UNIVERSIDAD DE COSTA RICA

SISTEMA DE ESTUDIOS DE POSGRADO

MÉTODOS DE REDUCCIÓN DE LA DIMENSIONALIDAD PARA VARIABLES SIMBÓLICAS DE TIPO INTERVALO

Tesis sometida a la consideración de la Comisión del Programa de Estudios de Posgrado en Matemática para optar al grado y título de Maestría Académica en Matemática, con énfasis en Matemática Aplicada.

JORGE ANDRÉS ARCE GARRO

Ciudad Universitaria Rodrigo Facio, Costa Rica

2018

Dedicatoria

A mi madre y hermana que han estado presentes en todo momento, a mi esposa por apoyarme en cada una de mis locuras y en especial a mi primo Kenneth (Q.d.D.g) que me enseñó a luchar hasta el último momento independientemente de que tan grande sea el problema.

Agradecimientos

Quiero agradecer al Prof. Oldemar Rodríguez por la ayuda que me brindó como director de esta Tesis y además por su valiosa ayuda para la implementación de las funciones en el paquete **RSDA** escrito en **R**.

A los profesores Javier Trejos y Álvaro Guevara por el tiempo dedicado a la lectura de mi tesis y por sus valiosas sugerencias.

"Esta tesis fue aceptada por la Comisión del Programa de Estudios de Posgrado en Matemática de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Académica en Matemática, con énfasis en Matemática Aplicada".

Doctor Santiago Cambronero Villalobos.
Representante del Decano
Sistema de Estudios de Posgrado

Doctor Oldemar Rodríguez Rojas.
Director de Tesis

Doctor Javier Trejos Zelaya.
Asesor

Doctor Álvaro Guevara Villalobos.
Asesor

Doctor Luis Barboza Chinchilla.
Representante del Director del Posgrado
Posgrado de Matemática

Jorge Andrés Arce Garro.
Candidato

Tabla de contenidos

Resumen

En [Cazes, P.; Chouakria, A.; Diday, E.; Schektman, Y. (1997)] se propone el método de centros para extender el conocido método de análisis de componentes principales a variables simbólicas de tipo intervalo. En este trabajo, los autores proponen utilizar el centro del hipercubo como punto base para llevar a cabo el análisis de componentes principales y luego proyectar todos los vértices de los hipercubos como individuos suplementarios del ACP de centros.

En esta investigación, se muestra que si desea maximizar la varianza de las proyecciones o minimizar las distancias entre los vértices y sus respectivas proyecciones, no necesariamente el centro del hipercubo es el mejor punto realizar el ACP. Se propone utilizar un algoritmo de optimización que maximice la varianza de las proyecciones (o que minimice las distancias al cuadrado de los vértices y sus respectivas proyecciones) que encuentre ese punto óptimo para realizar el ACP. Además se propone el algoritmo para generazalizar las curvas principales a variables de tipo intervalo. Todos los métodos que se han propuesto en esta tesis se pueden ejecutar en el paquete **RSDA**.

Lista de tablas

Lista de figuras

Abreviaturas

MSPE	Mean Squared Prediction Error
ACP	Análisis de Componentes Principales
MOD	Matriz Óptima respecto a la Distancia
MOI	Matriz Óptima respecto a la Inercia

Capítulo 1

Presentación

1.1. Introducción

El foco de esta tesis es generar nuevos modelos de reducción de la dimensionalidad, entre ellos una teoría para el análisis de superficies principales para variables simbólicas, específicamente para variables de tipo intervalo. La teoría de las curvas principales para datos clásicos fue desarrollada por Trevor Hastie en [Hastie, T. (1984)], para el caso de datos simbólicos no se ha realizado ningún trabajo relacionado a generalizar el trabajo de Hastie. Las superficies principales buscan la generalización del análisis de componentes principales, mediante un modelo no lineal, permitiendo una mejor representación (más cercana) del conjunto de datos. El análisis de datos simbólicos fue propuesto por Edwin Diday en [Diday, E. (1987)]. La diferencia principal entre un análisis de datos clásico y un análisis de datos simbólico es que en el caso clásico se toma una tabla de datos donde todas sus columnas son números reales, para los casos de variables cualitativas se construyen tablas de contigencia y matrices de Burt que son la transformación de una variable cualitativa en una tabla de datos con elementos en los números reales, mientras que en el análisis simbólico se utilizan conceptos, para comprender un concepto se presenta el siguiente ejemplo, se considera una tabla con datos de estudiantes de la UCR, cada individuo de una carrera se encuentra descrito por un conjunto de variables numéricas o nominales. Cada individuo (estudiante) es considerado individuo de primer orden. Para poder estudiar las carreras, consideradas

individuos de segundo orden, se puede describir como un resumen de los valores tomados de sus estudiantes, como intervalos, subconjuntos de valores, histogramas, distribuciones de probabilidad, entre otros, dependiendo de la variable que se desea utilizar. De esta manera, se obtiene una tabla de datos simbólicos donde cada fila contiene una descripción de la carrera y cada columna está asociada a una variable simbólica. Algunos tipos de variables simbólicas propuestos en [Billard. L.; Diday. E. (2006)] son: intervalos, conjuntos, distribuciones de probabilidad, entre otros. Otro modelo que se generalizará es el método de centros para el análisis de componentes principales para variables de tipo intervalo, la generalización consiste en buscar un conjunto de puntos que minimize la distancia de los vértices o máximice la varianza explicada en los primeros componentes principales.

1.2. Objetivo general

Crear modelos de reducción de la dimensionalidad para variables simbólicos de tipo intervalo.

1.3. Objetivos específicos

- Generalizar el modelo de superficies principales de datos clásicos a variables simbólicas de tipo intervalo.
- Crear el algoritmo de superficies principales para una matriz de tipo intervalo.
- Estudiar la convergencia del algoritmo anterior.
- Generalizar el método de centros para el análisis de componentes principales para variables de tipo intervalo.
- Implementar en el paquete **RSDA** creado en **R**, la función que calcule la superficie principal para variables de tipo intervalo y el ACP que generalice el método de centros.

- Realizar comparaciones entre las superficies principales y el análisis de componentes principales para variables de tipo intervalo, utilizando datos reales.

Capítulo 2

Marco Teórico

En este capítulo se hará un repaso de los conceptos de básicos de estadística con el fin de crear la definición de autoconsistencia, que es la base para la construcción de las curvas principales propuestas por [Hastie, T. (1984)]. Además se realiza una introducción a los conceptos de datos simbólicos y la creación del análisis de componentes principales para datos simbólicos de tipo intervalo.

2.1. La esperanza de una variable aleatoria

Si X es una variable aleatoria discreta, con valores positivos $\{x_1, x_2, \dots\}$ y p una medida de probabilidad, se define la esperanza o media de X, denotada $E(X)$, por

$$E(X) = \sum_{i=1}^{\infty} x_i p(x_i). \tag{2.1}$$

Por la descomposición de Hahn, se puede descomponer al conjunto $\{x_1, x_2, \dots\}$ en dos conjuntos A y B donde A consiste de todos los x_i no negativos y B de todos los x_i negativos. Si $\sum_{x_i \in A} x_i p(x_i) < \infty$ y $\sum_{x_i \in B} -x_i p(x_i) < \infty$, se define $E(X)$ sin ambigüedad por (2 2)

Esta nueva definición de (2.1) viene dada por (2.2), la cual es más general.

$$E(X) = \sum_{x_i \in A} x_i p(x_i) - \sum_{x_i \in B} x_i p(x_i). \tag{2.2}$$

Si $\sum_{x_i \in A} x_i p(x_i) = \infty$ o $\sum_{x_i \in B} -x_i p(x_i) = \infty$, $E(X)$ estaría indefinida.

Algunas propiedades de la esperanza de X son:

1. Si X es una constante, $\forall w, X(w) = c$, entonces

$$E(X) = c. \tag{2.3}$$

2. Si A es un conjunto, la función indicadora de A (χ_A) viene dada por

$$\chi_A(x) = \begin{cases} 1 & si\ x \in A \\ 0 & si\ x \notin A \end{cases},$$

entonces

$$E(\chi_A) = p(A). \tag{2.4}$$

3. Si X es un vector aleatorio $n - dimensional$, si g es una función real en $\mathbb{R}^n$ y si $E(|g(x)|) < \infty$, entonces se puede demostrar que (ver detalles en Bickel. P. J.; Doksum. K. A. (1977)])

$$E(g(X)) = \sum_{i=1}^{\infty} g(x_i) p(x_i). \tag{2.5}$$

4. Como consecuencia de este resultado, se tiene

$$E(|X|) = \sum_{i=1}^{\infty} |x_i| p(x_i). \tag{2.6}$$

5. Tomando $g(x) = \sum_{i=1}^{n} \alpha_i x_i$ se obtiene una relación fundamental

$$E(X) = \sum_{i=1}^{n} \alpha_i E(x_i); \tag{2.7}$$

si $\{\alpha_1, \ldots, \alpha_n\}$ son constantes y $E(|X_i|) < \infty, i = 1, \ldots, n$.

Si X es una variable aleatoria continua, es natural pensar en definir la esperanza por medio de una aproximación del caso discreto. Para realizar esta definición se utilizará integración de Lebesgue,

$$E(X) = \int_{-\infty}^{\infty} xp(x)dx, \tag{2.8}$$

el valor esperado o media de X está bien definida cuando $\int_0^{\infty} xp(x)dx$ y $\int_{-\infty}^{0} xp(x)dx$ son finitas. En otro caso $E(X)$ se encuentra indefinida.

Definición 2.1. *Una variable aleatoria X es llamada integrable si $E(|X|) < \infty$.*

Se puede probar que si X es un vector aleatorio continuo $k - dimensional$ y $g(X)$ es una variable aleatoria tal que

$$\int_{\mathbb{R}^n} |g(x)|p(x)dx < \infty,$$

entonces $E(g(X))$ existe y

$$E(g(X)) = \int_{\mathbb{R}^n} g(x)p(x)dx. \tag{2.9}$$

En el caso continuo las propiedades del valor esperado (2.3), (2.4), (2.5), (2.6) y (2.7) se encuentran bien definidas.

Para más detalles de las pruebas y más ejemplos se recomienda leer [Bickel. P. J.; Doksum. K. A. (1977)] Las definiciones y teoremas de esta subsección fueron tomados de [Bickel. P. J.; Doksum, K. A. (1977)].

2.1.1. Propiedades del valor esperado condicional

La distribución condicionada a un vector aleatorio Y dado $Z = z$ corresponde a una única medida de probabilidad p_z en $(\Omega, 2^{\Omega})$, específicamente, defina para, $A \in 2^{\Omega}$,

$$p_z(A) = p(A \mid Z = z) \text{ si } p_z > 0. \tag{2.10}$$

Ahora la distribución condicional de Y dado $Z = z$ es la misma que la distribución de Y si p_z es la medida de probabilidad en $(\Omega, 2^{\Omega})$. Por lo tanto, la esperanza condicionada es una esperanza ordinaria con respecto a la medida de probabilidad p_z. De esto se desprende que todas las propiedades del valor esperado (2 3), (2 4), (2.5), (2.6) y (2 7) se cumplen para la esperanza condicionada dado $Z = z$. Por lo tanto, para cualquier función real $r(Y)$ con $E(r(Y)) < \infty$,

$$E(r(Y) \mid Z = z) = \Sigma_y r(y) p(y \mid Z = z);$$

y

$$p(\alpha Y_1 + \beta Y_2 \mid Z = z) = \alpha p(Y_1 \mid Z = z) + \beta p(Y_2 \mid Z = z); \tag{2.11}$$

idénticamente en z para cualesquiera variables aleatorias Y_1, Y_2 tales que $E(|Y_1|), E(|Y_2|)$ son finitas. Debido a que la identidad es válida para toda z, se puede definir

$$p(\alpha Y_1 + \beta Y_2 \mid Z) = \alpha p(Y_1 \mid Z) + \beta p(Y_2 \mid Z). \tag{2.12}$$

Este proceso se puede repetir para cualquiera de (2 3), (2 4), (2 5), (2 6) y (2 7) para obtener propiedades análogas de la esperanza condicionada.

En dos casos especiales se puede calcular la esperanza condicionada inmediatamente. Si Y y Z son independientes y $E(|Y| < \infty)$, entonces

$$E(Y \mid Z) = E(Y). \tag{2.13}$$

Por otra parte para cualquier función real h con $E(|h(Y)|) < \infty$,

$$E(h(Z) \mid Z) = E(h(Z)). \tag{2.14}$$

La noción implícita en (2 14) dado $Z = z$, Z actúa como constante, utilizando esto, se tiene una relación que se llamará el teorema de sustitución para la esperanza condicional:

$$E(q(Y,Z)) \mid Z = z) = E(q(Y,z)) \mid Z = z). \tag{2.15}$$

Esto es válido para cualquier z tal que $p(z) > 0$ si $E(|q(Y,Z)|) < \infty$. Esto se deduce de las definiciones (2 10) que $\forall a$

$$\begin{aligned} E(q(Y,Z)) = a \mid Z = z) &= E(q(Y,Z)) = a, Z = z \mid Z = z) \\ &= E(q(Y,z)) = a \mid Z = z). \end{aligned} \tag{2.16}$$

Si se considera $q(Y,Z)) = r(Y)h(Z)$, donde $E(|r(Y)h(Z)|) < \infty$ se obtiene de (2 15) el siguiente resultado:

$$\begin{aligned} E(q(Y,Z)) \mid Z = z) &= E(r(Y)h(Z)) \mid Z = z) \\ &= h(z)E(r(Y)) \mid Z = z). \end{aligned} \tag{2.17}$$

Por lo tanto,

$$E(r(Y)h(Z)) \mid Z) = h(Z)E(r(Y)) \mid Z). \tag{2.18}$$

Otro resultado intuitivamente razonable es que la media de las medias condicionales es la media:

$$E(E(Y \mid Z)) = E(Y); \tag{2.19}$$

por lo tanto Y tiene esperanza finita, a esto se le llamará el teorema de doble esperanza.

Para más detalles de las pruebas y más ejemplos se recomienda leer [Bickel, P. J.; Doksum, K. A. (1977)]. Las definiciones y teoremas de esta subsección fueron tomados de [Bickel, P. J.; Doksum, K. A. (1977)].

2.2. Autoconsistencia de vectores aleatorios

Se quiere aproximar la distribución de un vector aleatorio X por medio de un vector aleatorio Y. Una forma de obtener qué tan bien explica Y a X es el error cuadrático medio (ECM) $E(\|X-Y\|^2)$. En términos del error cuadrático medio, la aproximación de X por Y puede ser mejorada por medio de $E(X \mid Y)$, para cualquier función g, $E(\|X-E(X \mid Y)\|^2) \leq E(\|X-g(Y)\|^2)$, tomado g igual a la identidad se tiene que $E(\|X-E(X \mid Y)\|^2) \leq E(\|X-Y\|^2)$. Entonces Y es óptimo local para aproximar X si $Y = E(X \mid Y)$, en este caso se llamará a Y autoconsistente para X.

Lema 2.1. *Si $E(Y)^2 < \infty$, entonces $\mu = E(Y)$, existe.*

La prueba de este lema se encuentra en [Bickel, P. J.; Doksum, K. A. (1977)].

Lema 2.2. *$E(Y)^2 < \infty$, sii $\forall c$, $E(Y-c)^2 < \infty$.*

Prueba.

$$\begin{aligned} E(Y-c)^2 &= E(Y^2 - 2Yc + c^2) \\ &= E(Y)^2 - 2E(Y)E(c) + E(c^2) \\ &= E(Y)^2 - 2cE(Y) + c^2. \end{aligned}$$

Si $E(Y)^2 < \infty, E(Y) < \infty$ implica $E(Y-c)^2 < \infty$. □

Lema 2.3. *Cuando $E(Y)^2 < \infty$ y c es una constante,*

$$E(Y-c)^2 = var(Y) + (c-\mu)^2.$$

Luego, $E(Y-c)^2 < \infty$ es minimizada de manera única por $c = \mu = E(Y)$.

La prueba de este lema se encuentra en [Bickel, P. J.; Doksum. K. A. (1977)].

Teorema 2.1. *Si Z es cualquier vector aleatorio y Y cualquier variable aleatoria, entonces para toda función g, $E(Y - g(Z))^2 = \infty$ o*

$$E(Y - \mu(Z))^2 \leq E(Y - g(Z))^2. \tag{2.20}$$

Prueba. Se supone que $E(Y - g(Z))^2 < \infty$. Se tiene que $E(Y - g(Z))^2 = E(Y - g(z) \mid z \in Z)^2$ por el teorema 2 15. Sea

$$\mu(Z) = E(Y \mid Z = z).$$

Como $g(z)$ es constante, por el lema 2 2

$$E((Y - g(z))^2 \mid Z = z) = E((Y - \mu(z))^2 \mid Z = z) + (g(z) - \mu(z))^2. \tag{2.21}$$

El mínimo de 2 21 se obtiene cuando $g(z) = \mu(z)$, por lo cual para toda función g, $E(Y - \mu(Z))^2 \leq E(Y - g(Z))^2$. □.

En el teorema 2.1, tomando la función $\forall z \in Z, g(z) = z$ se obtiene

$$E(Y - E(Y \mid Z))^2 \leq E(Y - Z)^2.$$

Definición 2.2. *Sean X y Y dos vectores aleatorios distribuidos de forma conjunta, Y es autoconsistente para X si $Y = E(X \mid Y)$ casi por doquier.*

Ejemplo 2.1. *Sea X_n que denota una secuencia de variables aleatorias independientes con media cero y sea $S_n = \sum_{i=1}^{n} X_i$. Entonces, para todo k,*

$$\begin{aligned} E(S_{n+k} \mid S_n) &= S_n + E(X_{n+1} + \cdots + X_{n+k} \mid S_n) \\ &= S_n + E(X_{n+1} + \cdots + X_{n+k}) \\ &= S_n. \end{aligned}$$

Para más detalles de las pruebas y más ejemplos se recomienda leer [Bickel, P. J.; Doksum, K. A. (1977)]. Las definiciones y teoremas de esta subsección fueron tomados de [Bickel, P. J.; Doksum, K. A. (1977)].

2.3. Análisis de componentes principales

Está subsección fue facilitada por el Doctor Oldemar Rodríguez, la cual forma parta de sus apuntes para el curso de introducción al análisis de datos (SP1346) de la UCR.

El Análisis de Componentes Principales (ACP) es una técnica proveniente del análisis exploratorio de datos cuyo objetivo es la síntesis de la información, o reducción de la dimensión (número de variables). Es decir, ante una tabla de datos con muchas variables, el objetivo será reducirlas a un menor número perdiendo la menor cantidad de información posible. El ACP es uno de los métodos más utilizados en Minería de Datos en países como Francia. Fue primeramente introducido por Pearson en 1901 y desarrollado independientemente en 1933 por Hotelling y la primera implementación computacional se dió en los años 60. Fue aplicado para analizar encuestas de opinión pública por Jean Pierre Pagès. Como ya se mencionó el objetivo es construir un pequeño número de nuevas variables (componentes), combinación lineal de las variables originales, en las cuales se concentre la mayor cantidad posible de información.

2.3.1. Los datos

X es una matriz de n filas (individuos) y m columnas (variables), el conjunto de matrices de n filas y m columnas se denotará $M_{n\times m}$, por lo que se puede indicar que $X \in M_{n\times m}$.

Se parte de una tabla de datos:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm} \end{pmatrix} \leftarrow \text{individuo } i \text{ ,}$$

que se puede transformar en la siguiente matriz de distancias:

$$D = \begin{pmatrix} d_{11} & \cdots & d_{1j} & \cdots & d_{1n} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ d_{i1} & \cdots & d_{ij} & \cdots & x_{in} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nj} & \cdots & d_{nn} \end{pmatrix} ,$$

dónde d_{ij} es la distancia del individuo i al individuo j.

El conjunto de individuos de una tabla de datos X se pueden ver como una nube de puntos en $\mathbb{R}^m$, como se ilustra en la Figura 2 1-a (esta figura fue facilitada por el profesor Dr. Javier Trejos, pertenece a sus láminas del curso SP1346 de la UCR), con su centro de gravedad localizado en el origen, y lo que se busca es un subespacio $q-$dimensional L de $\mathbb{R}^p$, usualmente un plano (ver Figura 2 1-b), tal que la proyección ortogonal de los n puntos sobre L (ver Figura 2 1-c) tienen inercia máxima, lo cual permitirá el estudio de relaciones, clases, etc. entre los individuos (filas) de la tabla de datos.

2.3.2. El problema

- Se trata de sintetizar los datos contenidos en una tabla de datos X en un conjunto más pequeño de nuevas variables $C^1, C^2, \ldots$ llamadas componentes principales, manteniendo la información escencial de X.

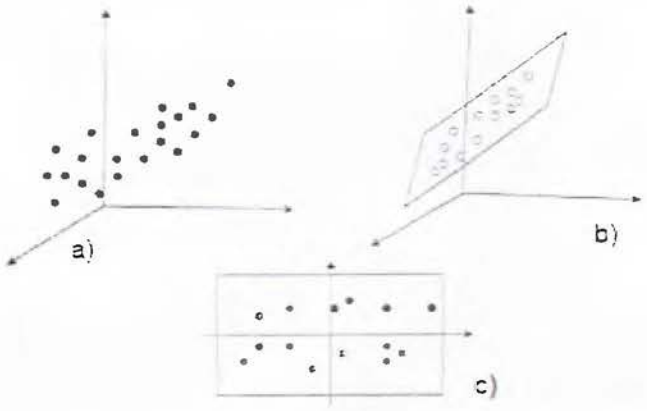


FIGURA 2.1: Proyección de los individuos en el plano de inercia **máxima.**

- Así, en la etapa 1 del algoritmo se encuentra una variable sintética C^1, la primera componente principal, la cual es combinación lineal de las variables originales X^j, es decir:

$$C^1 = a_{11}X^1 + \cdots + a_{1j}X^j + \cdots + a_{1m}X^m.$$

donde X^j es la columna j de X. Esto significa que el valor de C^1 para el individuo $i-$ésimo está dado por:

$$C_i^1 = a_{11}x_{i1} + \cdots + a_{1j}x_{ij} + \cdots + a_{1m}x_{im}.$$

- Generalmente esta primer componente principal, C^1, no es suficiente para condensar la información contenida en X, por lo que se construye una segunda componente principal C^2, luego una tercera C^3 y así sucesivamente.

- En general en la etapa k, se construye la componente principal $k-$ésima dada por:

$$C^k = a_{k1}X^1 + \cdots + a_{kj}X^j + \cdots + a_{km}X^m.$$

- Matricialmente se tiene que:

$$C^k = Xa^k,$$

donde:

$$a^k = \begin{pmatrix} a_{k1} \\ \vdots \\ a_{kj} \\ \vdots \\ a_{km} \end{pmatrix}.$$

- a^k se llama el **$k-$ésimo factor**.
- Los coeficientes a_{kj} constituyen un sistema de pesos para las variables, los cuales indican cuanto aporta cada variable a la construcción de la componente.
- Algunos coeficientes a_{kj} serán negativos y otros serán positivos. El valor de cada peso por sí solo no es importante, sino la relación con respecto a los otros pesos. Para evitar un problema de escalas se impone la siguiente restricción:

$$\sum_{j=1}^{m} (a_{kj})^2 = 1.$$

Estas nuevos componentes principales son calculados como una combinación lineal de las variables originales y además serán linealmente independientes. Un aspecto clave en ACP es la interpretación, ya que ésta no viene dada a priori, sino que será deducida tras observar la correlación de los componentes principales con las variables originales.

2.3.3. Cálculo de los factores y de las componentes principales

El ACP puede ser presentado tanto en el espacio de las variables como en el espacio de los individuos.

2.3.3.1. En el espacio de los individuos

- Se supondrá que las variables están centradas y reducidas.
- $V = \frac{1}{n}X^tX$ es la matriz de varianzas–covarianzas. Como las variables están centradas y reducidas entonces $V = R$, la matriz de correlaciones, pues:

$$v_{ij} = \text{cov}(X^i, X^j) = \frac{\text{cov}(X^i, X^j)}{\sigma_{X^i}\sigma_{X^j}} = R(X^i, X^j).$$

- Por lo tanto el espacio de las filas de X en $\mathbb{R}^m$ es el espacio de individuos cuyo origen será el centro de la nube de puntos.
- El objetivo del ACP es describir de manera sintética la nube de individuos.

Teorema 2.2. *En la etapa 1 de un ACP se calcula el eje D_1 que pasa por el origen para el cual la dispersión de la nube de puntos sea máxima, este eje D_1 pasa entonces lo más cerca posible de la nube de puntos, es decir, el promedio de las distancias al cuadrado de los n puntos de la nube y el eje D_1 es minimal.*

Sea a^1 es vector director normado (norma 1) del eje (recta) D_1 entonces*: a^1 es el vector propio asociado al valor propio más grande de la matriz de V de varianzas–covarianzas.*

Antes de probar el Teorema 2.2, se necesita primero el Lema 2.4 (el cual se va a asumir como válido, la demostración se encuentra en [Gallier. J.: Quaintance. J. (2018)])

Lema 2.4. *Sean A y B dos matrices cuadradas $m\times m$ simétricas y sea A una matriz definida positiva. Entonces el vector $y \in \mathbb{R}^n$ que resuelve el siguiente problema de optimización:*

$$\begin{cases} \text{máx}\, y^tBy \\ \textit{sujeto a } y^tAy = 1 \end{cases}$$

es el vector propio a^1 de $A^{-1}B$ de norma 1 asociado al valor propio más grande β_1.

Nota: Una matriz A es definida si para todo $u \in \mathbb{R}^m$ se tiene que $u^tAu > 0$.

Demostración 2.1. *Las coordenadas del individuos i-ésimo son:*

$$x_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{im}).$$

Además, se sabe que la proyección del individuo i sobre el eje D_1 es:

$$P(x_i, D_1) = \frac{\langle x_i, a^1 \rangle}{\|a^1\|} a^1,$$

donde $a^1 = (a_1^1, a_2^1, \ldots, a_m^1)$ (es vector director de norma 1 del eje D_1).

