

Universidad de Costa Rica

Facultad de Ciencias

Escuela de Matemática

Contribución al estudio de las métricas
en Análisis de Datos

Fabricio Bolaños Guerrero

Tesis sometida a la consideración
de la Facultad de Ciencias
para optar al grado de
Licenciatura en Matemática.

Ciudad Universitaria Rodrigo Facio

San José, Costa Rica

1996

Este trabajo fue aprobado por el siguiente Tribunal Examinador

Prof. Eduardo Piza Volio
Presidente del Tribunal Examinador

Prof. Javier Trejos Zelaya
Director de Tesis

Prof. Jorge Poltronieri Vargas
Lector

Prof. Jorge González Varela
Lector

Prof. Willian Castillo Elizondo
Miembro del Tribunal Examinador

Agradecimientos

Quiero agradecer al Prof. Javier Trejos por la ayuda que me brindó como director de esta Tesis, y por la paciencia que me tuvo.

A los lectores por el tiempo dedicado a la lectura de mi tesis y por sus valiosas sugerencias, y al Prof. Mario Villalobos por la ayuda en la implementación de los programas.

Gracias.

Dedicatoria

A mis padres: Gregorio y Elena por la confianza que siempre me han tenido y por darme el aliento para seguir adelante.

A Liz y Heinner por haberlos privado de muchos ratos felices en la preparaci3n de esta tesis.

Contents

1	Preliminares y Notaciones	1
1.1	Introducción	1
1.2	Producto interno o producto escalar	1
1.3	Repaso de algunos conceptos algebraicos.	3
1.4	La tabla de datos	5
1.5	La métrica de pesos en el espacio de variables	7
1.6	Análisis en Componentes Principales (A.C.P.)	8
1.6.1	Descripción del método	9
1.6.2	Las aplicaciones X y X^t	10
1.6.3	Inercia e información perdida	11
2	Estudio de las métricas	13
2.1	Introducción	13
2.2	Una parametrización de métricas en A.C.P.	13
2.3	Cualidades de las métricas	15
2.3.1	Cualidad tensorial	16
2.3.2	Cualidad algebraica: estrategias de agregación	18
2.3.3	Estrategia d	19
2.3.4	Estrategia Σ	20
2.3.5	Estrategia de Mahalanobis	20
2.3.6	Estrategia relacional	21
2.4	La métrica inducida por un A.C.P.	22
2.5	Otros Estudios	23

3	Generación de métricas útiles en análisis de datos	25
3.1	Introducción	25
3.2	Presentación de los generadores de métricas	26
3.2.1	Generadores primarios	26
3.2.2	Generadores espectrales	32
3.3	Generación proporcional	33
3.3.1	El efecto de asociación es estacionario	35
3.3.2	Proposición de un generador proporcional	37
3.4	Una extensión del generador espectral	39
3.4.1	Definición	40
3.4.2	ACP del triplete $(X, M(\gamma), D)$	42
4	Resultados Numéricos	47
4.1	Introducción	47
4.2	Generación primaria para A.C.P.	47
4.2.1	Descripción del programa	48
4.2.2	Resultados sobre las notas escolares	49

List of Figures

1.1	Esquema de Dualidad	11
4.1	Plano principal y círculo de correlaciones para la tabla de Notas cuando $\alpha = 0$, generador de V a I	52
4.2	Plano principal de la tabla de Notas cuando $\alpha = 6$, generador de V a I	52
4.3	Plano principal de la tabla de Notas cuando $\alpha = 12$, generador de V a I	52
4.4	Plano principal y círculo de correlaciones de Notas cuando $\alpha = 0$, de I a V^{-1}	54
4.5	Plano principal de Notas con $\alpha = 6$, de I a V^{-1}	54
4.6	Plano principal y círculo de correlaciones de Notas con $\alpha = 12$, de I a V^{-1}	54
4.7	Plano principal y círculo de correlaciones de Notas cuando $\alpha = 0$, de I a D_{1/σ^2}	56
4.8	Plano principal y círculo de correlaciones de Notas cuando $\alpha = 6$, de I a D_{1/σ^2}	56
4.9	Plano principal y círculo de correlaciones de Notas cuando $\alpha = 12$, de I a D_{1/σ^2}	56

List of Tables

2.1	Propiedades de algunas estrategias de agregación.	23
4.1	Tabla de datos de las notas escolares	49
4.2	Los valores propios y porcentajes de inercia explicada, para la tabla de Notas, generador de V a I	51
4.3	Las componentes principales, para la tabla de Notas, generador de V a I . . .	51
4.4	Correlaciones entre variables y componentes principales, para la tabla de Notas, generador de V a I	51
4.5	Los valores propios y porcentajes de inercia explicada, para la tabla de Notas, generador de I a V^{-1}	53
4.6	Las componentes principales, para la tabla de Notas, generador de I a V^{-1} . .	53
4.7	Correlaciones entre variables y componentes principales, para la tabla de Notas, generador de I a V^{-1}	53
4.8	Los valores propios y porcentajes de inercia explicada, para la tabla de Notas, generador de I a D_{1/σ^2}	55
4.9	Las componentes principales, para la tabla de Notas, generador de I a D_{1/σ^2} .	55
4.10	Las correlaciones entre las variables y las componentes principales, para la tabla de Notas, generador de I a D_{1/σ^2}	55

Introducción

Nos hemos planteado estudiar la dependencia de los resultados al cambiar las métricas para el método de Análisis en Componentes Principales (A.C.P.). En efecto, para este análisis el problema de la escogencia de la distancia no se ha resuelto todavía. Contrario por ejemplo a los métodos de Regresión Lineal y Discriminación Factorial, para los que se conoce la métrica optimal.

El estudio de la dependencia lo abordaremos principalmente a través de la generación de métricas, técnica planteada por el profesor Y. Schektman de la Universidad Paul Sabatier, de Toulouse, Francia.

Se estudiaron las propiedades teóricas de los generadores de métricas y se implementaron computacionalmente para observar su rendimiento. Además, se estudiaron los enfoques de otros autores alrededor del problema de la escogencia de la métrica.

Tomando en cuenta la importancia del Análisis de Datos con otras disciplinas científicas, consideramos que este estudio puede aportar alguna luz sobre la técnica de base del Análisis de Datos, como es el Análisis en Componentes Principales.

La presentación de nuestros resultados es como sigue:

- Repaso de conceptos algebraicos necesarios para el desarrollo del tema (Capítulo 1).
- Revisiones bibliográficas (Capítulo 2).
- Propuestas de generación de métricas (Capítulo 3).
- Revisión de resultados numéricos (Capítulos 4).
- Implementación computacional (Anexo).

Chapter 1

Preliminares y Notaciones

1.1 Introducción

El presente capítulo tiene por objetivo dejar establecidos para el lector, el lenguaje, las herramientas y las notaciones que se usarán a lo largo de este trabajo.

Entre las herramientas más importantes para la presente tesis están los conceptos básicos de los espacios euclídeos y, por supuesto, el Análisis en Componentes Principales (A.C.P.), por lo que también haremos un recordatorio de este tema.

En adelante, supondremos que los espacios vectoriales considerados son de dimensión finita y que el cuerpo de escalares es \mathbb{R} .

1.2 Producto interno o producto escalar

La idea de *longitud* y *ángulo* de vectores juegan un papel preponderante en este trabajo, debido a la interpretación estadística que tienen y a su uso en conceptos como el de distancia y correlación.

Como es sabido, estas dos nociones se pueden definir en términos de un producto interno, por lo que en esta sección se recordará la definición de los productos internos y consideraremos algunos ejemplos importantes.

Definición 1 *Sea E un espacio vectorial sobre \mathbb{R} . Un **producto interno** o **producto escalar** sobre E , es una función $M : E \times E \longrightarrow \mathbb{R}$ que cumple:*

- $M(u, v) = M(v, u)$ para todo $u, v \in E$.
- $M(u, v + w) = M(u, v) + M(u, w)$ para todo $u, v, w \in E$.
- $M(\lambda u, v) = M(u, \lambda v) = \lambda M(u, v)$ para todo $u, v \in E$ y todo $\lambda \in \mathbb{R}$.
- $M(u, u) \geq 0$ y $M(u, u) > 0$ si $u \neq 0 \forall u \in E$.

La última condición dice que es **definido positivo**. En ocasiones $M(u, v)$ se denota $\langle u, v \rangle$ o bien $\langle u, v \rangle_M$.

Observación 1 *Un producto interno sobre E es una forma bilineal sobre $E \times E$.*

Definición 2 *Si M es un producto interno sobre E , la forma cuadrática asociada a M , también denotada M , es la función definida por $M(v) = M(v, v)$, para cualquier $v \in E$.*

Nótese que la forma cuadrática M es tal que $M : E \rightarrow \mathbb{R}$.

Para el presente trabajo es de mucha importancia conocer cual es la matriz que se define por medio de un producto interno; para ello veamos la siguiente definición.

Definición 3 *Sea E un espacio vectorial con producto interno M . Sea $\{u_1, \dots, u_n\}$ una base de E ; la matriz de M con respecto a la base $\{u_1, \dots, u_n\}$ está dada por $[M]_{jk} = M(u_j, u_k)$.*

La matriz de un producto interno M respecto a la base canónica de E , en general la denotaremos también con M .

Para $x = \sum_{j=1}^n \alpha_j u_j$ y $y = \sum_{k=1}^n \beta_k u_k$, entonces se tiene:

$$\begin{aligned}
 M(x, y) &= M\left(\sum_{j=1}^n \alpha_j u_j, \sum_{k=1}^n \beta_k u_k\right) \\
 &= \sum_{j=1}^n \sum_{k=1}^n \alpha_j \beta_k M(u_j, u_k) \\
 &= \sum_{j=1}^n \sum_{k=1}^n \alpha_j \beta_k [M]_{jk} \\
 &= x^t M y.
 \end{aligned}$$

Observación 2 *La matriz del producto interno usual en la base canónica es la matriz identidad I .*

Definición 4 La **longitud** de un vector x la definimos como su norma:

$$\|x\|_M = \sqrt{M(x)} = \sqrt{M(x, x)} = \sqrt{x^t M x}.$$

Definición 5 El **ángulo** θ entre dos vectores x, y se define a partir de:

$$\cos \theta = \frac{M(x, y)}{\sqrt{M(x)M(y)}} = \frac{x^t M y}{\sqrt{(x^t M x)(y^t M y)}}.$$

Obsérvese que la desigualdad de Cauchy-Schwartz garantiza que $-1 \leq \cos \theta \leq 1$. Para más detalles ver [5, página 134.]

Definición 6 La **distancia** entre dos vectores x, y la definimos como

$$d_M(x, y) = \|x - y\|_M = \sqrt{M(x - y)}.$$

1.3 Repaso de algunos conceptos algebraicos.

Definición 7 Sea A un operador definido sobre E , el **espectro** de A es:

$\{\lambda \in \mathbb{R} : Ax = \lambda x, \text{ para } x \neq 0\}$. Un elemento λ del espectro de A se llama un **valor propio** de A .

Definición 8 Sea λ un valor propio de A , el **espacio propio** asociado a λ es:

$\{x \in E : Ax = \lambda x\}$. Un $x \in E, x \neq 0$ tal que $Ax = \lambda x$ se llama **vector propio** de A asociado a λ .

Teorema Espectral (Ver demostración en [11, páginas 331-332]) Sea E un espacio vectorial con un producto interno M y sea A un operador simétrico definido en E . Entonces existen escalares $\lambda_1, \dots, \lambda_r \in \mathbb{R}$ y operadores de proyección¹ P_1, \dots, P_r asociados a los espacios propios respectivos que cumplen:

1. $i \neq j \Rightarrow \lambda_i \neq \lambda_j$.
2. $P_j \neq 0$ y $P_i P_j = 0$ si $i \neq j$.

¹Recuérdese que un operador de proyección P_i es un operador lineal que cumple $P_i P_i = P_i$.

3. $\sum_{i=1}^r P_i = I$, el operador identidad.

4. $A = \sum_{i=1}^r \lambda_i P_i$.

Diremos que la descomposición espectral de A es: $A = \sum_{i=1}^r \lambda_i P_i$.

Definición 9 Sea A un operador que tiene la descomposición espectral $A = \sum_{i=1}^r \lambda_i P_i$. Supóngase que f es una función con valores reales que está definida en los $\lambda_1, \dots, \lambda_r$, entonces definimos la **función de operador** $f(A) = \sum_{i=1}^r f(\lambda_i) P_i$.

Ejemplo. Consideremos $f(x) = \sqrt{x}$ ($0 \leq x, x \in \mathbb{R}$), Si A es positivo definimos $f(A)$ como $A^{1/2} = \sqrt{A} = \sum_{i=1}^r \sqrt{\lambda_i} P_i$, donde $A = \sum_{i=1}^r \lambda_i P_i$.

Definición 10 Sea E un espacio vectorial con producto interno M y sea $x \in E$. Definimos la forma lineal M_x por :

$$\begin{aligned} M_x : E &\longrightarrow \mathbb{R} \\ y &\longmapsto M_x(y) = M(x, y) \end{aligned}$$

Luego también podemos definir un isomorfismo, denotado también M , como [5]:

$$\begin{aligned} M : E &\longrightarrow E^* \\ x &\longmapsto M_x \end{aligned}$$

Observación 3

1. La matriz asociada al isomorfismo anterior en la base canónica de E y la base dual de ésta, es la matriz M .
2. El isomorfismo M induce sobre el dual E^* el isomorfismo M^{-1} , asociado también a un producto interno definido positivo:

$$\begin{array}{ccc} & E & \\ & \updownarrow & \\ M & & M^{-1} \\ & E^* & \end{array}$$

Definición 11 Llamaremos **métrica** sobre un espacio vectorial E a un producto interno. Obsérvese que entonces una métrica M sobre E define:

- Una forma bilineal definida positiva $M : E \times E \longrightarrow \mathbb{R}$
- Una forma cuadrática $M : E \longrightarrow \mathbb{R}$.
- Un isomorfismo $M : E \longrightarrow E^*$.
- Una matriz simétrica definida positiva $M_{p \times p}$, si p es la dimensión de E .

Las definiciones y propiedades anteriores nos permiten definir toda la geometría que necesitamos en Análisis de Datos.

1.4 La tabla de datos

Supóngase que se tiene un conjunto Ω de n individuos, sobre cada uno de los cuales se miden p variables cuantitativas. Obsérvese que se está en presencia de p aplicaciones x^1, \dots, x^p :

$$\begin{aligned} x^j : \Omega &\longrightarrow \mathbb{R} \\ i &\longmapsto x^j(i) = x_i^j \end{aligned}$$

donde x_i^j denota la respuesta del individuo i a la variable j (o la medición de la variable j sobre el individuo i). Cada individuo $i \in \Omega$ está dotado de un peso $p_i > 0$ que mide la importancia relativa del individuo i en el análisis. Estos pesos cumplen:

$$\sum_{i=1}^n p_i = 1.$$

Al igual que con el producto interno, ahora es importante establecer cual es la matriz de datos o la tabla de datos, lo que se hará en la siguiente definición.

Definición 12 *La matriz X que resume todas las mediciones, es la **tabla de descripción de los datos** (individuos \times variables), llamada simplemente **tabla de datos**:*

$$X = X_{n \times p} = \begin{pmatrix} x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_i^1 & \dots & x_i^j & \dots & x_i^p \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_n^1 & \dots & x_n^j & \dots & x_n^p \end{pmatrix}.$$

El concepto de tabla de datos es de suma importancia en Análisis de Datos. Dependiendo de la naturaleza de esta matriz se usa cada una de las técnicas de análisis. Para nuestros propósitos, será suficiente considerar la matriz tal como la definimos anteriormente.

Observación 4

1. La i -ésima fila x_i de X representa al individuo i .

2. El vector $\begin{pmatrix} x_i^1 \\ \vdots \\ x_i^j \\ \vdots \\ x_i^p \end{pmatrix}$ representa en \mathbb{R}^p al individuo i .

3. La j -ésima columna x^j de X representa a la variable j .

4. El vector $\begin{pmatrix} x_1^j \\ \vdots \\ x_i^j \\ \vdots \\ x_n^j \end{pmatrix}$ representa en \mathbb{R}^n a la variable x^j .

Para nuestros propósitos necesitaremos medir las proximidades entre los individuos y entre las variables, definamos para ello los espacios vectoriales respectivos, con base en la observación anterior.

Definición 13 $E = \mathbb{R}^p$ lo llamaremos el **espacio de individuos**. $F = \mathbb{R}^n$ lo llamaremos el **espacio de variables**.

En general, dotaremos a estos espacios vectoriales de las bases canónicas usuales que denotaremos $\{e_1, \dots, e_p\}$ y $\{f_1, \dots, f_n\}$ respectivamente.

También serán de importancia los espacios duales E^* , F^* de cada uno de éstos. E^* y F^* están dotados de las bases duales usuales $\{e_1^*, \dots, e_p^*\}$ y $\{f_1^*, \dots, f_n^*\}$ tales que $e_i^*(e_j) = \delta_{ij}$ y $f_k^*(f_l) = \delta_{kl}$.

Definiendo sobre E una métrica M y sobre F una métrica D , entonces se ha dotado a todos los elementos de la tabla de datos X de una geometría, de acuerdo con los conceptos recordados en la sección 1.2. Por lo tanto, X define configuraciones de puntos en E y F asociadas a los conceptos de proximidad y ángulos. Con base en esto podemos enunciar la siguiente definición.

Definición 14 *Al triplete $\mathcal{N} = (X, M, D)$ lo llamamos **nube** de puntos, donde X es una matriz $n \times p$ que representa las observaciones de p variables cuantitativas sobre n individuos, M una matriz simétrica definida positiva $p \times p$ (que define la métrica euclídea en el espacio de individuos), D una matriz simétrica definida positiva $n \times n$ (que define la métrica euclídea en el espacio de variables).*

1.5 La métrica de pesos en el espacio de variables

Supondremos de ahora en adelante que las variables x^j están centradas, es decir:

$$\bar{x}^j = \sum_{i=1}^n p_i x_i^j = 0; \quad j = 1, \dots, p.$$

O sea, el centro de gravedad de la nube \mathcal{N} coincide entonces con el origen del espacio vectorial E ($\mathbf{g} = \mathbf{0}$).

La varianza de la j -ésima variable se escribe:

$$\text{var}(x^j) = \sum_{k=1}^n p_k (x_k^j)^2.$$

Igualmente, la covarianza entre las variables x^j y $x^{j'}$ se escribe:

$$\text{cov}(x^j, x^{j'}) = \sum_{k=1}^n p_k x_k^j x_k^{j'}.$$

En lo sucesivo, se provee al espacio de variables F , de la métrica euclídea de pesos D , definida por:

$$D(f_i, f_{i'}) = 0 \quad i \neq i'$$

$$D(f_i) = D(f_i, f_i) = \|f_i\|_D^2 = p_i.$$

La matriz asociada al isomorfismo D es la matriz diagonal de pesos:

$$D = \begin{pmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & p_n \end{pmatrix}.$$

De la definición de la métrica euclídea D , se deduce:

$$D(x, y) = \text{cov}(x, y)$$

$$D(x) = \|x\|_D^2 = \text{var}(x)$$

y la distancia entre los caracteres x y y en el sentido de la métrica D está dada por:

$$d(x, y) = \|x - y\|_D = \sqrt{\text{var}(x - y)}$$

Además se tienen las relaciones:

- $d^2(x, y) = \sum_{i=1}^n p_k (x_i - y_i)^2$.
- $d^2(x, y) = \text{var}(x) + \text{var}(y) - 2\text{cov}(x, y)$.

Si definimos θ como el ángulo entre las variables x , y , entonces:

$$\cos(\theta) = \frac{D(x, y)}{\sqrt{D(x)D(y)}} = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)\text{var}(y)}} = \rho(x, y).$$

Luego tenemos que el coseno del ángulo entre dos variables, es igual al coeficiente de correlación lineal.

1.6 Análisis en Componentes Principales (A.C.P.)

Este método es considerado como el método de base del Análisis de Datos [22], por lo que un breve repaso es de mucha importancia para el presente trabajo. El objetivo principal del A.C.P. es la descripción de la tabla de datos, presentada en la sección 1.4, donde p variables cuantitativas son observadas sobre n individuos; de modo que la información perdida sea mínima.

El caso $p = 2$ es bastante fácil de representar y se puede hacer la descripción con base en la representación en el plano.

En el caso $p = 3$, aunque ya no es tan fácil de representar los datos, todavía se pueden visualizar usando el espacio tridimensional.

Para $p \geq 4$ el estudio visual se hace imposible, de allí el interés del A.C.P., que consiste en encontrar los subespacios óptimos (planos en el caso usual), sobre los cuales proyectar la información para hacer las representaciones. Al efectuar esta proyección de los individuos se desea que la información perdida sea mínima.

