades para las: Columnas
Matriz de proximidad (Distancia de Manh

Matriz de proximidad (Distancia de Manhattan):

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0	46,000	58,000	89,000	139,000	153,000	141,000	171,000	209,000	201,000	149,000	172,000	137,000	148,000	171,000	185,000	198,000
2	46,000	0	46,000	59,000	119,000	147,000	161,000	191,000	217,000	215,000	171,000	196,000	161,000	180,000	185,000	219,000	232,000
3	58,000	46,000	0	73,000	139,000	175,000	179,000	217,000	233,000	243,000	193,000	218,000	183,000	194,000	221,000	243,000	256,000
4	89,000	59,000	73,000	0	76,000	134,000	164,000	190,000	208,000	214,000	178,000	201,000	164,000	189,000	200,000	242,000	255,000
5	139,000	119,000	139,000	76,000	0	120,000	172,000	146,000	188,000	176,000	156,000	181,000	152,000	173,000	176,000	222,000	235,000
6	153,000	147,000	175,000	134,000	120,000	0	154,000	124,000	156,000	114,000	136,000	95,000	90,000	145,000	136,000	184,000	221,000
7	141,000	161,000	179,000	164,000	172,000	154,000	0	152,000	146,000	150,000	98,000	133,000	142,000	139,000	158,000	200,000	195,000
8	171,000	191,000	217,000	190,000	146,000	124,000	152,000	0	122,000	64,000	124,000	103,000	114,000	111,000	122,000	182,000	199,000
9	209,000	217,000	233,000	208,000	188,000	156,000	146,000	122,000	0	104,000	138,000	137,000	158,000	141,000	152,000	222,000	231,000
10	201,000	215,000	243,000	214,000	176,000	114,000	150,000	64,000	104,000	0	114,000	91,000	122,000	133,000	136,000	208,000	219,000
11	149,000	171,000	193,000	178,000	156,000	136,000	98,000	124,000	138,000	114,000	0	87,000	92,000	73,000	72,000	124,000	129,000
12	172,000	196,000	218,000	201,000	181,000	95,000	133,000	103,000	137,000	91,000	87,000	0	67,000	96,000	89,000	135,000	154,000
13	137,000	161,000	183,000	164,000	152,000	90,000	142,000	114,000	158,000	122,000	92,000	67,000	0	79,000	94,000	122,000	159,000
14	148,000	180,000	194,000	189,000	173,000	145,000	139,000	111,000	141,000	133,000	73,000	96,000	79,000	0	59,000	99,000	110,000
15	171,000	185,000	221,000	200,000	176,000	136,000	158,000	122,000	152,000	136,000	72,000	89,000	94,000	59,000	0	96,000	111,000
16	185,000	219,000	243,000	242,000	222,000	184,000	200,000	182,000	222,000	208,000	124,000	135,000	122,000	99,000	96,000	0	39,000
17	198,000	232,000	256,000	255,000	235,000	221,000	195,000	199,000	231,000	219,000	129,000	154,000	159,000	110,000	111,000	39,000	0

ANÁLISIS DE COORDENADAS PRINCIPALES

Ej. de uso del ACoP - Especies de plantas en la Reserva Natural Steneryd:

Análisis de Coordenadas Principales:

Se detectó al menos un valor propio negativo, pero no se aplicó ninguna corrección.

Valores propios:

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17
Valor propio	100080,656	55710,764	19899,949	11819,752	8702,828	7317,958	3580,102	2861,856	1735,927	840,389	376,853	88,774	0,000	-278,927	-1441,396	-1968,883	-2893,544
Variabilidad	40,745	23,189	9,019	5,822	4,588	4,040	2,561	2,277	1,832	1,477	1,294	1,180	0,000	0,000	0,000	0,000	0,000
% acumulado	40,745	63,933	72,952	78,774	83,362	87,403	89,964	92,242	94,073	95,551	96,845	98,025	98,025	98,025	98,025	98,025	98,025

Procedimiento de Bryan Manly

	F1	F2
Valor propio	97.638,6	55.659,5
Varianza	47,3	27,0

Los dos primeros C.P. proporcionan una buena ordenación, representando el 74,3% de la varianza.

AMARN 2018 - IMFIA.FI.UDELAR -
Ing. Luis Silveira, Ph.D.