Entonces las coordenadas de la proyección del individuo x_i sobre el eje D_1 son:

$$\begin{aligned} C_{x_i}^1 &= \frac{\langle x_i, a^1 \rangle}{\|a^1\|} \\ &= a_1^1 x_{i1} + \cdots + a_2^1 x_{ij}, + \cdots +, a_m^1 x_{im} \\ &= X a^1. \end{aligned}$$

Se puede observar la proyección del individuo x_i en el gráfico 2.2,

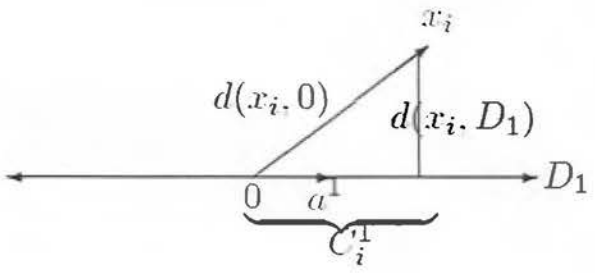


FIGURA 2.2: Proyección de un individuo sobre los ejes principales.

Usando el Teorema de Pitágoras, se deduce que:

$$d^2(x_i, 0) = (C_i^1)^2 + d^2(x_i, D_1),$$

por lo que sumando sobre i a ambos lados y multiplicando por $1/n$ se tiene que:

$$\frac{1}{n}\sum_{i=1}^{n} d^2(x_i, 0) = \frac{1}{n}\sum_{i=1}^{n}(C_i^1)^2 + \frac{1}{n}\sum_{i=1}^{n} d^2(x_i, D_1).$$

Como $\frac{1}{n}\sum_{i=1}^{n} d^2(i,0)$ es independiente del eje D_1 que se escoja, se deduce que es una cantidad constante. Por lo tanto maximizar $\frac{1}{n}\sum_{i=1}^{n}(C_i^1)^2$ es equivalente a minimizar $\frac{1}{n}\sum_{i=1}^{n} d^2(i, D_1)$.

Además es claro que:

$$\frac{1}{n}\sum_{i=1}^{n}(C_i^1)^2 = \frac{1}{n}(C^1)^t C^1 = \frac{1}{n}(a^1)^t X^t X a^1.$$

De esta manera el problema que se quiere resolver es:

$$\begin{cases} \text{máx } \frac{1}{n}(a^1)^t X^t X a^1 \\ \textit{sujeto a } (a^1)^t a^1 = 1 \textit{ (pues la norma de } a^1 \textit{ debe ser 1).} \end{cases}$$

Entonces aplicando el Lema anterior con $B = \frac{1}{n}X^t X$ y $A = I_{m\times m}$ se tiene que a^1 es el vector propio de norma 1 de la matriz $B = \frac{1}{n}X^t X$ asociado al valor propio más grande. □

Teorema 2.3. *En la etapa 2 de un ACP se calcula el eje D_2 que pasa por el origen para el cual la dispersión de la nube de puntos sea máxima, este eje D_2 pasa entonces lo más cerca posible de la nube de puntos, es decir, el promedio de las distancias al cuadrado de los n puntos de la nube y el eje D_2 es minimal.*

Sea a^2 el vector director normado (norma 1) del eje (recta) D_2 el cual será ortogonal al vector a^1 construido en la etapa 1, **entonces**: *Se tiene el siguiente problema de optimización:*

$$\text{máx} \quad \frac{1}{n}(a^2)^t X^t X a^2$$

$$\textit{sujeto a} \begin{cases} (a^2)^t a^2 = 1 \end{cases}$$

cuya solución es el vector propio asociado al segundo valor propio más grande de la matriz de V de varianzas-covarianzas.

Teorema 2.4. *En la etapa k de un ACP se calcula el eje D_k que pasa por el origen para el cual la dispersión de la nube de puntos sea máxima, este eje D_k pasa entonces lo más cerca posible de la nube de puntos, es decir, el promedio de las distancias al cuadrado de los n puntos de la nube y el eje D_k es minimal.*

Sea a^k es vector director normado (norma 1) del eje (recta) D_k el cual será ortogonal al vector a^r $\forall\, r < k$ construidos en las etapas $1, 2, \ldots, k-1$ **entonces**: *Se tiene el siguiente problema de optimización:*

$$\text{máx} \qquad \tfrac{1}{n}\left(a^k\right)^t X^t X a^k$$

$$\textit{sujeto a} \quad \begin{cases} \left(a^k\right)^t a^k = 1 \\ \left(a^k\right)^t a^k = 0 \text{ para } r = 1, 2, \ldots, k-1 \end{cases}$$

cuya solución es el vector propio asociado al $k-$ésimo valor propio más grande de la matriz de V de varianzas-covarianzas.

La prueba de los teoremas 2.3 y 2.4 son análogas a la prueba del teorema 2.2, por esto no se realizan estas demostraciones.

2.3.3.2. En el espacio de las variables

Teorema 2.5. *En la etapa 1 de un ACP se calcula una variable sintética (eje) C^1 que resuma lo mejor posible las variables originales, es decir, de tal manera que:*

$$\sum_{j=1}^{m} R^2(C^1, X^j) \text{ sea máxima.}$$

Entonces C^1 es el vector propio asociado al valor propio más grande λ_1 de la matriz $\frac{1}{n} X X^t$.

Demostración 2.2.

$$cov(C^1, X^j) = \frac{1}{n}(X^j)^t C^1 = \frac{1}{n}(C^1)^t X^j,$$

lo cual implica que:

$$cov^2(C^1, X^j) = \frac{1}{n^2}(C^1)^t X^j (X^j)^t C^1,$$

como $var(C^1) = \frac{1}{n}(C^1)^t C^1$ *y* $var(X^j) = 1$, *se tiene que:*

$$R^2(C^1, X^j) = \frac{cov^2(C^1, X^j)}{var(C^1)var(X^j)} = \frac{(C^1)^t X^j (X^j)^t C^1}{n(C^1)^t C^1},$$

entonces:

$$\sum_{j=1}^{m} R^2(C^1, X^j) = \frac{(C^1)^t \sum_{j=1}^{m} X^j (X^j)^t C^1}{n(C^1)^t C^1},$$

como $\sum_{j=1}^{m} X^j (X^j)^t = XX^t$, *se tiene que:*

$$\sum_{j=1}^{m} R^2(C^1, X^j) = \frac{(C^1)^t XX^t C^1}{n(C^1)^t C^1}.$$

De modo que maximizar $\sum_{j=1}^{m} R^2(C^1, X^j)$ *es equivalente a maximizar la siguiente expresión:*

$$\frac{(C^1)^t XX^t C^1}{n(C^1)^t C^1},$$

entonces, aplicando el lema anterior, C^1 *es el vector propio asociado al valor propio más grande* λ_1 *de la matriz* $\frac{1}{n}XX^t$. □

Teorema 2.6. *En la etapa k de un ACP se calcula una variable sintética (eje)* C^k *que resuma lo mejor posible las variables originales y que no esté correlacionada con las primeras*

$k-1$ componentes principales (variables sintéticas) ya calculadas, es decir, de tal manera que:

$$\text{máx} \qquad \sum_{j=1}^{m} R^2(C^k, X^j)$$

$$\textit{sujeto a} \quad R^2(C^k, C^r) = 0 \textit{ para } r = 1, 2, \ldots, k-1$$

Entonces: C^k es el vector propio de $\frac{1}{n}XX^t$ asociado al $k-$ésimo valor propio más grande.

La prueba del teorema 2.6 es análoga a la prueba del teorema 2.5, por esta razón se omite la prueba.

2.3.4. Equivalencia de los dos análisis – Relaciones de dualidad

Usualmente el número de variables es menor que el número de individuos, por eso se supone en adelante sin pérdidad de generalidad que $m < n$.

Teorema 2.7 (Relaciones de Dualidad). *1. Si v_k es el $k-$ésimo vector propio de norma 1 asociado a λ_k de la matriz $\frac{1}{n}XX^t$ entonces:*

$$u_k = \frac{X^t v_k}{\sqrt{n\lambda_k}},$$

es el $k-$ésimo vector propio de norma 1 asociado a λ_k de la matriz $\frac{1}{n}X^tX$.

2. Si u_k es el $k-$ésimo vector propio de norma 1 asociado a λ_k de la matriz $\frac{1}{n}X^tX$ entonces:

$$v_k = \frac{Xu_k}{\sqrt{n\lambda_k}},$$

es el $k-$ésimo vector propio de norma 1 asociado a λ_k de la matriz $\frac{1}{n}XX^t$.

Demostración 2.3. *1. Sea v_k el vector propio de norma 1 asociado a λ_k de la matriz $\frac{1}{n}XX^t$, entonces por definición se tiene que:*

$$\frac{1}{n}XX^tv_k = \lambda_k v_k,$$

multiplicando por X^t *a ambos lados por la izquierda se tiene que:*

$$\frac{1}{n}X^tXX^tv_k = \lambda_k X^tv_k,$$

lo cual es equivalente a:

$$\frac{1}{n}(X^tX)(X^tv_k) = \lambda_k(X^tv_k),$$

aplicando de nuevo la definición de valor propio se tiene que:

- λ_k *es un valor propio de la matriz* $\frac{1}{n}X^tX$.
- X^tv_k *es el vector propio de la matriz* $\frac{1}{n}X^tX$ *asociado al valor propio* λ_k.

Este vector propio X^tv_k *se debe normalizar, para esto:*

$$\|X^tv_k\|^2 = (X^tv_k)^t(X^tv_k) = v_k^tXX^tv_k = n\lambda_k v_k^tv_k = n\lambda_k,$$

entonces:

$$\|X^tv_k\| = \sqrt{n\lambda_k},$$

por lo que:

$$u_k = \frac{X^tv_k}{\sqrt{n\lambda_k}},$$

es un vector propio de norma 1 de la matriz $\frac{1}{n}X^tX$ *asociado al valor propio* λ_k.

2. *Análogo al anterior.*

□

2.3.5. Varianza explicada por cada eje

Teorema 2.8. *1.* $\frac{1}{n}X^tX$ *y* $\frac{1}{n}XX^t$ *tienen los mismos valores propios,* $\lambda_1, \lambda_2, \ldots, \lambda_m$.

2. Además el rango de ambas matrices es $n - m$ *y los últimos* $n - m$ *valores propios de* $\frac{1}{n}XX^t$ *son nulos.*

Demostración 2.4. *1. Sea* λ_k *el k-ésimo valor propio de la matriz* $\frac{1}{n}X^tX$*, entonces por definición se tiene que:*

$$\frac{1}{n}X^tXv_k = \lambda_k v_k,$$

multiplicando por X *a ambos lados se tiene que:*

$$\frac{1}{n}XX^tXv_k = \lambda_k Xv_k,$$

como se sabe que $Xv_k = C^k$ *(la componente k-ésima), entonces:*

$$\frac{1}{n}XX^tC^k = \lambda_k C^k,$$

lo cual implica que λ_k *el k-ésimo valor propio de la matriz* $\frac{1}{n}XX^t$*, asociado al vector propio* C^k.

2. Al suponer que $m < n$*, la matriz* $\frac{1}{n}XX^t$ *posee* $n-m$ *filas más que* $\frac{1}{n}X^tX$*, al poseer los mismos valores propios* $\lambda_1, \lambda_2, \ldots, \lambda_m$*, esto quiere decir que los restantes* $n-m$ *valores propios de* $\frac{1}{n}XX^t$ *son 0, ya que* $\frac{1}{n}XX^t$ *es semidefinida positiva y simétrica, por lo cual* $\forall i \in 1, 2, ..., n, 0 \leq \lambda_i$*, considere* C *la matriz cuyas columnas son los vectores propios de* $\frac{1}{n}XX^t$ *y* D *la matriz de tamaño* $n \times n$ *cuya diagonal esta formada por los valores propios de* $\frac{1}{n}XX^t$*, por lo que rango de* $\frac{1}{n}XX^t$ *es el mismo que el rango de* D*, pues* $\frac{1}{n}XX^t$ *es equivalente a* D*, como el rango de* D *es* $n - m$ *entonces* $\frac{1}{n}XX^t = n - m$.

□

Teorema 2.9. *La suma de los m valores propios de* $\frac{1}{n}X^tX$ *es igual al número de columnas* m *de la matriz* X*, es decir:*

$$\sum_{k=1}^{m}\lambda_k = m.$$

Demostración 2.5. *Del álgebra lineal se sabe que la suma de valores propios de una matriz es igual a la suma de los elementos de la diagonal de dicha matriz, es decir, es igual a la traza de la matriz. Además, como* X *está centrada y reducida* $\frac{1}{n}X^tX = R$*, de donde:*

$$\sum_{k=1}^{m}\lambda_k = Tr\left(\frac{1}{n}X^tX\right) = Tr(R).$$

entonces:

$$\sum_{k=1}^{m}\lambda_k = Tr(R) = Tr\begin{bmatrix} 1 & r_{12} & \cdots & r_{1m} \\ r_{21} & 1 & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & 1 \end{bmatrix}_{m\times m} = m.$$

□

El ACP tiene m etapas, en cada etapa se construye un resumen de la tabla X, menos interesante que el construido en la etapa anterior.

- ¿Cómo medir la calidad de la etapa k?
- En la etapa k, el criterio del ACP es maximizar:

$$\frac{1}{n}\sum_{i=1}^{n}(C_i^k)^2,$$

como:

$$\frac{1}{n}\sum_{i=1}^{n}(C_i^k)^2 = \frac{1}{n}(a^k)^tX^tXa^k = (a^k)^t\lambda_k a^k = \lambda_k.$$

- Entonces λ_k es la varianza explicada por el eje $k-$ésimo, es decir por C^k.
- Como:

$$\sum_{k=1}^{m} \lambda_k = m,$$

se tiene que:

$$\frac{\lambda_k}{m} = \% \text{ de la varianza explicada por el eje } C^k = \% \text{ de inercia.}$$

- Por ejemplo, la **inercia** explicada por el plano principal, ejes 1 y 2 es:

$$\frac{\lambda_1 + \lambda_2}{m}.$$

2.3.6. Gráficos y su interpretación

2.3.6.1. Representación de los individuos

Se debe recordar que para calcular las coordenadas de un individuo se tiene que (la matriz X se supone centrada y reducida):

- $C^s = Xa^s$ donde a^s es el vector propio de $R = \frac{1}{n}X^tX$ asociado a λ_s.
- De donde:

$$C_i^s = a_1^s X_{i1} + \cdots + a_j^s X_{ij} + \cdots + a_m^s X_{im},$$

es decir:

$$C_i^s = \sum_{j=1}^{m} X_{ij} a_j^s.$$

Análogamente:

- $C^r = Xa^r$ donde a^r es el vector propio de $R = \frac{1}{n}X^tX$ asociado a λ_r.
- De donde:

$$C_i^r = a_1^r X_{i1} + \cdots + a_j^r X_{ij} + \cdots + a_m^r X_{im},$$

es decir:

$$C_i^r = \sum_{j=1}^{m} X_{ij} a_j^r.$$

Gráficamente se ilustra como sigue:

- Así, dos individuos i y j cuyas proyecciones son cercanas son "semejantes" en la nube de puntos.
- Para proyectar un individuo en suplementario $s = (s_1, \ldots, s_m)$ simplemente se centra y reduce como si fuera la última fila de X, como sigue:

$$\tilde{s} = \left(\frac{s_1 - \bar{X}^1}{\sigma_1}, \ldots, \frac{s_m - \bar{X}^m}{\sigma_m} \right),$$

donde $\bar{X}^j$ es la media de la columna $j-$ésima de la matriz X. Entonces las coordenadas se calculan como sigue:

$$C_i^s = \sum_{j=1}^{m} \tilde{s}_j a_j^s.$$

2.3.6.2. Calidad de la representación de un individuo

- En el espacio de los individuos se tienen 2 bases ortonormales:

1. La base original, en la cual las coordenadas del individuo i son:

$$i = (X_{i1}, \ldots, X_{ij}, \ldots, X_{im}).$$

2. La base construida por los m factores, en la cual las coordenadas del individuo i son:

$$i = (C_i^1, \ldots, C_i^k, \ldots, C_i^m),$$

entonces la distancia del punto al origen se puede medir con ambas representaciones, lo que implica que:

$$\sum_{j=1}^{m}(X_{ij})^2 = \sum_{k=1}^{m}(C_i^k)^2.$$

- De modo que el individuo i tiene una buena representación en el eje r si $(C_i^r)^2$ tiene un valor importante respecto a la suma $\sum_{j=1}^{n}(X_{ij})^2$.
- Por lo que la calidad de la representación del individuo i sobre el eje r está dada por:

$$\frac{(C_i^r)^2}{\sum_{j=1}^{m}(X_{ij})^2} = \%\text{ del individuo } i \text{ representado en el eje } r.$$

- Lo anterior es útil para determinar que tan bien está representado un individuo en un eje o plano.

2.3.6.3. Las contribuciones de los individuos a la varianza total

- La varianza total de las componentes principales en la etapa r es igual a:

$$\frac{1}{n}\sum_{i=1}^{n}(C_i^r)^2 = \lambda_r.$$

- La parte de esta varianza explicada por el individuo i es:

$$\frac{1}{n}(C_i^r)^2.$$

- Entonces, la contribución del individuo i a la varianza total del eje r está dada por:

$$\frac{(C_i^r)^2}{n\lambda_r} = \%\text{ de contribución del individuo } i \text{ a la formación del eje } r.$$

- Lo anterior es útil para intepretar los ejes.

2.3.6.4. Representación de las variables

- La coordenada de la variable X^j sobre el eje r está dada por:

$$R(X^j, C^r),$$

que es el coeficiente de correlación entre la variable $j-$ésima y la componente principal $r-$ésima.

- Entonces las coordenadas de X^j sobre la base de componentes principales son:

$$(R(X^j, C^1), \ldots, R(X^j, C^s), \ldots, R(X^j, C^m)),$$

esto implica que:

$$\sum_{k=1}^{m} R^2(X^j, C^k) = 1.$$

- Por lo que si se usan solamente 2 componentes C^r y C^s se tiene que:

$$R^2(X^j, C^s) + R^2(X^j, C^r) \leqslant 1.$$

- Por esta razón las variables pueden ser representadas en un círculo de radio 1 como se ilustra a continuación:

Teorema 2.10 (Cálculo de las correlaciones).

$$\begin{pmatrix} R(X^1, C^r) \\ \vdots \\ R(X^j, C^r) \\ \vdots \\ R(X^m, C^r) \end{pmatrix} = \sqrt{\lambda_r} \cdot a^r = \begin{pmatrix} \sqrt{\lambda_r} a_1^r \\ \vdots \\ \sqrt{\lambda_r} a_j^r \\ \vdots \\ \sqrt{\lambda_r} a_m^r \end{pmatrix},$$

donde a^r es el $r-$ésimo vector propio de $R = \frac{1}{n} X^t X$ asociado a λ_r.

Demostración **2.6.** *Se sabe que:*

$$R(X^j, C^r) = \frac{cov(X^j, C^r)}{\sigma_{X^j} \sigma_{C^r}}.$$

Como la tabla X está reducida $\sigma_{X^j} = 1$. Además se sabe que la varianza del eje C^r es λ_r, es decir, $\sigma_{C^r} = \sqrt{\lambda_r}$, entonces se tiene que:

$$R(X^j, C^r) = \frac{cov(X^j, C^r)}{\sigma_{X^j} \sigma_{C^r}} = \frac{cov(X^j, C^r)}{\sqrt{\lambda_r}}.$$

Entonces:

$$\begin{pmatrix} R(X^1, C^r) \\ \vdots \\ R(X^j, C^r) \\ \vdots \\ R(X^m, C^r) \end{pmatrix} = \frac{1}{n\sqrt{\lambda_r}} X^t C^r = \frac{1}{n\sqrt{\lambda_r}} X^t X a^r = \frac{1}{\sqrt{\lambda_r}} \lambda_r a^r = \sqrt{\lambda_r} a^r.$$

□

- Por dualidad, en el espacio de las variables, para calcular las coordenadas (correlaciones) se podría diagonalizar la matriz $H = \frac{1}{n} X X^t$ (que es de tamaño $n \times n$) y proceder a calcular dichas coordenadas de manera completamente análoga al caso de los individuos.

Es decir, suponiendo que la matriz X está centrada y reducida, y si denotará por $Z = X^t$ entonces:

$R^s = Za^s$ donde a^s es el vector propio de $H = \frac{1}{n}XX^t$ asociado a λ_s, de donde:

$$R_i^s = a_1^s Z_{i1} + \cdots + a_j^s Z_{ij} + \cdots + a_n^s Z_{in},$$

es decir:

$$R_i^s = \sum_{j=1}^{n} Z_{ij} a_j^s.$$

- Calidad de representación de una variable.

  La calidad de la representación de una variable sobre el círculo de correlaciones, será también medida con el cuadrado del coseno del ángulo entre la variable y su proyección. Ahora bien, recuérdese que entre variables, el coseno es igual a una correlación, por lo que serán las correlaciones al cuadrado las que midan la calidad de la representación de las variables (para ver los detalles de estas propiedades se recomienda el libro [17]). Así la matriz de calidades de las variables $S \in M_{m\times m}$ se puede calcular como sigue:

$$S = \begin{pmatrix} R^2(X^1,C^1) & \cdots & R^2(X^1,C^r) & \cdots & R^2(X^1,C^m) \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ R^2(X^j,C^1) & \cdots & R^2(X^j,C^r) & \cdots & R^2(X^j,C^m) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R^2(X^m,C^1) & \cdots & R^2(X^m,C^r) & \cdots & R^2(X^m,C^m) \end{pmatrix}.$$

- Para proyectar una variable suplementaria:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

primero se centra y se reduce respecto a sí misma como sigue:

$$y^c = \begin{pmatrix} \frac{y_1 - \bar{y}}{\sigma_y} \\ \frac{y_2 - \bar{y}}{\sigma_y} \\ \vdots \\ \frac{y_n - \bar{y}}{\sigma_y}. \end{pmatrix},$$

y luego se calculan las correlaciones de y^c con las componentes principales, de manera análoga a proyectar una columna de X.

2.3.7. El algoritmo

Entrada: Las tabla de datos $X \in M_{n\times m}$.

Salida: La matriz de componentes principales $C \in M_{n\times m}$, la matriz de calidades de los individuos (cosenos cuadrados) $Q \in M_{n\times m}$, la matriz de coordenadas de las variables $T \in M_{m\times m}$, la matriz de calidades de las variables (cosenos cuadrados) $S \in M_{m\times m}$ y el vector de inercias de los ejes $I \in M_{1\times m}$.

Paso 1: Centrar y reducir la tabla de datos X.

Paso 2: Calcular la matriz de correlaciones $R \in M_{m\times m}$. R se puede calcular: $R = \frac{1}{n}X^tX$, o bien calculando todas las correlaciones.

Paso 3: Calcular los vectores y valores propios de la matriz $R \in M_{m\times m}$.

Paso 4: Ordenar de mayor a menor estos valores propios.

Paso 5: Si se denota por $\lambda_1, \lambda_2, \ldots, \lambda_m$ estos valores propios ordenados y por $v_1, v_2, \ldots, v_m$ los respectivos vectores propios, entonces se construye la matriz $V \in M_{m\times m}$ de la siguiente forma:

$$V = [v_1|v_2|\cdots|v_m].$$

Es decir, la matriz V tiene como columnas los vectores $v_1, v_2, \ldots, v_m$.

Paso 6: Calcular la matriz de componentes principales $C \in M_{n\times m}$:

$$C = X \cdot V.$$

Paso 7: Calcular la matriz de calidades de los individuos (cosenos cuadrados) $Q \in M_{n\times m}$, como sigue:

$$Q_{ir} = \frac{(C_{i,r})^2}{\sum_{j=1}^{m}(X_{ij})^2} \quad \text{para } i = 1, 2, \ldots, n; \;\; r = 1, 2, \ldots, m.$$

Paso 8: Calcule la matriz de coordenadas de las variables $T \in M_{m\times m}$, como sigue:

$$T = \begin{pmatrix} R(X^1, C^1) & \cdots & R(X^1, C^r) & \cdots & R(X^1, C^m) \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ R(X^j, C^1) & \cdots & R(X^j, C^r) & \cdots & R(X^j, C^m) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(X^m, C^1) & \cdots & R(X^m, C^r) & \cdots & R(X^m, C^m) \end{pmatrix}$$

$$= \begin{pmatrix} \sqrt{\lambda_1}v_{1,1} & \cdots & \sqrt{\lambda_r}v_{1,r} & \cdots & \sqrt{\lambda_m}v_{1,m} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \sqrt{\lambda_1}v_{j,1} & \cdots & \sqrt{\lambda_r}v_{j,r} & \cdots & \sqrt{\lambda_m}v_{j,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sqrt{\lambda_1}v_{m,1} & \cdots & \sqrt{\lambda_r}v_{m,r} & \cdots & \sqrt{\lambda_m}v_{m,m} \end{pmatrix}.$$

Paso 9: Calcule la matriz de calidades de las variables (cosenos cuadrados) $S \in M_{m\times m}$, como sigue:

$$S = \begin{pmatrix} \lambda_1(v_{1,1})^2 & \cdots & \lambda_r(v_{1,r})^2 & \cdots & \lambda_m(v_{1,m})^2 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \lambda_1(v_{j,1})^2 & \cdots & \lambda_r(v_{j,r})^2 & \cdots & \lambda_m(v_{j,m})^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda_1(v_{m,1})^2 & \cdots & \lambda_r(v_{m,r})^2 & \cdots & \lambda_m(v_{m,m})^2 \end{pmatrix}.$$

Paso 10: Calcule el vector **de inercias** de los ejes $I \in M_{1\times m}$, como sigue:

$$I = (100 \cdot \frac{\lambda_1}{m}, 100 \cdot \frac{\lambda_2}{m}, \ldots, 100 \cdot \frac{\lambda_m}{m}).$$

2.4. Curvas principales

La teoría de las curvas principales para datos clásicos fue desarrollada por Trevor Hastie en [Hastie, T. (1984)] (además expuestas en [Hastie, T.; Stuetzle, W. (1989)] y [Hastie, T.; Tibshirani, R.; Friedman, J. (2008)]), las curvas principales buscan una representación no lineal de los datos en un número menor de dimensiones.