1.6.1 Descripción del método

El A.C.P. se realiza normalmente sobre los datos centrados.

En el caso en que los datos de entrada no estén centrados, éstos se pueden centrar mediante la transformación $x_i^j := x_i^j - \bar{x}^j$.

Llamaremos el centro de gravedad de la nube \mathcal{N} al punto $g = (\bar{x}^1, \dots, \bar{x}^p)$, que en este caso es $(0, \dots, 0)$, por lo que el centro de gravedad de la nube \mathcal{N} está en el origen de coordenadas del espacio de individuos E .

La matriz de varianzas y covarianzas asociada a la tabla X , se denotará V y está dada por:

$$X^t D_p X = V = \begin{pmatrix} v_1 & \dots & v_{1j} & \dots & v_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ v_{j1} & \dots & v_j & \dots & v_{jp} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ v_{p1} & \dots & v_{pj} & \dots & v_p \end{pmatrix}$$

donde $v_{jk} = v_{kj} = \sum_{i=1}^n p_i x_i^j x_i^k = \text{cov}(x^j, x^k)$, y $v_j = v_{jj} = \text{var}(x^j)$.

De igual forma, los coeficientes de correlación lineal forman la matriz R , donde:

$$D_{1/\sigma} V D_{1/\sigma} = R = \begin{pmatrix} 1 & \dots & r_{1j} & \dots & r_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ r_{j1} & \dots & 1 & \dots & r_{jp} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ r_{p1} & \dots & r_{pj} & \dots & 1 \end{pmatrix}$$

con $D_{1/\sigma} = \text{diag}\left(\frac{1}{\sqrt{v_j}}\right)$ y

$$r_{jk} = r_{kj} = \rho(x^j, x^k) = \frac{\text{cov}(x^j, x^k)}{\sqrt{\text{var}(x^j)\text{var}(x^k)}}.$$

Denotemos D_{1/σ^2} la siguiente métrica sobre E :

$$D_{1/\sigma^2} = \begin{pmatrix} 1/v_1 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 1/v_j & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 1/v_p \end{pmatrix}_{p \times p}$$

representa la matriz diagonal de los inversos de las varianzas.

Si identificamos a un individuo con el vector $x_i = (x_i^1, \dots, x_i^p)^t \in \mathbb{R}^p$, descrito por sus mediciones, podemos calcular la distancia entre dos individuos x_i, x_j mediante:

$$d^2(x_i, x_j) = \|x_i - x_j\|_M^2 = (x_i - x_j)^t M (x_i - x_j) = \sum_{k=1}^p \sum_{l=1}^p m_{kl} (x_i^k - x_j^k)(x_i^l - x_j^l)$$

1.6.2 Las aplicaciones X y X^t .

Después de repasar algunos conceptos del Algebra Lineal podemos definir de manera natural las siguientes aplicaciones.

Definición 15

$$\begin{aligned} X : E^* &\longrightarrow F \\ e_j^* &\longmapsto x^j, \text{ la variable } j. \\ \\ X^t : F^* &\longrightarrow E \\ f_i^* &\longmapsto x_i, \text{ el individuo } i. \end{aligned}$$

La matriz de X en las respectivas bases canónicas es la tabla de datos.

Ahora podemos resumir toda la información en un diagrama conocido como esquema de dualidad [5], dado en la figura 1.1.

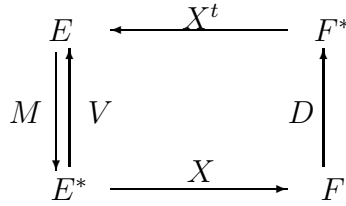


Figure 1.1: Esquema de Dualidad

1.6.3 Inercia e información perdida

Para medir la proximidad de un punto $a \in E$ a la nube de puntos $\mathcal{N} = (X, M, D)$, utilizamos la siguiente definición.

Definición 16 *La inercia de la nube $\mathcal{N} = (X, M, D)$ respecto al punto a , es:*

$$I_a = \sum_{i=1}^n p_i \|x_i - a\|_M^2$$

Se deduce el siguiente resultado $I_a = I_g + \|g - a\|_M^2$, conocido como teorema de Huygens. Como $I_a = I_g + \|g - a\|_M^2 \leq I_g \forall a$, la inercia es mínima cuando se calcula respecto al centro de gravedad.

Definición 17 *Sean S un subespacio de \mathbb{R}^p y S^\perp su complemento M -ortogonal, y sea $x_i = \alpha_i + \beta_i$, $\alpha_i \in S$, $\beta_i \in S^\perp$, el i -simo individuo. La inercia de la nube \mathcal{N} de individuos proyectada en S se define como:*

$$I_{S^\perp} = \sum_{i=1}^n p_i \|\alpha_i\|_M^2$$

El A.C.P. pretende encontrar el subespacio S de dimensión r de inercia I_{S^\perp} máxima (para tener la mejor representación). El resultado conocido del A.C.P establece que [5] el subespacio S de dimensión r de inercia proyectada máxima, es generado por los r vectores propios de VM asociados a los r valores propios mayores.

Si se denotan $\lambda_1, \dots, \lambda_p$ los valores propios de VM en orden decreciente, entonces se tiene:

$$I_{S^\perp} = \sum_{i=1}^r \lambda_i$$

Definición 18 *El porcentaje de inercia explicado por S es:*

$$\frac{\sum_{i=1}^r \lambda_i}{\text{Tr}(VM)}.$$

donde $\text{Tr}(VM) = \sum_{i=1}^p \lambda_i$ es la inercia de la nube.

Definición 19 *Al par (U, D_{λ_i}) lo llamaremos **la solución matricial** del A.C.P. de (X, M, D) si $U = [u_1, \dots, u_p]$ es una matriz de vectores propios u_j (por columna) asociados a los valores propios λ_j del operador VM , además los vectores propios son M -ortonormales y D_{λ_j} es la matriz con los valores propios en la diagonal.*

Note que matricialmente la definición anterior se expresa como $U^{-1}VMU = D_{\lambda_i}$ y $U^tMU = I$

Definición 20 *Al conjunto formado por los valores propios del operador VM lo llamaremos el espectro del A.C.P. de (X, M, D) .*

Chapter 2

Estudio de las métricas

2.1 Introducción

La escogencia de la métrica en Análisis en Componentes Principales juega un papel importante, ya que los resultados dependen de tal elección; por ejemplo, los porcentajes de inercia y las componentes principales. En el presente capítulo estudiaremos algunas propiedades que la métrica puede poseer, de acuerdo con el enfoque de algunos autores, especialmente el de M. Maurin [15, 16, 17, 18].

2.2 Una parametrización de métricas en A.C.P.

En la presente sección presentaremos una parametrización simultánea del triplete clásico del A.C.P. y de la solución del mismo, esta parametrización ofrece una nueva luz sobre algunos problemas en A.C.P.

Proposición 1 *Sea $\gamma_1, \dots, \gamma_p$ una lista de escalares positivos. Entonces existe una métrica M_γ tal que el espectro del A.C.P. de (X, M_γ, D) es $\{\gamma_1, \dots, \gamma_p\}$.*

DEMOSTRACIÓN:

Sea $Q = V^{1/2}O(D_{\gamma_i})^{-1/2}$ donde O es una matriz $p \times p$ ortogonal (i.e. $O^tO = I$.)

Sea $M_\gamma = (QQ^t)^{-1}$. Entonces (Q, D_{γ_i}) es una solución matricial del A.C.P. de (X, M_γ, D) .

- Verifiquemos primero que M_γ es una métrica.

La expresión M_γ se puede simplificar de la siguiente manera:

$$\begin{aligned} M_\gamma &= [(V^{1/2}O(D_{\gamma_i})^{-1/2})((D_{\gamma_i})^{-1/2}O^tV^{1/2})]^{-1} \\ &= (V^{1/2}OD_{\gamma_i}^{-1}O^tV^{1/2})^{-1} \\ &= V^{-1/2}OD_{\gamma_i}O^tV^{-1/2}. \end{aligned}$$

Es claro que M_γ es simétrica.

También es definida positiva: $x^tM_\gamma x \geq 0$ para cualquier vector p -dimensional x .

En efecto $x^tM_\gamma x = x^tV^{-1/2}OD_{\gamma_i}O^tV^{-1/2}x$. Sea $y = O^tV^{-1/2}x$

Por lo tanto $x^tM_\gamma x = y^tD_{\gamma_i}y \geq 0$ ya que D_{γ_i} es definida positiva, con igualdad sólo si $y = 0$.

Así $x^tM_\gamma x \geq 0$ con igualdad sólo si $x = 0$.

- Veamos que $Q^{-1}VM_\gamma Q = D_{\gamma_i}$.

En efecto se tiene

$$\begin{aligned} Q^{-1}VM_\gamma Q &= (D_{\gamma_i})^{1/2}O^tV^{-1/2}V(Q^t)^{-1}Q^{-1}Q \\ &= (D_{\gamma_i})^{1/2}O^tV^{-1/2}V^{1/2}V^{1/2}((D_{\gamma_i})^{-1/2}O^tV^{1/2})^{-1} \\ &= (D_{\gamma_i})^{1/2}O^tV^{1/2}V^{-1/2}O(D_{\gamma_i})^{1/2} \\ &= (D_{\gamma_i})^{1/2}O^tO(D_{\gamma_i})^{1/2} \\ &= (D_{\gamma_i})^{1/2}(D_{\gamma_i})^{1/2} \\ &= D_{\gamma_i} \end{aligned}$$

- Probemos ahora que $Q^tM_\gamma Q = I$.

Por la definición de M_γ tenemos:

$$\begin{aligned} Q^tM_\gamma Q &= Q^t(QQ^t)^{-1}Q \\ &= Q^tQ^{t-1}Q^{-1}Q \\ &= I. \end{aligned}$$

Por lo tanto, (Q, D_{γ_i}) es una solución matricial del A.C.P. de (X, M_γ, D) . ■

La proposición anterior dice, en el fondo, que prácticamente cualquier conjunto de p números positivos puede ser el espectro de algún A.C.P. Sin embargo, hay que tomar en cuenta que el interés de tener un conjunto de componentes principales para una tabla de datos, es su *interpretabilidad*, con el fin de que sean útiles para la descripción de la información contenida en la tabla.

En este sentido, dependerá de cada caso la interpretabilidad de las componentes principales asociadas a un espectro arbitrario.

Observación 5 *La métrica que establece la proposición anterior depende de una matriz ortogonal, luego puede haber varias métricas que conducen al mismo espectro.*

2.3 Cualidades de las métricas

El espacio de individuos $E = \mathbb{R}^p$ está dotado de la métrica euclídea clásica que se conoce como el **tensor métrico fundamental**, el cual es un tensor simétrico de orden 2, dos veces covariante.

Uno de los objetivos del cálculo tensorial consiste en establecer las reglas de transformación de coordenadas de tensores cuando realizamos un cambio lineal de variables del tipo $y = A^{-1}x$, donde A es una matriz $p \times p$ invertible.

Tomando en cuenta lo anterior, vamos a estudiar cómo se comporta la solución de un A.C.P. cuando efectuamos cambios lineales de variables.

2.3.1 Cualidad tensorial

Definición 21

1. Sean (X, M_X, D) una nube de puntos y $Y^t = A^{-1}X^t$ una transformación lineal de las variables. Decimos que una métrica M_Y es **tensorial** si cumple $M_Y = A^t M_X A$.
2. Si ponemos $Y^t = D^{-1}X^t$ con D diagonal y si $M_Y = D^t M_X D$ decimos que la métrica es **subtensorial**.

Obsérvese que al realizar $Y^t = A^{-1}X^t$ tendríamos dos nubes (X, M_X, D) y (Y, M_Y, D) . Por lo tanto, si efectuamos dos A.C.P. tendríamos dos soluciones. En la presente sección estudiaremos la relación entre estas soluciones.

Denotaremos u_{X_i} los vectores propios de $V_X M_X$ asociados a los valores propios λ_{X_i} , donde $V_X = X^t D X$, y u_{Y_i} los vectores propios de $V_Y M_Y$ asociados a los valores propios λ_{Y_i} , donde $V_Y = Y^t D Y = A^{-1} X^t D X (A^{-1})^t = A^{-1} V_X (A^{-1})^t$.

Proposición 2 Si hacemos una transformación de los datos $Y^t = A^{-1}X^t$, entonces la métrica M_Y es tensorial si y sólo si $u_{Y_i} = A^{-1}u_{X_i}$ y $\lambda_{X_i} = \lambda_{Y_i}$.

DEMOSTRACIÓN:

Supongamos que $u_{Y_i} = A^{-1}u_{X_i}$ y $\lambda_{X_i} = \lambda_{Y_i}$, probemos que M_Y es tensorial.

Tenemos $V_Y M_Y u_{Y_i} = \lambda_{Y_i} u_{Y_i}$, además suponemos que V_Y es invertible, se tiene entonces la serie de igualdades:

$$\begin{aligned}
 M_Y u_{Y_i} &= \lambda_{Y_i} (V_Y)^{-1} u_{Y_i} \\
 &= \lambda_{Y_i} (A^{-1} V_X (A^{-1})^t)^{-1} u_{Y_i} \\
 &= \lambda_{Y_i} A^t (V_X)^{-1} A u_{Y_i} \\
 &= A^t (V_X)^{-1} \lambda_{X_i} u_{X_i} \\
 &= A^t (V_X)^{-1} V_X M_X u_{X_i} \\
 &= A^t M_X u_{X_i} \\
 &= A^t M_X A u_{Y_i}
 \end{aligned}$$

Como $\{u_{Y_i}, i = 1, \dots, p\}$ forman una base del espacio de individuos $E = \mathbb{R}^p$, entonces $M_Y = A^t M_X A$.

Supongamos ahora que M_Y es tensorial. Tenemos por hipótesis $Y^t = A^{-1} X^t$, $V_Y = A^{-1} V_X (A^{-1})^t$ y $M_Y = A^t M_X A$.

Probemos que $u_{Y_i} = A^{-1} u_{X_i}$ y $\lambda_{Y_i} = \lambda_{X_i}$.

El operador $V_Y M_Y$ se escribe como

$$V_Y M_Y = A^{-1} V_X (A^{-1})^t A^t M_X A = A^{-1} V_X M_X A.$$

Así $V_Y M_Y A^{-1} u_{X_i} = A^{-1} V_X M_X A A^{-1} u_{X_i} = A^{-1} V_X M_X u_{X_i} = A^{-1} \lambda_{X_i} u_{X_i} = \lambda_{X_i} A^{-1} u_{X_i}$.

Por lo tanto $A^{-1} u_{X_i}$ son los vectores propios de $V_Y M_Y$ asociados a los valores propios λ_{X_i} . Luego $u_{Y_i} = A^{-1} u_{X_i}$ y $\lambda_{Y_i} = \lambda_{X_i}$ ■

Proposición 3 *Sea u_{X_i}, λ_{X_i} la solución del A.C.P. de (X, M_X, D) , con M_X tensorial. Entonces $M_X u_X Q u_X^t M_X$ es una métrica tensorial, para toda Q matriz simétrica definida positiva que no dependa de los datos.*

DEMOSTRACIÓN:

Supóngase que hacemos la transformación de datos $Y^t = A^{-1} X^t$. Por hipótesis sabemos que $M_Y = A^t M_X A$, pues M_X es tensorial.

Por la proposición 2, debemos probar que si \tilde{u}_{X_i} y $\tilde{\lambda}_{X_i}$ es la solución matricial del A.C.P. de $(X, M_X u_X Q u_X^t M_X, D)$ y si \tilde{u}_{Y_i} y $\tilde{\lambda}_{Y_i}$ es la solución matricial del A.C.P. de $(X, M_Y u_Y Q u_Y^t M_Y, D)$, entonces $\tilde{\lambda}_{X_i} = \tilde{\lambda}_{Y_i}$ y $\tilde{u}_{Y_i} = A^{-1} \tilde{u}_{X_i}$.

Verifiquemos primero que $M_Y u_Y Q u_Y^t M_Y = A^t M_X u_X Q u_X^t M_X A$.

Sustituyendo se tiene:

$$\begin{aligned} M_Y u_Y Q u_Y^t M_Y &= A^t M_X A u_Y Q u_Y^t A^t M_X A \\ &= A^t M_X u_X Q u_X^t M_X A. \end{aligned}$$

Ahora, sabemos que $V_Y M_Y u_Y Q u_Y^t M_Y \tilde{u}_{Y_i} = \tilde{\lambda}_{Y_i} \tilde{u}_{Y_i}$.

Reemplazando obtenemos la siguiente serie de igualdades:

$$\begin{aligned} \tilde{\lambda}_{Y_i} \tilde{u}_{Y_i} &= Y^t D Y A^t M_X u_X Q u_X^t M_X A \tilde{u}_{Y_i} \\ &= A^{-1} X^t D X (A^{-1})^t A^t M_X u_X Q u_X^t M_X A \tilde{u}_{Y_i} \\ &= A^{-1} X^t D X M_X u_X Q u_X^t M_X A \tilde{u}_{Y_i} \end{aligned}$$

Luego,

$$\begin{aligned} X^t D X M_X u_X Q u_X^t M_X A u_{\tilde{Y}_i} &= A \tilde{\lambda}_{Y_i} u_{\tilde{Y}_i} \\ &= \tilde{\lambda}_{Y_i} A u_{\tilde{Y}_i}. \end{aligned}$$

Luego $A u_{\tilde{Y}_i}$ es vector propio de $X^t D X M_X u_X Q u_X^t M_X$ asociado al valor propio $\tilde{\lambda}_{Y_i}$. Pero sabemos que la solución del A.C.P de $(X, M_X u_X Q u_X^t M_X, D)$ es $u_{\tilde{X}_i}$ y $\tilde{\lambda}_{X_i}$.

Entonces tenemos $A u_{\tilde{Y}_i} = u_{\tilde{X}_i}$ y $\tilde{\lambda}_{Y_i} = \tilde{\lambda}_{X_i}$. ■

Las proposiciones anteriores dan algunas propiedades interesantes cuando se hace una transformación lineal de los datos.

Estas propiedades permiten deducir los elementos principales del A.C.P., así como establecer ciertas condiciones sobre las métricas.

En la siguiente sección veremos algunas propiedades relacionadas con el caso en que se tengan tablas de datos divididas en bloques.

2.3.2 Calidad algebraica: estrategias de agregación

Esta cualidad interviene cuando tomamos el problema del análisis de varias nubes de datos $(X_{h(n,p_h)}, M_h, D)$ para $h = 1 \dots H$, con X_h una matriz $n \times p_h$.

Observe que son varios bloques de variables, cada uno con p_h variables medidas sobre los mismos n individuos. Además cada bloque está dotado de una métrica M_h llamada métrica intra.

El espacio de individuos E , que se obtiene al unir los bloques, está asociado a la tabla concatenada $X = (X_1 | \dots | X_H)$ y es de dimensión $\sum_h p_h$.

Definición 22 *Una estrategia de agregación o protocolo de agregación de métricas intra es una regla de construcción de la métrica $M(M_h)$ de E a partir de las métricas intra.*

Observación 6 *La concatenación de tablas es una ley de composición que posee las propiedades elementales de asociatividad y conmutatividad:*

- $X = ((X_1, X_2), X_3) = (X_1, (X_2, X_3))$

- $X = (X_1, X_2) = (X_2, X_1)$ mediante una permutación sobre las columnas, en el sentido de que el espacio E engendrado es el mismo.

Definición 23

1. Una estrategia de agregación es **conmutativa** si aplica sobre las líneas y columnas de la métrica resultante la misma permutación que sobre las columnas de (X_1, X_2)
2. Una estrategia de agregación es **asociativa** si respeta una regla análoga:
 $M((M_1, M_2), M_3) = M(M_1, (M_2, M_3))$, donde los paréntesis indican el orden de prioridad en que se agregan las métricas.

Veamos a continuación algunas estrategias de agregación clásicas.

2.3.3 Estrategia d

Definición 24 La estrategia d consiste en yuxtaponer en matrices de bloques diagonales, las diferentes métricas intra: $M_d = \text{diag}(M_h)$

La métrica resultante está determinada por las métricas intra.

La dependencia eventual en función de los datos, en general no se puede hacer por la intermediación de métricas intra. Es asociativa y conmutativa.

Proposición 4 La estrategia diagonal d es subtensorial si cada métrica intra M_h es subtensorial.

DEMOSTRACIÓN:

Sea (U, D_{γ_i}) la solución matricial del A.C.P. de (X, M, D) . Si hacemos la transformación de variables $Y^t = D_{\lambda}^{-1}X^t$, con D_{λ} una matriz diagonal, entonces si $w_i = (D_{\lambda})^{-1}u_i$, por la proposición 2 hay que probar que w_i es vector propio de $V_Y M_Y$ donde $V_Y = Y^t D Y$ y $M_Y = D_{\lambda}^t M D_{\lambda}$.

Luego el resultado se reduce a un cálculo:

$$\begin{aligned} Y^t D Y D_{\lambda}^t M D_{\lambda} w_i &= Y^t D Y D_{\lambda}^t M D_{\lambda} (D_{\lambda})^{-1} u_i \\ &= Y^t D Y D_{\lambda}^t M u_i \end{aligned}$$

$$\begin{aligned}
&= D_\lambda^{-1} X^t D X (D_\lambda^{-1})^{-1} D_\lambda^t M u_i \\
&= D_\lambda^{-1} X^t D X M u_i \\
&= D_\lambda^{-1} V M u_i \\
&= D_\lambda^{-1} \gamma_i u_i \\
&= \gamma_i D_\lambda^{-1} u_i \\
&= \gamma_i w_i. \quad \blacksquare
\end{aligned}$$

2.3.4 Estrategia Σ

Definición 25 *La estrategia Σ consiste en tomar la yuxtaposición de métricas de Mahalanobis intra en bloques diagonales $M_\Sigma = \text{diag}(V_h^{-1})$ una matriz diagonal por bloques.*

Esta estrategia respeta el carácter tensorial de cada métrica, ya que cada una es tensorial por la proposición 6. La métrica resultante está enteramente determinada por los datos. Esta estrategia es asociativa y conmutativa en relación con la operación de yuxtaposición de bloques diagonales. Véase que es un caso particular de la estrategia d cuando las métricas intra son las métricas de Mahalanobis.

Corolario 5 *La estrategia Σ es subtensorial.*

DEMOSTRACIÓN:

Es obvio de la proposición anterior ya que cada métrica V_h^{-1} es subtensorial. \blacksquare

2.3.5 Estrategia de Mahalanobis

Definición 26 *La estrategia M_a de Mahalanobis consiste en tomar por métrica global M_{M_a} a V^{-1} , donde V es la matriz de varianzas y covarianzas asociada a la tabla X .*

Esta estrategia está enteramente determinada por las tablas X_h en el sentido de que no depende en ningún modo de las métricas intra. Es asociativa y conmutativa.

Proposición 6 *La estrategia M_a de Mahalanobis es tensorial.*

DEMOSTRACIÓN:

Tenemos el triple (X, V_X^{-1}, D) y consideramos $Y^t = A^{-1}X^t$. Calculemos ahora $V_Y = Y^t D Y = A^{-1}X^t D X (A^{-1})^t = A^{-1}V_X(A^{-1})^t$. Aquí $M_Y = V_Y^{-1} = (A^{-1}V_X(A^{-1})^t)^{-1} = A^t V_X^{-1} A$. Luego $M_Y = A^t M_X A$. Por lo tanto M_Y es tensorial. ■

2.3.6 Estrategia relacional

Definición 27 *La estrategia relacional Rel consiste en $M_R = D_B V D_B^t$, donde*

$B_h = M_h(V_h M_h)^{-1/2}$ para $h = 1, \dots, H$ y D_B es la matriz en bloques diagonales de las B_h para $h = 1, \dots, H$.

Esta métrica fue introducida por Schektman y es presentada en [25], la métrica resultante depende de los datos y de las métricas intra. Por la fórmula de la definición es evidentemente conmutativa y asociativa.

Proposición 7 *Si cada métrica intra es subtensorial, la estrategia relacional Rel es subtensorial.*

DEMOSTRACIÓN:

A partir de $Y_h^t = D_h^{-1}X_h^t$ y $V_{Y_h} = D_h^{-1}V_{X_h}(D_h^{-1})^t$ y la hipótesis de cada una de las métricas es subtensorial se tiene :

$$M_{Y_h} = D_h^t M_{X_h} D_h.$$

Veamos primero que $V_{Y_h} M_{Y_h} = D_h^{-1}V_{X_h} M_{X_h} D_h$ ya que al sustituir obtenemos :

$$V_{Y_h} M_{Y_h} = D_h^{-1}V_{X_h}(D_h^{-1})^t D_h^t M_{X_h} D_h = D_h^{-1}V_{X_h} M_{X_h} D_h.$$

Usando

$$(V_{Y_h} M_{Y_h})^{-1/2} (V_{Y_h} M_{Y_h})^{-1/2} = (V_{Y_h} M_{Y_h})^{-1}.$$

obtenemos :

$$(V_{Y_h} M_{Y_h})^{-1/2} = D_h^{-1} (V_{X_h} M_{X_h})^{-1/2} D_h.$$

Probemos ahora que $B_{Y_h} = D_h^t B_{X_h} D_h$. Por definición tenemos:

$$\begin{aligned}
 B_{Y_h} &= M_{Y_h} (V_{Y_h} M_{Y_h})^{-1/2} \\
 &= D_h^t M_{X_h} D_h (D_h^{-1} (V_{X_h} M_{X_h})^{-1/2} D_h) \\
 &= D_h^t M_{X_h} (V_{X_h} M_{X_h})^{-1/2} D_h \\
 &= D_h^t B_{X_h} D_h.
 \end{aligned}$$

Obsérvese que : $Diag(B_Y) = Diag(D)^t Diag(B_X) Diag(D)$

Ahora, si hacemos la transformación de datos $Y^t = Diag(D)^{-1} X^t$ podemos notar que

$$V_Y = Diag(D)^{-1} V_X Diag^{-1}(D)^t$$

ya que:

$$V_Y = Y^t D Y = Diag(D)^{-1} X^t D X Diag^{-1}(D)^t = Diag(D)^{-1} V_X Diag^{-1}(D)^t.$$

Por último veamos que la métrica M_{Rel_Y} es subtensorial:

$$\begin{aligned}
 M_{Rel_Y} &= Diag(B_Y) V_Y Diag(B_Y)^t \\
 &= Diag(D)^t Diag(B_X) Diag(D) Diag(D)^{-1} V_X Diag^{-1}(D)^t Diag(D)^t Diag(B_X)^t Diag(D) \\
 &= Diag(D)^t Diag(B_X) V_X Diag(B_X)^t Diag(D) \\
 &= Diag(D)^t M_{Rel_X} Diag(D). \quad \blacksquare
 \end{aligned}$$

En la tabla 2.1 se presenta un resumen de las cualidades algebraicas para cada estrategia de agregación.

2.4 La métrica inducida por un A.C.P.

Cuando se hace un Análisis en Componentes Principales, se interpreta la posición de los individuos de acuerdo con sus coordenadas en los planos principales. En ocasiones, ocurre que las posiciones de algunas parejas de individuos son muy similares y es difícil hacer una

	Tensorial	Subtensorial	Asociativa	Commutativa
d	No	Sí	Sí	Sí
Σ	No	Sí	Sí	Sí
M_a	Sí	Sí	Sí	Sí
Rel	No	Sí	Sí	Sí

Table 2.1: Propiedades de algunas estrategias de agregación.

separación de los individuos en clases. En tales casos, se recomienda [20] usar una *distancia inducida* por el A.C.P., esto es, que el cálculo de la distancia euclídea pondere las respectivas coordenadas por el valor propio correspondiente. Así, la distancia entre los puntos x, y inducida por el A.C.P. del triplete (X, M, D) es :

$$\sum_{j=1}^p \lambda_j (x^j - y^j)^2$$

donde λ_j son los valores propios de VM , ordenados en forma decreciente .

Este procedimiento se puede iterar, es decir, se puede calcular el A.C.P. sobre la tabla deducida de las coordenadas de los individuos sobre el espacio principal, usando la distancia inducida por el primer A.C.P. Si se reitera este proceso, E. Piza [20] se cuestiona sobre la estabilidad del procedimiento y muestra que la única métrica estable es la de Mahalanobis.

2.5 Otros Estudios

Existen otros autores que también han hecho estudios sobre la influencia de la métrica en Análisis de Datos. Estos trabajos no tienen una relación directa con nuestro tema de estudio, sin embargo, los citamos a continuación.

Por ejemplo, J.P. Benzécri et al. [2] estudian la estabilidad de espacios principales cuando se tienen distancias relacionadas por las desigualdades:

$$\forall x, y \in E : k \|x - y\|_M^2 \leq \|x - y\|_N^2 \leq K \|x - y\|_M^2$$

para dos métricas M y N sobre el espacio de individuos E . Se deducen relaciones análogas para los valores propios de VM y VN , y para las inercias. Además encuentran una relación

importante para el ángulo entre espacios principales de los diferentes A.C.P. Esta relación muestra que hay estabilidad de la solución, si hay una separación importante entre los valores propios de VM respecto a K/k .

Por otra parte, J.P. Kazmierczak [13] se preocupa de definir una generalización apropiada de la distancia de chi-cuadrado con el fin de satisfacer un principio de equivalencia distribucional en sentido amplio, el cual generaliza el principio de equivalencia distribucional usual del análisis de correspondencias.

Chapter 3

Generación de métricas útiles en análisis de datos

3.1 Introducción

Uno de los problemas abiertos en ciertos métodos de Análisis de Datos es el de la escogencia de la métrica. En la práctica generalmente la selección consiste en la métrica asociada a la matriz identidad o a la métrica diagonal de las inversas de las varianzas. Este es el caso, por ejemplo, para la métrica en el espacio de individuos en Análisis en Componentes Principales (A.C.P.) o en Clasificación Automática para el Método de Nubes Dinámicas. Por otro lado, hay métodos para los que se conoce la métrica óptima según algún criterio (por ejemplo, en discriminación factorial y en regresión lineal), pero los problemas numéricos ligados al cálculo de la métrica impiden encontrar soluciones confiables en ciertos casos. Aparte del uso de semiproductos escalares [26], se pueden abordar estos problemas mediante la generación de métricas o de productos escalares [3, 24]. Esta consiste en construir una sucesión finita (M_r) de matrices asociadas a las métricas.

En este capítulo se presentan diferentes proposiciones para generar métricas definidas por medio de productos escalares, principalmente los generadores primarios y espectrales. Se estudian algunas propiedades que ayudarían a simplificar la utilización de ciertas técnicas de análisis de datos.

En 3.2 presentaremos las definiciones de productos escalares primarios y espectrales. En

3.3 estudiaremos las propiedades de una nueva familia de productos escalares que los llamaremos proporcionales. Finalmente, en 3.4 presentaremos una extensión de los generadores espectrales desarrollados por S. Labrèche.

3.2 Presentación de los generadores de métricas

La generación de métricas consiste en construir una sucesión finita (M_r) de matrices de producto escalar sobre un espacio vectorial. Consideraremos el caso de la generación de métricas sobre el espacio de individuos $E = \mathbb{R}^p$, aunque por supuesto el mismo principio se puede aplicar al espacio de variables. Identificaremos una métrica y su matriz en la base canónica de E ; así, M_r denotará tanto una métrica como una matriz simétrica definida positiva.

Sean \mathcal{M} el conjunto de las métricas sobre E y $\bar{\mathcal{M}} = \{\bar{M}_1, \dots, \bar{M}_\nu\} \subset \mathcal{M}$ una familia dada de ν métricas sobre E . Sea r en un conjunto finito, estudiamos métricas de la forma:

$$M_r = w_r \cdot \bar{\mathcal{M}} = \sum_{\ell=1}^{\nu} w_\ell(r) \bar{M}_\ell$$

donde $w_r = (w_1(r), \dots, w_\nu(r))^t \in \mathbb{R}^\nu$.

3.2.1 Generadores primarios

La generación primaria de métricas consiste en construir una sucesión finita (M_r) de m métricas, tal que $M_r = w_r \cdot \bar{\mathcal{M}} = w_1(r) \bar{M}_1 + w_2(r) \bar{M}_2$, donde $r \in \{0, 1, \dots, m\}$, $\bar{\mathcal{M}} = \{\bar{M}_1, \bar{M}_2\}$ y $w_r = (w_1(r), w_2(r))^t \in (\mathbb{R}^+)^2$, con $\bar{M}_1 = M_0$ y $\bar{M}_2 = M_m$.

Eventualmente se puede proponer generar $M_r = (w_r \cdot \bar{\mathcal{M}})^{-1}$. Se definen los generadores primarios de métricas siguientes:

(a) $M_r = w_1(r) \bar{M}_1 + w_2(r) \bar{M}_2$,

donde $w_1(r) = \frac{m-r}{m} \in [0, 1]$ y $w_2(r) = \frac{r}{m} \in [0, 1]$.

(b) $M_r = [w_1(r) \bar{M}_1^{-1} + w_2(r) \bar{M}_2^{-1}]^{-1}$,

donde $w_1(r) = \frac{m-r}{m} \in [0, 1]$ y $w_2(r) = \frac{r}{m} \in [0, 1]$.

$$(c) \quad M_r = w_1(r)\bar{M}_1 + w_2(r)\bar{M}_2,$$

$$\text{donde } w_2(r) = \text{sen}\left(\frac{\pi r}{2m}\right) \text{ y } w_1(r) = 1 - w_2(r).$$

$$(d) \quad M_r = \left[w_1(r)\bar{M}_1^{-1} + w_2(r)\bar{M}_2^{-1} \right]^{-1},$$

$$\text{donde } w_2(r) = \text{sen}\left(\frac{\pi r}{2m}\right) \text{ y } w_1(r) = 1 - w_2(r).$$

Nótese que $w_1(r) + w_2(r) = 1$, y que los generadores (c) y (d) permiten un acercamiento “lento” a M_m . En general, se tomarán $M_0, M_m \in \{V, V^{-1}, I_p, D_{1/\sigma^2}\}$, donde V es la matriz de covarianzas de las variables observadas x^1, \dots, x^p , que suponemos invertible, I_p la matriz identidad de orden p y $D_{1/\sigma^2} = \text{diag}(1/\sigma_j^2)$ la matriz diagonal de las inversas de las varianzas de las variables.

La generación primaria de métricas se ha usado habitualmente [1], [4], [8], [28], [27], [29] para buscar un producto escalar que optimice un criterio dado, con el fin de que sea bien adaptado al problema.

Para esta tesis, nosotros hemos estudiado las propiedades de estos generadores y hemos obtenido algunos resultados interesantes, que presentamos a continuación.

Proposición 8 *Si $\bar{M}_1, \bar{M}_2 \in \{V, V^{-1}, I_p\}$, entonces los vectores propios de las métricas primarias definidas en (a), (b), (c) y (d) son idénticos a los de V y los valores propios correspondientes se deducen de los de V .*

DEMOSTRACIÓN:

En efecto, sea $V = \sum_{j=1}^p \lambda_j P_j$ la descomposición espectral de V , donde P_j es el operador de proyección I_p -ortogonal sobre el subespacio propio asociado a λ_j .

$$\text{Se tiene } V^{-1} = \sum_{j=1}^p \frac{1}{\lambda_j} P_j \text{ y } I_p = \sum_{j=1}^p P_j.$$

Si M_r está definido por (a) entonces:

$$M_r = \sum_{j=1}^p \left[\frac{m-r}{m} \mu_j^1 + \frac{r}{m} \mu_j^2 \right] P_j$$

$$\text{donde } \mu_j^k = \begin{cases} \lambda_j & \text{si } \bar{M}_k = V \\ 1/\lambda_j & \text{si } \bar{M}_k = V^{-1} \\ 1 & \text{si } \bar{M}_k = I_p \end{cases} \text{ son los valores propios respectivos de } \bar{M}_k \text{ para } k = 1, 2.$$

También para (b), (c) y (d) tendremos respectivamente:

$$\begin{aligned} M_r &= \sum_{j=1}^p \left[\frac{m-r}{m\mu_j^1} + \frac{r}{m\mu_j^2} \right]^{-1} P_j \\ M_r &= \sum_{j=1}^p [\mu_j^1 + \text{sen}(\frac{\pi r}{2m})(\mu_j^2 - \mu_j^1)] P_j \\ M_r &= \sum_{j=1}^p \left[\frac{1}{\mu_j^1} + \text{sen}(\frac{\pi r}{2m}) \left(\frac{1}{\mu_j^2} - \frac{1}{\mu_j^1} \right) \right]^{-1} P_j \end{aligned}$$

Luego, todos los productos escalares tienen los mismos vectores propios que V y los valores propios correspondientes figuran en la descomposición espectral asociada. ■

Proposición 9 Si $\bar{M}_1 = D_{1/\sigma^2}$, $\bar{M}_2 = V^{-1}$ y si se usan los generadores definidos en (a) o (c), los vectores propios de VM_r son iguales a los de VD_{1/σ^2} .

DEMOSTRACIÓN:

Para mostrar este resultado, consideremos un vector propio u de VD_{1/σ^2} asociado al valor propio λ , entonces para (a) tenemos:

$$\begin{aligned} VM_r u &= V \left(\frac{m-r}{m} D_{1/\sigma^2} + \frac{r}{m} V^{-1} \right) u \\ &= \left(\frac{m-r}{m} V D_{1/\sigma^2} + \frac{r}{m} I_p \right) u \\ &= \left(\frac{m-r}{m} \lambda + \frac{r}{m} \right) u, \end{aligned}$$

luego, u es vector propio de VM_r asociado al valor propio $\frac{m-r}{m} \lambda + \frac{r}{m}$.

Para (c) tenemos:

$$\begin{aligned} VM_r u &= V(D_{1/\sigma^2} - \text{sen}(\pi r/2m) D_{1/\sigma^2} + \text{sen}(\pi r/2m) V^{-1}) u \\ &= (V D_{1/\sigma^2} + \text{sen}(\pi r/2m)(I_p - V D_{1/\sigma^2})) u \\ &= (\lambda + \text{sen}(\pi r/2m)(1 - \lambda)) u, \end{aligned}$$

luego, u es vector propio de VM_r asociado al valor propio $\lambda + \text{sen}(\pi r/2m)(1 - \lambda)$. ■

Proposición 10 Si $\lambda_1, \dots, \lambda_p$ son los valores propios de V , asociados a los vectores propios u_1, \dots, u_p , entonces tenemos:

1. Si $\bar{M}_1 = V$, $\bar{M}_2 = I$ y si se usan los generadores definidos en (a) o (c), los vectores propios de VM_r son iguales a los de V y los valores propios respectivos son: $(\frac{m-r}{m})\lambda_i^2 + \frac{r}{m}\lambda_i$ y $(1 - \text{sen}(\frac{\pi r}{2m}))\lambda_i^2 + \text{sen}(\frac{\pi r}{2m})\lambda_i$.
2. Si $\bar{M}_1 = I$, $\bar{M}_2 = V^{-1}$ y si se usan los generadores definidos en (a) o (c), si $r < m$ los vectores propios de VM_r son iguales a los de V y los valores propios respectivos son: $(\frac{m-r}{m})\lambda_i + \frac{r}{m}$ y $(1 - \text{sen}(\frac{\pi r}{2m}))\lambda_i + \text{sen}(\frac{\pi r}{2m})$.

DEMOSTRACIÓN:

Veamos la demostración del primer apartado. Para el caso (a) tenemos

$$\begin{aligned}
 VM_r u_i &= V[\frac{m-r}{m}V + \frac{r}{m}I]u_i \\
 &= V[\frac{m-r}{m}Vu_i + \frac{r}{m}u_i] \\
 &= V[\frac{m-r}{m}\lambda_i u_i + \frac{r}{m}u_i] \\
 &= \frac{m-r}{m}\lambda_i Vu_i + \frac{r}{m}Vu_i \\
 &= \frac{m-r}{m}\lambda_i^2 u_i + \frac{r}{m}\lambda_i u_i \\
 &= [\frac{m-r}{m}\lambda_i^2 + \frac{r}{m}\lambda_i]u_i.
 \end{aligned}$$

Para el caso (c) tenemos similarmente

$$\begin{aligned}
 VM_r u_i &= V[(1 - \text{sen}(\frac{\pi r}{2m}))V + \text{sen}(\frac{\pi r}{2m})I]u_i \\
 &= V[(1 - \text{sen}(\frac{\pi r}{2m}))Vu_i + \text{sen}(\frac{\pi r}{2m})u_i] \\
 &= V[(1 - \text{sen}(\frac{\pi r}{2m}))\lambda_i u_i + \text{sen}(\frac{\pi r}{2m})u_i] \\
 &= (1 - \text{sen}(\frac{\pi r}{2m}))\lambda_i Vu_i + \text{sen}(\frac{\pi r}{2m})Vu_i \\
 &= (1 - \text{sen}(\frac{\pi r}{2m}))\lambda_i^2 u_i + \text{sen}(\frac{\pi r}{2m})\lambda_i u_i \\
 &= [(1 - \text{sen}(\frac{\pi r}{2m}))\lambda_i^2 + \text{sen}(\frac{\pi r}{2m})\lambda_i]u_i.
 \end{aligned}$$

La demostración del segundo apartado es análoga. ■

Observación 7 Las proposiciones 8, 9 y 10 son útiles cuando, para cada métrica M_r , sea necesario obtener los elementos propios. Estos serán calculados una sola vez durante la diagonalización de V o la de VD_{1/σ^2} .