Se considera un conjunto de n observaciones y dos variables aleatorias x y y, al realizar un gráfico de dispersión se desea estudiar la relación entre una variable y la otra, una forma trivial es considerar $f(x) = \bar{y}$. Otra forma de poder aproximar esta relación por medio de una función lineal que minimice la suma de cuadrados de la diferencia entre $f(x)$ e y (ver figura 2.3), el primer componente principal minimiza la distancia entre el punto (x, y) y $[(x, y)]_v$ donde v es el primer componente principal y $[(x, y)]_v$ es la coordenada del punto $[(x, y)]$ en el vector v (ver figura 2.3). La regresión lineal fue generalizada, incluyendo funciones no lineales de x, utilizando una variedad de funciones suaves como por ejemplo splines. La idea es buscar la mejor función no lineal que minimice la suma de cuadrados de la diferencia entre $f(x)$ e y (ver figura 2.3 3). Por último las superficies principales buscan generalizar los componentes principales, a través de funciones no lineales (ver figura 2.3 4).

2.4.1. Curvas principales de una distribución de probabilidad

Sea h una distribución de probabilidad suave en p dimensiones, sea $f(\lambda)$ un vector de p dimensiones donde $f_i : \Lambda \to \mathbb{R}$ con $\Lambda \subset \mathbb{R}, \forall i \leq p$, las funciones $f_i(\lambda)$ son llamadas coordenadas. Si $\forall i$, $f_i(\lambda)$ son suaves , entonces f se llama curva suave. La longitud de una curva f de λ_0 a λ_1 es dada por $l = \int_{\lambda_0}^{\lambda_1} \|f'(z)\| dz$, si $\|f'(z)\| = 1$ entonces $l = \int_{\lambda_0}^{\lambda_1} \|f'(z)\| dz = \lambda_1 - \lambda_0$, esta situación es la deseable, ya que la longitud del arco es igual a la diferencia entre λ_0 y λ_1. Ahora bien la f debe ser monótona, en caso que f no sea monótona se puede modificar las funciones (coordenadas) por medio de un reordamiento para que f sea monótona.

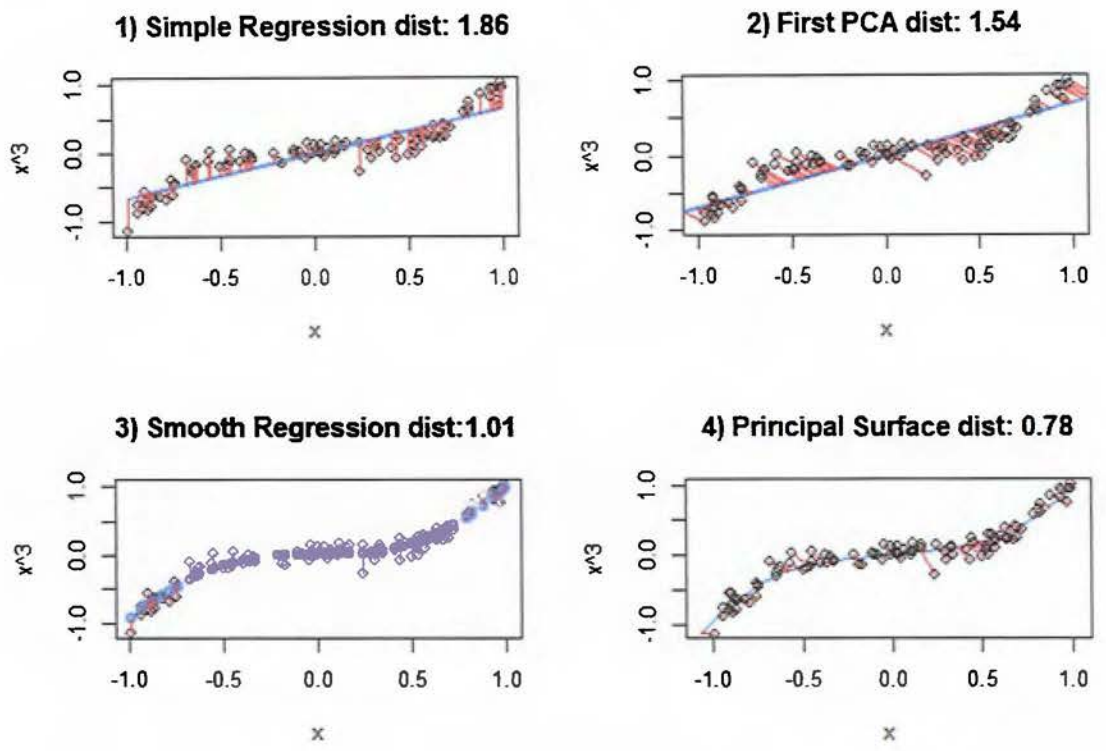


FIGURA 2.3: Diferencia entre los métodos de ajuste.

El vector $f'(\lambda)$ es tangente a $f(\lambda)$, se le llama vector de velocidad, si $\|f'(z)\| = 1$ es llamada curva parametrizada de velocidad-unidad ([Hastie, T. (1984)]) Según [Hastie, T. (1984)] siempre se podrá reparametrizar cualquier curva suave con $\|f'(\lambda)\| > 0$ para hacer una curva velocidad-unidad. Si v es un vector con $\|v\| = 1$, entonces $f(\lambda) = v_0 + \lambda v$ es una recta de velocidad-unidad. Pero esta parametrización no es única: $l^*(\lambda) = u + av + \lambda v$, en este caso $l^*(\lambda) = l(\lambda)$ asumiendo $\langle a, v\rangle = 0$. El vector f'' es llamado aceleración de la curva $f(\lambda)$, para una curva velocidad-unidad $\langle f'', f'\rangle = 0$. En este caso $f''/\|f''\|$ es llamado normal principal de $f(\lambda)$.

Sea X un vector aleatorio de $\mathbb{R}^p$, con densidad h y segundo momento finito. Sin pérdida de generalidad se asume $E(X) = 0$. Sea f una curva suave de velocidad-unidad en $\mathbb{R}^p$ parametrizada sobre $\Lambda \subset \mathbb{R}$ sobre un intervalo cerrado (posiblemente infinito), que no se interseca consigo misma ($\lambda_1 \neq \lambda_2 \Rightarrow f(\lambda_1) \neq f(\lambda_2)$) y tiene largo finito dentro de cualquier

bola finita en $\mathbb{R}^p$.

Sea

$$\forall x \in \mathbb{R}^p, D(x) = \inf_{\lambda \in \Lambda} \mathrm{d}(x, f(\lambda)); \tag{2.22}$$

donde

$$d(x, f(\lambda)) = \|x - f(\lambda)\|; \tag{2.23}$$

es la distancia usual entre vectores en $\mathbb{R}^p$. Sea

$$M(\mathrm{x}) = \{\lambda \mid \mathrm{d}(x, f(\lambda)) = D(x)\}. \tag{2.24}$$

Como Λ es compacto, $M(x)$ no es vacío. $\mathrm{d}(x, f(\lambda))$ es continua pues f es continua, $M^c(x)$ es abierto, entonces $M(x)$ es cerrado.

Por lo cual

$$\lambda_f : \mathbb{R}^p \to \mathbb{R}, \text{con } \lambda_f = \sup M(x). \tag{2.25}$$

El índice de proyección λ_f de x es el valor λ para el cual λ es más cercano a x. Para ver la prueba de que λ_f está bien definida y es medible se encuentra en [Hastie, T. (1984)].

Definición 2.3. *La curva f es llamada autoconsitente o una curva principal de h si $E(X \mid \lambda_f = \lambda) = f(\lambda)$ para c.p.d. λ.*

Para cualquier valor λ se toman todas las observaciones donde $f(\lambda)$ es el punto más cercano en la curva (h). Si $f(\lambda)$ es el promedio de las observaciones y si esto es para todo λ, entonces f es llamada curva principal.

2.4.2. Curvas principales y componentes principales

Proposición 2.1. *Si una recta $l(\lambda) = u_0 + \lambda v_0$ es autoconsistente y $u_0 \perp v_0$ entonces $l(\lambda)$ es un componente principal.*

Antes de probar la proposición 2.1, se debe definir la proyección ortogonal de un vector sobre otro.

Definición 2.4. *Sean $v, w \in \mathbb{R}^p$ la proyección ortogonal de w sobre v se define por* $Proy_v w = \frac{w^t v}{\|v\|^2} v$.

Prueba. Primero se probará que si $l(\lambda)$ contiene al origen o lo que es equivalente que $u_0 = 0$

$$\begin{aligned} 0 &= E(X) \\ &= E_\lambda E(X \mid \lambda_f = \lambda) \\ &= E_\lambda(u_0 + \lambda v_0) \\ &= u_0 + \bar{\lambda} v_0. \end{aligned}$$

Como $u_0 \perp v_0$ entonces se tiene que $u_0 = 0$ por lo cual $l(\lambda) = \lambda v_0$.

Ahora se debe probar que v_0 es un vector propio de la matriz de covarianza de X, sea Σ la matriz de covarianza de X:

$$\begin{aligned} \Sigma v_0 &= E(XX^t)v_0 \\ &= E_\lambda E(XX^t v_0 \mid \lambda_f = \lambda) \\ &= E_\lambda E(XX^t v_0 \mid X^t v_0 = \lambda) \\ &= E_\lambda E(\lambda X \mid X^t v_0 = \lambda) \\ &= E_\lambda \lambda E(X \mid \lambda_f = \lambda) \\ &= E_\lambda \lambda^2 v_0. \end{aligned}$$

□

2.4.3. Algoritmo curvas principales

El algoritmo 1 fue presentado por Trevor Hastie en [Hastie. T. (1984)].

Algoritmo 1 Curva principal

Entradas: X es una matriz de $n \times p$, TOL es la tolerancia de las variaciones entre iteraciones y N es el número máximo de iteraciones

Salidas: f es la curva principal de X

1: Sea h la densidad continua de probabilidad de X

2: $f^{(0)}(\lambda) = v\lambda$ donde v es el primer componente principal de X. Tome $\lambda_0(x) = \lambda_{f^{(0)}}(x)$

3: **while** $|D^2(h, f^{(j)}) - D^2(h, f^{(j-1)})| >$ TOL and $j < N$ **do**

4: Sea $f^{(j)}(\lambda) = E(X \mid \lambda_{j-1}(X) = \lambda)$

5: $\lambda_j(x) = \lambda_{f^{(j)}}(x)$

6: $D^2(h, f^{(j)}) = E_{\lambda^{(j)}} E(\|X - f(\lambda_j(X))\|^2 \mid \lambda_j(X))$

7: $j = j + 1$, $f = f^{(j)}$

8: **end while**

9: **return** f

Para más detalles de las pruebas y más ejemplos se recomienda leer las tesis doctoral del profesor Trevor Hastie [Hastie, T. (1984)]. Las definiciones y teoremas para esta subsección fueron tomadas de [Hastie, T. (1984)].

2.5. Datos simbólicos

El análisis de datos simbólicos fue propuesto por Edwin Diday en [Diday, E. (1987)]. La diferencia principal entre un análisis de datos clásico y un análisis de datos simbólico es que en el caso clásico se toma una tabla de datos donde todas sus columnas son números reales, para los casos de variables cualitativas se construyen tablas de contigencia y matrices de Burt que son la transformación de una variable cualitativa en una tabla de datos con elementos en los números reales, mientras que en el análisis simbólico se utilizan conceptos, para comprender un concepto se presenta el siguiente ejemplo, se considera una tabla con datos de estudiantes de la UCR, cada individuo de una carrera se encuentra descrito por un conjunto de variables numéricas o nominales. Cada individuo (estudiante) es considerado

individuo de primer orden. Para poder estudiar las carreras, consideradas individuos de segundo orden, se puede describir como un resumen de los valores tomados de sus estudiantes, como intervalos, subconjuntos de valores, histogramas, distribuciones de probabilidad, entre otros, dependiendo de la variable que se desea utilizar. De esta manera, se obtiene una tabla de datos simbólicos donde cada fila contiene una descripción de la carrera y cada columna está asociada a una variable simbólica. Algunos tipos de variables simbólicas propuestos en [Billard, L.; Diday, E. (2006)] son: intervalos, conjuntos, distribuciones de probabilidad, entre otros.

Sea $X = \left(x_{(1)}, x_{(2)}, \ldots, x_{(m)}\right)$ denota $m-$variables aleatorias donde $x_{(j)}$ es la j-ésima variable para $j = 1, 2, \ldots, m$. Sea X_i la i-ésima observación de los datos de la matriz X donde $i = 1, 2, \ldots, n$. La notación encerrada en paréntesis $\left(x_{(j)}\right)$ significa el índice de la variable y la notación sin paréntesis X_i representa el índice de la observación. Utilizando esta notación, la matriz de datos X se puede expresar como un vector de variables o un vector de observaciones respectivamente

$$X = \left[x_{(1)}, x_{(2)}, \ldots, x_{(m)}\right] = \left[X_1, X_2, \ldots, X_n\right]^t.$$

Además, la variable aleatoria X_{ij} representa la $j-$ésima variable y la $i-$ésima observación, la letra minúscula x_{ij} es un valor clásico (número real) y ξ_{ij} denota un valor simbólico de la varible X_{ij}.

Sea $\mathfrak{X}_{(j)}$ el dominio de $X_{(j)}$ y $\mathfrak{X} = \mathfrak{X}_{(1)} \times \mathfrak{X}_{(2)} \times \ldots \mathfrak{X}_{(m)}$ el dominio de $X = \left[x_{(1)}, x_{(2)}, \ldots, x_{(m)}\right]$.

Definición 2.5. *Todo punto* $x = \left(X_{(1)}, X_{(2)}, \ldots, X_{(m)}\right) \in \mathfrak{X}$ *es llamado vector de descripción.*

Definición 2.6. *Todo conjunto* $D \subseteq \mathfrak{X}$ *tal que* $D = \left(D_{(1)}, D_{(2)}, \ldots, D_{(m)}\right)$ *donde* $D_{(j)} \subseteq \mathfrak{X}_{(j)}$ *es llamado conjunto descripción.*

Definición 2.7. *Sean* $A \subseteq D$ *y* $B \subseteq D$ *dos conjuntos de descripción y* $x \in \mathfrak{X}$. *Se define la regla de dependencia lógica* v *como*

$$v : [x \in A] \Rightarrow [x \in B].$$

De manera equivalente, v es un mapeo de $\mathfrak{X}$ en $\{0,1\}$ tal que

$$v(x) = \begin{cases} 1 & \text{si } x \in (A \cap B) \text{ o } x \notin A \\ 0 & \text{en otro caso.} \end{cases} \tag{2.26}$$

El conjunto de todas las reglas de dependencias lógicas v definida en $\mathfrak{X}$ es denotado $V_{\mathfrak{X}}$

Definición 2.8. *La descripción virtual de un vector de descripción d, $vir(d)$, es el conjunto de todos los vectores de descripcion x que satisfacen todas las reglas $v \in \mathfrak{X}$. Esto es,*

$$\forall v \in V_{\mathfrak{X}}, vir(d) = \{x \in D \mid v(x) = 1\}. \tag{2.27}$$

2.5.1. Variables simbólicas de tipo intervalo

El foco de esta tesis es el método de superficies principales para variables simbólicos de tipo intervalo, sea $X_i, i = 1, ..., n$, una muestra aleatoria. Sea la $j-$ ésima variable, $x_{(j)}$, una variable de tipo intervalo, entonces una realización ξ_{ij} de X_{ij} toma un valor de intervalo $[a_{ij}, b_{ij}]$. Sea W un punto de $X_{(j)}$, asumiendo que W es uniformemente distribuida sobre el intervalo $X_{ij} = [a_{ij}, b_{ij}]$ para todos vectores de descripción individual $x \in vir(d_i)$. Entonces, para todo $\xi \in \mathbb{R}$,

$$P\{W \leq \xi \mid x \in vir(d_i)\} = \begin{cases} 0, & \xi < a_{ij} \\ \frac{\xi - a_{ij}}{b_{ij} - a_{ij}}, & a_{ij} \leq \xi < b_{ij} \\ 1, & \xi \leq b_{ij} \end{cases}. \tag{2.28}$$

En [Billard, L.; Diday, E. (2006)] se define la media y la varianza para variables de tipo intevalo de la siguiente forma:

Definición 2.9. *Para una variable aleatoria de tipo intervalo, la media muestral simbólica esta dada por:*

$$\bar{W} = \tfrac{1}{2n} \textstyle\sum_{i=1}^{n} (a_{ij} + b_{ij}), \tag{2.29}$$

y la varianza muestral simbólica esta dada por:

$$\sigma^2 = \frac{1}{3n}\sum_{i=1}^{n}(a_{ij}^2 + a_{ij}b_{ij} + b_{ij}^2) - \frac{1}{4n^2}\left[\sum_{i=1}^{n}(a_{ij} + b_{ij})\right]^2. \tag{2.30}$$

2.5.2. Análisis de componentes principales para variables de tipo intervalo

Uno de los métodos más conocidos en el análisis de datos clásico es el ACP, el cual resuelve de una manera lineal el problema de reducción de la dimensionalidad (más detalles en la sección 2.3). Diversos autores proponen dos ACP para variables de tipo intervalos , un método de vértices y método de centros ([Billard, L.; Diday, E. (2006)] , [Cazes, P.; Chouakria, A.; Diday, E.; Schektman, Y. (1997)], [Chouakria, A.; Billard, L.; Diday, E. (2011)] y [Rodríguez, O. (2000)]), el método de vértices consiste en construir una matriz con todos los vértices de los hipercubos que se desean analizar y a esta aplicarle el ACP clásico, por su parte el método de centros considera una matriz con todos los centros de los hipercubos que se desean analizar, y se proyectan los vértices de cada hipercubo como elementos suplementarios. La dualidad para el método de centros es propuesta por el profesor Oldemar Rodríguez en su tesis doctoral [Rodríguez, O. (2000)]. Antes de explicar estos métodos, se define la matriz de datos de tipo intervalo. Sea X una matriz de $n \times m$ datos. Entonces:

$$X = \begin{bmatrix} \xi_{11} & \xi_{12} & \xi_{13} & \cdots & \xi_{1m} \\ \xi_{21} & \xi_{22} & \xi_{23} & \cdots & \xi_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \xi_{n1} & \xi_{n2} & \xi_{n3} & \cdots & \xi_{nm} \end{bmatrix}. \tag{2.31}$$

Si X es una matriz de datos tipo intervalo, entoces la matriz X de la ecuación (2.31) tiene la siguiente forma,

$$X = \begin{bmatrix} [a_{11}, b_{11}] & [a_{12}, b_{12}] & [a_{13}, b_{13}] & \cdots & [a_{1m}, b_{1m}] \\ [a_{21}, b_{21}] & [a_{22}, b_{22}] & [a_{23}, b_{23}] & \cdots & [a_{2m}, b_{2m}] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ [a_{n1}, b_{n1}] & [a_{n2}, b_{n2}] & [a_{n3}, b_{n3}] & \cdots & [a_{nm}, b_{nm}] \end{bmatrix}. \tag{2.32}$$

donde $a_{ij} \leq b_{ij}$ para todo $i = 1, 2, \ldots, n$ y $j = 1, 2, \ldots, m$.

Una matriz de intervalos se puede definir como un subconjunto de $M_{n\times m}$, de la siguiente forma:

Definición 2.10. *Sea X una matriz de intervalos,*

$$X = \{Z \in M_{n\times m} \mid \forall i \in \{1, 2, ..., n\}, \forall j \in \{1, 2, ..., m\}, Z_{ij} \in [a_{ij}, b_{ij}]\}.$$

2.5.2.1. Método de vértices

Para el análisis de componentes principales realizado en este método, la matriz X es transformada en una matriz de vértices X^v.

Antes de definir la matriz de vértices (X^v) se da una idea de la construcción de esta, por medio del siguiente ejemplo, sea A una matriz de intervalos,

$$A = \begin{bmatrix} [1,3] & [4,6] \\ [4,4] & [2,5] \\ [2,2] & [1,1] \end{bmatrix}. \tag{2.33}$$

En la figura , se grafica cada una de las filas de matriz A.

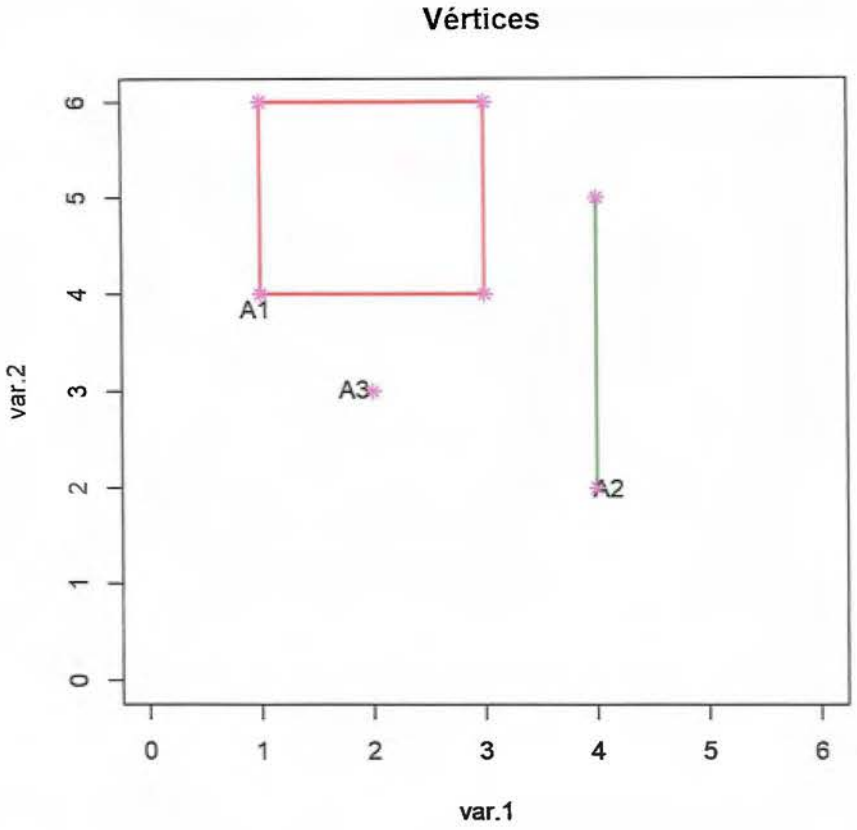


FIGURA 2.4: Objetos simbólicos de tipo intervalo y sus vértices.

- $A_1 = ([1,3],[4,6])$, los dos intervalos de A_1 son no triviales, por lo cual la matriz de vértices está formada por todas las combinaciones de los extremos de los intervalos $(2 * 2 = 2^2 = 4)$.

$$A_1^v = \begin{bmatrix} 1 & 4 \\ 1 & 6 \\ 3 & 4 \\ 3 & 6 \end{bmatrix}.$$

- $A_2 = ([4,4],[2,5])$, uno de los intervalos de A_2 es trivial, por lo cual la matriz de vértices está formada por todas las combinaciones de los extremos de los intervalos

$(2 * 1 = 2^1 = 2)$.

$$A_2^v = \begin{bmatrix} 4 & 2 \\ 4 & 5 \end{bmatrix}.$$

- $A_3 = ([2,2], [1,1])$, los dos intervalos de A_3 son triviales (es un punto en $\mathbb{R}^2$), por lo cual la matriz de vértices está formada por todas las combinaciones de los extremos de los intervalos $(1 * 1 = 1)$.

$$A_3^v = \begin{bmatrix} 2 & 1 \end{bmatrix}.$$

- La matriz de todos los vértices de A es

$$A_v = \begin{bmatrix} \begin{bmatrix} 1 & 4 \\ 1 & 6 \\ 3 & 4 \\ 3 & 6 \end{bmatrix} \\ \begin{bmatrix} 4 & 2 \\ 4 & 5 \end{bmatrix} \\ \begin{bmatrix} 2 & 1 \end{bmatrix} \end{bmatrix}.$$

A partir del ejemplo anterior se generaliza la matriz de vértices para m dimensiones. Sea X la matriz definida en (2 32) Entonces, para $i = 1, \ldots, n$,

$$X_i = ([a_{i1}, b_{i1}], [a_{i2}, b_{i2}], [a_{i3}, b_{i3}], \ldots, [a_{im}, b_{im}]).$$

Define la matriz de vértices para la observación i como,

$$X_i^v = \begin{bmatrix} a_{i1} & a_{i2} & \ldots & a_{im} \\ a_{i1} & a_{i2} & \ldots & b_{im} \\ \vdots & \vdots & \vdots & \vdots \\ b_{i1} & b_{i2} & \ldots & a_{im} \\ b_{i1} & b_{i2} & \ldots & b_{im} \end{bmatrix}. \tag{2.34}$$

Se puede notar, que todo elemento de X_i^v es un punto en $\mathbb{R}^m$, i.e., X_i^v es una matriz clásica (real). Para toda observación X_i, X_i^v es una matriz de $2^{m_i} \times p$ donde m_i es el número de variables en observación i tal que

$$a_{ij} \neq b_{ij}.$$

Toda fila de X_i^v representan las coordenadas de un vértice del hipercubo formado por la observación i en un espacio $m-$dimensional. Un intervalo es llamado trivial si $a_{ij} = b_{ij}$. Así, si $[a_{ij}, b_{ij}]$ es trivial para todo $j = 1, 2, \ldots, p$, entonces $X_i^v = [a_{i1}, a_{i2}, \ldots, a_{im}]$ se reduce a un punto en $\mathbb{R}^m$. La matriz de vértices para todo el conjunto X es

$$X^v = \begin{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ b_{11} & b_{12} & \ldots & b_{1m} \end{bmatrix} \\ \vdots \; \vdots \; \vdots \; \vdots \\ \begin{bmatrix} a_{i1} & a_{i2} & \ldots & a_{im} \\ \vdots & \vdots & \vdots & \vdots \\ b_{i1} & b_{i2} & \ldots & b_{im} \end{bmatrix} \\ \vdots \; \vdots \; \vdots \; \vdots \\ \begin{bmatrix} a_{n1} & a_{n2} & \ldots & a_{nm} \\ \vdots & \vdots & \vdots & \vdots \\ b_{n1} & b_{n2} & \ldots & b_{nm} \end{bmatrix} \end{bmatrix}. \tag{2.35}$$

Si $m_i = m$ para todo i, X^v tiene dimensión $n2^m \times m$. Lo siguiente es aplicar el ACP clásico en X^v. El $k-$ésimo componente principal de X^v es

$$Y_k^v = X^v w_k^v;$$

donde w_k^v es el vector propio correspondiente al $k-$ésimo valor propio de la matriz de varianza y covarianza de X^v. El $k-$ésimo componente principal para observación i basado en este método es $Y_{ik}^v = \left[y_{ik}^{lo}, y_{ik}^{up}\right]$ donde

$$y_{ik}^{lo} = \min_{\eta \in L_i} y_{\eta k}^{v}; \tag{2.36}$$

$$y_{ik}^{up} = \max_{\eta \in L_i} y_{\eta k}^{v}; \tag{2.37}$$

donde L_i es un conjunto de filas en X^v que pertenecen a la observación i. Esto es, para $N_i = 2^{m_i}$,

$$L_i = \left\{ \sum_{m=1}^{i-1} N_m + 1, \sum_{m=1}^{i-1} N_m + 2, \ldots, \sum_{m=1}^{i-1} N_m + N_i \right\}. \tag{2.38}$$

Se pueden encontrar expresiones para (2 36) y (2 37) de la siguiente forma:

$$y_{ik}^{lo} = \sum_{j \in J_c^-} (b_{ij} - \bar{X}^v{}_{(j)}) w_{kj}^v + \sum_{j \in J_c^+} (a_{ij} - \bar{X}^v{}_{(j)}) w_{kj}^v;$$

$$y_{ik}^{up} = \sum_{j \in J_c^-} (a_{ij} - \bar{X}^v{}_{(j)}) w_{kj}^v + \sum_{j \in J_c^+} (b_{ij} - \bar{X}^v{}_{(j)}) w_{kj}^v;$$

donde $J_c^- = \{j | w_{kj}^v < 0\}$ y $J_c^+ = \{j | w_{kj}^v \geq 0\}, \bar{X}^v{}_{(j)}$ es la media de la $j-$ésima columna.