Proposición 11

1. Si $\bar{M}_1 = V$, $\bar{M}_2 = I$ y si se usan los generadores definidos en (a) o (c), las correlaciones entre las variables y las componentes principales de los respectivos A.C.P. son las mismas.
2. Si $\bar{M}_1 = I$, $\bar{M}_2 = V^{-1}$, si se usan los generadores definidos en (a) o (c) y si $r < m$, entonces las correlaciones entre las variables y las componentes principales de los respectivos A.C.P. son las mismas que las correlaciones mencionadas en el inciso 1.

DEMOSTRACIÓN:

Sean $\lambda_1, \dots, \lambda_p$ los valores propios de V asociados a los vectores propios u_1, \dots, u_p .

Para la demostración de la primera parte, en el caso (a) sea $C_1^k = XIu_k$. Veamos que:

$$\begin{aligned} C_\alpha^k &= X[(1 - \alpha)V + \alpha I]u_k = (1 - \alpha)XVu_k + \alpha Xu_k \\ &= [(1 - \alpha)\lambda_k + \alpha]Xu_k = [(1 - \alpha)\lambda_k + \alpha]C_1^k. \end{aligned}$$

Con base en lo anterior escribimos luego:

$$\begin{aligned} \rho(x^j, C_\alpha^k) &= \frac{\text{cov}(x^j, C_\alpha^k)}{\|x^j\|_D \|C_\alpha^k\|_D} \\ &= \frac{\text{cov}(x^j, [(1 - \alpha)\lambda_k + \alpha]C_1^k)}{\|x^j\|_D \|[(1 - \alpha)\lambda_k + \alpha]C_1^k\|_D} \\ &= \frac{[(1 - \alpha)\lambda_k + \alpha]\text{cov}(x^j, C_1^k)}{\|x^j\|_D [(1 - \alpha)\lambda_k + \alpha]\|C_1^k\|_D} \\ &= \frac{\text{cov}(x^j, C_1^k)}{\|x^j\|_D \|C_1^k\|_D} \\ &= \rho(x^j, C_1^k). \end{aligned}$$

La demostración del caso (c), así como de la segunda parte son análogas. ■

La proposición anterior es muy importante para el análisis, ya que de ella se deduce que cuando se efectúa una generación primaria de V a I o de I a V^{-1} (sin llegar a esta última), las

variables no se pueden interpretar de otra forma, es decir la elección de la métrica (primaria) es independiente de los resultados, para las variables.

Hacemos notar que esta proposición nació después que se implementaron los programas ya que a la luz de los resultados para estas generaciones, fue posible percibir que las correlaciones eran las mismas.

Proposición 12 *Al aplicar el generador primario definido en (a) la inercia de la nube de puntos $\mathcal{N}_r = (X, M_r, D)$ es:*

1. Si $\bar{M}_1 = V$ y $\bar{M}_2 = I_p$ entonces $I(\mathcal{N}_r) = (1 - \alpha) \sum_{k=1}^p \sum_{j=1}^p cov^2(x^j, x^k) + \alpha \sum_{j=1}^p var(x^j)$.
2. Si $\bar{M}_1 = I_p$ y $\bar{M}_2 = V^{-1}$ entonces $I(\mathcal{N}_r) = (1 - \alpha) \sum_{j=1}^p var(x^j) + \alpha p$.
3. Si $\bar{M}_1 = I_p$ y $\bar{M}_2 = D_{1/\sigma^2}$ entonces $I(\mathcal{N}_r) = (1 - \alpha) \sum_{j=1}^p var(x^j) + \alpha p$.

donde $\alpha = \frac{r}{m}$.

DEMOSTRACIÓN:

Veamos la demostración del inciso 1:

$$\begin{aligned}
 I(\mathcal{N}_r) &= traza(VM_r) \\
 &= traza\{V[(1 - \alpha)V + \alpha I_p]\} \\
 &= (1 - \alpha)traza(V^2) + \alpha traza(V) \\
 &= (1 - \alpha) \sum_{k=1}^p \sum_{j=1}^p cov^2(x^j, x^k) + \alpha \sum_{j=1}^p var(x^j).
 \end{aligned}$$

Las demostraciones de los incisos 2 y 3 son análogos. ■

Corolario 13 *Al aplicar el generador primario definido en (c) la inercia de la nube de puntos $\mathcal{N}_r = (X, M_r, D)$ es:*

1. Si $\bar{M}_1 = V$ y $\bar{M}_2 = I_p$ entonces

$$I(\mathcal{N}_r) = (1 - \text{sen}(\frac{\pi r}{2m})) \sum_{k=1}^p \sum_{j=1}^p cov^2(x^j, x^k) + \text{sen}(\frac{\pi r}{2m}) \sum_{j=1}^p var(x^j).$$

2. Si $\bar{M}_1 = I_p$ y $\bar{M}_2 = V^{-1}$ entonces

$$I(\mathcal{N}_r) = (1 - \text{sen}(\frac{\pi r}{2m})) \sum_{j=1}^p \text{var}(x^j) + \text{sen}(\frac{\pi r}{2m}) p.$$

3. Si $\bar{M}_1 = I_p$ y $\bar{M}_2 = D_{1/\sigma^2}$ entonces

$$I(\mathcal{N}_r) = (1 - \text{sen}(\frac{\pi r}{2m})) \sum_{j=1}^p \text{var}(x^j) + \text{sen}(\frac{\pi r}{2m}) p.$$

3.2.2 Generadores espectrales

Sean $\{u_j/j = 1, \dots, p\}$ una base I_p -ortonormada de E , $\bar{\mathcal{M}} = \{u_j u_j^t / j = 1, \dots, p\}$ y $w_r = (w_1(r) \cdots w_p(r))^t \in (\mathbb{R}^+)^p$. Los generadores espectrales de métricas construyen una sucesión finita (M_r) tal que

$$M_r = w_r \cdot (\bar{M}_1 \cdots \bar{M}_p) = \sum_{j=1}^p w_j(r) u_j u_j^t$$

donde $\bar{M}_j = u_j u_j^t$ es el operador de proyección ortogonal sobre Δu_j . Así, $\{u_j\}$ constituye una base de vectores propios ortonormales de M_r asociados a los valores propios $\{w_j(r)\}$. En general, se tomará por $\{u_j/j = 1, \dots, p\}$ una base I_p -ortonormada de vectores propios de V .

La generación espectral de métricas consiste en encontrar conjuntos de reales positivos $\{w_j(r)\}$ de tal manera que se mejore un criterio hasta que haya convergencia. Nótese que algunas métricas primarias definidas anteriormente también son espectrales.

El principio de la generación espectral es diferente del de la generación primaria. En este último caso, se genera un número m fijado a priori de combinaciones de métricas fijas (\bar{M}_1 y \bar{M}_2) utilizando una fórmula dada. Así, en una aplicación en Análisis de Datos, se retendrá la métrica de la familia así construida que optimice un criterio dado. En la generación espectral, en cambio, se da solamente la *forma* de las métricas y el número de métricas a generar no se fija de antemano. Dada una métrica inicial M_0 correspondiente a $w_1(0), \dots, w_p(0)$, se trata de encontrar heurísticas para calcular los $w_j(r)$ con tal de optimizar el criterio dado.

Una propiedad interesante que poseen las métricas espectrales es la siguiente.

Proposición 14 Si $M_0 = V$ con descomposición espectral $V = \sum_{j=1}^p w_j(0) u_j u_j^t$ entonces $\forall k \in \{1, \dots, p\}$ y $\forall r$, u_k es vector propio de $V M_r$ asociado al valor propio $w_k(r) w_k(0)$.

DEMOSTRACIÓN:

En efecto, $\forall k \in \{1, \dots, p\}$ y $\forall r$ tenemos:

$$\begin{aligned}
 VM_r u_k &= V\left(\sum_{j=1}^p w_j(r) u_j u_j^t\right) u_k \\
 &= \sum_{j=1}^p w_j(r) V u_j \delta_{jk} \\
 &= w_k(r) V u_k \\
 &= w_k(r) w_k(0) u_k \quad \blacksquare
 \end{aligned}$$

En la sección siguiente, se presenta una familia de métricas espectrales, que hemos llamado proporcionales. Se estudian algunas propiedades de estos generadores y proponemos un algoritmo iterativo para generar métricas proporcionales.

3.3 Generación proporcional

La idea de proponer un generador proporcional nos fue sugerida por los estudios de C. Fabre que planteó la siguiente definición.

Definición 28 Sea $\{\theta_{jk}/j, k = 1, \dots, p; j \neq k\}$ una familia de ángulos expresados en radianes, se define \mathcal{C}_θ como el conjunto de métricas M sobre E tales que $\cos_M(e_j, e_k) = \cos \theta_{jk}$ para todo $j, k \in \{1, \dots, p\}, j \neq k$, donde $\{e_j/j = 1, \dots, p\}$ es la base canónica de E . Si $\{v_j/j = 1, \dots, p\}$ es una base I_p -ortonormada de E , se define S_v como el conjunto de las métricas espectrales M sobre E tales que $M = \sum_{j=1}^p \lambda_j v_j v_j^t$, donde $\lambda_1, \dots, \lambda_p \in \mathbb{R}^+$.

Observación 8 Para un conjunto de θ_{jk} dados, \mathcal{C}_θ puede ser vacío. Sin embargo, en las heurísticas de generación de métricas que proponemos, los θ_{jk} serán deducidos de un producto escalar dado. Por lo tanto, estaremos en el caso $\mathcal{C}_\theta \neq \emptyset$.

Proposición 15 Si $M_0 \in \mathcal{C}_\theta \cap S_v$, entonces toda M en \mathcal{C}_θ tal que

$$\|e_2\|_M = \frac{\|e_2\|_{M_0}}{\|e_1\|_{M_0}} \|e_1\|_M, \dots, \|e_p\|_M = \frac{\|e_p\|_{M_0}}{\|e_1\|_{M_0}} \|e_1\|_M$$

pertenece a S_v .

La demostración de esta propiedad se puede encontrar en [8].

Si $[M_r]_{jk}$ denota la entrada (j, k) de M_r , las igualdades anteriores son equivalentes a:

$$\frac{[M_r]_{22}}{[M_r]_{11}} = \frac{[M_0]_{22}}{[M_0]_{11}}, \dots, \frac{[M_r]_{pp}}{[M_r]_{11}} = \frac{[M_0]_{pp}}{[M_0]_{11}} \quad (3.1)$$

Este resultado muestra que las métricas de \mathcal{C}_θ que preservan ciertas proporciones relativas a los elementos de la diagonal de su matriz, son espectrales. Es por ello que hemos llamado **métricas proporcionales** a las métricas espectrales que satisfacen las igualdades 3.1. La generación de tales métricas es llamada **generación proporcional**.

La generación proporcional consistirá entonces en dar una métrica inicial M_0 que definirá los cosenos de los ángulos (θ_{jk}) y considerar una descomposición espectral de M_0 . Para generar M_r , bastará definir $[M_r]_{11} = \|e_1\|_{M_r}^2$ según algún criterio y se calcularán los otros elementos diagonales a partir de las igualdades 3.1, los valores de los elementos fuera de la diagonal se deducirán de:

$$[M_r]_{jk} = \sqrt{[M_r]_{jj}[M_r]_{kk}} \cos \theta_{jk}, \quad j \neq k. \quad (3.2)$$

Observación 9 *Nótese que la métrica inicial M_0 fija los conjuntos \mathcal{C}_θ y S_v . De acuerdo con las igualdades 3.1, los cocientes entre los elementos diagonales y el primer elemento de la diagonal son los mismos para las métricas M y M_0 . Estos cocientes pueden ser definidos sobre cualquier otro elemento de la diagonal; en este caso, éste sería el elemento que habría que definir según algún criterio. Finalmente, los cocientes en 3.1 son independientes de M_0 ; las proporciones son idénticas con cualquier elemento de la sucesión generada $(M_r)_r$.*

Un asunto importante para el uso de los generadores proporcionales, es la escogencia de la métrica inicial M_0 . Se puede pensar en utilizar los generadores primarios de métricas para escoger la métrica inicial de un generador proporcional: la métrica primaria que optimice el criterio será usada como M_0 . Otra cuestión importante es garantizar que $\mathcal{C}_\theta \cap S_v \neq \emptyset$, es decir, la existencia de métricas espectrales cuyos ángulos entre los elementos de $\{e_j\}$ sean los ángulos dados. Este asunto ha sido abordado en [27] para el caso de dos dimensiones.

Nosotros abordamos el estudio en tres dimensiones pero no obtuvimos ningún resultado interesante.

Para aproximar la distancia de Mahalanobis con un generador espectral, Fabre [8] propuso utilizar un generador parecido al que nosotros hemos llamado proporcional.

A continuación se estudia el comportamiento del efecto de asociación, relacionado con la métrica de Mahalanobis, cuando se usa un generador proporcional.

3.3.1 El efecto de asociación es estacionario

Supóngase dado un conjunto de p variables cuantitativas $\{x^1, \dots, x^p\}$.

Definición 29 *Dada una métrica M en el espacio de individuos, el efecto de asociación de las x^j y $x_M^j = XM(e_j)$ se define como*

$$EA_M[x^j; x^k] = \rho(x_M^j, x^k).$$

Definición 30 *Se dice que una métrica M elimina el efecto de asociación si:*

$$EA_M[x^j; x^k] = 0, \text{ para todo } j \neq k$$

Se puede probar [8] que la métrica de Mahalanobis V^{-1} elimina el efecto de asociación. En un caso en que V^{-1} sea mal condicionada numéricamente, puede ser interesante aplicar un generador espectral “que se acerque” a V^{-1} (o eventualmente a la inversa generalizada de Moore-Penrose de V). En tal caso se construiría una sucesión (M_r) tal que el efecto de asociación disminuya hasta que éste sea lo más cercano a cero. La proposición 16 establece que la generación proporcional no sirve en este caso pues el efecto de asociación es estacionario.

Proposición 16 *Dados los conjuntos \mathcal{C}_θ y S_v , si M_r y M_{r+1} son métricas en $\mathcal{C}_\theta \cap S_v$ tales que M_{r+1} se deduce por un generador proporcional a partir de M_r , entonces para todo $j, k \in \{1, \dots, p\}, j \neq k$, se tiene*

$$EA_{M_r}[x^j; x^k] = EA_{M_{r+1}}[x^j; x^k]$$

DEMOSTRACIÓN:

Vamos a demostrar que para cualquier escogencia de $[M_{r+1}]_{11}$ nosotros tenemos que:

$$\rho(x_{M_r}^j, x^k) = \rho(x_{M_{r+1}}^j, x^k).$$

De las igualdades 3.1 y 3.2 tenemos $\forall j \in \{1, \dots, p\}$:

$$\begin{aligned} x_{M_r}^j &= X M_r(e_j) \\ &= \sum_{l=1}^p [M_r]_{lj} x^l \\ &= [M_r]_{jj} x^j + \sum_{l \neq j} \sqrt{[M_r]_{ll} [M_r]_{jj}} \cos \theta_{lj} x^l \end{aligned}$$

Por otra parte tenemos:

$$\begin{aligned} x_{M_{r+1}}^j &= X M_{r+1}(e_j) \\ &= \sum_{l=1}^p [M_{r+1}]_{lj} x^l \\ &= [M_{r+1}]_{jj} x^j + \sum_{l \neq j} \sqrt{[M_{r+1}]_{ll} [M_{r+1}]_{jj}} \cos \theta_{lj} x^l \\ &= \frac{[M_{r+1}]_{11}}{[M_r]_{11}} [M_r]_{jj} x^j + \frac{[M_{r+1}]_{11}}{[M_r]_{11}} \sum_{l \neq j} \sqrt{[M_r]_{ll} [M_r]_{jj}} \cos \theta_{lj} x^l \\ &= \frac{[M_{r+1}]_{11}}{[M_r]_{11}} ([M_r]_{jj} x^j + \sum_{l \neq j} \sqrt{[M_r]_{ll} [M_r]_{jj}} \cos \theta_{lj} x^l) \\ &= \frac{[M_{r+1}]_{11}}{[M_r]_{11}} x_{M_r}^j \end{aligned}$$

Luego tenemos $\forall k \neq j$:

$$\begin{aligned} \rho(x_{M_{r+1}}^j, x^k) &= \frac{\text{cov}(x_{M_{r+1}}^j, x^k)}{\sqrt{\text{var } x_{M_{r+1}}^j \text{ var } x^k}} \\ &= \frac{\frac{[M_{r+1}]_{11}}{[M_r]_{11}} \text{cov}(x_{M_r}^j, x^k)}{\sqrt{\left(\frac{[M_{r+1}]_{11}}{[M_r]_{11}}\right)^2 \text{var } x_{M_r}^j \text{ var } x^k}} \\ &= \frac{\text{cov}(x_{M_r}^j, x^k)}{\sqrt{\text{var } x_{M_r}^j \text{ var } x^k}} \\ &= \rho(x_{M_r}^j, x^k) \quad \blacksquare \end{aligned}$$

De este resultado podemos concluir que la generación proporcional de productos escalares no tiene interés cuando queremos aproximar la matriz V^{-1} .

En la siguiente sección propondremos un generador proporcional de productos escalares espectrales.

3.3.2 Proposición de un generador proporcional

Para definir un generador proporcional, es necesario definir la primera entrada de la diagonal de la matriz de las métricas a generar $[M_{r+1}]_{11}$. Todos los otros elementos de M_{r+1} se deducen de las fórmulas dadas por las igualdades 3.1 y 3.2.

A continuación proponemos una definición de un generador proporcional y estudiamos algunas de sus propiedades.

Definición 31 Sean dados $M_0 \in \mathcal{C}_\theta \cap S_v$ y $\delta \in \mathbb{R}^+$. A partir de la métrica M_r , se define M_{r+1} por $[M_{r+1}]_{11} = (\|e_1\|_{M_r} \pm \delta)^2$.

Observación 10 La escogencia entre $\|e_1\|_{M_r} + \delta$ y $\|e_1\|_{M_r} - \delta$ se hace aleatoriamente; δ puede tomarse como un “pequeño” porcentaje de $[M_0]_{11}$. Es entonces poco probable encontrar, para una iteración, el caso $\|e_1\|_{M_r} = \delta$ tal que $[M_{r+1}]_{11} = (\|e_1\|_{M_r} - \delta)^2 = 0$. Véase además que \mathcal{C}_θ y S_v son fijados por los cosenos $\cos_{M_0}(e_j, e_k)$ y la descomposición espectral de M_0 ; se tiene entonces $M_0 \in \mathcal{C}_\theta \cap S_v$ por la escogencia de M_0 .

Proposición 17 Si M_{r+1} es definido por el generador proporcional dado por la definición 31, entonces $M_{r+1} = (1 \pm \frac{\delta}{\|e_1\|_{M_r}})^2 M_r$.

DEMOSTRACIÓN:

Por la definición 31, tenemos inmediatamente:

$$[M_{r+1}]_{11} = \left(1 \pm \frac{\delta}{\|e_1\|_{M_r}}\right)^2 [M_r]_{11}. \quad (3.3)$$

Calculemos ahora los otros elementos de la matriz de M_{r+1} . Por las igualdades 3.1 y 3.3, para todo $j \in \{2, \dots, p\}$ se tiene:

$$\begin{aligned} [M_{r+1}]_{jj} &= \frac{[M_0]_{jj}}{[M_0]_{11}} [M_{r+1}]_{11} \\ &= \left(1 \pm \frac{\delta}{\|e_1\|_{M_r}}\right)^2 \frac{[M_r]_{jj}}{[M_r]_{11}} [M_r]_{11} \\ &= \left(1 \pm \frac{\delta}{\|e_1\|_{M_r}}\right)^2 [M_r]_{jj}. \end{aligned}$$

Por otra parte, según la igualdad 3.2, para todo $j, k \in \{1, \dots, p\}$ $j \neq k$ se tiene:

$$\begin{aligned} [M_{r+1}]_{jk} &= \sqrt{[M_{r+1}]_{jj}[M_{r+1}]_{kk} \cos \alpha_{jk}} \\ &= \left(1 \pm \frac{\delta}{\|e_1\|_{M_r}}\right)^2 \sqrt{[M_r]_{jj}[M_r]_{kk} \cos \alpha_{jk}} \\ &= \left(1 \pm \frac{\delta}{\|e_1\|_{M_r}}\right)^2 [M_r]_{jk}. \quad \blacksquare \end{aligned}$$

De esta proposición se deduce inmediatamente:

Corolario 18

1. M_{r+1} es una semi-métrica.
2. Si u es un vector propio de M_r asociado a λ , entonces también es vector propio de M_{r+1} , asociado al valor propio $\left(1 \pm \frac{\delta}{\|e_1\|_{M_r}}\right)^2 \lambda$.