Para detalles ver [Billard, L.; Diday, E. (2006)].

2.5.2.2. Método de centros

Sea la matriz de centros correspondiente a la matriz de datos definida en la (2 32)

$$X^c = \begin{bmatrix} X_{11}^c & X_{12}^c & \ldots & X_{1m}^c \\ X_{21}^c & X_{22}^c & \ldots & X_{2m}^c \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1}^c & X_{n2}^c & \ldots & X_{nm}^c \end{bmatrix}. \tag{2.39}$$

donde

$$X_{ij}^{c} = \frac{a_{ij} + b_{ij}}{2}. \tag{2.40}$$

Note que para $i = 1, \ldots, n$ y para $j = 1, \ldots, m$ $X_{ij}^{c} \in \mathbb{R}$. Entonces la matriz de centros (2.39) es una matriz clásica ($X^{c} \in M_{n \times m}$). En este caso se realiza, el ACP clásico sobre la matriz de centros X^{c}. El resultado $k-$ésimo componente principal de centros es

$$Y_{(k)}^{c} = X^{c} w_{k}^{c}; \tag{2.41}$$

donde w_{k}^{c} es el vector propio correspondiente al $k-$ésimo valor propio de la matriz de covarianza de la matriz X^{c} de la ecuación (2.39) Para las observaciones $i = 1, \ldots, n$, la $k-$ésimo componente principal para una variable de tipo intervalo es construida de la siguiente forma. Sea $Y_{ik}^{c} = \left[y_{ik}^{lo}, y_{ik}^{up}\right]$ un componente principal para una variable de tipo intervalo. Entonces, el punto inicial y el punto final son construidos de la siguiente forma,

$$y_{ik}^{lo} = \sum_{j \in J_{c}^{-}} (b_{ij} - \bar{X}_{(j)}) w_{kj}^{c} + \sum_{j \in J_{c}^{+}} (a_{ij} - \bar{X}_{(j)}) w_{kj}^{c}; \tag{2.42}$$

$$y_{ik}^{up} = \sum_{j \in J_{c}^{-}} (a_{ij} - \bar{X}_{(j)}) w_{kj}^{c} + \sum_{j \in J_{c}^{+}} (b_{ij} - \bar{X}_{(j)}) w_{kj}^{c}; \tag{2.43}$$

donde $J_{c}^{-} = \{j | w_{kj}^{c} < 0\}$ y $J_{c}^{+} = \{j | w_{kj}^{c} \geq 0\}$.

Las fórmulas (2.42) y (2.43), fueron propuestas por [Cazes. P.: Chouakria, A.: Diday. E.: Schektman. Y. (1997)].

2.5.2.3. Dualidad del ACP de centros

Sea Z una matriz de intervalos, que tiene la siguiente forma:

$$\forall j, \forall i, Z_{ij} = \left[\frac{a_{ij} - \bar{X}^{c}{}_{(j)}}{\sqrt{(n)}\sigma_{(j)}}, \frac{b_{ij} - \bar{X}^{c}{}_{(j)}}{\sqrt{(n)}\sigma_{(j)}}\right]$$

con $\bar{X}^c_{(j)}$ y $\sigma_{(j)}$ son respectivamente el promedio y desviación estándar de la columna j.

Teorema 2.11. *Si se proyecta el hipercubo definido por la j-ésima columna de Z en el i-ésimo componente principal (en la dirección de w_i), entonces se tiene que el mínimo y el máximo valor están dados por (2.44) y (2.45) respectivamente:*

$$\underline{r_{ij}} = \sum_{k=1, w_{kj}<0}^{m} \overline{z}^c_{ki} w_{kj} + \sum_{k=1, w_{kj}>0}^{m} \underline{z}^c_{ki} w_{kj}, \tag{2.44}$$

$$\overline{r_{ij}} = \sum_{k=1, w_{kj}<0}^{m} \underline{z}^c_{ki} w_{kj} + \sum_{k=1, w_{kj}>0}^{m} \overline{z}^c_{ki} w_{kj}. \tag{2.45}$$

El teorema 2.11 fue propuesto por el profesor Oldemar Rodríguez en [Rodríguez, O. (2000)], en el capítulo 3 se generaliza dicho teorema a cualquier punto que pertenezca a la matriz de intervalos.

2.6. Métodos de optimización

La teoría expuesta en esta sección fue tomada del libro [Nocedal, J.; Wright, S. (1999)], esta teoría es necesaria para realizar las optimizaciones que se proponen en el capítulo 3

Dada una función $f : \mathbb{R}^n \to \mathbb{R}$, 2 veces diferenciable, cuya segunda derivada es continua, se desea encontrar $x \in \mathbb{R}^n$, que sea la solución de:

$$\min_{x \in \mathbb{R}^n} f(x).$$

Encontrar una solución del problema corresponde a encontrar un punto x^* que cumple:

$$\forall x \in \mathbb{R}^n, f(x^*) \leq f(x).$$

Este punto x^* se denomina mínimo global de f sobre $\mathbb{R}^n$.

Definición 2.11. *Un punto x^* es un mínimo local de f en $A \subset \mathbb{R}^n$ si:*

$$\forall x \in A, f(x^*) \leq f(x).$$

Teorema 2.12. *Sea $f : \mathbb{R}^n \to \mathbb{R}$, 2 veces diferenciable, cuya segunda derivada es continua, un punto x^* es mínimo local de f si se verifican las siguientes condiciones:*

1. *x^* es un punto estacionario de f, esto es $\nabla f(x^*) = \vec{0}$, donde $\nabla f(x^*)$ es el gradiente de f en x^*.*
2. *La matriz hessiana $\nabla^2 f(x^*)$ de f en x^* es definida positiva.*

La demostración del teorema 2.12 se encuentra en las notas del curso MA0450 impartido por el profesor William Ugalde, [Ugalde, W. (2009)].

Para obtener el mínimo local se debe resolver un sistema de ecuaciones no lineales, en general complejo. Surge de esta forma la necesidad de aplicar métodos numéricos para encontrar soluciones aproximadas del sistema. Se consideran métodos numéricos llamados de descenso, que pueden ser clasificados en distintos grupos (ver [Nocedal, J.; Wright, S. (1999)]).

- **Métodos descenso**
  - Primer orden
    - Método del gradiente y variantes
    - Método del gradiente conjugado
    - Método de Fletcher-Reeves
  - Segundo orden: Método de Newton
  - Tercer orden
    - DFP: Davidon - Fletcher - Powell
    - BFGS

2.6.1. Método de Newton

El método de Newton es un método numérico que se utiliza para encontrar ceros de una función. Este método se puede aplicar a la búsqueda de los ceros del gradiente de una función $f : \mathbb{R}^n \to \mathbb{R}$ (puntos estacionarios de f); de esta forma, es utilizado para hallar mínimos globales de f.

2.6.2. Métodos de optimización Cuasi-Newton

En cada iteración del método de Newton es necesario calcular la inversa de la matriz hessiana de f en x_k de manera exacta, lo que es costoso computacionalmente, $O(n^3)$ operaciones aritméticas. Por esto razón, se propone un método iterativo de la forma:

$$x^{k+1} = x^k - t^k S^k g^k, \text{ con } g^k = \nabla f(x^k), \tag{2.46}$$

donde S^k es una matriz que aproxima a $(\nabla^2 f(x^k))^{-1}$ y $0 \leq t^k$ minimiza f sobre $x^k - \lambda g^k$ para $0 \leq \lambda$ (paso exacto o aproximado).

Se presentan 2 métodos que permiten construir iterativamente la matriz S^k de manera que se verifiquen las siguientes condiciones:

1. Definida Positiva:
   Si S^k es definida positiva implica S^{k+1} también lo es.

2. Aproximación de la matriz hessiana inversa de f:
   Si $x^k \to x^*$, $S^k \to (\nabla^2 f(x^k))^{-1}$ para $k \to \infty$.

2.6.3. Algoritmo BFGS

El algoritmo BFGS fue propuesto por Broyden, Fletcher, Goldfarb y Shanno actualmente, se considera la fórmula más efectiva y eficaz de actualización Cuasi-Newton; el éxito de

este algoritmo depende de la aproximación de la inversa de la verdadera matriz hessiana. Los costos computacionales se ven disminuidos entre un método Newton ($O(n^3)$) y BFGS ($O(n^2)$), este método solamente necesita multiplicaciones entre vectores y matrices.

En el algoritmo BFGS, la aproximación de la matriz hessiana puede basarse en el historial completo de los gradientes, en cuyo caso se denomina BFGS, o puede basarse solo en los gradientes de m más recientes, en cuyo caso se conoce como BFGS de memoria limitada, abreviada como L-BFGS. La ventaja de L-BFGS es que solo requiere conservar los m gradientes más recientes, donde $m \leq n$, lo cual es un requisito de almacenamiento mucho más pequeño que $\frac{n(n+1)}{2}$ elementos necesarios para almacenar la totalidad de una estimación de la matriz hessiana, como se requiere con BFGS, donde n es la dimensión del problema. A diferencia de BFGS (completo), la estimación de la matriz hessiana nunca se almacena explícitamente en L-BFGS; más bien, los cálculos que se requerirían con la estimación de la hessiana se realizan sin formarlo explícitamente. L-BFGS se usa en lugar de BFGS para problemas muy grandes (cuando n es muy grande), pero podría no funcionar tan bien como BFGS. Por lo tanto, BFGS se prefiere a L-BFGS cuando se pueden cumplir los requisitos de memoria de BFGS.

El siguiente pseucódigo fue tomado de [Nocedal, J.: Wright, S. (1999)].

Paso 1: Seleccionar un punto inicial $x^0 \in \mathbb{R}^n$

Inicializar $S_0 = I_n$, I_n : matriz identidad de tamaño n

$k = 0$

Paso 2: Calcular $g^k = \nabla f(x^k)$

Si $\|g^k\| \approx 0 \Rightarrow$ PARE

Si no, calcular $x^{k+1} = x^k - t^k S^k g^k$

donde t^kse escoge según la regla de Goldstein y seguir a Paso 3

Paso 3: Calcular:

$p^k = x^{k+1} - x^k$

$q^k = g^{k+1} - g^k$

$$S^{k+1} = S^k + \left[\frac{(q^k)^t S^k q^k}{(p^k)^t q^k}\right] \frac{p^k (p^k)^t}{(p^k)^t q^k} - \frac{\left[p^k (q^k)^t (S^k)^t + S^k q^k (p^k)^t\right]}{(p^k)^t q^k}$$

$k = k + 1$ volver al Paso 2.

Capítulo 3

Métodos de Reducción de la Dimensionalidad para Variables de Tipo Intervalo

En este capítulo se realiza la creación de los conceptos y teoremas (generalización de las ecuaciones (2.42), (2.43) y del teorema 2.11) necesarios (vistos en el capítulo 2) para la construcción del mejor ACP. Se quiere mejorar el método de centros ([Billard, L.; Diday, E. (2006)] , [Cazes, P.; Chouakria, A.; Diday, E.; Schektman, Y. (1997)], [Chouakria, A.; Billard, L.; Diday, E. (2011)] y [Rodríguez, O. (2000)]), aplicando algoritmos de optimización, se busca el mejor punto en algún sentido (minimizar la distancia de los individuos suplementarios, el punto que posee más varianza en las primeras componentes) y a partir de este proyectar los vértices como elementos suplementarios. Además se busca generalizar la teoría de dualidad propuesta en [Rodríguez, O. (2000)], por otra parte se propone un algoritmo para la construcción de la superficies principales de tipo intervalo.

3.1. Componentes principales para datos de tipo intervalo: Método del mejor punto

Sea X una matriz de variables de tipo intervalo de tamaño $n \times m$, sea $Z \in X$ (definición 2.10) se realiza un ACP a la matriz Z. El $k-$ésimo componente principal de Z para la observación ξ_u con $k = 1, \ldots, s < m$, $i = 1, \ldots, n$,

$$y_{ik}^{Z} = \sum_{j=1}^{m} (Z_{jk} - \bar{Z}_{(j)}) w_{k_j}^{Z}, \tag{3.1}$$

con $\bar{Z}_{(j)}$ la media de la variable $z_{(j)}$

$$\bar{Z}_{(j)} = \frac{1}{n} \sum_{i=1}^{n} Z_{ij}, \tag{3.2}$$

con $w_k^Z = (w_{v_1}^Z, \ldots, w_{k_m}^Z)$ es el $k-$ésimo vector propio asociado a la matriz de varianza y covarianza de Z.

Sea $\beta(Z) = \{w_1^Z, \ldots, w_m^Z\}$ una base ortonormal de $\mathbb{R}^m$.

Sea ξ_i la $i-$ésima observación de X con $i = 1, \ldots, n$, se considera (2.34) la matriz de vértices, los vértices ahora son elementos suplementarios para el ACP de Z.

Se define la matriz de vértices suplementaria como:

$$\widetilde{\mathbf{X}}_i^v(Z) = \begin{bmatrix} \frac{a_{i1} - \bar{Z}_{(1)}}{\sqrt{n}\sigma_{(1)}} & \frac{a_{i2} - \bar{Z}_{(2)}}{\sqrt{n}\sigma_{(2)}} & \cdots & \frac{a_{ip} - \bar{Z}_{(m)}}{\sqrt{n}\sigma_{(m)}} \\ \frac{a_{i1} - \bar{Z}_{(1)}}{\sqrt{n}\sigma_{(1)}} & \frac{a_{i2} - \bar{Z}_{2}}{\sqrt{n}\sigma_{(2)}} & \cdots & \frac{a_{ip} - \bar{Z}_{(m)}}{\sqrt{n}\sigma_{(m)}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{b_{i1} - \bar{Z}_{(1)}}{\sqrt{n}\sigma_{(1)}} & \frac{b_{i2} - \bar{Z}_{(2)}}{\sqrt{n}\sigma_{(2)}} & \cdots & \frac{b_{ip} - \bar{Z}_{(m)}}{\sqrt{n}\sigma_{(m)}} \\ \frac{b_{i1} - \bar{Z}_{(1)}}{\sqrt{n}\sigma_{(1)}} & \frac{b_{i2} - \bar{Z}_{(2)}}{\sqrt{n}\sigma_{(2)}} & \cdots & \frac{b_{ip} - \bar{Z}_{(m)}}{\sqrt{n}\sigma_{(m)}} \end{bmatrix}, \tag{3.3}$$

con $\sigma_{(j)}$ la desviación estándar de $Z_{(j)}$.

Para simplificar se denota cada fila de la matriz $\widetilde{\mathbf{X}}_i^v(Z)$ de la siguiente forma

$$\widetilde{x_{i_tj}^v}(Z),$$

con $t = 1, \ldots, 2^{m_i}$, m_i es el número de intervalos no triviales y $j = 1, \ldots . m$.

Las coordenadas se obtienen de la siguiente forma:

$$C^k(x_{i_j}^v) = \sum_{h=1}^{m} \widetilde{x_{i_{jh}}^v}(Z) w_{k_h}, \tag{3.4}$$

con $j = 1, \ldots, 2^{m_i}$, m_i es el número de intervalos no triviales.

Las coordenadas en el $k-$ésimo componente para un objeto simbólico vienen dadas por:

$$\widetilde{Y_{ik}^v} = \widetilde{y_{ik}} = [\widetilde{y_{ik}^{az}}, \widetilde{y_{ik}^{bz}}] \text{ con } k = 1, \ldots, s < m, \tag{3.5}$$

con

$$\widetilde{y_{ik}^{az}} = \min_{j=1,\ldots,2^{m_i}} C^k(x_{i_j}^v), \tag{3.6}$$

$$\widetilde{y_{ik}^{bz}} = \max_{j=1,\ldots,2^{m_i}} C^k(x_{i_j}^v). \tag{3.7}$$

Se propone el teorema 3 1 el cual generaliza las fórmulas (2 42) y (2 43)

Teorema 3.1 (Generalización de coordenas individuos). *Las coordenadas de* $\widetilde{Y_{ik}^{vz}}$ *se pueden encontrar de la siguiente forma:*

$$\widetilde{y_{ik}^{az}} = \sum_{j \in J_z^-} (b_{ij} - \overline{Z}_{(j)}) w_{k_j}^Z + \sum_{j \in J_z^-} (a_{ij} - \overline{Z}_{(j)}) w_{k_j}^Z.$$

$$\widetilde{y_{ik}^{bz}} = \sum_{j \in J_z^-} (a_{ij} - \overline{Z}_{(j)}) w_{k_j}^Z + \sum_{j \in J_z^+} (b_{ij} - \overline{Z}_{(j)}) w_{k_j}^Z,$$

donde $J_Z^- = \{j|w_{kj}^v < 0\}$ *y* $J_Z^+ = \{j|w_{kj}^v \geq 0\}, \overline{Z}_{(j)}$ *es el promedio de la* $j-$*ésima columna.*

Demostración 3.1. *Sea* $Z \in X$, *entonces*

$$\forall j,\ \forall i,\ Z_{ij} \in X_{ij} = [a_{ij}, b_{ij}]\,. \tag{3.8}$$

Como a_{ij} *y* b_{ij} *son elementos suplementarios del ACP de* Z *lo primero que se debe realizar es centrarlos, respecto a las columnas (variables) de* Z

$$\forall i,\ \forall j,\ a_{ij} - \overline{Z}_{(j)},\ b_{ij} - \overline{Z}_{(j)}. \tag{3.9}$$

De (3 8) y (3 9) se tiene que

$$\forall i,\ \forall j,\ z_{ij} - \overline{Z}_{(j)} \in X_{ij} - \overline{Z}_{(j)} = \left[a_{ij} - \overline{Z}_{(j)}, b_{ij} - \overline{Z}_{(j)}\right]. \tag{3.10}$$

- *Caso 1:* $\forall j,\ w_{kj}^Z > 0$, *se tiene*

$$\forall j,\ \forall i,\ w_{kj}^Z \left(z_{ij} - \overline{Z}_{(j)}\right) \in w_{kj}^Z \left[a_{ij} - \overline{Z}_{(j)}, b_{ij} - \overline{Z}_{(j)}\right]$$

$$\Rightarrow\ \textstyle\sum_{j=1}^{p} \left(a_{ij} - \overline{Z}_{(j)}\right) w_{kj}^Z \leq \sum_{j=1}^{p} \left(z_{ij} - \overline{Z}_{(j)}\right) w_{kj}^Z \leq \sum_{j=1}^{p} \left(b_{ij} - \overline{Z}_{(j)}\right) w_{kj}^Z.$$

- *Caso 2:* $\forall j,\ w_{kj}^Z < 0$, *se tiene*

$$\forall j,\ \forall i,\ w_{kj}^Z \left(z_{ij} - \overline{Z}_{(j)}\right) \in w_{kj}^Z \left[b_{ij} - \overline{Z}_{(j)}, a_{ij} - \overline{Z}_{(j)}\right]$$

$$\Rightarrow\ \textstyle\sum_{j=1}^{p} \left(b_{ij} - \overline{Z}_{(j)}\right) w_{kj}^Z \leq \sum_{j=1}^{p} \left(z_{ij} - \overline{Z}_{(j)}\right) w_{kj}^Z \leq \sum_{j=1}^{p} \left(a_{ij} - \overline{Z}_{(j)}\right) w_{kj}^Z.$$

- ***Caso 3:*** *Sean* $J_Z^- = \{j|w_{kj} < 0\}$ *y* $J_Z^+ = \{j|w_{kj} \geq 0\}$.
  Por ***caso 1*** *aplicado a* J_Z^+*:*

$$\sum_{j \in J_Z^+} \left(a_{ij} - \overline{Z}_{(j)}\right) w_{kj}^Z \leq \sum_{j \in J_Z^+} \left(z_{ij} - \overline{Z}_{(j)}\right) w_{kj}^Z \leq \sum_{j \in J_Z^+} \left(b_{ij} - \overline{Z}_{(j)}\right) w_{kj}^Z. \tag{3.11}$$

Por ***caso 2*** *aplicado a* J_v^-*:*

$$\sum_{j\in J_Z^+}\left(b_{ij}-\overline{Z}_{(j)}\right)w_{k_j}^Z \leq \sum_{j\in J_Z^+}\left(z_{ij}-\overline{Z}_{(j)}\right)w_{k_j}^Z \leq \sum_{j\in J_Z^+}\left(a_{ij}-\overline{Z}_{(j)}\right)w_{k_j}^Z. \tag{3.12}$$

De (3 11) y (3 12) se obtiene lo siguiente

$$\sum_{j\in J_Z^-}\left(b_{ij}-\overline{Z}_{(j)}\right)w_{k_j}^Z+\sum_{j\in J_Z^+}\left(a_{ij}-\overline{Z}_{(j)}\right)w_{k_j}^Z$$

$$\leq \sum_{j=1}^{n}\left(z_{ij}-\overline{Z}_{(j)}\right)w_{k_j}^Z$$

$$\leq \sum_{j\in J_Z^-}\left(a_{ij}-\overline{Z}_{(j)}\right)w_{k_j}^Z+\sum_{j\in J_Z^+}\left(b_{ij}-\overline{Z}_{(j)}\right)w_{k_j}^Z.$$

Se tiene que

$$\widetilde{y_{ik}^{aZ}}=\sum_{j\in J_Z^-}\left(b_{ij}-\overline{Z}_{(j)}\right)w_{k_j}^Z+\sum_{j\in J_Z^+}\left(a_{ij}-\overline{Z}_{(j)}\right)w_{k_j}^Z,$$

$$\widetilde{y_{ik}^{bZ}}=\sum_{j\in J_Z^-}\left(a_{ij}-\overline{Z}_{(j)}\right)w_{k_j}^Z+\sum_{j\in J_Z^+}\left(b_{ij}-\overline{Z}_{(j)}\right)w_{k_j}^Z.$$

□

El siguiente teorema da una forma de encontrar las coordenadas en el espacio de variables, esto es una relación de dualidad. Se necesita centrar y estandarizar Z:

$$\widetilde{z}_{ij}=\frac{z_{ij}-\bar{Z}_{(j)}}{\sqrt{n}\sigma_{(j)}}.$$

A partir de ahora se trabaja con la matriz $\widetilde{Z}=\widetilde{z}_{ij}$, $\forall i$, $\forall i$. Se denota $\widetilde{z}^j$ la $j-$ésima columna de $\widetilde{Z}$, se tiene que $(\widetilde{z}^j)^t\cdot\widetilde{z}^i=R(i,j)\leq 1$, con R la matriz de correlaciones, la matriz de intervalos centrados y estandarizados respecto a Z es:

$$\widetilde{X}(Z) = \begin{bmatrix} \left[\frac{a_{11}-\bar{Z}_{(1)}}{\sqrt{n}\sigma_{(1)}}, \frac{b_{11}-\bar{Z}_{(1)}}{\sqrt{n}\sigma_{(1)}}\right] & \left[\frac{a_{12}-\bar{Z}_{(2)}}{\sqrt{n}\sigma_{(2)}}, \frac{b_{12}-\bar{Z}_{(2)}}{\sqrt{n}\sigma_{(2)}}\right] & \cdots & \left[\frac{a_{1p}-\bar{Z}_{(m)}}{\sqrt{n}\sigma_{(m)}}, \frac{b_{1p}-\bar{Z}_{(m)}}{\sqrt{n}\sigma_{(m)}}\right] \\ \vdots & \vdots & \ddots & \vdots \\ \left[\frac{a_{i1}-\bar{Z}_{(1)}}{\sqrt{n}\sigma_{(1)}}, \frac{b_{i1}-\bar{Z}_{(1)}}{\sqrt{n}\sigma_{(1)}}\right] & \left[\frac{a_{i2}-\bar{Z}_{(2)}}{\sqrt{n}\sigma_{(2)}}, \frac{b_{i2}-\bar{Z}_{(2)}}{\sqrt{n}\sigma_{(2)}}\right] & \cdots & \left[\frac{a_{ip}-\bar{Z}_{(m)}}{\sqrt{n}\sigma_{(m)}}, \frac{b_{im}-\bar{Z}_{(m)}}{\sqrt{n}\sigma_{(m)}}\right] \\ \vdots & \vdots & \ddots & \vdots \\ \left[\frac{a_{n1}-\bar{Z}_{(1)}}{\sqrt{n}\sigma_{(1)}}, \frac{b_{n1}-\bar{Z}_{(1)}}{\sqrt{n}\sigma_{(1)}}\right] & \left[\frac{a_{n2}-\bar{Z}_{(2)}}{\sqrt{n}\sigma_{(2)}}, \frac{b_{n2}-\bar{Z}_{(2)}}{\sqrt{n}\sigma_{(2)}}\right] & \cdots & \left[\frac{a_{np}-\bar{Z}_{(m)}}{\sqrt{n}\sigma_{(m)}}, \frac{b_{np}-\bar{Z}_{(m)}}{\sqrt{n}\sigma_{(m)}}\right] \end{bmatrix}. \tag{3.13}$$

Para facilitar la notación a partir de ahora

$$(\widetilde{\mathrm{X}}(Z))_{ij} = \left[\frac{a_{ij}-\bar{Z}_{(j)}}{\sqrt{n}\sigma_{(j)}}, \frac{b_{ij}-\bar{Z}_{(j)}}{\sqrt{n}\sigma_{(j)}}\right] = [a_{ij}^Z, b_{ij}^Z].$$

La matriz $\widetilde{Z}\widetilde{Z}^t$ es simétrica y semidefinida positiva, entonces todos sus vectores propios son ortogonales y sus valores propios son reales no negativos. Se denota $w_1^Z, w_2^Z, \ldots, w_s^Z$ los s vectores propios asociados de $\widetilde{Z}\widetilde{Z}^t$ a los valores propios $\lambda_1, \lambda_2, \ldots, \lambda_s \geq 0$. Se denota por $W(Z) = [w_1^Z | w_2^Z | \ldots | w_s^Z]$ una matriz de tamaño $n \times s$ que tiene como columnas los vectores propios de $\widetilde{Z}\widetilde{Z}^t$. Se puede calcular las coordenadas de las variables en el círculo de correlación de la siguiente forma $\widetilde{Z}^t W$, entonces se puede calcular la $i-$ésima columna de Z sobre la $j-$ésimo componente principal (en dirección w_j^Z) por:

$$r_{ij}^Z = \sum_{k=1}^{m} \widetilde{z}_{ki} w_{kj}^Z. \tag{3.14}$$

Se propone el teorema que generaliza el teorema de dualidad para el ACP de centros (teorema 2 11) .