Proposición 19 Para todo $r \in \mathbb{N}$, existe $m_r \in \mathbb{Z}$ tal que $\|e_1\|_{M_r} = \|e_1\|_{M_0} + m_r \delta$. Además, se tiene: $M_r = \left(1 + \frac{m_r \delta}{\|e_1\|_{M_0}}\right)^2 M_0$.

DEMOSTRACIÓN:

Por inducción sobre r :

- Sea $r = 1$. Se define $m_r = \pm 1$.

Es claro que por la definición 31 se tiene $\|e_1\|_{M_1} = \|e_1\|_{M_0} \pm \delta = \|e_1\|_{M_0} + m_1 \delta$.

Por la proposición 17 se tiene:

$$M_1 = (1 \pm \delta / \|e_1\|_{M_0})^2 M_0 = (1 + m_1 \delta / \|e_1\|_{M_0})^2 M_0$$

- Supóngase que el resultado es cierto para r .

Existe entonces $m_r \in \mathbb{Z}$ tal que:

$$\|e_1\|_{M_{r+1}} = \|e_1\|_{M_r} \pm \delta = \|e_1\|_{M_0} + m_r \delta \pm \delta.$$

Sea $m_{r+1} = m_r \pm 1$. Se tiene entonces que $\|e_1\|_{M_{r+1}} = \|e_1\|_{M_0} + m_{r+1} \delta$.

Por otra parte,

$$\begin{aligned}
M_{r+1} &= \left(1 \pm \frac{\delta}{\|e_1\|_{M_r}}\right)^2 M_r = \left(1 \pm \frac{\delta}{\|e_1\|_{M_0} + m_r \delta}\right)^2 \left(1 + \frac{m_r \delta}{\|e_1\|_{M_0}}\right)^2 M_0 \\
&= \left(\frac{\|e_1\|_{M_0} + m_r \delta \pm \delta}{\|e_1\|_{M_0} + m_r \delta}\right)^2 \left(\frac{\|e_1\|_{M_0} + m_r \delta}{\|e_1\|_{M_0}}\right)^2 M_0 = \left(1 + \frac{(m_r \pm 1)\delta}{\|e_1\|_{M_0}}\right)^2 M_0 \\
&= \left(1 + \frac{m_{r+1}\delta}{\|e_1\|_{M_0}}\right)^2 M_0 \quad \blacksquare
\end{aligned}$$

Corolario 20

1. $M_r = \frac{[M_r]_{11}}{[M_0]_{11}} M_0.$

2. Además, si u es un vector propio de M_0 asociado al valor propio λ , entonces para todo r , u es vector propio de M_r asociado al valor propio $\frac{[M_r]_{11}}{[M_0]_{11}} \lambda.$

Observación 11 De acuerdo con el corolario anterior, en caso de usar el generador proporcional propuesto para hacer un ACP, entonces se conservarán las mismas componentes principales, pero modificándose la forma de la nube de puntos por un factor de $[M_r]_{11}/[M_0]_{11}.$ De allí el poco interés en la utilización de este tipo de generador en el ACP. Además, el generador propuesto no convergería debido a que la inercia de la nube de puntos no sería acotada. En efecto, la inercia de la nube $\mathcal{N}_r = (X, M_r, D)$ sería:

$$\begin{aligned}
I(\mathcal{N}_r) &= \text{traza}(VM_r) \\
&= \frac{[M_r]_{11}}{[M_0]_{11}} \text{traza}(VM_0) \\
&= \frac{\|e_1\|_{M_0} + m_r \delta}{[M_0]_{11}} \text{traza}(VM_0),
\end{aligned}$$

donde $m_r \in \mathbb{Z}.$

3.4 Una extensión del generador espectral

En esta sección, abordamos los trabajos de S. Labrèche sobre los generadores espectrales, en que se definen conjuntos infinitos de métricas. Para toda selección de la métrica dentro de uno de estos conjuntos, los coeficientes de correlaciones lineales entre las variables estudiadas y las componentes principales, son invariantes, siendo fija la métrica en el espacio de variables. De esta propiedad, una proposición para la selección de elementos interesantes será formulada.

3.4.1 Definición

El generador de métricas definido enseguida es de alguna manera la extensión del generador espectral propuesto por el Profesor Y. Schektman y desarrollado por su equipo de investigación. En particular, estas ideas se deben principalmente a S. Labrèche [14] y nosotros hemos completado algunas demostraciones.

Definición 32 *El generador espectral está definido por:*

$$M(\gamma) = \sum_{j=1}^p \gamma(j) u_j u_j^t$$

donde $\{u_j ; j = 1, \dots, p\}$ es el conjunto de vectores propios ortonormados de la matriz V de varianzas-covarianzas de las variables $\{x^j ; j = 1, \dots, p\}$ (rango $V = p$).

Observación 12 *Obsérvese que este generador es un caso particular del generador espectral definido en la sección 3.2.2.*

Definición 33 *Sea M una métrica en el espacio de individuos E . Podemos definir $\{(\lambda_j, c_j, C^j) ; j = 1, \dots, s\}$ el conjunto de elementos propios del A.C.P. del triplete (X, M, D) , donde $s = \text{rango} X$; c_j es el vector propio asociado al valor propio λ_j de VM . C^j es la componente principal respectiva.*

Los valores propios de VM están ordenados en forma decreciente ($\lambda_1 \geq \dots \lambda_j \geq \dots \geq \lambda_s$). El conjunto de vectores propios puede completarse por una base $\{c_j ; j = s + 1, \dots, p\}$ del núcleo de VM tal que el conjunto $\mathcal{B} = \{c_j ; j = 1, \dots, p\}$ sea una base M -ortonormada de E .

Para toda métrica M de E , un conjunto (\mathcal{M}_M) de métricas es asociado, estas métricas son definidas por:

$$\mathcal{M}_M = \{M(\gamma), \gamma \in (\mathbb{R}^p)^+\}$$

donde:

- $M(\gamma) = M C D_\gamma C^t M$,
- C es la matriz $p \times p$ cuya j -ésima columna es el vector c_j de la base \mathcal{B} ,

- la j -ésima componente $\gamma(j)$ del vector γ es estrictamente positiva,
- D_γ es una matriz diagonal: $D_\gamma(j, j) = \gamma(j)$.

Se puede decir que las $\gamma(j)$ ponderan la influencia de cada componente principal C^j en el cálculo de las distancias entre los puntos del espacio de individuos. En efecto, si consideramos dos puntos $x_i, x_{i'}$ asociados a dos individuos, tenemos:

$$\begin{aligned} \|x_i - x_{i'}\|_{M(\gamma)}^2 &= [\sum_{j=1}^s (C^j(i) - C^j(i'))(c_j)^t] M(\gamma) [\sum_{j=1}^s (C^j(i) - C^j(i'))c_j] \\ &= \sum_{j=1}^s \gamma(j) [C^j(i) - C^j(i')]^2. \end{aligned}$$

Denotando I_p la métrica usualmente asociada a la matriz identidad de orden p , tenemos la propiedad:

Proposición 21

- a) $M \in \mathcal{M}_M$.
- b) Si $M', M'' \in \mathcal{M}_M$ entonces $M' + M'' \in \mathcal{M}_M$.
- c) Si $M', M'' \in \mathcal{M}_{I_p}$ entonces $M' \cdot M'' \in \mathcal{M}_{I_p}$.
- d) Si $M', M'' \in \mathcal{M}_{I_p}$ entonces $M' \cdot M'' = M'' \cdot M'$.
- e) Si $N \in \mathcal{M}_{I_p}$ entonces $N^{-1} \in \mathcal{M}_{I_p}$.
- f) Si $s = p$ entonces $V \in \mathcal{M}_{I_p}$.

DEMOSTRACIÓN:

- a) Para demostrar a) es suficiente con encontrar D_γ tal que $MCD_\gamma C^t M = M$. Multiplicando $MCD_\gamma C^t M$ a la izquierda por C^t y a la derecha por C , obtenemos:

$$\begin{aligned} MCD_\gamma C^t M &= M \\ \Leftrightarrow C^t MCD_\gamma C^t M C &= C^t M C \\ \Leftrightarrow D_\gamma &= I_p \end{aligned}$$

ya que la matriz C es M -ortogonal. Por lo tanto, basta tomar $D_\gamma = I_p$.

b) Sean $M' = MCD_{\gamma'}C^tM$ y $M'' = MCD_{\gamma''}C^tM$, entonces:

$$\begin{aligned} M' + M'' &= MCD_{\gamma'}C^tM + MCD_{\gamma''}C^tM \\ &= MC(D_{\gamma'} + D_{\gamma''})C^tM \\ &= MCD_{\gamma'+\gamma''}C^tM \end{aligned}$$

Luego, $M' + M'' \in \mathcal{M}_M$.

c) Sean $M' = CD_{\gamma'}C^t$ y $M'' = CD_{\gamma''}C^t$, entonces:

$$\begin{aligned} M' \cdot M'' &= CD_{\gamma'}C^t \cdot CD_{\gamma''}C^t \\ &= C(D_{\gamma'}D_{\gamma''})C^t \\ &= CD_{\gamma' \cdot \gamma''}C^t \end{aligned}$$

ya que en este caso C es I_p ortogonal. Luego, $M' \cdot M'' \in \mathcal{M}_M$.

d) Sean $M' = CD_{\gamma'}C^t$ y $M'' = CD_{\gamma''}C^t$, entonces:

$$\begin{aligned} M' \cdot M'' &= CD_{\gamma' \cdot \gamma''}C^t \\ &= CD_{\gamma'' \cdot \gamma'}C^t \\ &= M'' \cdot M'. \end{aligned}$$

e) Sea $N = CD_{\gamma}C^t$, es claro que $N^{-1} = C(D_{\gamma})^{-1}C^t = CD_{1/\gamma}C^t$.

f) Si $M = I_p$ y si $s = p$, entonces V puede descomponerse de la siguiente manera:

$$V = \sum_{j=1}^p \lambda_j c_j c_j^t = CD_{\gamma}C^t, \quad \gamma(j) = \lambda_j$$

Luego V es un elemento de \mathcal{M}_{I_p} . ■

3.4.2 ACP del triplete $(X, M(\gamma), D)$

Los elementos principales del A.C.P. del triplete $(X, M(\gamma), D)$ se deducen de los del A.C.P. del triplete (X, M, D) después de algunas operaciones elementales. Los ejes principales son invariantes y los coeficientes de correlación lineal entre las componentes principales y las variables x^j son independientes del vector γ .

Definición 34 Definimos $\{(\lambda_j^\gamma, c_j^\gamma, C_\gamma^j) ; j = 1, s\}$ como el conjunto de elementos principales del A.C.P del triplete $(X, M(\gamma), D)$.

Podemos enunciar la siguiente propiedad:

Proposición 22

1. $\forall j : c_j^\gamma = [\gamma(j)]^{-\frac{1}{2}} c_j$ es vector propio de $VM(\gamma)$ asociado al valor propio $\lambda_j^\gamma = \gamma(j)\lambda_j$.
2. $\forall j : C_\gamma^j = [\gamma(j)]^{\frac{1}{2}} C^j$.
3. $\forall j, k : \rho(x^k, C_\gamma^j) = \rho(x^k, C^j)$.

DEMOSTRACIÓN:

1. La afirmación 1) se puede verificar de la siguiente serie de igualdades:

$$\begin{aligned}
 VM(\gamma)c_j^\gamma &= VMCD_\gamma C^t M[\gamma(j)]^{-\frac{1}{2}} c_j \\
 &= [\gamma(j)]^{-\frac{1}{2}} VMCD_\gamma e_j \\
 &= [\gamma(j)]^{-\frac{1}{2}} VM C \gamma(j) e_j \\
 &= [\gamma(j)]^{-\frac{1}{2}} \gamma(j) VM c_j \\
 &= [\gamma(j)]^{-\frac{1}{2}} \gamma(j) \lambda_j c_j \\
 &= \gamma(j) \lambda_j [\gamma(j)]^{-\frac{1}{2}} c_j \\
 &= \gamma(j) \lambda_j c_j^\gamma.
 \end{aligned}$$

2. Utilizando la definición de las componentes principales para demostrar 2), obtenemos:

$$\begin{aligned}
 C_\gamma^j &= XM(\gamma)c_j^\gamma = [\gamma(j)]^{-\frac{1}{2}} XMCD_\gamma C^t M c_j \\
 &= [\gamma(j)]^{-\frac{1}{2}} \gamma(j) XMC e_j = [\gamma(j)]^{\frac{1}{2}} XMC e_j = [\gamma(j)]^{\frac{1}{2}} C^j.
 \end{aligned}$$

3. Por último de la definición del coeficiente de correlación, se deduce:

$$\begin{aligned}
 \rho(x^k, C_\gamma^j) &= \frac{\text{cov}(x^k, C_\gamma^j)}{\sqrt{\text{var}(x^k)\text{var}(C_\gamma^j)}} = \frac{(x^k)^t DC_\gamma^j}{\|x^k\|_D \|C_\gamma^j\|_D} \\
 &= \frac{[\gamma(j)]^{\frac{1}{2}} (x^k)^t DC^j}{[\gamma(j)]^{\frac{1}{2}} \|x^k\|_D \|C^j\|_D} = \frac{(x^k)^t DC^j}{\|x^k\|_D \|C^j\|_D} = \rho(x^k, C^j). \quad \blacksquare
 \end{aligned}$$

Observación 13

1. El conjunto $\{\lambda_j^\gamma ; j = 1, p\}$ de valores propios no es necesariamente ordenado.
2. Según el valor de $\gamma(j)$, la nube de puntos es “comprimida” o “dilatada” a lo largo del eje principal Δ_j generado por el vector principal c_j .

Con base en la proposición anterior y utilizando nuestros resultados numéricos del capítulo siguiente descubrimos la siguiente propiedad:

Proposición 23

1. Si $\bar{M}_1 = V$, $\bar{M}_2 = I$ y si se usan los generadores definidos en (a) o (c), las componentes principales restantes se pueden calcular por: $C_\alpha^k = \sqrt{[(1 - \alpha)\lambda_k + \alpha]} C_1^k$.
2. Si $\bar{M}_1 = I$, $\bar{M}_2 = V^{-1}$, si se usan los generadores definidos en (a) o (c) y si $r < m$, entonces las componentes principales restantes se pueden calcular por: $C_\alpha^k = \sqrt{[(1 - \alpha) + \frac{\alpha}{\lambda_k}]} C_1^k$.

DEMOSTRACIÓN: De la proposición 11 se tiene que los valores propios de la matriz son, para el caso 1 con los generadores primarios dados en (a): $\gamma_k = (1 - \alpha)\lambda_k + \alpha$, luego por la proposición anterior se tiene que:

$$C_\alpha^k = \sqrt{\gamma_k} C_1^k = \sqrt{[(1 - \alpha)\lambda_k + \alpha]} C_1^k$$

La demostración de los generadores primarios dados en (c) así como la demostración del caso 2, son análogas. ■

Otra propiedad interesante con respecto a los ejes principales se puede enunciar aquí; la demostración de esta propiedad se puede encontrar en [14].

Proposición 24 Para todo vector A de componentes no nulas de $[\mathbb{R}^p]^+$, existe al menos un elemento $M(\gamma)$ de \mathcal{M}_M tal que la j -ésima componente de A es un valor propio del A.C.P. del triplete $(X, M(\gamma), D)$. Además el eje principal asociado puede ser escogido en el conjunto $\{\Delta_j ; j = 1, \dots, s\}$ generados por los vectores principales c_j .

Esta propiedad da bastante libertad en A.C.P. Sin embargo, esta libertad puede no tener interés, porque se pueden valorizar dos ejes principales que expliquen porcentajes de inercia tan elevados como se desee, sin tener necesariamente un buen resumen sobre las variables (correlaciones débiles entre los componentes principales asociadas y las variables a explicar).

Observación 14 *Una propiedad equivalente fue demostrada en el capítulo 1 de la demostración de esta propiedad se deduce que:*

Para todo vector A de $[\mathbb{R}^p]^+$ ($\text{rang}(V) = p$), existe un conjunto de métricas M_γ , tal que las componentes de A sean los elementos principales del A.C.P del triplete (X, M_γ, D) . Las métricas están dadas por la fórmula siguiente:

$$M_\gamma = V^{-\frac{1}{2}} O D_\gamma O^t V^{-\frac{1}{2}}$$

donde O es una matriz ortogonal ($OO^t = O^tO = I$).

Corolario 25 $M_\gamma = \mathcal{M}_{V^{-1}}$.

DEMOSTRACIÓN: Tenemos

$$\begin{aligned} M \in \mathcal{M}_{V^{-1}} &\Leftrightarrow M = V^{-1} C D_\gamma C^t V^{-1} \\ &\Leftrightarrow M = V^{-\frac{1}{2}} V^{-\frac{1}{2}} C D_\gamma C^t V^{-\frac{1}{2}} V^{-\frac{1}{2}} \\ &\Leftrightarrow M = V^{-\frac{1}{2}} B D_\gamma B^t V^{-\frac{1}{2}} \\ &\Leftrightarrow M \in M_\gamma. \end{aligned}$$

Tomando $B = V^{-\frac{1}{2}} C$, luego tenemos que probar que B es una matriz ortogonal, para esto considere:

$$\begin{aligned} B^t B &= C^t (V^{-\frac{1}{2}})^t V^{-\frac{1}{2}} C \\ &= C^t V^{-\frac{1}{2}} V^{-\frac{1}{2}} C \\ &= C^t V^{-1} C \\ &= I. \end{aligned}$$

Lo anterior porque la matriz C es M -ortogonal y en este caso $M = V^{-1}$. ■

Chapter 4

Resultados Numéricos

4.1 Introducción

En el presente capítulo presentaremos los resultados de aplicar un generador primario de métricas al Análisis en Componentes Principales.

El generador primario implementado fue el descrito en el capítulo 3, que tiene la forma:

$$M_\alpha = (1 - \alpha)\bar{M}_1 + \alpha\bar{M}_2$$

donde, $\bar{M}_1, \bar{M}_2 \in \{I_p, V, V^{-1}, D_{1/\sigma^2}\}$, $\alpha = \frac{r}{m}$, $r \in \{1, \dots, m\}$ y m es el número de métricas a generar escogido por el usuario.

4.2 Generación primaria para A.C.P.

En vista de que la escogencia de la métrica en A.C.P. es un problema abierto, se pensó aplicar los generadores primarios definidos y estudiados en el capítulo 3 para determinar una métrica útil al usar esta técnica. Los generadores primarios permitieron explorar el uso de combinaciones de métricas clásicas, dando alguna luz sobre este asunto.

Se aplicaron los generadores sobre diversas tablas de datos. Aquí presentamos los resultados obtenidos sobre las siguientes tablas: las notas escolares, los peces de Amiard y los gastos del Estado de Francia.

Entre las otras tablas analizadas, estuvieron la sociomatriz de Thomas y una tabla de preferencias.

4.2.1 Descripción del programa

El lenguaje usado fue Pascal y el compilador fue Turbo Pascal versión 7. El programa se implementó en una computadora AT 486-DX 33 MHz.

Los datos fueron almacenados, en la gran mayoría, en tipos `array`, ya sea en matrices o vectores.

Los principales procedimientos fueron:

- **Jacobi**, que es el que se encarga de diagonalizar una matriz simétrica, encontrando los valores propios y los vectores propios respectivos. Este procedimiento fue tomado de [21], y es indispensable en la solución de cualquier A.C.P. ya que permite encontrar los elementos propios del mismo.
- **Inversa**, como su nombre lo dice este procedimiento encuentra la inversa de una matriz. Se utiliza cuando el usuario desea efectuar el A.C.P con la métrica V^{-1} o una combinación lineal de ésta con otra métrica.
- **CalculaMalaMedio**, cuya función principal es la de encontrar la raíz cuadrada de una matriz.
- **DefineMetrica**, lo que hace es definir la métrica con la cual se va a realizar el A.C.P.; por supuesto, esta métrica es escogida por el usuario.
- **GeneraMetricas**, como su nombre lo dice se encarga de generar las métricas primarias escogidas por el usuario.
- **EscDatos**, procedimiento que escribe los datos obtenidos en un archivo. Este procedimiento fue implementado para que dicho archivo sea generado en \LaTeX .