Teorema 3.2 (Dualidad ACP Simbólico). *Si se proyecta el hipercubo definido por la j-ésima columna de $\widetilde{\mathbf{X}}(Z)$ en el i-ésimo componente principal (en la dirección de w_i), entonces se tiene que el mínimo y el máximo valor están dados por (3 15) y (3 16) respectivamente:*

$$\underline{r_{ij}} = \sum_{k=1, v_{kj}<0}^{m} b_{ki}^Z v_{kj} + \sum_{k=1, v_{kj}>0}^{m} a_{ki}^Z v_{kj}, \tag{3.15}$$

$$\overline{r_{ij}} = \sum_{k=1,v_{kj}<0}^{m} a_{ki}^{Z} v_{kj} + \sum_{k=1,v_{kj}>0}^{m} b_{ki}^{Z} v_{kj}. \tag{3.16}$$

Demostración 3.2. *Sea* $\widehat{z}_j = (\widehat{z}_{1j}, \widehat{z}_{2j}, \ldots, \widehat{z}_{mj}) \in Z_H^j$, *el hipercubo definido por la* $j-$*ésima columna de* $\widetilde{\mathbf{X}}(Z)$, *entonces* $\widehat{z}_{ij} \in [a_{ij}^{Z}, b_{ij}^{Z}]$ *para todo* $i = 1, 2, \ldots, n$ *y* $j = 1, 2, \ldots, s$. *Se denota por* $p\widehat{z}_{ij}$ *la proyección de* $\widehat{z}_j$ *en el eje principal con dirección* v_i.

Dado que $\widehat{z}_{ij} \in [a_{ij}^{Z}, b_{ij}^{Z}]$ *se puede obtener:*

$$a_{ki}^{Z} v_{kj} \leq \widehat{z}_{ki} v_{kj} \leq b_{ki}^{Z} v_{kj} \ \ si\ v_{kj} \geq 0, \tag{3.17}$$

$$a_{ki}^{Z} v_{kj} \geq \widehat{z}_{ki} v_{kj} \geq b_{ki}^{Z} v_{kj} \ \ si\ v_{kj} \leq 0. \tag{3.18}$$

Por definición de $p\widehat{z}_{ij} = \sum_{k=1}^{m} \widehat{z}_{ki} v_{kj}$ *entonces:*

$$p\widehat{z}_{ij} = \sum_{k=1}^{m} \widehat{z}_{ki} v_{kj} = \sum_{k=1,v_{kj}>0}^{m} \widehat{z}_{ki} v_{kj} + \sum_{k=1,v_{kj}<0}^{m} \widehat{z}_{ki} v_{kj}.$$

Utilizando (3.17) y (3.18) se obtiene:

$$p\widehat{z}_{ij} \leq \sum_{k=1,v_{kj}<0}^{m} a_{ki}^{Z} v_{kj} + \sum_{k=1,v_{kj}>0}^{m} b_{ki}^{Z} v_{kj} = \overline{r_{ij}},$$

de manera análoga:

$$p\widehat{z}_{ij} \geq \sum_{k=1,v_{kj}<0}^{m} b_{ki}^{Z} v_{kj} + \sum_{k=1,v_{kj}>0}^{m} a_{ki}^{Z} v_{kj} = \underline{r_{ij}}.$$

□

Se ha probado que $p\widehat{z}_{ij} \in [\underline{r_{ij}}, \overline{r_{ij}}]$ además que $\underline{r_{ij}}$, $\overline{r_{ij}}$ son la combinación de las proyecciones de los vértices del hipercubo. Se ha probado que el valor $\underline{r_{ij}}$ y $\overline{r_{ij}}$ pueden obtenerse por las fórmulas (3 15) y (3 16) respectivamente.

Se puede realizar relaciones de dualidad entre los vectores propios de $\widetilde{Z}\widetilde{Z}^t$ y $\widetilde{Z}^t\widetilde{Z}$, ambas matrices tienen los mismos s valores propios positivos $\lambda_1^Z, \lambda_2^Z, \ldots, \lambda_s^Z$ y si se denota por $u_1^Z, u_2^Z, \ldots, u_s^Z$ los primeros s vectores propios de $\widetilde{Z}^t\widetilde{Z}$, entonces las relaciones entre los vectores propios de $\widetilde{Z}\widetilde{Z}^t$ y $\widetilde{Z}^t\widetilde{Z}$ se pueden observar en las fórmulas (3 19) y (3 20)

$$u_\ell^Z = \frac{\widetilde{Z}^t v_\ell^Z}{\sqrt{\lambda_\ell}} \text{ para } \ell = 1, 2, \ldots, s. \tag{3.19}$$

$$v_\ell^Z = \frac{\widetilde{Z} u_\ell^Z}{\sqrt{\lambda_\ell}} \text{ para } \ell = 1, 2, \ldots, s. \tag{3.20}$$

Se ha desarrollado una teoría en la cual se desarrolla un ACP para cada $Z \in X$, la idea es buscar la matriz Z^* que sea optima en algún sentido:

- Minimizar la distancia al cuadrado de los vértices a los ejes principales de Z.
- Maximizar la varianza en los primeros componentes.

3.1.1. Minimizar la distancia al cuadrado de los vértices a los ejes principales de Z

Sea X una matriz de intervalos de $n \times m$, sea $Z \in X$, sea

$$\beta(Z) = \{w_1^Z, \ldots, w_s^Z\},$$

con $s \leq m$ con w_i^Z vectores propios de la matriz de varianza covarianza de Z. Si se considera X^v la matriz de vértices de X, sea

$$N = \sum_{i=1}^{n} 2^{m_i},$$

con m_i el número de intervalos no triviales para la observación ξ_i.

Se considera la matriz de vértices centrada y estandarizada respecto a Z, que posee la siguiente forma:

$$\widetilde{X^v}(Z) = \begin{bmatrix} \begin{bmatrix} \frac{a_{11}-\bar{Z}_{(1)}}{\sqrt{n}\sigma_{(1)}} & \frac{a_{12}-\bar{Z}_{(2)}}{\sqrt{n}\sigma_{(2)}} & \cdots & \frac{a_{1p}-\bar{Z}_{(m)}}{\sigma_{(m)}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{b_{11}-\bar{Z}_{(1)}}{\sqrt{n}\sigma_{(1)}} & \frac{b_{12}-\bar{Z}_{(2)}}{\sqrt{n}\sigma_{(2)}} & \cdots & \frac{b_{1m}-\bar{Z}_{(m)}}{\sqrt{n}\sigma_{(m)}} \end{bmatrix} \\ \vdots \; \vdots \; \vdots \; \vdots \\ \begin{bmatrix} \frac{a_{i1}-\bar{Z}_{(1)}}{\sqrt{n}\sigma_{(1)}} & \frac{a_{i2}-\bar{Z}_{(2)}}{\sqrt{n}\sigma_{(2)}} & \cdots & \frac{a_{im}-\bar{Z}_{(m)}}{\sqrt{n}\sigma_{(m)}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{b_{i1}-\bar{Z}_{(1)}}{\sqrt{n}\sigma_{(1)}} & \frac{b_{i2}-\bar{Z}_{(2)}}{\sqrt{n}\sigma_{(2)}} & \cdots & \frac{b_{ip}-\bar{Z}_{(m)}}{\sqrt{n}\sigma_{(m)}} \end{bmatrix} \\ \vdots \; \vdots \; \vdots \; \vdots \\ \begin{bmatrix} \frac{a_{n1}-\bar{Z}_{(1)}}{\sqrt{n}\sigma_{(1)}} & \frac{a_{n2}-\bar{Z}_{(2)}}{\sqrt{n}\sigma_{(2)}} & \cdots & \frac{a_{np}-\bar{Z}_{(m)}}{\sqrt{n}\sigma_{(m)}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{b_{n1}-\bar{Z}_{(1)}}{\sqrt{n}\sigma_{(1)}} & \frac{b_{n2}-\bar{Z}_{(2)}}{\sqrt{n}\sigma_{(2)}} & \cdots & \frac{b_{np}-\bar{Z}_{(m)}}{\sqrt{n}\sigma_{(m)}} \end{bmatrix} \end{bmatrix}. \tag{3.21}$$

Se define la **función distancia** como:

$$d(x, y) = \|x - y\|, \forall x, y \in \mathbb{R}^m \tag{3.22}$$

con $\|.\|$ es la norma euclídea ($\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}$).

Sea $\varphi(Z) : X \to \mathbb{R}^+ \cup \{0\}$ una función definida de la siguiente forma:

$$\varphi(Z) = \sum_{i=1}^{N} \|\widetilde{X_i^v}(Z) - Pr_{\beta(Z)}(\widetilde{X_i^v}(Z))\|^2. \tag{3.23}$$

Para calcular el valor de **la función** (3.23) se propone el algoritmo 2.

Algoritmo 2 Cálculo de φ

Entradas: X es una matriz de intervalos de $n \times p$, $Z \in X$, s número de componentes principales

Salidas: $\varphi(Z)$

1: Aplicar el ACP a Z

2: $\beta = \{w_1, \ldots, w_s\}$ con $s \leq p$ con w_i vectores propios de la matriz de varianza covarianza de Z

3: Calcular la matriz de vértices de X (X^v)

4: Calcular la matriz de vértices centrada y estandarizada X respecto a Z $(\widetilde{X^v}(Z))$

5: $\varphi(Z) = \sum_{i=1}^{N} \|\widetilde{X_i^v}(Z) - Pr_{\beta(Z)}(\widetilde{X_i^v}(Z))\|^2$

6: return $\varphi(Z)$

Como $Z \in X$ y X es la unión finita de conjuntos compactos y la $\varphi(Z)$ es una función cóncava entonces posee un máximo y un mínimo, en este caso se desea obtener la matriz Z que minimize la distancia a la matriz de vértices X^v. El problema que se desea resolver es el siguiente:

$$\text{Minimizar} \quad \varphi(Z) = \sum_{i=1}^{N} \|\widetilde{X}_i^v(Z) - Pr_{\beta(Z)}(\widetilde{X}_i^v(Z))\|^2$$

$$\text{sujeto a} \quad \begin{cases} a_{11} \leq z_{11} \leq b_{11} \\ \vdots \\ a_{1j} \leq z_{1j} \leq b_{1j} \\ \vdots \\ a_{1p} \leq z_{1m} \leq b_{1m}. \\ \vdots \\ a_{ij} \leq z_{ij} \leq b_{ij} \\ \vdots \\ a_{n1} \leq z_{n1} \leq b_{n1} \\ \vdots \\ a_{np} \leq z_{nm} \leq b_{nm} \end{cases} \tag{3.24}$$

La ecuación (3.24) se puede escribir de manera matricial de la siguiente forma:

$$\begin{aligned} &\text{mín} \quad \varphi(Z) \\ &s.a. \quad Z \in X. \end{aligned} \tag{3.25}$$

Definición 3.1. *La matriz $Z \in X$ que resuelve el problema (3.25), se llama la matriz de óptima respecto a la distancia y se denota Z^{φ}.*

Definición 3.2. *Se define el vector $\overline{Z^{\varphi}} \in \mathbb{R}^m$ de la siguiente forma: $\overline{Z^{\varphi}}(j) = \overline{Z^{\varphi}_{(j)}}$, es el promedio de la $j-$ésima columna de Z^{φ}.*

Definición 3.3. *Se define el vector $\sigma_{Z^{\varphi}} \in \mathbb{R}^m$ de la siguiente forma: $\sigma_{Z^{\varphi}}(j) = \sigma_{Z^{\varphi}_{(j)}}$, es la desviación estándar de la $j-$ésima columna de Z^{φ}.*

Para realizar la optimización se propone el algoritmo 3:

Algoritmo 3 Mejor Matriz respecto a distancias de los vértices

Entradas: X es una matriz simbólica de intervalos de $n \times m$, $Z \in X$, s número de componentes prinicipales, TOL es la tolerancia de la variación por iteraciones y N es el número máximo de iteraciones

Salidas: $\widetilde{Y^{vz^{\varphi}}}$

1: Considere $Z = X^c$ la matriz de centros 2.39, como valor inicial

2: Obtener Z^{φ} por medio de AlgoritmoOptimizacion ($valorinicial = Z$, $funcion = \varphi(Z)$, TOL, N)

3: Obtener $\widetilde{Y^{vz^{\varphi}}}$ aplicando teorema 3.1

4: return $\widetilde{Y^{vz^{\varphi}}}$

3.1.2. Maximizar la varianza en los primeros componentes

Sea X una matriz de intervalos de $n \times m$, sea $Z \in X$, sea

$$\beta(Z) = \{w_1^Z, \ldots, w_s^Z\},$$

con $s \leq m$ con w_i^Z vectores propios de la matriz de varianza covarianza de Z y $\lambda(Z) = \lambda_1^Z, \ldots, \lambda_s^Z$ el conjunto de valores propios asociados de la matriz de varianza covarianza de Z, se define la función

$$\Lambda(Z, s) : X \times \mathbb{N} \to \mathbb{R}^+,$$

de la siguiente forma:

$$\Lambda(Z, s) = \sum_{i=1}^{s} \lambda_i^Z. \tag{3.26}$$

Para calcular el valor de la función (3.26) se propone el algoritmo 4

Algoritmo 4 Cálculo de Λ

Entradas: X es una matriz simbólica de tipo intervalo $n \times m$, $Z \in X$, s número de componentes prinicipales

Salidas: $\Lambda(Z, s)$

1: Aplicar el ACP a Z

2: $\lambda(Z) = \lambda_1^Z, \ldots, \lambda_s^Z$ el conjunto de valores propios asociados de la matriz de varianza covarianza de Z

3: $\Lambda(Z, s) = \sum_{i=1}^{s} \lambda_i^Z$

4: return $\Lambda(Z, s)$

Como $Z \in X$ y X es la unión finita de conjuntos compactos, s es el número de componentes principales y la $\Lambda(Z, s)$ es una función continua entonces posee un máximo y un mínimo, en este caso se desea obtener la matriz Z que maximice la varianza acumulada en las primeras s componentes principales. El problema que se desea resolver es el siguiente.

$$\text{Maximizar} \quad \Lambda(Z,s) = \sum_{i=1}^{s} \lambda_i^Z$$

$$\text{sujeto a} \quad \begin{cases} a_{11} \leq z_{11} \leq b_{11} \\ \vdots \\ a_{1j} \leq z_{1j} \leq b_{1j} \\ \vdots \\ a_{1m} \leq z_{1m} \leq b_{1m}. \\ \vdots \\ a_{ij} \leq z_{ij} \leq b_{ij} \\ \vdots \\ a_{n1} \leq z_{n1} \leq b_{n1} \\ \vdots \\ a_{nm} \leq z_{nm} \leq b_{nm} \end{cases} \tag{3.27}$$

El problema (3.27) se puede escribir de manera matricial de la siguiente forma:

$$\begin{aligned} &\text{máx} \quad \Lambda(Z,s) \\ &s.a. \quad Z \in X. \end{aligned} \tag{3.28}$$

Definición 3.4. *La matriz $Z \in X$ que resuelve el problema (3.28) se llama matriz de óptima respecto a la varianza y se denota Z^Λ.*

Definición 3.5. *Se define al vector $\overline{Z^\Lambda} \in \mathbb{R}^m$ de la siguiente forma: $\overline{Z^\Lambda}(j) = \overline{Z^\Lambda_{(j)}}$, es el promedio de la $j-$ésima columna de Z^Λ.*

Definición 3.6. *Se define al vector $\sigma_{Z^\Lambda} \in \mathbb{R}^m$ de la siguiente forma: $\sigma_{Z^\varphi}(j) = \sigma_{Z^\Lambda_{(j)}}$, es la desviación estándar de la $j-$ésima columna de Z^Λ.*

Para realizar la optimización se propone el siguiente algoritmo:

Algoritmo 5 Mejor Matriz respecto a la varianza

Entradas: X es una matriz simbólica de intervalos de $n \times m$, $Z \in X$, s número de componentes prinicipales.

Salidas: $Y^{\tilde{V}_{Z^\wedge}}$

1: Considere $Z = X^c$ la matriz de centros 2.39, como valor inicial

2: Obtener $Z^\wedge$ por medio de AlgoritmoOptimizacion $(valorinicial = Z, funcion = \Lambda(Z, s))$

3: Obtener $\widetilde{Y^{v_{Z^\wedge}}}$ Aplicar teorema 3.1

4: return $\widetilde{Y^{v_{Z^\wedge}}}$

3.2. Curvas principales para variables simbólicas de tipo intervalo

Sea $\hat{X}$ una matriz de datos de tipo intervalo (definida en 2.32) con tamaño $n \times m$

$$\hat{X} = \begin{bmatrix} [a_{11}, b_{11}] & [a_{12}, b_{12}] & [a_{13}, b_{13}] & \dots & [a_{1p}, b_{1m}] \\ [a_{21}, b_{21}] & [a_{22}, b_{22}] & [a_{23}, b_{23}] & \dots & [a_{2p}, b_{2m}] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ [a_{n1}, b_{n1}] & [a_{n2}, b_{n2}] & [a_{n3}, b_{n3}] & \dots & [a_{nm}, b_{nm}] \end{bmatrix}. \tag{3.29}$$

Se define la matriz de vértices para la observación i como,

$$X_i^v = \begin{bmatrix} a_{i1} & a_{i2} & \dots & a_{im} \\ a_{i1} & a_{i2} & \dots & b_{im} \\ \vdots & \vdots & \vdots & \vdots \\ b_{i1} & b_{i2} & \dots & a_{im} \\ b_{i1} & b_{i2} & \dots & b_{im} \end{bmatrix}. \tag{3.30}$$

Se puede notar que todo elemento de X_i^v es un punto, i.e., X_i^v es una matriz clásica (elementos que son números reales). Para toda observación X_i, X_i^v es una matriz de $2^{m_i} \times m$

donde m_i es el número de variables en observación i tal que

$$a_{ij} \neq b_{ij}.$$

Toda fila de X_i^v representa las coordenadas de un vértice del hipercubo formado por la observación i en un espacio $m-$dimensional. Un intervalo es llamado trivial si $a_{ij} = b_{ij}$. Así, si $[a_{ij}, b_{ij}]$ es trivial para todo $j = 1, 2, \ldots, m$, entonces $X_i^v = [a_{i1}, a_{i2}, \ldots, a_{im}]$ se reduce a un punto en $\mathbb{R}^m$. La matriz de vértices para todo el conjunto $\hat{X}$ es

$$X^v = \begin{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ b_{11} & b_{12} & \ldots & b_{1p} \end{bmatrix} \\ \vdots \; \vdots \; \vdots \; \vdots \\ \begin{bmatrix} a_{i1} & a_{i2} & \ldots & a_{ip} \\ \vdots & \vdots & \vdots & \vdots \\ b_{i1} & b_{i2} & \ldots & b_{ip} \end{bmatrix} \\ \vdots \; \vdots \; \vdots \; \vdots \\ \begin{bmatrix} a_{n1} & a_{n2} & \ldots & a_{np} \\ \vdots & \vdots & \vdots & \vdots \\ b_{n1} & b_{n2} & \ldots & b_{np} \end{bmatrix} \end{bmatrix}. \tag{3.31}$$

Si $m_i = m$ para todo i, X^v tiene dimensión $n2^m \times m$.

X^v es una matriz de datos clásicos de dimensión $\sum_{i=1}^{n} 2^{m_i} \times p$ donde m_i es el número de variables en observación i tal que $a_{ij} \neq b_{ij}$, sean $f_{(1)}(\lambda), \ldots, f_{(n)}(\lambda)$ las curvas principales de X^v con dominio $\Lambda \subset \mathbb{R}$ y considere Δ_i el conjunto de índices en X^v identificando los vértices del hipercubo X_i^v, Δ_i representa las filas de X_i^v que describen el objeto simbólico X_i. Para cualquier fila $k_i = 1, \ldots, 2^{m_i}$, sea f_{vik_i} el valor de la curva principal $f_{vi}(\lambda)$, $v = 1, \ldots, n$, para cada fila k_i, entonces, la k-ésima curva principal para los vértices del intervalo está dada por

$$\hat{f}_{vi}^{V} = \hat{f}_{vi} = [\min_{k_i \in \Delta_i} f_{vik_i}, \max_{k_i \in \Delta_i} f_{vik_i}]. \tag{3.32}$$

Notación: Sea $\hat{\mathbf{X}}$ denota una matriz de datos de tipo intervalo con tamaño $n \times p$, $\hat{f}$ denota la curva principal de $\hat{X}$.

Para el cálculo de la superficie principal simbólica se propone el algoritmo 6.

Algoritmo 6 Curvas principales para datos de tipo intervalo.

Entradas: X una matriz de datos simbólicos de tipo intervalo $n \times p$, TOL es la tolerancia de las variaciones y N el máximo número de iteraciones

Salidas: $\hat{f}$ la matriz de $n \times p$ curvas principales de X

1: Calcular X^v la matriz de vértices de X

2: Sea h la densidad de probabilidad de X^v

3: $f^{(0)}(\lambda) = v\lambda$ donde v es el primer componente principal de X^v. Tome $\lambda_0(x) = \lambda_{f^{(0)}}(x)$

4: **while** $|D^2(h, f^{(j)}) - D^2(h, f^{(j-1)})| > \texttt{TOL}$ and $j < N$ **do**

5: Sea $f^{(j)}(.) = E(X \mid \lambda_{j-1}(X) = .)$

6: $\lambda_j(x) = \lambda_{f^{(j)}}(x)$

7: $D^2(h, f^{(j)}) = E_{\lambda^{(j)}} E(\|X - f(\lambda_j(X))\|^2 \mid \lambda_j(X))$

8: $j = j + 1$, $f = f^{(j)}$

9: end while

10: $\hat{f}_{iv} = [\min_{k_i \in \Delta_i} f_{vik_i}, \max_{k_i \in \Delta_i} f_{vik_i}]$ for $i = 1, \ldots, n$ and $v = 1, \ldots, p$

11: **return** $\hat{f}$

Capítulo 4

Análisis Experimental

4.1. Comparación de ACP simbólicos para datos de tipo intervalo

4.1.1. Datos de aceite

Los datos de aceite fueron propuestos por el profesor Ichino en [Ichino. M. (1994)]. Cada fila en la tabla describe cuatro variables de tipo intervalo:

- gravedad específica (GRA)
- puntos de congelamiento (FRE)
- índice de yodo (IOD)
- saponificación (SAP).

Los individuos son los siguientes aceites:

- Linaza (L)
- Perilla (P)

- Semilla de Algodón (Co)
- Sésamo o ajonjolí (S)
- Camelia (Ca)
- Oliva (O)
- Res (B)
- Cerdo (H).

	I GRA	I FRE	I IOD	I SAP
L	[0.93,0.94]	[-27,-18]	[170,204]	[118,196]
P	[0.93,0.94]	[-5,-4]	[192,208]	[188,197]
Co	[0.92,0.92]	[-6,-1]	[99,113]	[189,198]
S	[0.92,0.93]	[-6,-4]	[104,116]	[187,193]
Ca	[0.92,0.92]	[-25,-15]	[80,82]	[189,193]
O	[0.91,0.92]	[0,6]	[79,90]	[187,196]
B	[0.86,0.87]	[30,38]	[40,48]	[190,199]
H	[0.86,0.86]	[22,32]	[53,77]	[190,202]

TABLA 4.1: Datos de aceite propuestos por el profesor Ichino.

4.1.1.1. Matriz óptima respecto a la distancia (MOD)

La matriz que resuelve el problema (3.25). expuesto en la sección 3.1.1 es:

	GRA	FRE	IOD	SAP
L	0.94	-27.00	204.00	118.00
P	0.94	-5.00	208.00	197.00
Co	0.92	-6.00	99.00	198.00
S	0.93	-6.00	116.00	193.00
Ca	0.92	-25.00	80.00	193.00
O	0.92	6.00	79.00	196.00
B	0.86	38.00	40.00	199.00
H	0.86	32.00	53.00	202.00

TABLA 4.2: Matriz Z^{φ}.

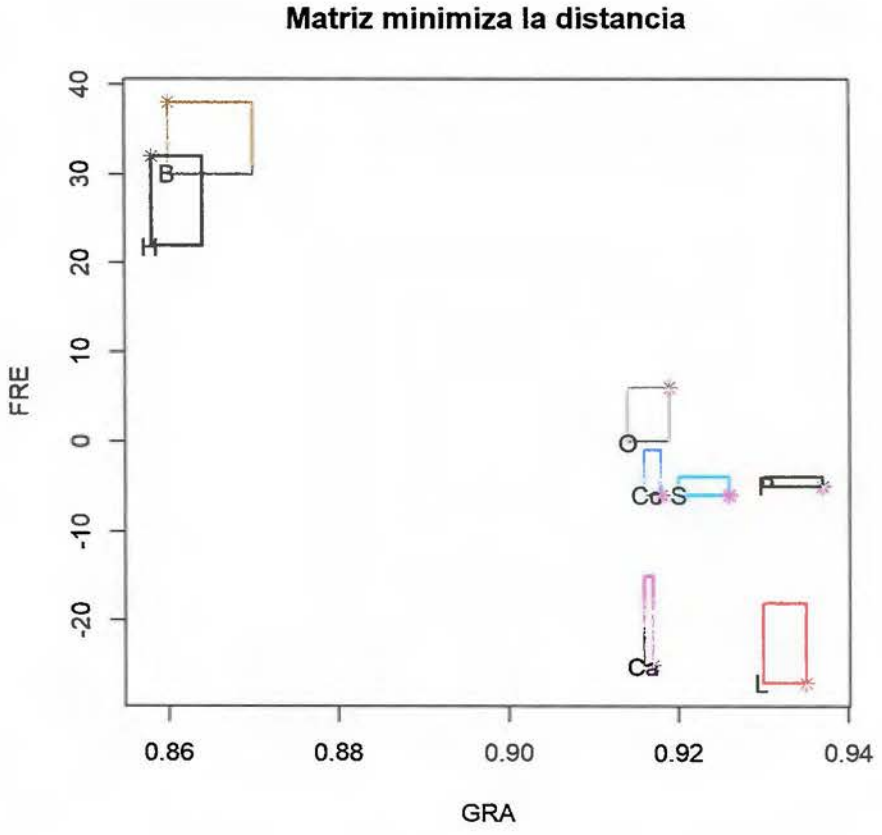


FIGURA 4.1: Matriz que minimiza la distancia de los vértices como elementos suplementarios.

La varianza explicada por el primer componente principal del método de minimización de distancia es de $\frac{\lambda_1}{\sum \lambda_i} = 73.98\,\%$. De igual manera se puede tomar más componentes principales, si se utilizan 2 o 3 componentes principales, la varianza acumulada será de 90.28 % y 98.63 % respectivamente. En la tabla 4.3, se pueden observar todos los valores propios para este caso.

	Valor propio	Porcentaje de varianza	Porcentaje acumulado de varianza
λ_1	2.97	73.98 %	73.98 %
λ_2	0.65	16.30 %	90.28 %
λ_3	0.33	8.35 %	98.63 %
λ_4	0.05	1.37 %	100.00 %

TABLA 4.3: Valores propios para el ACP de Z^{φ}.

Los vectores propios se muestran en la tabla 4.4,

	comp 1	comp 2	comp 3	comp 4
GRA	0.53	0.47	0.03	0.70
FRE	-0.53	-0.27	0.58	0.55
IOD	0.52	-0.08	0.77	-0.36
SAP	-0.42	0.83	0.25	-0.25

TABLA 4.4: Vectores propios para el ACP de Z^{φ}.

Obteniendo los valores y vectores propios se verifica a modo de ejemplo las coordenadas del individuo L para el primer componente prinicipal $[1.2, 3.05]$, las coordenadas en este ACP simbólico vienen dadas por el teorema 3.1.

Para esta verificación se realizan los siguientes pasos:

1. Se obtienen las coordenadas de los vértices del objeto simbólico L.

	comp 1	comp 2	comp 3	comp 4
L.1	2.66	-1.60	-0.60	0.11
L.2	2.75	-1.52	-0.60	0.23
L.3	2.45	-1.71	-0.36	0.34
L.4	2.54	-1.63	-0.36	0.46
L.5	2.96	-1.64	-0.16	-0.09
L.6	3.05	-1.56	-0.16	0.03
L.7	2.74	-1.75	0.07	0.13
L.8	2.83	-1.67	0.08	0.25
L.9	1.42	0.88	0.15	-0.64
L.10	1.51	0.96	0.15	-0.52
L.11	1.20	0.77	0.39	-0.42
L.12	1.29	0.85	0.39	-0.30
L.13	1.71	0.84	0.59	-0.85
L.14	1.80	0.92	0.59	-0.73
L.15	1.50	0.73	0.82	-0.62
L.16	1.59	0.81	0.83	-0.50

TABLA 4.5: Coordenadas de los vértices (individuos suplementarios) para el ACP de Z^{φ}.

2. Obtener el valor mínimo y máximo para el primer componente 1.20 y 3.05 respectivamente.

3. Obtener $\overline{Z^{\varphi}} = (0.9087, 0.875, 109.875, 187.00)$.

4. Obtener $\sigma_{Z^{\varphi}} = (0.02957, 22.21732, 59.79849, 26.22975)$.

5. Considerar el primer vector propio de la matriz de varianza covarianza de $w_1^{Z^{\varphi}} = (0.5287, -0.5266, 0.5166, -0.4198)$.

6. Por Teorema 3.1 se tiene que:

$$y_{11}^{\tilde{a}_{Z^\varphi}} = \underbrace{\left(\frac{-18-0.875}{22.21732}\right)(-0.5266)+\left(\frac{196-187.00}{26.22975}\right)(-0.4198)}_{\sum_{j\in J_Z^-}\left(b_{ij}-\overline{Z}_{(j)}\right)w_{kj}^Z} + \underbrace{\left(\frac{0.93-0.9087}{0.03}\right)(0.02957)+\left(\frac{170-109.875}{59.79849}\right)(0.5166)}_{\sum_{j\in J_Z^+}\left(a_{ij}-\overline{Z}_{(j)}\right)w_{kj}^Z}$$

$$= 1.20.$$

$$y_{11}^{\tilde{b}_{Z^\varphi}} = \underbrace{\left(\frac{0.94-0.9087}{0.02957}\right)(0.5287)+\left(\frac{204-109.875}{59.79849}\right)(0.5166)}_{\sum_{j\in J_Z^+}\left(b_{ij}-\overline{Z}_{(j)}\right)w_{kj}^Z} + \underbrace{\left(\frac{-27-0.875}{22.21732}\right)(-0.5266)+\left(\frac{118-187}{26.22975}\right)(-0.4198)}_{\sum_{j\in J_Z^-}\left(a_{ij}-\overline{Z}_{(j)}\right)w_{kj}^Z}$$

$$= 3.05.$$

7. De acuerdo al item 2 y al 6 las coordenadas para el objeto simbólico L son iguales, verificando a manera de ejemplo con un caso particular el teorema 3 1.

Las coordenadas en el ACP de Z^φ se encuentran en la tabla 4 6.

	I comp 1	I comp 2	I comp 3	I comp 4
L	[1.2,3.05]	[-1.75,0.96]	[-0.6,0.83]	[-0.85,0.46]
P	[1.04,1.48]	[0.29,0.73]	[0.93,1.26]	[-0.33,0.04]
Co	[-0.1,0.32]	[0.2,0.6]	[-0.29,0.11]	[-0.12,0.22]
S	[0.17,0.52]	[0.23,0.56]	[-0.25,0.03]	[0,0.33]
Ca	[0.15,0.49]	[0.41,0.68]	[-1.04,-0.71]	[-0.36,-0.04]
O	[-0.44,0.03]	[0.05,0.5]	[-0.42,-0.03]	[0.14,0.56]
B	[-2.55,-1.97]	[-1.05,-0.5]	[-0.15,0.26]	[-0.18,0.4]
H	[-2.38,-1.63]	[-1.05,-0.42]	[-0.2,0.5]	[-0.63,0.03]

TABLA 4.6: Coordenadas de los individuos suplementarios (vértices) para el ACP de Z^φ.

La calidad de la representación de cada individuo en cada eje principal viene dada por el coseno cuadrado, se realiza un ejemplo con el objeto simbólico L, los datos de los cosenos cuadrados los podemos observar en la tabla 4.7.

	comp 1	comp 2	comp 3	comp 4
L.1	0.71	0.25	0.04	0.00
L.2	0.74	0.22	0.03	0.01
L.3	0.66	0.32	0.01	0.01
L.4	0.68	0.28	0.01	0.02
L.5	0.76	0.24	0.00	0.00
L.6	0.79	0.21	0.00	0.00
L.7	0.71	0.29	0.00	0.00
L.8	0.74	0.26	0.00	0.01
L.9	0.62	0.24	0.01	0.13
L.10	0.65	0.27	0.01	0.08
L.11	0.61	0.25	0.06	0.07
L.12	0.63	0.28	0.06	0.03
L.13	0.62	0.15	0.07	0.15
L.14	0.65	0.17	0.07	0.11
L.15	0.58	0.14	0.18	0.10
L.16	0.61	0.16	0.17	0.06

TABLA 4.7: Cosenos cuadrados de los vértices de L en el ACP de Z^{φ}.

Los cosenos cuadrados de los vértices del objeto simbólico L son:

- Primer componente principal: $[0.58, 0.79]$.
- Segundo componente principal: $[0.14, 0.32]$.
- Tercer componente principal: $[0, 0.18]$.
- Cuarto componente principal: $[0, 0.15]$.

Los cosenos cuadrados para los demás objetos simbólicos se detallan en la tabla 4.8.

	I comp 1	I comp 2	I comp 3
B	[0.83,0.99]	[0.01,0.16]	[0,0.02]
Ca	[0.02,0.16]	[0.15,0.43]	[0.49,0.74]
Co	[0,0.56]	[0.32,0.97]	[0,0.49]
H	[0.75,0.98]	[0,0.2]	[0,0.07]
L	[0.44,0.55]	[0.23,0.52]	[0.01,0.2]
O	[0,0.4]	[0,0.8]	[0,0.71]
P	[0.3,0.64]	[0.03,0.3]	[0.29,0.47]
S	[0.05,0.8]	[0.11,0.86]	[0,0.31]

TABLA 4.8: Cosenos cuadrados de los individuos suplementarios para el ACP de Z^{φ}.

Los individuos mejor representados (cuyo mínimo es mayor que 0.5) en el primer componente principal de Z^{φ} son B y H.

Para los demás consideramenos más dimensiones para que puedan ser representados de mejor manera. En la tabla 4.9 se representantan la calidades de los individuos en 2 y 3 dimensiones.

	Componentes 1 y 2	Componentes 1,2 y 3
B	[0.96,1]	[0.96,1]
Ca	[0.24,0.46]	[0.88,0.99]
Co	[0.4,1]	[0.69,1]
H	[0.89,1]	[0.92,1]
L	[0.71,0.98]	[0.85,1]
O	[0.01,0.92]	[0.12,0.98]
P	[0.51,0.7]	[0.98,1]
S	[0.59,1]	[0.75,1]

TABLA 4.9: Calidades de los individuos suplementarios para el ACP de Z^{φ}.

Con dos componentes principales los individuos que se encuentran bien representados (cuyo mínimo es mayor que 0.69) son B, H y L.

Los objetos simbólicos Co y Ca necesitan tres componentes principales para poder representarse de buena manera, además el individuo O necesita más de tres componentes principales para estar bien representado.

La figura 4 2 muestra los individuos en el primer plano principal, mientras que la figura 4 3 muestra el círculo de correlaciones.

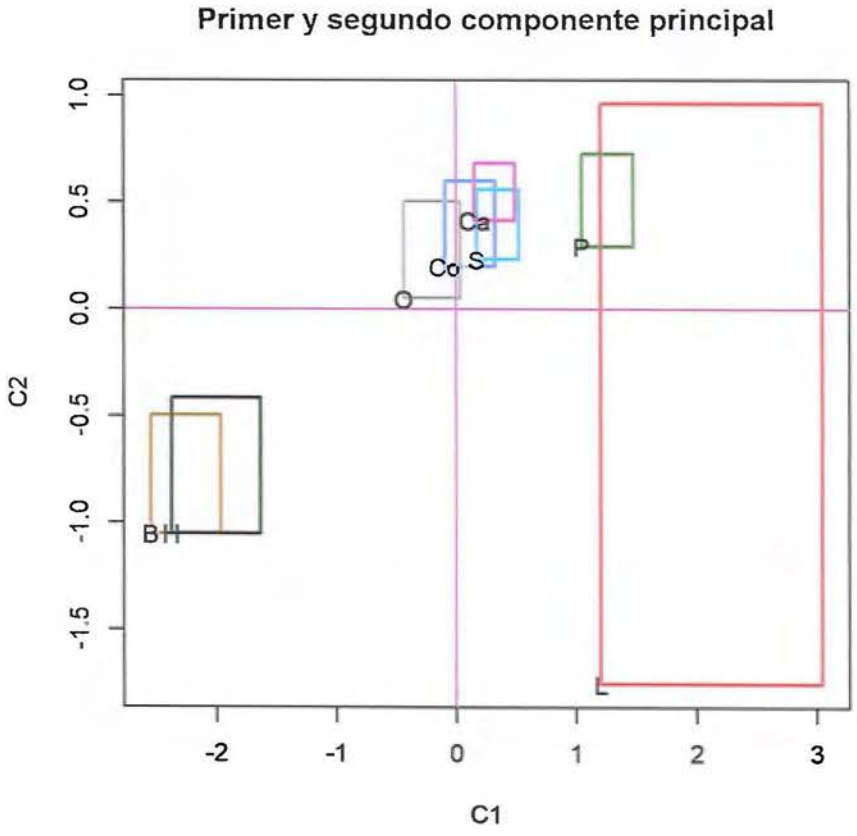


FIGURA 4.2: Primer y segundo componente principal de Z^{φ}.

En la figura 4 2 se puede notar una segmentación de los aceites vegetales (Co, Ca, P, O, L y B) y animales (B y H).

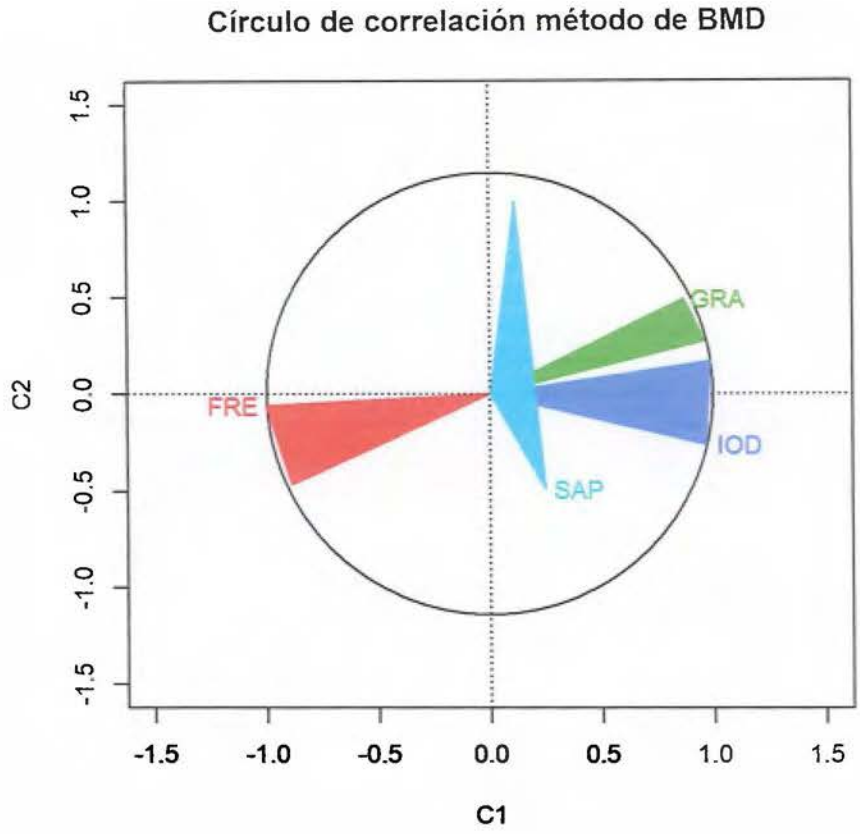


FIGURA 4.3: Círculo de correlaciones del primer y segundo componente principal de Z^{φ}.

En la figura 4 3 puede apreciarse:

- Las variables *IOD*, *SAP* y *GRA* poseen correlación positiva con el primer componente principal.
- Las variables *SAP* y *GRA* poseen una alta correlación con el primer componente principal.
- La variable *FRE* posee una alta correlación ngativa con el primer componente principal.
- Las variables *FRE* y *GRA* se encuentran correlacionadas negativamente.

- La variable SAP se correlaciona de mejor manera con el segundo componente principal.

Si se realiza un análisis de dualidad sobre el primer plano principal, utilizando las figuras 4.2 y 4.3, se puede notar:

- Los aceites animales se encuentran el tercer cuadrante según la figura 4.2, esto quiere decir que estos se congelan a una mayor temperatura.
- Los aceites vegetales (Ca, S y P) poseen una mayor gravedad específica y necesitan de una menor temperatura para congelarse.

4.1.1.2. Matriz óptima respecto a la varianza (MOV)

La matriz que resuelve el problema 3.28 es

	GRA	FRE	IOD	SAP
L	0.93	-27.00	190.21	167.02
P	0.94	-4.00	195.23	188.00
Co	0.92	-2.79	105.77	192.32
S	0.92	-6.00	111.34	193.00
Ca	0.92	-15.00	80.00	189.00
O	0.91	0.00	90.00	196.00
B	0.87	32.85	44.50	195.92
H	0.86	32.00	60.58	190.00

TABLA 4.10: Matriz Z^{A}.

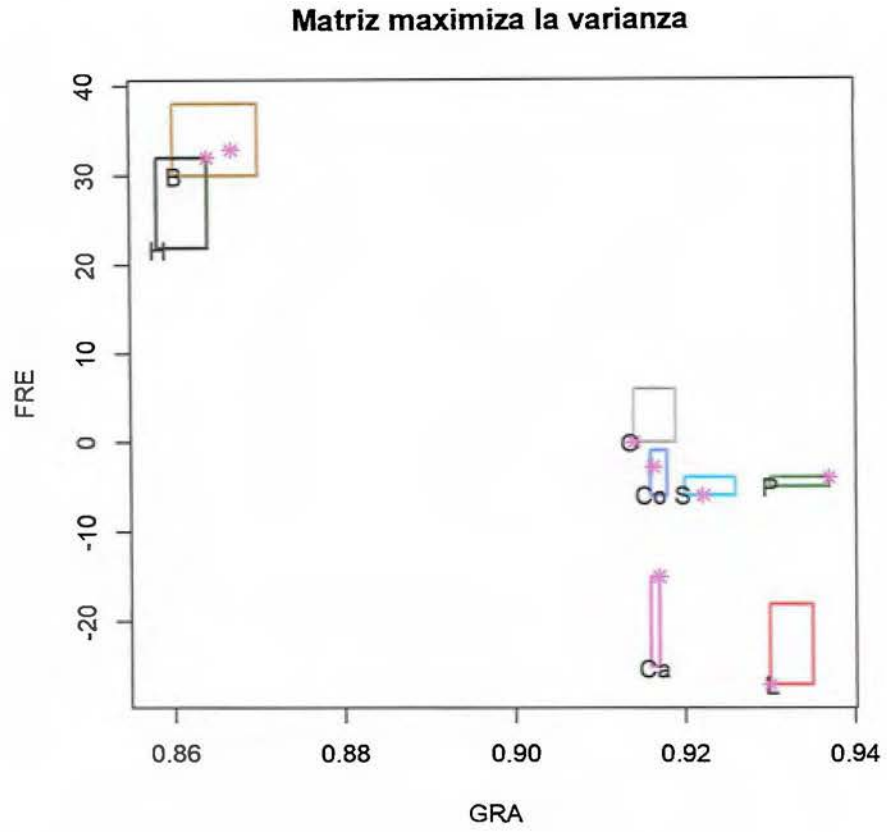


FIGURA 4.4: Matriz con mejor varianza.

La varianza explicada por el primer componente principal del método de maximización de la varianza es de $\frac{\lambda_1}{\sum \lambda_i} = 76.79\,\%$. De igual manera se puede tomar más componentes principales, si se utiliza 2 o 3 componentes principales, la varianza acumulada será de 93.11 % y 100.00 % respectivamente. En la tabla 4.11, se pueden observar todos los valores propios para este caso.

	Valor propio	Porcentaje de varianza	Porcentaje acumulado de varianza
λ_1	3.07	76.79 %	76.79 %
λ_2	0.65	16.32 %	93.11 %
λ_3	0.28	6.89 %	100.00 %
λ_4	0.00	0.00 %	100.00 %

TABLA 4.11: Valores propios para el ACP de Z^{Λ}.

Los vectores propios se muestran en la tabla 4.12.

	comp 1	comp 2	comp 3	comp 4
GRA	0.52	0.52	0.06	0.68
FRE	-0.53	-0.25	0.61	0.54
IOD	0.52	-0.11	0.75	-0.39
SAP	-0.42	0.81	0.24	-0.32

TABLA 4.12: Vectores propios para el ACP de Z^{Λ}.

Las coordenadas en el ACP de Z^{Λ} se encuentran en la tabla 4.13.

	I comp 1	I comp 2	I comp 3	I comp 4
L	[1.21,5.69]	[-6.11,1.43]	[-1.93,1.03]	[-1.17,2.33]
P	[1,1.77]	[0.22,1.24]	[1.02,1.55]	[-0.64,0.02]
Co	[-0.34,0.41]	[0.18,1.15]	[-0.36,0.25]	[-0.36,0.27]
S	[0.12,0.7]	[0.11,0.84]	[-0.33,0.09]	[-0.09,0.43]
Ca	[0.09,0.59]	[0.42,0.94]	[-1.22,-0.77]	[-0.46,0]
O	[-0.67,0.14]	[-0.08,0.95]	[-0.52,0.09]	[0,0.71]
B	[-3.14,-2.21]	[-1.21,-0.05]	[-0.2,0.43]	[-0.4,0.47]
H	[-3.04,-1.83]	[-1.23,0.18]	[-0.27,0.73]	[-0.99,0.06]

TABLA 4.13: Coordenadas de los individuos suplementarios para el ACP de Z^{Λ}.

Los cosenos cuadrados para este ACP vienen dados en la tabla 4.14.

	I comp 1	I comp 2	I comp 3
B	[0.82,1]	[0,0.17]	[0,0.02]
Ca	[0.01,0.16]	[0.14,0.51]	[0.46,0.76]
Co	[0,0.62]	[0.27,0.94]	[0,0.48]
H	[0.71,0.97]	[0,0.23]	[0,0.07]
L	[0.36,0.51]	[0.23,0.53]	[0.01,0.18]
O	[0,0.4]	[0,0.88]	[0,0.61]
P	[0.24,0.68]	[0.01,0.33]	[0.28,0.45]
S	[0.03,0.85]	[0.04,0.95]	[0,0.28]

TABLA 4.14: Cosenos cuadrados de los individuos suplementarios para el ACP de Z^{Λ}.

Los individuos mejor representados (cuyo mínimo es mayor que 0.5) en el primer componente principal de Z^{Λ} son B y H.

Para los demás consideramenos más dimensiones para que puedan ser representados de mejor manera. En la tabla 4.15 se representan la calidades de los individuos en 2 y 3 dimensiones.

	Componentes 1 y 2	Componentes 1,2 y 3
B	[0.97,1]	[0.97,1]
Ca	[0.24,0.52]	[0.91,1]
Co	[0.39,0.99]	[0.63,1]
H	[0.85,1]	[0.87,1]
L	[0.66,0.93]	[0.8,0.96]
O	[0,0.99]	[0.07,1]
P	[0.47,0.72]	[0.92,1]
S	[0.56,0.99]	[0.69,1]

TABLA 4.15: Calidades de los individuos suplementarios para el ACP de Z^{Λ}.

Con dos componentes principales los individuos que se encuentran bien representados (cuyo mínimo es mayor que 0.69) son B y H.

Los objetos simbólicos Co y P necesitan tres componentes principales para poder representarse de buena manera, además el individuo O necesita más de tres componentes principales para estar bien representado.

La figura 4 5 muestra los individuos en el primer plano principal, mientras que la figura 4 6 muestra el círculo de correlaciones.