4.2.2 Resultados sobre las notas escolares

Se analizó, en primer lugar, una tabla de datos que muestra las notas obtenidas por 10 estudiantes en 5 materias en una escala de 0 a 10. Estos datos se muestran en la tabla 4.1

Estudiante	Matemáticas	Ciencias	Español	Historia	Ed. Física
Lucía	7.0	6.5	9.2	8.6	8.0
Pedro	7.5	9.4	7.3	7.0	7.0
Carmen	7.6	9.2	8.0	8.0	7.5
Luis	5.0	6.5	6.5	7.0	9.0
Andrés	6.0	6.0	7.8	8.9	7.3
Ana	7.8	9.6	7.7	8.0	6.5
Carlos	6.3	6.4	8.2	9.0	7.2
José	7.9	9.7	7.5	8.0	6.0
Sonia	6.0	6.0	6.5	5.5	8.7
María	6.8	7.2	8.7	9.0	7.0

Table 4.1: Tabla de datos de las notas escolares

Se aplicaron generadores primarios de la forma $M_\alpha = (1-\alpha)\bar{M}_1 + \alpha\bar{M}_2$, para los siguientes casos:

- a) $\bar{M}_1 = V$ y $\bar{M}_2 = I_5$
- b) $\bar{M}_1 = I_5$ y $\bar{M}_2 = V^{-1}$
- c) $\bar{M}_1 = I_5$ y $\bar{M}_2 = D_{1/\sigma^2}$

De esta forma, se generan métricas de V a V^{-1} y de V a D_{1/σ^2} , pasando por I_5 y una serie de posiciones intermedias.

En los ejemplos presentados aquí, usamos un número de $m = 13$ métricas en los generadores, desde $r = 0$ hasta $r = 12$.

A continuación presentamos los resultados obtenidos. Estos resultados son comentados, globalmente, al final de esta sección.

Puede observarse que para algunos generadores se presenta un único círculo de correlaciones, debido a que gracias a la proposición 11 las correlaciones entre las variables y las componentes principales son las mismas, si se usan esos generadores.

Generación $V \rightarrow I_5$

En la tabla 4.2 se muestra un resumen conteniendo los valores propios y los porcentajes de inercia explicada para 5 de las 13 métricas generadas.

Así mismo, la tabla 4.3 contiene las componentes principales correspondientes y la tabla 4.4 las correlaciones entre las componentes principales y las variables originales.

$i =$	$\alpha = 0/12(M = V)$		$\alpha = 3/12$		$\alpha = 6/12$		$\alpha = 9/12$		$\alpha = 12/12(M = I)$	
	λ_i	% inercia	λ_i	% inercia	λ_i	% inercia	λ_i	% inercia	λ_i	% inercia
1	12.2382	78.6983	10.0532	76.7806	7.8682	73.9769	5.6833	69.4884	3.4983	61.1442
2	3.2163	20.6828	2.8606	21.8477	2.5049	23.5507	2.1491	26.2772	1.7934	31.3458
3	0.0733	0.4716	0.1227	0.9372	0.1721	1.6179	0.2214	2.7076	0.2708	4.7334
4	0.0228	0.1469	0.0549	0.4195	0.0870	0.8180	0.1191	1.4559	0.1512	2.6418
5	0.0001	0.0004	0.0020	0.0151	0.0039	0.0365	0.0058	0.0709	0.0077	0.1348

Table 4.2: Los valores propios y porcentajes de inercia explicada, para la tabla de Notas, generador de V a I .

Individuo	$\alpha = 0/12$		$\alpha = 3/12$		$\alpha = 6/12$		$\alpha = 9/12$		$\alpha = 12/12$	
	C^1	C^2	C^1	C^2	C^1	C^2	C^1	C^2	C^1	C^2
x_1	-1.4303	2.1183	-1.2964	1.9977	-1.1469	1.8694	-0.9747	1.7315	-0.7647	1.5818
x_2	3.1214	-1.8641	2.8291	-1.7580	2.5028	-1.6451	2.1271	-1.5238	1.6689	-1.3920
x_3	2.9519	-0.4011	2.6754	-0.3783	2.3669	-0.3540	2.0116	-0.3279	1.5782	-0.2995
x_4	-4.8761	-1.7680	-4.4194	-1.6674	-3.9098	-1.5602	-3.3229	-1.4452	-2.6070	-1.3202
x_5	-2.6911	1.7887	-2.4390	1.6869	-2.1578	1.5785	-1.8338	1.4621	-1.4388	1.3357
x_6	4.3915	-0.5197	3.9802	-0.4901	3.5212	-0.4586	2.9926	-0.4248	2.3479	-0.3881
x_7	-1.6716	2.0341	-1.5151	1.9183	-1.3403	1.7951	-1.1391	1.6627	-0.8937	1.5189
x_8	4.9562	-0.5698	4.4920	-0.5373	3.9740	-0.5028	3.3775	-0.4658	2.6498	-0.4255
x_9	-4.9183	-2.9240	-4.4577	-2.7575	-3.9436	-2.5804	-3.3516	-2.3901	-2.6296	-2.1834
x_{10}	0.1664	2.1056	0.1508	1.9857	0.1334	1.8582	0.1134	1.7212	0.0890	1.5723

Table 4.3: Las componentes principales, para la tabla de Notas, generador de V a I .

Variable	$\alpha = 0/12$	
	$r(C^1, x^j)$	$r(C^2, x^j)$
x^1	0.9438	0.0448
x^2	0.9468	-0.2952
x^3	0.2561	0.8776
x^4	0.2489	0.9414
x^5	-0.8556	-0.3049

Table 4.4: Correlaciones entre variables y componentes principales, para la tabla de Notas, generador de V a I .

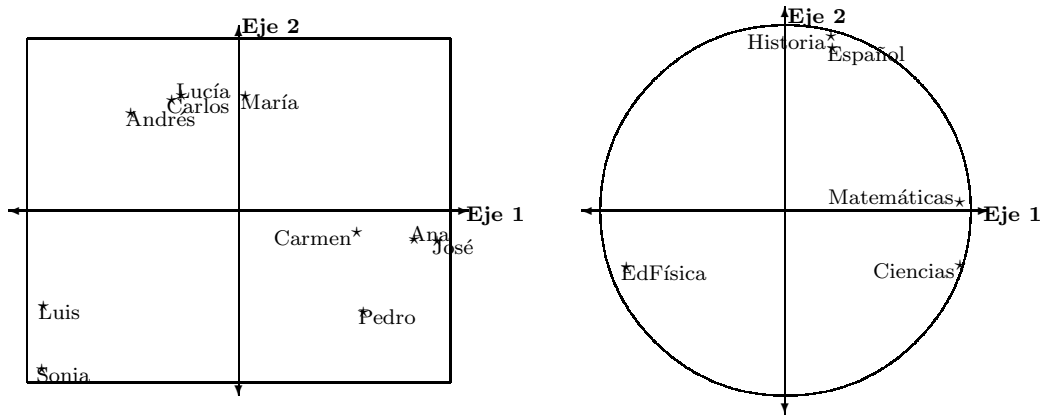


Figure 4.1: Plano principal y círculo de correlaciones para la tabla de Notas cuando $\alpha = 0$, generador de V a I .

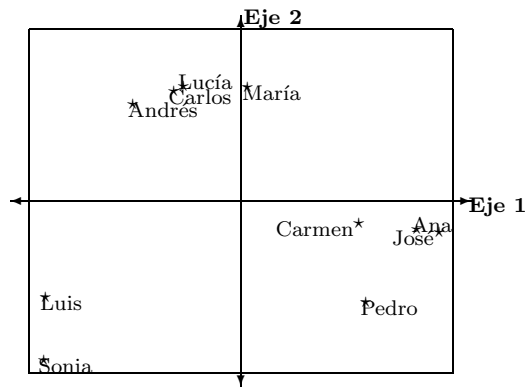


Figure 4.2: Plano principal de la tabla de Notas cuando $\alpha = 6$, generador de V a I .

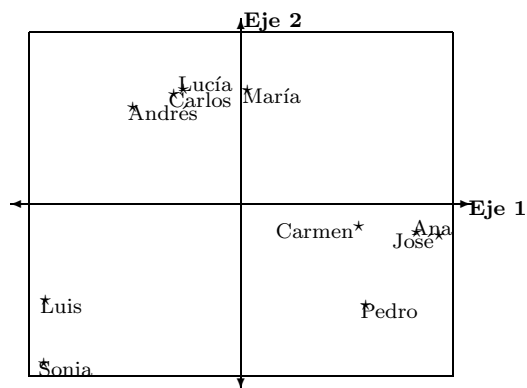


Figure 4.3: Plano principal de la tabla de Notas cuando $\alpha = 12$, generador de V a I .

Generación $I_5 \rightarrow V^{-1}$

En la tabla 4.5 se muestra un resumen conteniendo los valores propios y los porcentajes de inercia explicada para 5 de las 13 métricas generadas.

Así mismo, la tabla 4.6 contiene las componentes principales correspondientes y la tabla 4.7 las correlaciones entre las componentes principales y las variables originales.

$i =$	$\alpha = 3/12$		$\alpha = 6/12$		$\alpha = 9/12$		$\alpha = 11/12$		$\alpha = 12/12(M = V^{-1})$	
	λ_i	% inercia	λ_i	% inercia	λ_i	% inercia	λ_i	% inercia	λ_i	% inercia
1	2.8737	51.8626	2.2492	41.9564	1.6246	31.3604	1.2082	23.8768	1.0000	20.0001
2	1.5951	28.7863	1.3967	26.0546	1.1984	23.1327	1.0661	21.0690	1.0000	20.0000
3	0.4531	8.1774	0.6354	11.8531	0.8177	15.7847	0.9392	18.5615	1.0000	20.0000
4	0.3634	6.5577	0.5756	10.7369	0.7878	15.2072	0.9293	18.3645	1.0000	20.0000
5	0.2558	4.6161	0.5039	9.3990	0.7519	14.5150	0.9173	18.1282	1.0000	20.0000

Table 4.5: Los valores propios y porcentajes de inercia explicada, para la tabla de Notas, generador de I a V^{-1} .

Individuo	$\alpha = 3/12$		$\alpha = 6/12$		$\alpha = 9/12$		$\alpha = 11/12$		$\alpha = 12/12(M = V^{-1})$	
	C^1	C^2	C^1	C^2	C^1	C^2	C^1	C^2	C^1	C^2
x_1	-0.6931	1.4917	-0.6132	1.3959	-0.5211	1.2930	-0.4494	1.2196	1.0797	0.6323
x_2	1.5126	-1.3127	1.3381	-1.2284	1.1373	-1.1378	0.9807	-1.0732	-0.8918	0.7435
x_3	1.4304	-0.2824	1.2655	-0.2643	1.0755	-0.2448	0.9275	-0.2309	0.4227	-2.0030
x_4	-2.3629	-1.2451	-2.0904	-1.1651	-1.7766	-1.0792	-1.5321	-1.0179	-2.1524	-0.6222
x_5	-1.3040	1.2596	-1.1536	1.1787	-0.9805	1.0918	-0.8455	1.0298	0.9040	-0.6443
x_6	2.1280	-0.3660	1.8826	-0.3425	1.6000	-0.3172	1.3798	-0.2992	0.2863	-0.5645
x_7	-0.8100	1.4324	-0.7166	1.3404	-0.6090	1.2416	-0.5252	1.1711	0.7659	0.2943
x_8	2.4017	-0.4012	2.1247	-0.3755	1.8058	-0.3478	1.5573	-0.3281	0.4120	-0.2169
x_9	-2.3833	-2.0591	-2.1085	-1.9268	-1.7920	-1.7848	-1.5454	-1.6834	-1.1975	0.5023
x_{10}	0.0806	1.4828	0.0713	1.3875	0.0606	1.2852	0.0523	1.2122	0.3712	1.8784

Table 4.6: Las componentes principales, para la tabla de Notas, generador de I a V^{-1} .

Variable	$\alpha = i/12 \ i = 0, \dots, 11$		$\alpha = 12/12$	
	$r(C^1, x^j)$	$r(C^2, x^j)$	$r(C^1, x^j)$	$r(C^2, x^j)$
x^1	0.9438	0.0448	0.4869	-0.0745
x^2	0.9468	-0.2952	0.0714	-0.2841
x^3	0.2561	0.8776	0.8434	0.2693
x^4	0.2489	0.9414	0.8117	0.0559
x^5	-0.8556	-0.3049	-0.5882	-0.0448

Table 4.7: Correlaciones entre variables y componentes principales, para la tabla de Notas, generador de I a V^{-1} .

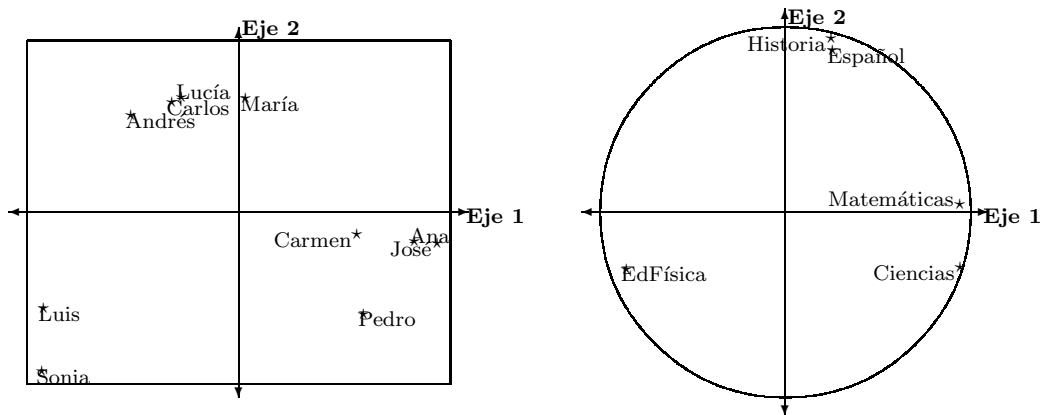


Figure 4.4: Plano principal y círculo de correlaciones de Notas cuando $\alpha = 0$, de I a V^{-1} .

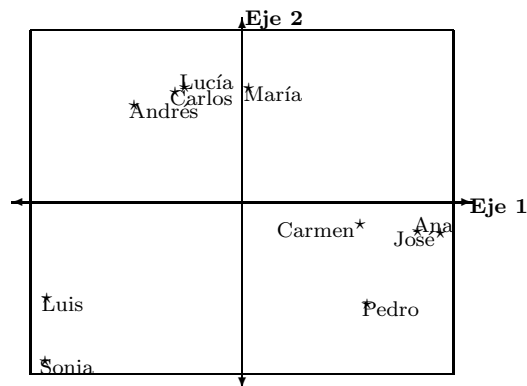


Figure 4.5: Plano principal de Notas con $\alpha = 6$, de I a V^{-1} .

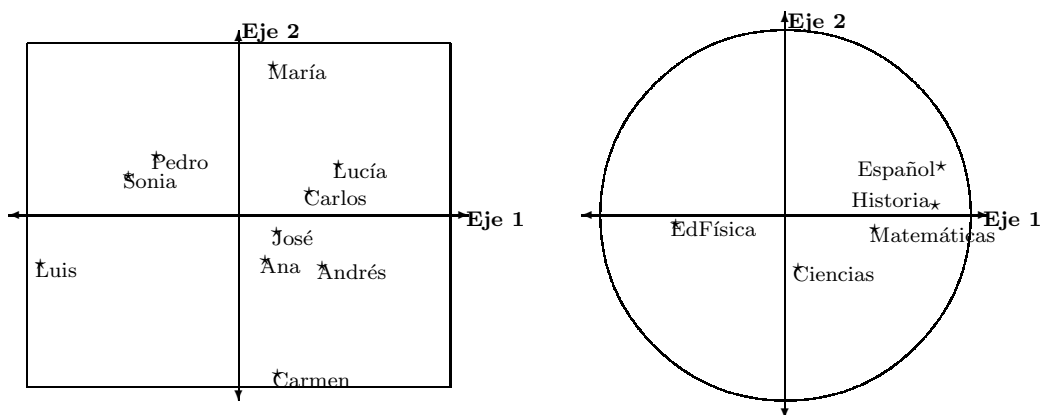


Figure 4.6: Plano principal y círculo de correlaciones de Notas con $\alpha = 12$, de I a V^{-1} .

Generación $I_5 \rightarrow D_{1/\sigma^2}$

En la tabla 4.8 se muestra un resumen conteniendo los valores propios y los porcentajes de inercia explicada para 5 de las 13 métricas generadas.

Así mismo, la tabla 4.9 contiene las componentes principales correspondientes y la tabla 4.10 las correlaciones entre las componentes principales y las variables originales.

$i =$	$\alpha = 3/12$		$\alpha = 6/12$		$\alpha = 9/12$		$\alpha = 12/12$	
	λ_i	% inercia	λ_i	% inercia	λ_i	% inercia	λ_i	% inercia
1	3.2965	59.4921	3.1227	58.2517	2.9862	57.6447	2.8932	57.8650
2	1.7955	32.4035	1.7737	33.0873	1.7197	33.1968	1.6287	32.5730
3	0.2912	5.2552	0.3107	5.7956	0.3292	6.3548	0.3466	6.9319
4	0.1497	2.7023	0.1451	2.7069	0.1365	2.6344	0.1226	2.4523
5	0.0081	0.1469	0.0085	0.1586	0.0088	0.1693	0.0089	0.1778

Table 4.8: Los valores propios y porcentajes de inercia explicada, para la tabla de Notas, generador de I a D_{1/σ^2} .

Individuo	$\alpha = 3/12$		$\alpha = 6/12$		$\alpha = 9/12$		$\alpha = 12/12$	
	C^1	C^2	C^1	C^2	C^1	C^2	C^1	C^2
x_1	-0.5521	1.6748	-0.2926	1.7486	0.0081	1.7855	0.3231	1.7725
x_2	1.4668	-1.4743	1.2261	-1.5567	0.9512	-1.6196	0.6654	-1.6387
x_3	1.4534	-0.3651	1.3135	-0.4344	1.1600	-0.4924	1.0025	-0.5157
x_4	-2.7469	-1.1704	-2.9020	-0.9394	-3.0522	-0.6265	-3.1721	-0.2628
x_5	-1.2377	1.3726	-1.0049	1.4028	-0.7474	1.4076	-0.4889	1.3654
x_6	2.2264	-0.5301	2.0801	-0.6971	1.9047	-0.8718	1.7086	-1.0217
x_7	-0.6932	1.5449	-0.4577	1.5550	-0.1955	1.5323	0.0676	1.4623
x_8	2.5385	-0.6061	2.3976	-0.8226	2.2197	-1.0579	2.0119	-1.2759
x_9	-2.7148	-2.0197	-2.8249	-1.8019	-2.9430	-1.5354	-3.0420	-1.2549
x_{10}	0.2596	1.5733	0.4647	1.5457	0.6945	1.4782	0.9239	1.3694

Table 4.9: Las componentes principales, para la tabla de Notas, generador de I a D_{1/σ^2} .

Var	$\alpha = 3/12$		$\alpha = 6/12$		$\alpha = 9/12$		$\alpha = 12/12$	
	$r(C^1, x^j)$	$r(C^2, x^j)$	$r(C^1, x^j)$	$r(C^2, x^j)$	$r(C^1, x^j)$	$r(C^2, x^j)$	$r(C^1, x^j)$	$r(C^2, x^j)$
x^1	0.948	-0.027	0.945	-0.121	0.928	-0.231	0.895	-0.345
x^2	0.917	-0.370	0.871	-0.459	0.805	-0.557	0.722	-0.648
x^3	0.327	0.861	0.415	0.830	0.514	0.781	0.610	0.717
x^4	0.320	0.915	0.407	0.876	0.505	0.819	0.599	0.748
x^5	-0.881	-0.228	-0.904	-0.128	-0.917	-0.008	-0.913	0.119

Table 4.10: Las correlaciones entre las variables y las componentes principales, para la tabla de Notas, generador de I a D_{1/σ^2} .