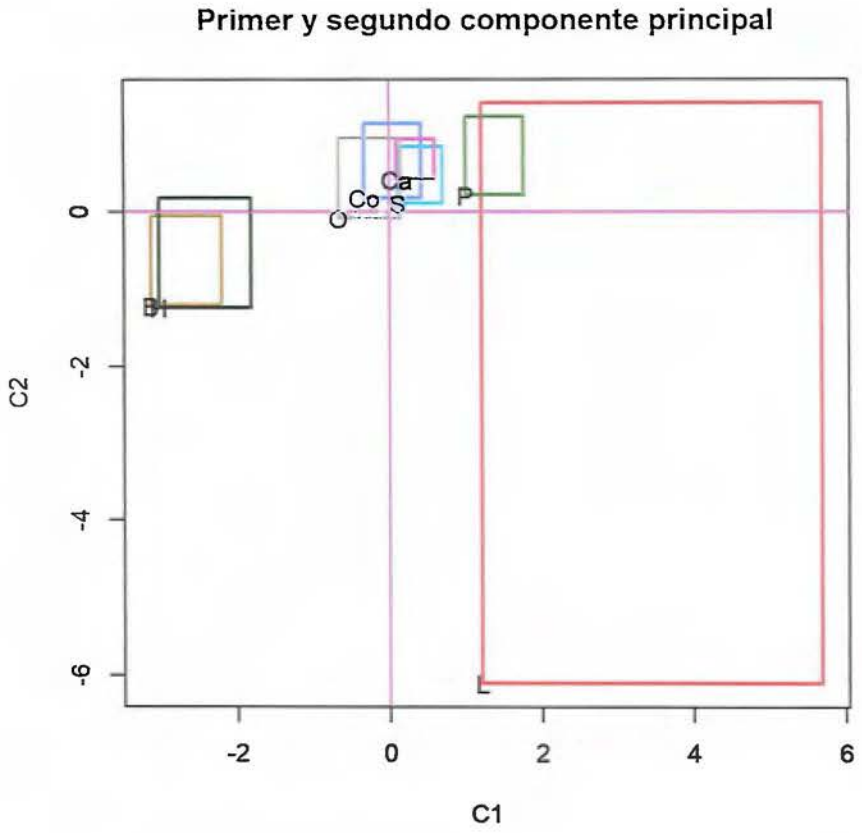


FIGURA 4.5: Primer y segundo componente principal de Z^{Λ}.

En la figura 4 5 se puede notar una segmentación de los aceites vegetales (Co, Ca, P, O, L y B) y animales (B y H).

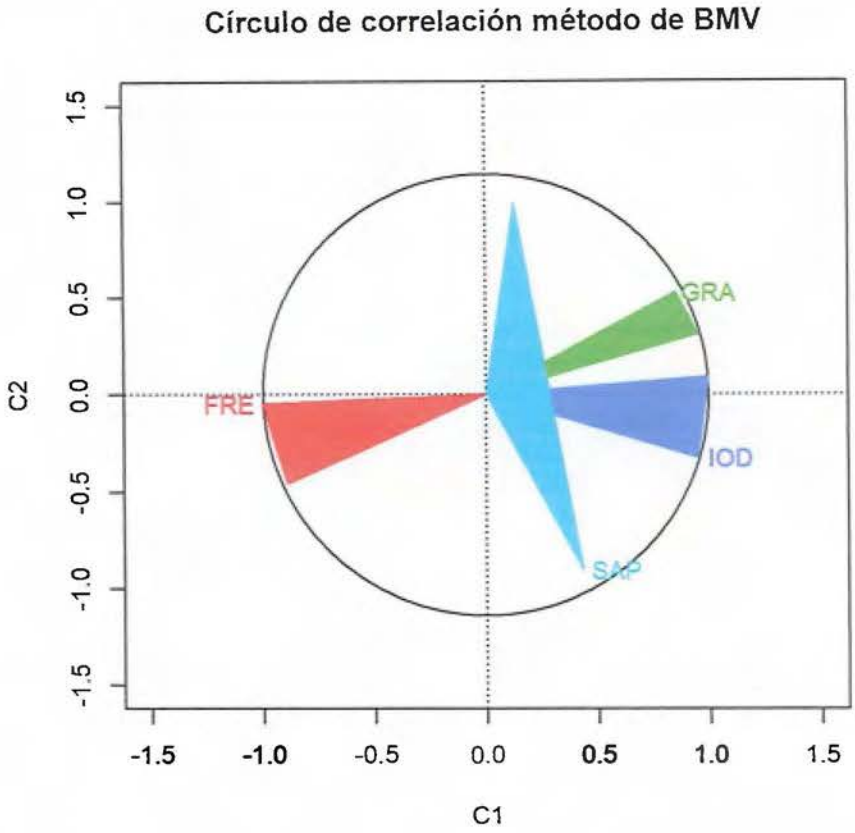


FIGURA 4.6: Círculo de correlaciones del primer y segundo componente principal de Z^{Λ}.

En la figura 4 6 puede apreciarse:

- Las variables *IOD*, *SAP* y *GRA* poseen correlación positiva con el primer componente principal.
- Las variables *SAP* y *GRA* poseen una alta correlación con el primer componente principal.
- La variable *FRE* posee una alta correlación ngativa con el primer componente principal.
- Las variables *FRE* y *GRA* se encuentran correlacionadas negativamente.

- La variable *SAP* se correlaciona de mejor manera con el segundo componente principal.

Si se realiza un análisis de dualidad sobre el primer plano principal, utilizando las figuras 4.5 y 4.6, se puede notar:

- Los aceites animales se encuentran el tercer cuadrante según la figura 4.5, esto quiere decir que estos se congelan a una mayor temperatura.
- Los aceites vegetales (*Ca*, *S* y *P*) se encuentran en el poseen una mayor gravedad específica y necesitan de una menor temperatura para congelarse.

4.1.1.3. Método de centros (CM)

La varianza explicada por el primer componente principal del método de centros es de $\frac{\lambda_1}{\sum \lambda_i} =$ 76.79 %. De igual manera se puede tomar más componentes principales, si se utilizan 2 o 3 componentes principales, la varianza acumulada será de 89.77 % y 98.73 % respectivamente. En la tabla 4.16, se pueden observar todos los valores propios para este caso.

	Valor propio	Porcentaje de varianza	Porcentaje acumulado de varianza
λ_1	2.98	74.60 %	74.60 %
λ_2	0.61	15.17 %	89.77 %
λ_3	0.36	8.95 %	98.73 %
λ_4	0.05	1.27 %	100.00 %

TABLA 4.16: Valores propios para el ACP de centros.

Los vectores propios se muestran en la tabla 4.17.

	comp 1	comp 2	comp 3	comp 4
GRA	0.53	0.45	-0.04	0.71
FRE	-0.53	-0.27	0.54	0.60
IOD	0.51	-0.04	0.81	-0.30
SAP	-0.43	0.85	0.23	-0.21

TABLA 4.17: Vectores propios para el ACP de centros.

Las coordenadas en el ACP de centros se encuentran en la tabla 4.18.

	I comp 1	I comp 2	I comp 3	I comp 4
L	[1.21,5.69]	[-6.11,1.43]	[-1.93,1.03]	[-1.17,2.33]
P	[1,1.77]	[0.22,1.24]	[1.02,1.55]	[-0.64,0.02]
Co	[-0.34,0.41]	[0.18,1.15]	[-0.36,0.25]	[-0.36,0.27]
S	[0.12,0.7]	[0.11,0.84]	[-0.33,0.09]	[-0.09,0.43]
Ca	[0.09,0.59]	[0.42,0.94]	[-1.22,-0.77]	[-0.46,0]
O	[-0.67,0.14]	[-0.08,0.95]	[-0.52,0.09]	[0,0.71]
B	[-3.14,-2.21]	[-1.21,-0.05]	[-0.2,0.43]	[-0.4,0.47]
H	[-3.04,-1.83]	[-1.23,0.18]	[-0.27,0.73]	[-0.99,0.06]

TABLA 4.18: Coordenadas de los individuos suplementarios para el ACP de centros.

Los cosenos cuadrados para este ACP vienen dados en la tabla 4.19.

	I comp 1	I comp 2	I comp 3
B	[0.83,0.99]	[0.01,0.16]	[0,0.02]
Ca	[0.02,0.16]	[0.15,0.43]	[0.49,0.74]
Co	[0,0.56]	[0.32,0.97]	[0,0.49]
H	[0.75,0.98]	[0,0.2]	[0,0.07]
L	[0.44,0.55]	[0.23,0.52]	[0.01,0.2]
O	[0,0.4]	[0,0.8]	[0,0.71]
P	[0.3,0.64]	[0.03,0.3]	[0.29,0.47]
S	[0.05,0.8]	[0.11,0.86]	[0,0.31]

TABLA 4.19: Cosenos cuadrados de los individuos suplementarios para el ACP de centros.

Los individuos mejor representados (cuyo mínimo es mayor que 0.5) en el primer componente principal de centros son B y H.

Para los demás consideramenos más dimensiones para que puedan ser representados de mejor manera. En la tabla 4.20 se representantan la calidades de los individuos en 2 y 3 dimensiones.

	Componentes 1 y 2	Componentes 1,2 y 3
B	[0.96,1]	[0.96,1]
Ca	[0.24,0.46]	[0.88,0.99]
Co	[0.4,1]	[0.69,1]
H	[0.89,1]	[0.92,1]
L	[0.71,0.98]	[0.85,1]
O	[0.01,0.92]	[0.12,0.98]
P	[0.51,0.7]	[0.98,1]
S	[0.59,1]	[0.75,1]

TABLA 4.20: Calidades de los individuos suplementarios para el ACP de centros.

Con dos componentes principales los individuos que se encuentran bien representados (cuyo mínimo es mayor que 0.69) son B, H y L.

Los objetos simbólicos Ca y P necesitan tres componentes principales para poder representarse de buena manera, además el individuo O necesita más de tres componentes principales para estar bien representado.

La figura 4.7 muestra los individuos en el primer plano principal, mientras que la figura 4.8 muestra el círculo de correlaciones.

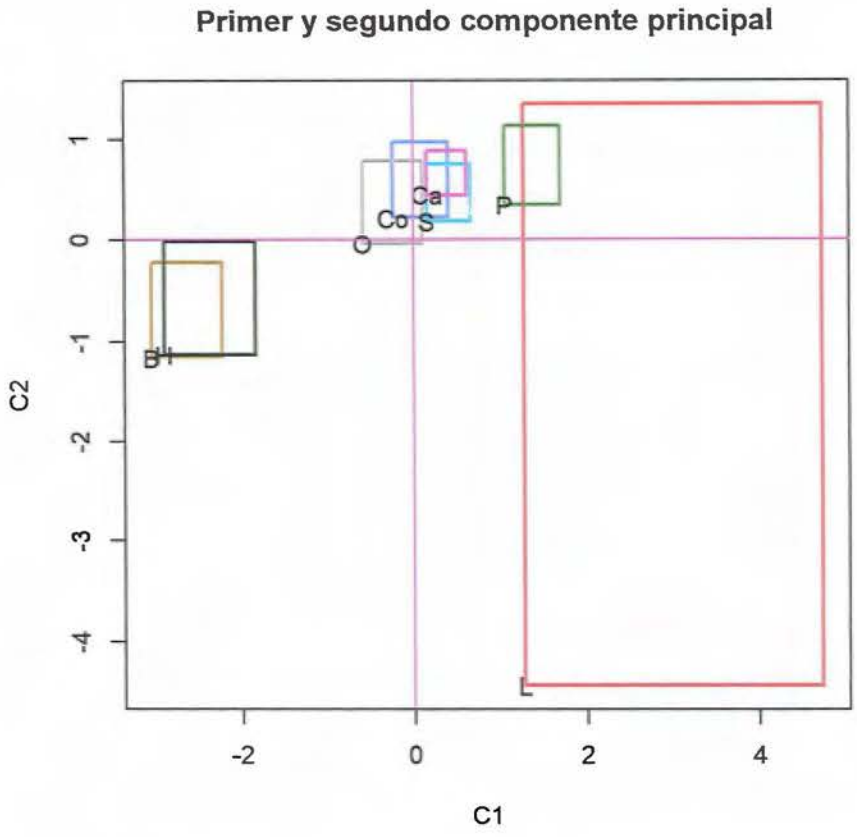


FIGURA 4.7: Primer y segundo componente principal de centros.

En la figura 4.7 se puede notar una segmentación de los aceites vegetales (Co, Ca, P, O, L y B) y animales (B y H).

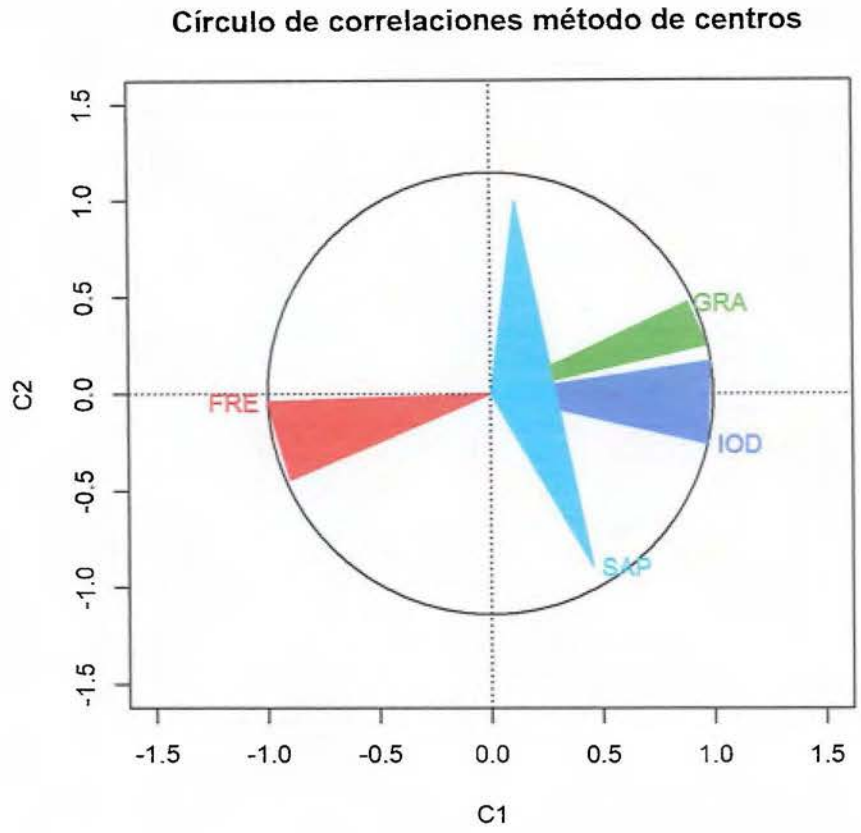


FIGURA 4.8: Círculo de correlaciones del primer y segundo componente principal de centros.

En la figura 4 8 puede apreciarse:

- Las variables *IOD*, *SAP* y *GRA* poseen correlación positiva con el primer componente principal.
- Las variable *IOD* y *GRA* poseen una alta correlación con el primer componente principal.
- La variable *FRE* posee una alta correlación negativa con el primer componente principal.
- Las variables *FRE* y *GRA* se encuentran correlacionadas negativamente.

- La variable *SAP* se correlaciona de mejor manera con el segundo componente principal.

Si se realiza un análisis de dualidad sobre el primer plano principal, se utilizan las figuras 4.7 y 4.8 se puede notar:

- Los aceites animales se encuentran el tercer cuadrante según la figura 4.7, esto quiere decir que estos se congelan a una mayor temperatura.
- Los aceites vegetales (*Ca*, *S* y *P*) se encuentran en el poseen una mayor gravedad específica y necesitan de una menor temperatura para congelarse.

4.1.1.4. Método de vértices (VM)

La varianza explicada por el primer componente principal del método de centros es de $\frac{\lambda_1}{\sum \lambda_i} =$ 67.76 %. De igual manera se puede tomar más componentes principales, si se utilizan 2 o 3 componentes principales, la varianza acumulada será de 88.02 % y 97.77 % respectivamente. En la tabla 4.21, se pueden observar todos los valores propios para este caso.

	Valor propio	Porcentaje de varianza	Porcentaje acumulado de varianza
λ_1	2.71	67.76 %	67.76 %
λ_2	0.81	20.26 %	88.02 %
λ_3	0.39	9.75 %	97.77 %
λ_4	0.09	2.23 %	100.00 %

TABLA 4.21: Valores propios para el ACP de vértices.

Los vectores propios se muestran en la tabla 4.22.

	comp 1	comp 2	comp 3	comp 4
GRA	0.57	0.29	-0.14	0.76
FRE	-0.55	-0.19	0.56	0.59
IOD	0.52	0.07	0.81	-0.26
SAP	-0.33	0.94	0.10	-0.10

TABLA 4.22: Vectores propios para el ACP de vértices.

Las coordenadas en el ACP de vértices se encuentran en la tabla 4.23.

	I comp 1	I comp 2	I comp 3	I comp 4
L	[1.46,3.51]	[-3.02,1.07]	[-0.41,0.8]	[-0.76,0.25]
P	[1.26,1.75]	[0.38,0.94]	[0.92,1.28]	[-0.11,0.25]
Co	[-0.05,0.43]	[0.13,0.67]	[-0.42,0]	[-0.07,0.26]
S	[0.26,0.66]	[0.11,0.51]	[-0.39,-0.09]	[0.06,0.38]
Ca	[0.25,0.65]	[0.24,0.55]	[-1.26,-0.91]	[-0.47,-0.1]
O	[-0.45,0.09]	[-0.09,0.49]	[-0.56,-0.15]	[0.19,0.62]
B	[-2.94,-2.26]	[-0.88,-0.24]	[0,0.45]	[-0.21,0.4]
H	[-2.73,-1.88]	[-0.83,-0.04]	[-0.01,0.74]	[-0.67,-0.01]

TABLA 4.23: Coordenadas de los individuos para el ACP de vértices.

Los cosenos cuadrados para este ACP vienen dados en la tabla 4.24.

	I comp 1	I comp 2	I comp 3
B	[0.88,0.99]	[0.01,0.09]	[0,0.02]
Ca	[0.05,0.19]	[0.05,0.18]	[0.64,0.84]
Co	[0,0.67]	[0.12,0.96]	[0,0.63]
H	[0.8,0.98]	[0,0.1]	[0,0.08]
L	[0.47,0.72]	[0.17,0.53]	[0,0.13]
O	[0,0.33]	[0,0.53]	[0.05,0.74]
P	[0.46,0.69]	[0.04,0.22]	[0.23,0.37]
S	[0.19,0.82]	[0.04,0.53]	[0.02,0.38]

TABLA 4.24: Cosenos cuadrados de los individuos para el ACP de vértices.

Los individuos mejor representados (cuyo mínimo es mayor que 0.5) en el primer componente principal de centros son B y H.

Para los demás consideramenos más dimensiones para que puedan ser representados de mejor manera. En la tabla 4.25 se representantan la calidades de los individuos en 2 y 3 dimensiones.

	Componentes 1 y 2	Componentes 1,2 y 3
B	[0.96,1]	[0.98,1]
Ca	[0.15,0.28]	[0.9,0.99]
Co	[0.2,0.98]	[0.63,1]
H	[0.88,0.99]	[0.92,1]
L	[0.82,1]	[0.91,1]
O	[0,0.62]	[0.15,0.9]
P	[0.62,0.75]	[0.98,1]
S	[0.51,0.94]	[0.72,0.99]

TABLA 4.25: Calidades de los individuos para el ACP de vértices.

Con dos componentes principales los individuos que se encuentran bien representados (cuyo mínimo es mayor que 0.69) son B, H y L.

Los objetos simbólicos Ca y P necesitan tres componentes principales para poder representarse de buena manera, además el individuo O necesita más de tres componentes principales para estar bien representado.

La figura 4.9 muestra los individuos en el primer plano principal.

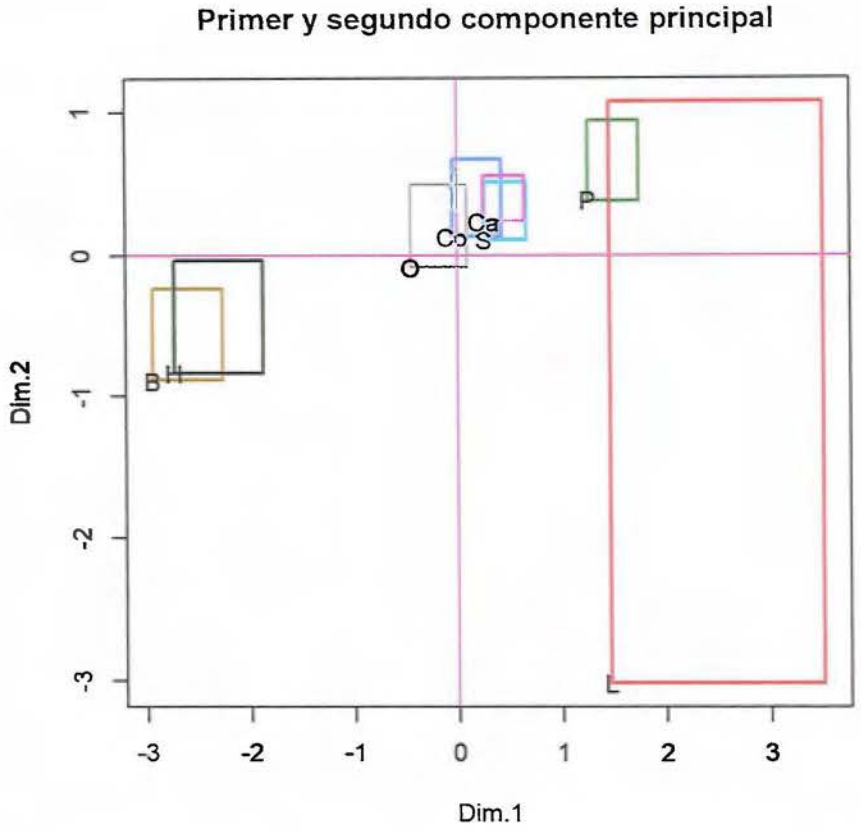


FIGURA 4.9: Primer y segundo componente principal de vértices.

En la figura 4 9 se puede notar una segmentación de los aceites vegetales (*Co*, *Ca*, *P*, *O*, *L* y *B*) y animales (*B* y *H*).

4.1.1.5. Comparación de métodos

En esta sección se realizará una comparación de los métodos antes expuestos, utilizando los datos de la tabla 4 1.

4.1.1.5.1 Varianza

En la tabla 4.26, se muestra la varianza acumulada en cada uno de los componentes principales, el método de vértices es el que presenta una menor varianza en los primeros tres componentes principales 97.77 %, mientras que el ACP de maximización de varianza obtiene un 100 % de la varianza en los primeros tres componentes principales.

	Vértices	Centros	Z^{φ}	Z^{Λ}
comp 1	67.76 %	74.60 %	73.98 %	76.79 %
comp 2	88.02 %	89.77 %	90.28 %	93.11 %
comp 3	97.77 %	98.73 %	98.63 %	100.00 %
comp 4	100.00 %	100.00 %	100.00 %	100.00 %

TABLA 4.26: Comparación de la varianza para diferentes ACP, datos de aceite.

Independiente del componente principal la matriz Z^{Λ} es mejor (respecto a la varianza acumulada) en todos los componentes.

4.1.1.5.2 Distancias

En la tabla 4.27, se muestran las distancias de los vértices a cada uno de los componentes principales, para este caso la mayor distancia (982.12) se alcanza en la matriz Z^{Λ} y la menor distancia (368.51) se obtiene con Z^{φ}.

Vértices	Centros	Z^{φ}	Z^{Λ}
512.00	706.80	368.51	982.12

TABLA 4.27: Comparación de la distancia para diferentes ACP, datos de aceite.

4.1.1.5.3 Cosenos cuadrados

En las tablas 4.28 y 4.29 se muestran los valores de los cosenos cuadrados para cada individuo, se puede notar que entre los diversos tipos de ACP existe consistencia en estos

resultados, esto quiere decir que un individuo se encuentra muy bien representado (mayor 80 %) en un tipo de ACP, en los demás también se encuentra muy bien representado.

1. **Primer componente principal.**

	Vértices	Centros	Z^{φ}	Z^{Λ}
B	[0.88,0.99]	[0.83,0.99]	[0.83,0.95]	[0.82,1]
Ca	[0.05,0.19]	[0.02,0.16]	[0.03,0.14]	[0.01,0.16]
Co	[0,0.67]	[0,0.56]	[0,0.52]	[0,0.62]
H	[0.8,0.98]	[0.75,0.98]	[0.73,0.94]	[0.71,0.97]
L	[0.47,0.72]	[0.44,0.55]	[0.58,0.79]	[0.36,0.51]
O	[0,0.33]	[0,0.4]	[0,0.41]	[0,0.4]
P	[0.46,0.69]	[0.3,0.64]	[0.41,0.62]	[0.24,0.68]
S	[0.19,0.82]	[0.05,0.8]	[0.11,0.69]	[0.03,0.85]

TABLA 4.28: Comparación del coseno cuadrado (primer componente principal) para diferentes ACP, datos de aceite.

Los individuos muy bien representados en el primer componente principal, son los aceites de origen animal B y H, mientras que el aceite L solamente se encuentra bien representado en el ACP de Z^{φ}, para los datos de aceite el ACP que cuenta con mayor número de individuos bien representados es el ACP de Z^{φ}.

2. **Segundo componente principal.**

	Vértices	Centros	Z^{φ}	Z^{Λ}
B	[0.01,0.09]	[0.01,0.16]	[0.05,0.17]	[0,0.17]
Ca	[0.05,0.18]	[0.15,0.43]	[0.17,0.36]	[0.14,0.51]
Co	[0.12,0.96]	[0.32,0.97]	[0.4,0.95]	[0.27,0.94]
H	[0,0.1]	[0,0.2]	[0.04,0.21]	[0,0.23]
L	[0.17,0.53]	[0.23,0.52]	[0.14,0.32]	[0.23,0.53]
O	[0,0.53]	[0,0.8]	[0.01,0.68]	[0,0.88]
P	[0.04,0.22]	[0.03,0.3]	[0.03,0.18]	[0.01,0.33]
S	[0.04,0.53]	[0.11,0.86]	[0.26,0.75]	[0.04,0.95]

TABLA 4.29: Comparación del coseno cuadrado (segundo componente principal) para diferentes ACP, datos de aceite.

Para el segundo componente principal no existe ningún individuo bien representado.

4.1.1.5.4 Calidades

La calidad de un individuo en p componentes principales (no necesariamente consecutivos) es la suma de sus cosenos cuadrados en dichos p componentes principales.