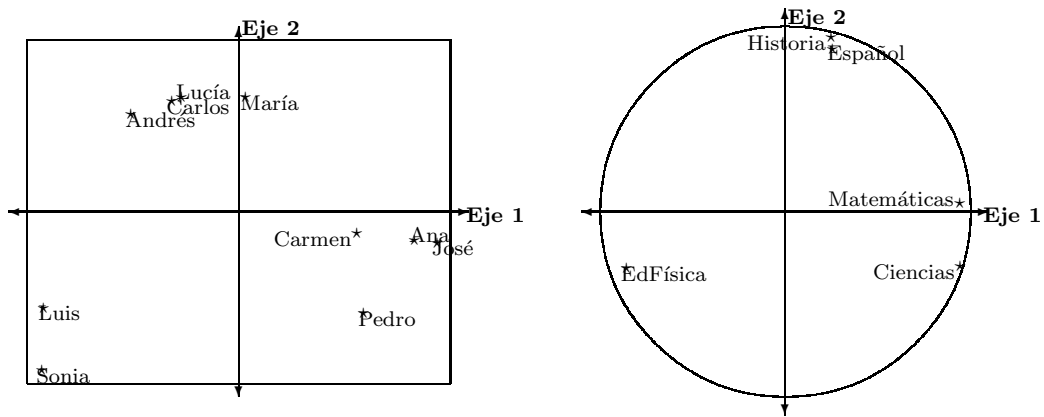


Figure 4.7: Plano principal y círculo de correlaciones de Notas cuando $\alpha = 0$, de I a D_{1/σ^2} .

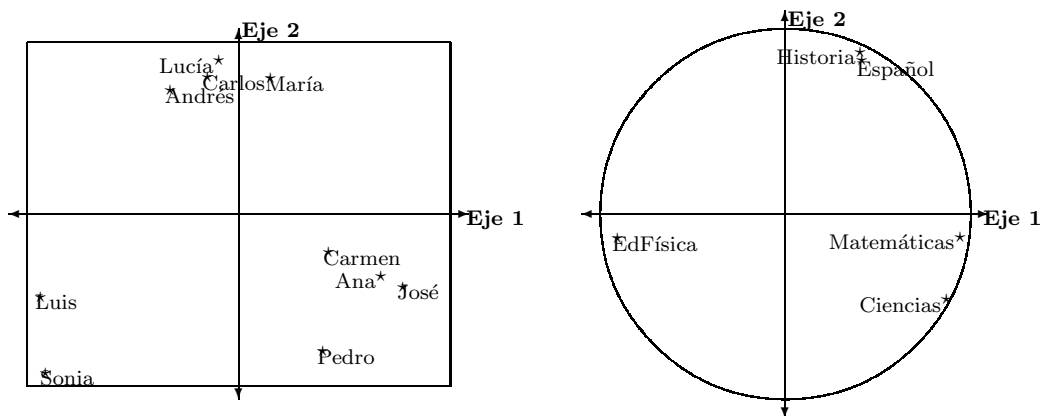


Figure 4.8: Plano principal y círculo de correlaciones de Notas cuando $\alpha = 6$, de I a D_{1/σ^2} .

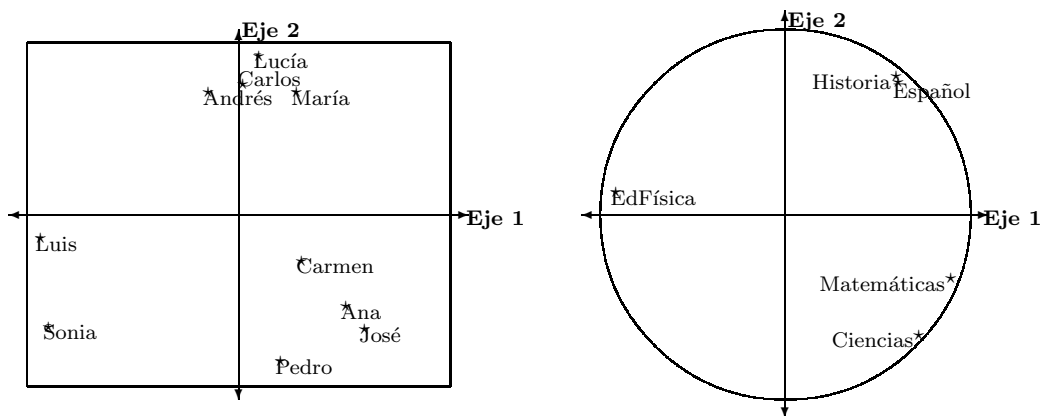


Figure 4.9: Plano principal y círculo de correlaciones de Notas cuando $\alpha = 12$, de I a D_{1/σ^2} .

Comentarios

Puede notarse que la generación de V a I_5 prácticamente los resultados son los mismos para todas las métricas de la familia generada. Esta estabilidad se sigue observando cuando se pasa de I_5 hacia V^{-1} , justo antes de llegar a esta última.

En general se observa lo siguiente:

- En el círculo de correlaciones, se nota que la primera componente principal hace una oposición entre la variable “Educación Física” y las materias científicas “Matemáticas” y “Ciencias”, mientras que la segunda componente principal sigue la tendencia de las humanísticas “Historia” y “Español”.
- En el plano principal se aprecian tres grupos de estudiantes: Luis y Sonia (los fuertes en Ed. Física); Pedro, Carmen, Ana y José (los buenos en materias científicas) y Andrés, Carlos, Lucía y María (los buenos en materias humanísticas).

Cuando $M = V^{-1}$ se observan algunos cambios importantes:

- En el círculo de correlaciones, se nota que deja de tener importancia la oposición “Educación Física vs Materias científicas”. Recuérdese que en este caso cualquier vector de \mathbb{R}^p es vector propio de VM ya que $VM = I_5$, y que la componente principal asociada tendrá varianza 1 y una inercia explicada de $1/5$.
- En el plano principal, se observa una “mezcla” de los individuos que anteriormente formaban grupos bien diferenciados.

De los resultados de la generación de I_5 a D_{1/σ^2} , se observa que:

- En el círculo de correlaciones las materias académicas hacen una rotación de unos 30 grados, quedando cuando $M = D_{1/\sigma^2}$ una clara oposición, para la primera componente principal, entre Educación Física y las materias académicas (ya no son sólo las científicas), y para la segunda componente principal, entre las materias científicas y las humanísticas.
- En el plano principal se mantienen los mismos tres grupos de individuos anteriormente interpretados.

Podemos concluir que los resultados mejor interpretables, a la luz de la tabla de datos, se obtienen para cuando $M = D_{1/\sigma^2}$, esto es, cuando se estandarizan las variables y se usa la métrica usual. Además se ve que la interpretación es difícil cuando $M = V^{-1}$.

Conclusiones generales y recomendaciones

El presente estudio sobre las métricas en Análisis de Datos nos ha permitido profundizar en el conocimiento de las propiedades que tienen las distancias y su influencia en los resultados de un análisis, que en nuestro caso fue el Análisis en Componentes Principales (A.C.P.). En efecto, los trabajos de M. Maurin, de Y. Schektman y de S. Labrèche dan propiedades importantes acerca de las métricas y muestran cómo los resultados de un A.C.P. pueden ser influenciados por un cambio de métrica.

El trabajo desarrollado nos permitió profundizar sobre todo en la generación primaria de métricas, campo en el que pudimos encontrar importantes propiedades que permiten simplificar los análisis. Por ejemplo, la propiedad 11 que establece que las correlaciones entre las componentes principales y las variables observadas, en algunos casos de generación primaria, son idénticas; y la propiedad 23 que establece que las componentes principales de un A.C.P. obtenidas con un generador primario, se pueden deducir mediante una fórmula sencilla a partir de las componentes principales de un A.C.P. con métrica identidad. Además, se encontraron otras propiedades interesantes acerca de la inercia de la nube de puntos (12 y su corolario) y los elementos principales del A.C.P. (9 y 10).

Una parte importante de este trabajo fue la implementación computacional, pues no sólo permitió probar las ideas desarrolladas respecto a los generadores primarios en el capítulo 3, sino que también permitió encontrar propiedades teóricas, como las proposiciones 11 y 23. En efecto, fue a la luz de los resultados obtenidos sobre diversas tablas de datos que estas propiedades fueron sugeridas, y posteriormente se estableció su validez general.

Los resultados obtenidos sobre las tablas de datos analizadas, muestran que la métrica

que ofrece ms calidad de interpretacin, tanto en el plano principal como en el c'rculo de correlaciones, es D_{1/σ^2} , que es la mtrica que nosotros sugerimos emplear despues de finalizar el presente trabajo de investigacin. Las mtricas primarias diferentes de D_{1/σ^2} dieron en todos los casos resultados ms pobres y ms dif'ciles de interpretar.

Se resolvi el problema de la generacin primaria de mtricas en A.C.P., cuando se parte de un conjunto de mtricas de la familia $\{V, V^{-1}, I, D_{1/\sigma^2}\}$. Sin embargo, creemos que esta l'nea de investigacin tiene an un gran campo, ya que puede estudiarse el uso de los generadores espectrales en A.C.P., o bien, en general, el uso de generadores de mtricas, como el propuesto, en algunas tcnicas de Anlisis de Datos donde tambien el problema de la escogencia de la distancia est abierto; entre estas ltimas tcnicas se encuentra, por ejemplo, el mtodo de Nubes Dinmicas para Clasificacin Automtica. As' mismo, la generacin de mtricas puede servir en el caso en que se tengan tablas de datos particionadas y que se empleen las mtricas relacionales propuestas por Y. Schektman.

Anexo: Programa

```
PROGRAM ACP;
USES CRT;

CONST
  NmaxInd = 28; {Numero maximo de Individuos}
  NmaxCar = 28; {Numero maximo de Variables}

TYPE
  Mat_Dat_T = ARRAY[1..NmaxInd+1,1..NmaxCar] OF SINGLE;
  Mat_C_n_T = ARRAY[1..NmaxInd,1..NmaxInd] OF SINGLE;
  Mat_C_p_T = ARRAY[1..NmaxCar,1..NmaxCar] OF SINGLE;
  Vec_np_T = ARRAY[1..NmaxInd] OF SINGLE;
  Vec_Er_T = ARRAY[1..NmaxInd] OF SINGLE;
  VectorE_T = ARRAY[1..NmaxInd] OF INTEGER;

VAR
  X,
  Mat_Comp, Metrica, M1,
  M_ala_medio,
  Mat_Vec,
  D, V : Mat_Dat_T ; {Matriz de Datos iniciales}
  Vec, vec1, vec2,
  vp1, vp2,
  vcirc1, vcirc2 : Vec_np_T ;
  {Vec es el vector de valores propios}
  {vec1 lo uso para llevar el % de inercia}
  {vec2 lo uso para llevar el % de inercia acumulada}
  {vp1 y vp2 son las componentes principales}
  {vcirc1 y vcirc2 son las correlaciones de las variables}
  itera : INTEGER;
  n, p, opcion, a1, a2, niter, h : BYTE;
  lambda : SINGLE;
  ArchE, ArchS : TEXT;
  letra1, letra2, letra3, namefile : STRING;

{Este procedimiento se encarga de leer la tabla de datos, que debe de tener
en las 2 primeras entradas la dimension de la matriz filas (n), columnas (p)}
```

```

PROCEDURE LeeDatos(VAR n,p: BYTE;
                  VAR Mat: Mat_Dat_T);
VAR
  i,j : BYTE;
BEGIN
  WRITELN('De el nombre del archivo donde esta la tabla de Datos.txt');
  READ(namefile);
  ASSIGN(ArchE,namefile+'.TXT');
  RESET(ArchE);
  READLN(ArchE,n,p);
  FOR i:=1 TO n DO BEGIN
    FOR j:=1 TO p-1 DO BEGIN
      READ(ArchE,Mat[i,j]);
    END;
    READLN(ArchE,Mat[i,p]);
  END;
  CLOSE(ArchE);
END;

{Este procedimiento escribe los resultados de los A.C.P. con la mtrica
escogida por el usuario, el archivo en el cual se escriben los datos se
genera en \Latex automticamente.}
PROCEDURE EscDatos( n,p : BYTE;
                  Mat : Mat_Dat_T;
                  V   : Mat_Dat_T;
                  ite : BYTE);

VAR
  j,i : BYTE;           {en pala2 guardo la metrica con la cual trabajo}
  pala,pala1,pala2 :STRING; {en pala guardo h y en pala1 guardo (niter-1)}
BEGIN
  IF ite<10 THEN
    STR(ite:1,pala)
  ELSE
    STR(ite:2,pala);
  IF niter<11 THEN
    STR((niter-1):1,pala1)
  ELSE
    STR((niter-1):2,pala1);
  ASSIGN(ArchS,namefile+pala+'.tex');
  REWRITE(ArchS);
  WRITELN(ArchS);
  IF opcion=5 THEN BEGIN
    pala2:='(1-' +pala+'/' +pala1+')'+letra1+ ' + ('+ pala+'/' +pala1+')'+letra2;
  END

```

```

ELSE BEGIN
    pala2:= letra3;
END;
WRITELN(Archs,'\begin{table}');
WRITELN(Archs,'\centering');
WRITELN(ArchS,'\begin{tabular}{|c|c|c|} ');
IF opcion=5 THEN BEGIN
    WRITE(ArchS,'\multicolumn{3}{c}{Datos de '+ namefile + ' obtenidos con la metrica:  '});
    WRITELN(ArchS,pala2 + ' } \\ \hline');
END
ELSE BEGIN
    WRITE(ArchS,'\multicolumn{3}{c}{Datos de ' + namefile + ' obtenidos con la metrica:  '});
    WRITELN(ArchS,pala2 + ' } \\ \hline');
END;
WRITE(ArchS,'Valor propio &');
WRITE(ArchS,' ':5);
WRITE(ArchS,'% de inercia &');
WRITE(ArchS,' ':3);
WRITELN(ArchS,'Inercia acumulada \\ \hline');
FOR j:=1 TO p DO BEGIN
    WRITE(ArchS,'$ ',vec[j]:7:4); WRITE(ArchS,' $ & $ ');
    WRITE(ArchS,vec1[j]:7:4); WRITE(ArchS,' $ & $ ');
    WRITE(ArchS,vec2[j]:7:4); WRITELN(ArchS,' $ \\ \hline ');
END;
WRITELN(ArchS,'\end{tabular}');
WRITELN(Archs,'\caption{Valores propios, \% inercia, e inercia acumulada}');
WRITE(ArchS,'%\label{' + namefile + 'ValProp'+pala2);
WRITELN(Archs,}')');
WRITELN(Archs,'\end{table}');
WRITELN(ArchS);
WRITELN(Archs,'\begin{table}');
WRITELN(Archs,'\centering');
WRITELN(ArchS,'\begin{tabular}{|c|c|c|} ');
IF opcion=5 THEN BEGIN
    WRITE(ArchS,'\multicolumn{3}{c}{Datos de '+ namefile + ' obtenidos con la metrica:  '});
    WRITELN(ArchS,pala2 + ' } \\ \hline ');
END
ELSE BEGIN
    WRITE(ArchS,'\multicolumn{3}{c}{Datos de '+ namefile+ ' obtenidos con la metrica:  '});
    WRITELN(ArchS,pala2 + ' } \\ \hline ');
END;
WRITELN(ArchS,'Individuo & $c_1$ & $c_2$ \\ \hline ');
FOR j:=1 TO n DO BEGIN
    WRITE(ArchS,'$x_{ '); WRITE(ArchS,j:1); WRITE(ArchS,}')$ & $');
    WRITE(ArchS,vp1[j]:7:4); WRITE(ArchS,'$ & $');

```

```

WRITE(ArchS, vp2[j]:7:4); Writeln(ArchS, ' $ \\ \hline ');
END;
Writeln(ArchS, '\end{tabular}');
Writeln(ArchS, '\caption{Las componentes principales}');
WRITE(ArchS, '%\label{'+namefile+'Compo}');
Writeln(ArchS, pala2 + ' ');
Writeln(ArchS, '\end{table}');
Writeln(ArchS);
Writeln(ArchS);
Writeln(ArchS);
Writeln(ArchS, '\begin{table}');
Writeln(ArchS, '\centering');
Writeln(ArchS, '\begin{tabular}{|c|c|c|} ');
IF opcion=5 THEN BEGIN
WRITE(ArchS, '\multicolumn{3}{c}{Datos de ' + namefile+ ' obtenidos con la metrica: ');
Writeln(ArchS, pala2 + ' } \\ \hline');
END
ELSE BEGIN
WRITE(ArchS, '\multicolumn{3}{c}{Datos de ' + namefile+ ' obtenidos con la metrica: ');
Writeln(ArchS, pala2 + ' } \\ \hline ');
END;
Writeln(ArchS, 'var & $r(c_1;x^j)$ & $r(c_2;x^j) $ \\ \hline');
FOR j:=1 TO p DO BEGIN
WRITE(ArchS, '$x^{j}$'); WRITE(ArchS, j:1); WRITE(ArchS, '$ & $');
WRITE(ArchS, vcirc1[j]:7:4); WRITE(ArchS, '$ & $');
WRITE(ArchS, vcirc2[j]:7:4); Writeln(ArchS, ' $ \\ \hline ');
END;
Writeln(ArchS, '\end{tabular}');
Writeln(ArchS, '\caption{Correlaciones entre $x^j$ y $c_i$}');
WRITE(ArchS, '%\label{'+namefile+'Corr}');
Writeln(ArchS, pala2 + ' ');
Writeln(ArchS, '\end{table}');
CLOSE(ArchS);
END;

{El siguiente procedimiento multiplica 2 matrices.}
PROCEDURE Multiplica2( A : Mat_Dat_T;
m,n : BYTE;
B : Mat_Dat_T;
p : BYTE;
Var C: Mat_Dat_T);
VAR
i,j,k : BYTE;
BEGIN
FOR i:=1 TO m DO BEGIN

```

```

FOR j:=1 TO p DO BEGIN
  C[i,j] := 0;
  FOR k:=1 TO n DO BEGIN
    C[i,j] := C[i,j] + A[i,k]*B[k,j];
  END;
END;
END;
END;

{Este procedimiento multiplica 3 matrices}
PROCEDURE Multiplica3( A   : Mat_Dat_T;
                      m,n : BYTE;
                      B   : Mat_Dat_T;
                      p   : BYTE;
                      C   : Mat_Dat_T;
                      q   : BYTE;
                      Var D: Mat_Dat_T);

VAR
  Temp : Mat_Dat_T;
BEGIN
  Multiplica2(A,m,n,B,p,Temp);
  Multiplica2(Temp,m,p,C,q,D);
END;

{Este procedimiento se encarga de definir la matriz de pesos D.}
PROCEDURE Inicia( n : BYTE;
                 VAR D: Mat_Dat_T);

VAR
  i,j : BYTE;
BEGIN
  FOR i:=1 TO n DO BEGIN
    FOR j:=1 TO n DO BEGIN
      D[i,j]:=0;
    END;
    D[i,i]:=1/n;
  END;
END;

{Como su nombre lo dice el siguiente procedimiento centra la matriz de Datos
$$ para realizar el A.C.P. con los datos centrados}
PROCEDURE Centra(n,p:BYTE;D:Mat_Dat_T; VAR Mat : Mat_Dat_T);

VAR
  i,j,i1: BYTE;
BEGIN
  FOR j:=1 TO p DO BEGIN      {Mat[n+1,..] es el centro de la nube}

```

```

Mat[n+1,j]:=D[1,1]*Mat[1,j];
FOR i:=2 TO n DO BEGIN
  Mat[n+1,j]:=Mat[n+1,j]+D[i,i]*Mat[i,j];
END;
FOR i:=1 TO n DO BEGIN
  Mat[i,j]:=Mat[i,j]-Mat[n+1,j];
END
END;
END;

```

{Este procedimiento lo que hace es calcular la matriz de varianzas y covarianzas de la tabla de datos}

```

PROCEDURE CalcMatVar(n,p : BYTE;
                    D : Mat_Dat_T;
                    X : Mat_Dat_T;
                    VAR V : Mat_Dat_T);
VAR
  i,j,k: BYTE;
BEGIN
  FOR j:=1 TO p DO BEGIN
    FOR k:=j TO p DO BEGIN
      V[j,k]:= D[1,1]*X[1,j]*X[1,k];
      FOR i:=2 TO n DO BEGIN
        V[j,k]:= V[j,k]+D[i,i]*X[i,j]*X[i,k]
      END;
      V[k,j]:=V[j,k];
    END;
  END;
END;

```

{El siguiente procedimiento calcula la inversa de una matriz.}

```

PROCEDURE Inversa(
  {+}      n : BYTE;
  {+}      B : Mat_Dat_T;
  {-} VAR  A : Mat_Dat_T);
VAR
  big,dum,pivin : REAL;
  i,icol,
  irow,
  j,k,
  l,ll : BYTE;
  indxc,
  indxr,
  ipiv : VectorE_T;
  d,aux : REAL;

```

```

BEGIN
  A := B;
  FOR j:=1 TO n DO BEGIN
    ipiv[j] := 0;
  END;
  FOR i:=1 TO n DO BEGIN
    big := 0;
    FOR j:= 1 TO n DO BEGIN
      IF (ipiv[j]<>1) THEN BEGIN
        FOR k:=1 TO n DO BEGIN
          IF (ipiv[k]=0) THEN BEGIN
            IF ABS(A[j,k]) >= big THEN BEGIN
              big := ABS(A[j,k]);
              irow := j;
              icol := k;
            END;
          END
        ELSE BEGIN
          IF ipiv[k] > 1 THEN BEGIN
            WRITELN(' PAUSE 1 EN GAUSSJMATRIX SINGULAR');
            READLN;
          END;
        END;
      END;
    END;
  END;
  ipiv[icol]:=ipiv[icol]+1;
  IF (irow<>icol) THEN BEGIN
    FOR l:=1 TO n DO BEGIN
      dum := A[irow,l];
      A[irow,l]:= A[icol,l];
      A[icol,l]:= dum;
    END;
  END;
  indxr[i]:=irow;
  indxc[i]:=icol;
  IF A[icol,icol] = 0.0 THEN BEGIN
    WRITELN(' PAUSE 2 EN GAUSSJ MATRIX SINGULAR');
    READLN;
  END;
  pivinv:=1/A[icol,icol];
  A[icol,icol]:=1;
  FOR l:=1 TO n DO BEGIN
    A[icol,l]:=A[icol,l]*pivinv;
  END;

```



```

FOR l1:=1 TO n DO BEGIN
  IF l1<>icol THEN BEGIN
    dum := A[l1,icol];
    A[l1,icol]:=0;
    FOR l:=1 TO n DO BEGIN
      A[l1,l] := A[l1,l]-A[icol,l]*dum;
    END;
  END;
END;
END;
FOR l:=n DOWNTO 1 DO BEGIN
  IF (indxr[l]<>indxc[l]) THEN BEGIN
    FOR k:=1 TO n DO BEGIN
      dum := A[k,indxr[l]];
      A[k,indxr[l]]:= A[k,indxc[l]];
      A[k,indxc[l]]:= dum;
    END;
  END;
END;
END;

```

{Este procedimiento se encarga de encontrar los valores propios de una matriz simetrica "a" y los vectores propios respectivos; los vectores propios son almacenados en la variable "v"(matriz) y los valores propios estan almacenados en la variable "d"(vector) }

{!!! Cuidado la matriz simetrica "a" la devuelve con ceros sobre la diagonal}

```

PROCEDURE Jacobi(Var a : Mat_Dat_T ; {p x p }
                n : BYTE;
                Var d : Vec_np_t ;
                Var v : Mat_Dat_T ; {p x p}
                Var nrot : INTEGER);

```

```

LABEL 99;
CONST nmax = 100;
VAR j,iq,ip,i: INTEGER;
    tresh, theta,tau,t,sm,s,h,g,c: REAL;
    b,z: ARRAY[1..nmax] OF REAL;
BEGIN
  FOR ip:= 1 TO n DO BEGIN
    FOR iq:= 1 TO n DO BEGIN
      v[ip,iq]:=0;
    END;
    v[ip,ip]:=1;
  END;
  FOR ip:= 1 TO n DO BEGIN

```

```

b[ip]:= a[ip,ip];
d[ip]:= b[ip];
z[ip]:= 0;
END;
nrot:=0;
FOR i:= 1 TO 50 DO BEGIN
  sm:= 0;
  FOR ip:= 1 TO n-1 DO BEGIN
    FOR iq:= ip+1 TO n DO BEGIN
      sm:=sm + abs(a[ip,iq]);
    END;
  END;
  IF sm = 0 THEN BEGIN
    GOTO 99;
  END;
  IF i<4 THEN
    tresh:=0.2*sm/sqr(n)
  ELSE
    tresh:=0;
  FOR ip:= 1 TO n-1 DO BEGIN
    FOR iq:= ip+1 TO n DO BEGIN
      g:= 100*abs(a[ip,iq]);
      IF ((i>4) AND ((abs(d[ip])+g)=abs(d[ip]))
        AND ((abs(d[iq])+g)=abs(d[iq]))) THEN
        a[ip,iq]:= 0
      ELSE BEGIN
        IF (abs(a[ip,iq]) > tresh) THEN BEGIN
          h:=d[iq]-d[ip];
          IF (abs(h)+g) = abs(h) THEN BEGIN
            t:=a[ip,iq]/h
          END
        ELSE BEGIN
          theta :=0.5*h/a[ip,iq];
          t:= 1/(abs(theta)+sqrt(1+sqr(theta)));
          IF theta < 0 THEN BEGIN
            t:=-1*t;
          END;
        END;
        c:=1/sqrt(1+sqr(t));
        s:=t*c;
        tau:=s/(1+c);
        h:=t*a[ip,iq];
        z[ip]:=z[ip]-h; z[iq]:=z[iq]+h;
        d[ip]:=d[ip]-h; d[iq]:=d[iq]+h;
        a[ip,iq]:=0;
      END;
    END;
  END;

```

```

FOR j:=1 TO ip-1 DO BEGIN
    g:=a[j,ip]; h:= a[j,iq];
    a[j,ip]:= g-s*(h+g*tau);
    a[j,iq]:= h+s*(g-h*tau);
END;
FOR j:=ip+1 TO iq-1 DO BEGIN
    g:=a[ip,j]; h:= a[j,iq];
    a[ip,j]:= g-s*(h+g*tau);
    a[j,iq]:= h+s*(g-h*tau);
END;
FOR j:=iq+1 TO n DO BEGIN
    g:=a[ip,j]; h:= a[iq,j];
    a[ip,j]:= g-s*(h+g*tau);
    a[iq,j]:= h+s*(g-h*tau);
END;
FOR j:=1 TO n DO BEGIN
    g:=v[j,ip]; h:= v[j,iq];
    v[j,ip]:= g-s*(h+g*tau);
    v[j,iq]:= h+s*(g-h*tau);
END;
nrot := nrot +1;
END;
END;{ELSE}
END;
END;
END;
FOR ip:=1 TO n DO BEGIN
    b[ip]:= b[ip]+ z[ip];
    d[ip]:= b[ip];
    z[ip]:=0;
END;
WRITELN('Pausa en el procedimiento Jacobi');
WRITELN('50 iteraciones no bastan'); readln;
99: END;

```

{Este procedimiento ordena los valores propios de la diagonalizacin de VM, con los porcentajes de inercia respectivos.}

```

PROCEDURE OrdEleProp(VAR d:Vec_np_T;
                    VAR v:Mat_Dat_T;
                    n:BYTE); {dimension del espacio "p"}
VAR
    i,j,k : BYTE;
    p      : SINGLE;
BEGIN
    FOR i:=1 TO n-1 DO BEGIN

```

```

k:=i;
p:=d[i];
FOR j:=i+1 TO n DO BEGIN
  IF d[j]>=p THEN BEGIN
    k:=j;
    p:=d[j];
  END;
END;
IF k<>i THEN BEGIN
  d[k]:=d[i];
  d[i]:=p;
  FOR j:=1 TO n DO BEGIN
    p:=v[j,i];
    v[j,i]:=v[j,k];
    v[j,k]:=p;
  END;
END;
END;
END;

```

{Este procedimiento calcula los porcentajes de inercia as como la inercia acumulada.}

```

PROCEDURE InerPorc(d          : Vec_np_T;
                  VAR vec_iner,vec_iner_acum: Vec_np_T;
                  n:BYTE);
  {d es el vector de los valores propios}

```

```

VAR
  inercia : SINGLE;
  i,j     : BYTE;
BEGIN
  inercia:=0;
  FOR i:=1 TO n DO BEGIN
    inercia:=inercia + d[i];
  END;
  FOR i:=1 TO n DO BEGIN
    vec_iner[i]:= d[i]*100/inercia;
  END;
  vec_iner_acum[1]:=vec_iner[1];
  FOR i:=2 TO n DO BEGIN
    vec_iner_acum[i]:=vec_iner_acum[i-1] + vec_iner[i];
  END;
END;

```

{Este procedimieto calcula para una matriz A una matriz B tal que $B*B = A$ }

```

PROCEDURE CalculaM_ala_medio(A      : Mat_Dat_T;

```

```

                VAR B   : Mat_Dat_T;
                z      : BYTE ) ;

VAR
  Ct,C,Diag : Mat_Dat_T;
  i,j,dim   : BYTE;
  vector    : Vec_np_T;
  nraiz     : INTEGER;
BEGIN
dim:=p;
Jacobi(A,dim,vector,C,nraiz);
FOR i:=1 TO dim DO BEGIN
  FOR j:=1 TO dim DO BEGIN
    Ct[i,j]:=C[j,i];
  END;
END;
FOR i:=1 TO dim DO BEGIN
  FOR j:=1 TO dim DO BEGIN
    Diag[i,j]:=0;
  END;
END;
FOR i:=1 TO dim DO BEGIN
  IF (z=0) THEN Diag[i,i]:=sqrt(vector[i])
  ELSE Diag[i,i]:=1/vector[i]
END;
Multiplica3(C,dim,dim,Diag,dim,Ct,dim,B);
END;

{Este procedimiento define la matriz  $M^{(1/2)}VM^{(1/2)}$ , que en el caso trivial
es la matriz de correlaciones R, ademas la diagonaliza, ordena los vectores
propios y define los porcentajes de inercia}
PROCEDURE Diagonalizar (A   : Mat_Dat_T;
                      dim : BYTE;
                      B   : Mat_Dat_T);

VAR
  C : Mat_Dat_T;
BEGIN
  Multiplica3(A,dim,dim,B,dim,A,p,C);
  Jacobi(C,dim,vec,Mat_vec,itera);
  OrdEleProp(Vec,Mat_vec,dim);
  InerPorc(Vec,vec1,vec2,dim);
END;

{Este procedimiento guarda las componentes principales en 2 variables c1 y c2}
PROCEDURE CompPrinc(A           :Mat_Dat_T;

```

```

                VAR c1,c2   :Vec_np_T;
                n           :BYTE);
VAR
    i: BYTE;
BEGIN
    FOR i:=1 TO n DO BEGIN
        c1[i]:=A[i,1];
        c2[i]:=A[i,2];
    END;
END;

{Este procedimiento da las correlaciones entre las variables y las
componentes principales.}
PROCEDURE Circ_corr(VAR vcirculo1,vcirculo2 : Vec_np_T);
VAR
    Xt,
    Mat_Circ_corr  : Mat_Dat_T;
    i,j : BYTE;
BEGIN
    FOR i:=1 TO p DO BEGIN
        FOR j:=1 TO n DO BEGIN
            Xt[i,j]:=X[j,i]/n;
        END;
    END;
    Multiplica2(Xt,p,n,Mat_Comp,p,Mat_Circ_corr);
    CompPrinc(Mat_Circ_corr,vcirculo1,vcirculo2,p);
    FOR i:=1 TO p DO BEGIN
        vcirculo1[i] :=vcirculo1[i]/SQRT(V[i,i]*Vec[1]);
        vcirculo2[i] :=vcirculo2[i]/SQRT(V[i,i]*Vec[2]);
    END;
END;

Procedure EscogeMet(VAR resp:BYTE);
BEGIN
WRITELN(' Seleccione la metrica con la cual desea que se efectue el A.C.P');
WRITELN('1 La matriz de varianzas. ');
WRITELN('2 La inversa de la matriz de varianzas. ');
WRITELN('3 La matriz identidad. ');
WRITELN('4 La matriz diagonal de las inversas de las varianzas de las variables');
WRITELN('5 Efectuar una generacion primaria con dos de las anteriores');
READLN(resp);
END;

{Este procedimiento define una mtrica seleccionada por el usuario.}
PROCEDURE DefMetrica(numero :BYTE);

```

```

                VAR A      :Mat_Dat_T;
                VAR letra:STRING);

VAR k,l : BYTE;
BEGIN
CASE numero OF
1: BEGIN
    letra:='$V$';
    FOR k:=1 TO p DO BEGIN
        FOR l:=1 TO p DO BEGIN
            A[k,l]:=V[k,l];
        END;
    END;
END;
2: BEGIN
    letra:='$V^{-1}$';
    Inversa(p,V,A);
END;
3: BEGIN
    letra:='$I_{p}$';
    FOR k:=1 TO p DO BEGIN
        FOR l:=1 TO p DO BEGIN
            A[k,l]:=0;
        END;
    END;
    FOR k:=1 TO p DO BEGIN
        A[k,k]:=1;
    END;
END;
4: BEGIN
    letra:='$D_{1/\sigma}$';
    FOR k:=1 TO p DO BEGIN
        FOR l:=1 TO p DO BEGIN
            A[k,l]:=0;
        END;
    END;
    FOR k:=1 TO p DO BEGIN
        A[k,k]:=1/V[k,k];
    END;
END;
END;  (*CASE*)
END;

```

{Este procedimiento define las 2 mtricas para hacer la generacin primaria.}

```

PROCEDURE DEFMETRICAS(a,b:BYTE);
BEGIN

```

```

DEFMETRICA(a,M1,letra1);
DEFMETRICA(b,D,letra2);
END;

{Este procedimiento se encarga de generar las metricas cuando el usuario
desea hacer una generacin primaria de mtricas. }
PROCEDURE GENERAMETRICA(x:BYTE);
VAR i,j:BYTE;
BEGIN
FOR i:=1 TO p DO BEGIN
  FOR j:=1 TO p DO BEGIN
    METRICA[i,j]:=(1-x*lambda)*M1[i,j] + x*lambda*D[i,j];
  END;
END;
END;
END;

{Este es el begin del programa principal.}
BEGIN {PRINCIPAL}
  clrscr;
  LeeDatos(n,p,X);
  Inicia(n,D);
  Centra(n,p,D,X);
  CalcMatVar(n,p,D,X,V);
  EscogeMet(opcion);
  IF opcion = 5 THEN BEGIN
    WRITELN(' Seleccione la metrica con la cual desea que se comience el A.C.P');
    WRITELN('1 La matriz de varianzas. ');
    WRITELN('2 La inversa de la matriz de varianzas. ');
    WRITELN('3 La matriz identidad. ');
    WRITELN('4 La matriz diagonal de las inversas de las varianzas de las variables');
    READLN(a1);
    WRITELN(' Seleccione la metrica con la cual desea que se termine el A.C.P');
    WRITELN('1 La matriz de varianzas. ');
    WRITELN('2 La inversa de la matriz de varianzas. ');
    WRITELN('3 La matriz identidad. ');
    WRITELN('4 La matriz diagonal de las inversas de las varianzas de las variables');
    READLN(a2);
    WRITELN('Cuantas iteraciones desea hacer');
    READLN(niter);
    DefMetricas(a1,a2);
    lambda:=1/(niter-1);
    FOR h:=0 TO (niter-1) DO BEGIN
      GENERAMETRICA(h); { CalculaLambda( {1M1+(1-l)M2 --> Metrica} }
      CalculaM_ala_medio(Metrica,M_ala_medio,0);
      Diagonalizar(M_ala_medio,p,V);
    END;
  END;
END;

```



```
    Multiplica3(X,n,p,M_ala_medio,p,Mat_vec,p,Mat_Comp);
    CompPrinc(Mat_Comp,vp1,vp2,n); {Para escribir las componentes en vp1,vp2}
    Circ_Corr(vcirc1,vcirc2);
    EscDatos(n,p,X,V,h);
END;
END
ELSE BEGIN
    DefMetrica(opcion,Metrica,letra3);
    CalculaM_ala_medio(Metrica,M_ala_medio,0);
    Diagonalizar(M_ala_medio,p,V);
    Multiplica3(X,n,p,M_ala_medio,p,Mat_vec,p,Mat_Comp);
    CompPrinc(Mat_Comp,vp1,vp2,n); {Para escribir las componentes en vp1,vp2}
    Circ_Corr(vcirc1,vcirc2);
    EscDatos(n,p,X,V,1);
END;
END.
```

Bibliography

- [1] Abdesselam, R.; Schektman, Y. (1989) "Dissymmetrical association analysis between two qualitative variables," *Data Analysis, Learning Symbolic and Numeric Knowledge*, E. Diday (ed.), INRIA - Nova Science, New York, pp. 39-46.
- [2] Benzécri, J.P. (1988) "Sur la stabilité des sous-espaces principaux d'inertie sous un changement de métrique", *Les Cahiers de l'Analyse des Données*. Vol XIII(2): 197-203.
- [3] Besse, P. (1988) "Spline functions and optimal metric in linear principal component analysis." *Component and Correspondence Analysis: Dimension Reduction by Functional Approximation* J.L.A. van Rijkevorsel and J. van Leeuw (eds.), John Wiley & Sons, Chichester, pp. 81-101.
- [4] Cadet, O. (1989) "A method for analysing multidimensional experimental data," *Data Analysis, Learning Symbolic and Numeric Knowledge* E. Diday (ed.), INRIA - Nova Science, New York, pp. 87-94.
- [5] Cailliez, F.; Pagès, J.P. (1976) *Introduction à l'Analyse de Données*. SMASH, Paris.
- [6] Diday, E. et al. (1982) *Eléments d'Analyse de Données*. Dunod, Paris.
- [7] Espinoza J.L. et al. (1988) *Clasificación Automática*. Tesis de licenciatura, Universidad de Costa Rica, San Pedro.
- [8] Fabre, C. (1986) *Contribution à la Protection des Méthodes Relationnelles*. Tesis de doctorado de tercer ciclo. Universidad Paul Sabatier, Toulouse.
- [9] Govaert, G. (1983) *Classification Automatique et Distances Adaptatives*. Tesis de doctorado de tercer ciclo. Universidad de París VI.

- [10] Guevara, R. (1980) *Tópicos de Análisis de Datos*. Tesis de licenciatura. Universidad de Costa Rica, San Pedro.
- [11] Hoffman, K.; Kunze, R. (1980) *Algebra Lineal*. Prentice Hall. México.
- [12] Jambu, M. (1989) *Exploration Informatique et Statistique des Données*. Dunod, París.
- [13] Kazmierczak, J.B. (1987) “Sur l’usage d’un principe d’invariance pour aider au choix d’une métrique.” *Statistique et Analyse des Données*, vol.12(3): 37-57.
- [14] Labrèche, S. (1995) “Selección de métricas en A.C.P. y en clasificación.” *IX Simposio Métodos Matemáticos Aplicados a las Ciencias*, J. Trejos (ed.), U.C.R.-I.T.C.R., Turrialba, pp 87-96.
- [15] Maurin, M. (1986) “Métriques et agrégations de métriques en ACP.” En: *Data Analysis and Informatics IV* E. Diday et al. (eds.), Elsevier Sc. Publ. (North-Holland), Amsterdam, pp. 415-420.
- [16] Maurin, M. (1987) “Sur les qualités des métriques en A.C.P.” En: *Statistique et Analyse des Données*, vol.12(3): 58-74.
- [17] Maurin, M. (1991) “Un paramétrage des métriques en ACP, applications et conséquences.” *XXIII Journées de Statistique*, ASU, Estrasburgo, pp. 358-360.
- [18] Maurin, M. (1992) “La recherche de métriques optimales en ACP.” *Distancia’92*, G. Le Calvé & S. Joly (eds.), Rennes.
- [19] Maurin, M. *La non unicité des solutions en ACP et son application au changement de métriques*. Metron, Vol XLVII, # 1-4.
- [20] Piza, E. (1989) “La métrica inducida por un Análisis en Componentes Principales.” *Ciencia y Tecnología*, Vol XIII(2): 167-179.
- [21] Press, W. et al (1986) *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press. Cambridge.
- [22] Saporta, G. (1978) *Théories et Méthodes de la Statistique*. Technip, París.

- [23] Saporta, G. (1990) *Probabilités, Analyse des données et Statistique*. Technip, Paris.
- [24] Schektman, Y. (1983) *Quelques réflexions sur les métriques. Une approche originale pour la protection des méthodes relationnelles et l'étude de la régression multidimensionnelle*. Nota interna, Laboratorio MLAD, Universidad Paul Sabatier, Toulouse.
- [25] Schektman, Y. (1984) "Métriques et analyse des données, contributions et synthèses," *II Colloque Franco-Soviétique de Statistique*, CNET, Paris.
- [26] Trejos, J. (1993) *Contribution à l'Acquisition Automatique de Connaissances à Partir de Données Qualitatives*. Tesis de doctorado, Universidad Paul Sabatier, Toulouse.
- [27] Trejos, J. (1994) *Génération de produits scalaires pour l'Analyse des Données*. Nota interna, PIMAD, Escuela de Matemática, Universidad de Costa Rica. 31pp.
- [28] Trejos, J.; Bolaños, F. (1995) "Generación de métricas útiles en análisis de datos," *IX Simposio Métodos Matemáticos Aplicados a las Ciencias*, J. Trejos (ed.), U.C.R.-I.T.C.R., Turrialba, pp 173-182.
- [29] Troupé, M. (1994) *Contribution à la Protection de la Régression Multiple Multidimensionnelle et à la Génération des Règles Prévisionnelles*. Tesis de doctorado, Universidad Paul Sabatier, Toulouse.