1. Componentes principales 1 y 2.

	Vértices	Centros	Z^{φ}	Z^{Λ}
B	[0.96,1]	[0.96,1]	[0.97,1]	[0.97,1]
Ca	[0.15,0.28]	[0.24,0.46]	[0.27,0.4]	[0.24,0.52]
Co	[0.2,0.98]	[0.4,1]	[0.47,0.98]	[0.39,0.99]
H	[0.88,0.99]	[0.89,1]	[0.9,1]	[0.85,1]
L	[0.82,1]	[0.71,0.98]	[0.72,1]	[0.66,0.93]
O	[0,0.62]	[0.01,0.92]	[0.1,0.79]	[0,0.99]
P	[0.62,0.75]	[0.51,0.7]	[0.51,0.69]	[0.47,0.72]
S	[0.51,0.94]	[0.59,1]	[0.65,1]	[0.56,0.99]

TABLA 4.30: Comparación de las coordenadas en el primer plano principal, para diferentes ACP, datos de aceite.

Los únicos individuos muy bien representados en el primer plano principal (para todos los ACP) son B y H, las calidades para el individuo B se encuentran entre $[0.97, 1]$ para todos los tipos de ACP, mientras que para el individuo H ronda entre $[0.85, 1]$ para todos los tipos de ACP.

2. **Componentes principales 1, 2 y 3.**

	Vértices	Centros	Z^{φ}	Z^{Λ}
B	[0.98,1]	[0.96,1]	[0.97,1]	[0.97,1]
Ca	[0.9,0.99]	[0.88,0.99]	[0.92,1]	[0.91,1]
Co	[0.63,1]	[0.69,1]	[0.66,1]	[0.63,1]
H	[0.92,1]	[0.92,1]	[0.91,1]	[0.87,1]
L	[0.91,1]	[0.85,1]	[0.85,1]	[0.8,0.96]
O	[0.15,0.9]	[0.12,0.98]	[0.14,0.93]	[0.07,1]
P	[0.98,1]	[0.98,1]	[0.97,1]	[0.92,1]
S	[0.72,0.99]	[0.75,1]	[0.73,1]	[0.69,1]

TABLA 4.31: Comparación de las calidades para los primeros tres componentes principales, para diferentes ACP, datos de aceite.

Los individuos muy bien representados en los componentes principales 1, 2 y 3 (para todos los ACP) son *B*, *H*, *Ca*, *L* y *P*, para los aceites vegetales es necesario tener al menos estos tres componentes principales, esto con el fin de obtener una mejor interpretación.

4.1.1.5.5 Coordenadas

En la figura 4.10 se muestra el primer plano principal, de acuerdo con el análisis de la varianza acumulada (subsección 4.1.1.5.1) la varianza oscila entre 88.02 % y 93.11 %, se puede observar lo siguiente:

1. En todos los análisis existe una segmentación entre los aceites vegetales y animales.
2. El ACP de Z^{φ} es el que más se parece al método de vértices.
3. El único aceite vegetal bien representado (análisis de calidades 4.1.1.5.4) es *L*, el cual es mejor representado por el método de vértices y después por el método Z^{φ}.

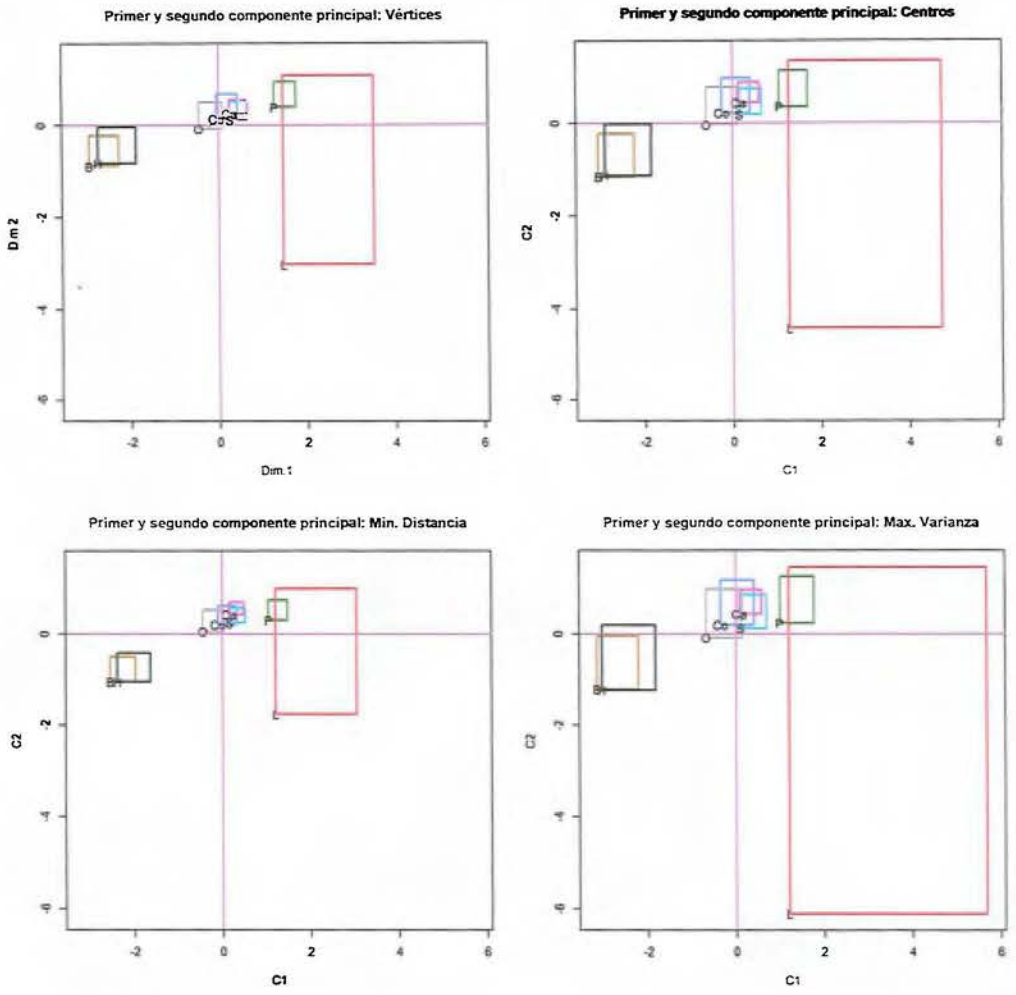


FIGURA 4.10: Comparación de ACP: datos de aceite.

4.1.2. Datos reconocimiento facial

Los datos de reconocimiento facial fueron propuestos por el profesor Edwin Diday en [Chouakria, A.; Billard, L.; Diday, E. (2011)]. Cada fila en la tabla es descrita por seis variables de tipo intervalo, las cuales son ilustradas en la figura 4 11.

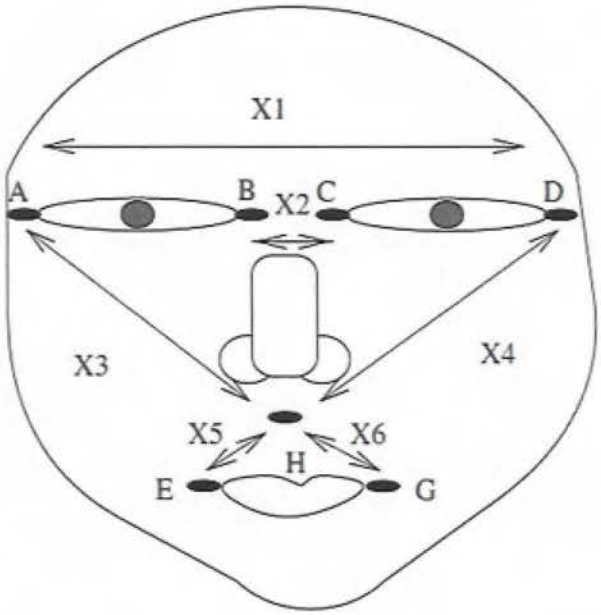


FIGURA 4.11: Descripción de las variables para reconocimiento facial.

La matriz de intervalos para reconocimiento facial es la tabla 4 32.

	AD	BC	AH	DH	EH	GH
HUS1	[168.9,172.84]	[58.55,63.39]	[102.83,106.53]	[122.38,124.52]	[56.73,61.07]	[60.44,64.54]
HUS2	[169.8,175.03]	[60.21,64.38]	[102.94,108.71]	[120.24,124.52]	[56.73,62.37]	[60.44,66.84]
HUS3	[168.8,175.15]	[61.4,63.51]	[104.35,107.45]	[120.93,125.18]	[57.2,61.72]	[58.14,67.08]
INC1	[155.3,160.45]	[53.15,60.21]	[95.88,98.49]	[91.68,94.37]	[62.48,66.22]	[58.9,63.13]
INC2	[156.3,161.31]	[51.09,60.07]	[95.77,99.36]	[91.21,96.83]	[54.92,64.2]	[54.41,61.55]
INC3	[154.5,160.31]	[55.08,59.03]	[93.54,98.98]	[90.43,96.43]	[59.03,65.86]	[55.97,65.8]
ISA1	[164,168]	[55.01,60.03]	[120.28,123.04]	[117.52,121.02]	[54.38,57.45]	[50.8,53.25]
ISA2	[163,170]	[54.04,59]	[118.8,123.04]	[116.67,120.24]	[55.47,58.67]	[52.43,55.23]
ISA3	[164,169.01]	[55,59.01]	[117.38,123.11]	[116.67,122.43]	[52.8,58.31]	[52.2,55.47]
JPL1	[167.1,171.19]	[61.03,65.01]	[118.23,121.82]	[108.3,111.2]	[63.89,67.88]	[57.28,60.83]
JPL2	[169.1,173.18]	[60.07,65.07]	[118.85,120.88]	[108.98,113.17]	[62.63,69.07]	[57.38,61.62]
JPL3	[169,170.11]	[59.01,65.01]	[115.88,121.38]	[110.34,112.49]	[61.72,68.25]	[59.46,62.94]
KHA1	[149.3,155.54]	[54.15,59.14]	[111.95,115.75]	[105.36,111.07]	[54.2,58.14]	[48.27,50.61]
KHA2	[149.3,155.32]	[52.04,58.22]	[111.2,113.22]	[105.36,111.07]	[53.71,58.14]	[49.41,52.8]
KHA3	[150.3,157.26]	[52.09,60.21]	[109.04,112.7]	[104.74,111.07]	[55.47,60.03]	[49.2,53.41]
LOT1	[152.6,157.62]	[51.35,56.22]	[116.73,119.67]	[114.62,117.41]	[55.44,59.55]	[53.01,56.6]
LOT2	[154.6,157.62]	[52.24,56.32]	[117.52,119.67]	[114.28,117.41]	[57.63,60.61]	[54.41,57.98]
LOT3	[154.8,157.81]	[50.36,55.23]	[117.59,119.75]	[114.04,116.83]	[56.64,61.07]	[55.23,57.8]
PHI1	[163.1,167.07]	[66.03,68.07]	[115.26,119.6]	[116.1,121.02]	[60.96,65.3]	[57.01,59.82]

Tabla 4.32 – Continúa en la siguiente página

Tabla 4.32 – **Continúa de la página anterior**

	AD	BC	AH	DH	EH	GH
PHI2	[164,168.03]	[65.03,68.12]	[114.55,119.6]	[115.26,120.97]	[60.96,67.27]	[55.32,61.52]
PHI3	[161,167]	[64.07,69.01]	[116.67,118.79]	[114.59,118.83]	[61.52,68.68]	[56.57,60.11]
ROM1	[167.2,171.24]	[64.07,68.07]	[123.75,126.59]	[122.92,126.37]	[51.22,54.64]	[49.65,53.71]
ROM2	[168.2,172.14]	[63.13,68.07]	[122.33,127.29]	[124.08,127.14]	[50.22,57.14]	[49.93,56.94]
ROM3	[167.1,171.19]	[63.13,68.03]	[121.62,126.57]	[122.58,127.78]	[49.41,57.28]	[50.99,60.46]

TABLA 4.32: Datos de reconocimiento facial.

4.1.2.1. Matriz óptima respecto a la distancia (MOD)

La matriz que resuelve el problema 3.25 es:

	AD	BC	AH	DH	EH	GH
HUS1	172.84	63.39	102.83	124.52	56.73	64.54
HUS2	175.03	64.38	102.94	124.52	62.37	66.84
HUS3	175.15	63.51	104.35	125.18	61.72	67.08
INC1	155.26	53.15	95.88	91.68	66.22	63.13
INC2	156.26	51.09	95.77	91.21	64.20	61.55
INC3	154.47	55.08	93.54	90.43	65.86	65.80
ISA1	168.00	55.01	123.04	121.02	54.38	50.80
ISA2	170.00	54.04	123.04	120.24	55.47	52.43
ISA3	169.01	55.00	123.11	122.43	52.80	52.20
JPL1	171.19	65.01	121.82	108.30	67.88	60.83
JPL2	173.18	65.07	120.88	108.98	69.07	61.62
JPL3	170.11	65.01	121.38	110.34	68.25	62.94
KHA1	149.34	54.15	111.95	105.36	54.20	48.27
KHA2	149.34	52.04	111.20	105.36	53.71	49.41
KHA3	150.33	52.09	109.04	104.74	55.47	49.20
LOT1	152.64	51.35	119.67	117.41	55.44	53.01
LOT2	154.64	52.24	119.67	117.41	57.63	54.41
LOT3	154.83	50.36	119.75	116.83	56.64	55.23
PHI1	167.07	68.07	119.60	121.02	65.30	59.82
PHI2	168.03	68.12	119.60	120.97	67.27	61.52
PHI3	161.01	69.01	118.79	118.83	68.68	60.11
ROM1	171.24	68.07	126.59	126.37	51.22	49.65
ROM2	172.14	68.07	127.29	127.14	50.22	49.93
ROM3	171.19	68.03	126.57	127.78	49.41	50.99

TABLA 4.33: Matriz Z^{φ} para datos faciales.

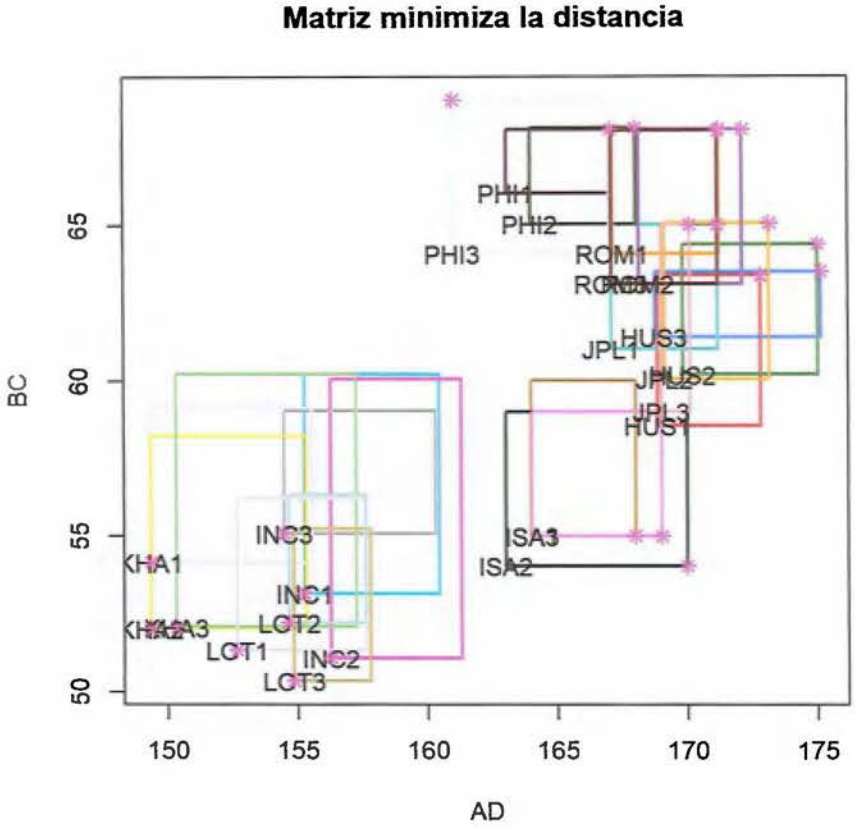


FIGURA 4.12: Matriz con mejor distancia.

La varianza explicada por el primer componente principal del método de maximización de la varianza es de $\frac{\lambda_1}{\sum \lambda_i} = 45.16\,\%$. De igual manera se puede tomar más componentes principales, si se utilizan 2 o 3 componentes principales, la varianza acumulada será de 83.21 % y 92.34 % respectivamente. En la tabla 4.34, se pueden observar todos los valores propios para este caso.

	Valor propio	Porcentaje de varianza	Porcentaje acumulado de varianza
λ_1	2.71	45.16 %	45.16 %
λ_2	2.28	38.05 %	83.21 %
λ_3	0.55	9.13 %	92.34 %
λ_4	0.25	4.11 %	96.45 %
λ_5	0.19	3.14 %	99.59 %
λ_6	0.02	0.41 %	100.00 %

TABLA 4.34: Valores propios para el ACP de Z^{φ}.

Los vectores propios se muestran en la tabla 4.43.

	comp 1	comp 2	comp 3	comp 4	comp 5	comp 6
AD	0.43	0.40	-0.29	0.22	-0.70	-0.21
BC	0.39	0.43	0.20	-0.77	0.13	0.10
AH	0.50	-0.17	0.64	0.36	-0.07	0.42
DH	0.56	0.02	-0.32	0.28	0.65	-0.29
EH	-0.26	0.52	0.55	0.26	0.14	-0.51
GH	-0.21	0.60	-0.24	0.28	0.21	0.65

TABLA 4.35: Vectores propios para el ACP de Z^{φ}.

Las coordenadas en el ACP de Z^{φ} se encuentran en la tabla 4.36.

	I comp 1	I comp 2	I comp 3	I comp 4	I comp 5	I comp 6
HUS1	[-0.32,0.73]	[0.39,1.67]	[-1.89,-0.8]	[-0.5,0.67]	[-0.16,0.62]	[-0.76,0.38]
HUS2	[-0.41,1]	[0.5,2.16]	[-2,-0.5]	[-0.64,0.77]	[-0.44,0.67]	[-0.88,0.76]
HUS3	[-0.28,0.98]	[0.36,2.06]	[-1.87,-0.55]	[-0.58,0.6]	[-0.45,0.76]	[-1.01,0.69]
INC1	[-3.29,-2.1]	[-0.13,1.29]	[-0.67,0.43]	[-1.31,0.1]	[-0.94,0]	[-0.63,0.5]
INC2	[-3.25,-1.45]	[-1.27,1.01]	[-1.42,0.46]	[-1.79,0.29]	[-1.4,-0.04]	[-1.05,0.96]
INC3	[-3.47,-1.81]	[-0.62,1.48]	[-1.21,0.57]	[-1.58,0.05]	[-1.14,0.25]	[-1.03,1.1]
ISA1	[0.37,1.35]	[-1.44,-0.42]	[-0.42,0.49]	[-0.27,0.81]	[-0.62,0.16]	[-0.59,0.27]
ISA2	[0.04,1.25]	[-1.3,-0.08]	[-0.57,0.55]	[-0.13,1.08]	[-0.76,0.28]	[-0.62,0.42]
ISA3	[0.08,1.43]	[-1.44,-0.1]	[-0.9,0.5]	[-0.28,1]	[-0.74,0.33]	[-0.69,0.64]
JPL1	[-0.37,0.65]	[0.47,1.62]	[0.33,1.37]	[-0.38,0.68]	[-0.85,-0.06]	[-0.57,0.48]
JPL2	[-0.33,0.84]	[0.42,1.88]	[0.08,1.32]	[-0.34,0.94]	[-1.01,-0.05]	[-0.74,0.56]
JPL3	[-0.49,0.65]	[0.47,1.85]	[-0.15,1.16]	[-0.36,1.01]	[-0.66,-0.04]	[-0.51,0.76]
KHA1	[-1.32,-0.05]	[-2.29,-1.09]	[-0.19,0.99]	[-1.24,0.01]	[-0.4,0.69]	[-0.72,0.35]
KHA2	[-1.55,-0.25]	[-2.31,-0.94]	[-0.42,0.76]	[-1.13,0.25]	[-0.38,0.76]	[-0.66,0.5]
KHA3	[-1.73,-0.13]	[-2.13,-0.48]	[-0.48,0.94]	[-1.36,0.38]	[-0.53,0.79]	[-0.97,0.42]
LOT1	[-0.89,0.19]	[-1.82,-0.57]	[-0.35,0.71]	[-0.17,0.99]	[0.08,0.95]	[-0.4,0.68]
LOT2	[-0.81,0.06]	[-1.36,-0.36]	[-0.14,0.69]	[0.03,1]	[0.17,0.86]	[-0.29,0.61]
LOT3	[-0.93,-0.01]	[-1.48,-0.4]	[-0.26,0.68]	[0.15,1.21]	[0.11,0.8]	[-0.26,0.66]
PHI1	[0.1,1.13]	[0.38,1.39]	[-0.07,1.01]	[-0.86,0.04]	[-0.05,0.81]	[-0.61,0.46]

Tabla 4.36 – Continúa en la siguiente página

Tabla 4.36 – Continúa de la página anterior

	I comp 1	I comp 2	I comp 3	I comp 4	I comp 5	I comp 6
PHI2	[-0.12,1.23]	[0.2,1.77]	[-0.24,1.24]	[-0.97,0.33]	[-0.25,0.84]	[-1,0.63]
PHI3	[-0.26,1.02]	[0.18,1.72]	[0.06,1.41]	[-1,0.32]	[-0.16,0.94]	[-0.83,0.51]
ROM1	[1.56,2.56]	[-1.16,-0.01]	[-0.49,0.49]	[-1.02,0.03]	[-0.53,0.3]	[-0.43,0.62]
ROM2	[1.34,2.71]	[-1.24,0.56]	[-0.87,0.67]	[-1.04,0.45]	[-0.57,0.44]	[-0.71,1.01]
ROM3	[1.06,2.65]	[-1.24,0.88]	[-1.11,0.67]	[-1.11,0.58]	[-0.56,0.68]	[-0.63,1.47]

TABLA 4.36: Coordenadas de los individuos suplementarios para el ACP de Z^{φ}.

Los cosenos cuadrados para este ACP vienen dados en la tabla 4.37.

	I comp 1	I comp 2	I comp 3
HUS1	[0,0.17]	[0.07,0.66]	[0.29,0.89]
HUS2	[0,0.27]	[0.12,0.8]	[0.12,0.83]
HUS3	[0,0.28]	[0.07,0.74]	[0.19,0.74]
INC1	[0.69,0.96]	[0,0.2]	[0,0.05]
INC2	[0.38,0.98]	[0,0.21]	[0,0.24]
INC3	[0.6,0.9]	[0,0.24]	[0,0.16]
ISA1	[0.1,0.79]	[0.13,0.78]	[0,0.13]
ISA2	[0,0.72]	[0.01,0.83]	[0,0.25]
ISA3	[0.01,0.77]	[0.01,0.79]	[0,0.32]
JPL1	[0,0.18]	[0.16,0.76]	[0.06,0.6]
JPL2	[0,0.31]	[0.15,0.78]	[0,0.47]
JPL3	[0,0.24]	[0.2,0.9]	[0,0.4]
KHA1	[0,0.34]	[0.42,0.94]	[0,0.18]
KHA2	[0.02,0.47]	[0.37,0.9]	[0,0.12]
KHA3	[0.01,0.62]	[0.12,0.85]	[0,0.19]
LOT1	[0,0.26]	[0.32,0.97]	[0,0.22]
LOT2	[0,0.28]	[0.14,0.84]	[0,0.31]
LOT3	[0,0.29]	[0.15,0.82]	[0,0.27]
PHI1	[0.01,0.65]	[0.1,0.79]	[0,0.48]
PHI2	[0,0.74]	[0.03,0.89]	[0,0.57]
PHI3	[0,0.6]	[0.04,0.77]	[0,0.6]
ROM1	[0.68,0.97]	[0,0.25]	[0,0.05]
ROM2	[0.66,0.96]	[0,0.26]	[0,0.15]
ROM3	[0.46,0.92]	[0,0.28]	[0,0.22]

TABLA 4.37: Cosenos cuadrados de los individuos suplementarios para el ACP de Z^{φ}.

Los individuos mejor representados (cuyo mínimo es mayor que 0.5) en el primer componente principal de Z^{φ} son *ROM1* y *ROM2*.

Para los demás considerarenos más dimensiones para que puedan ser representados de mejor manera. En la tabla 4.38 se representan la calidades de los individuos en 2 y 3 dimensiones.

	Componentes 1 y 2	Componentes 1,2 y 3
HUS1	[0.07,0.69]	[0.74,0.99]
HUS2	[0.13,0.83]	[0.75,1]
HUS3	[0.11,0.76]	[0.62,1]
INC1	[0.74,1]	[0.74,1]
INC2	[0.42,0.98]	[0.47,0.99]
INC3	[0.6,0.98]	[0.62,0.99]
ISA1	[0.56,0.99]	[0.57,1]
ISA2	[0.2,0.97]	[0.23,0.99]
ISA3	[0.23,0.98]	[0.3,0.99]
JPL1	[0.16,0.8]	[0.52,1]
JPL2	[0.18,0.85]	[0.36,0.98]
JPL3	[0.22,0.9]	[0.37,1]
KHA1	[0.6,0.99]	[0.66,0.99]
KHA2	[0.67,0.98]	[0.7,1]
KHA3	[0.4,0.98]	[0.43,0.99]
LOT1	[0.38,0.97]	[0.46,0.97]
LOT2	[0.18,0.88]	[0.3,0.92]
LOT3	[0.22,0.87]	[0.35,0.88]
PHI1	[0.37,0.88]	[0.45,0.99]
PHI2	[0.26,0.9]	[0.38,0.99]
PHI3	[0.11,0.94]	[0.42,0.95]
ROM1	[0.8,0.99]	[0.81,1]
ROM2	[0.71,0.98]	[0.76,0.99]
ROM3	[0.48,0.96]	[0.6,0.98]

TABLA 4.38: Calidades de los individuos suplementarios para el ACP de Z^{φ}.

Con dos componentes principales los individuos que se encuentran bien representados (cuyo mínimo es mayor que 0.69) son *ROM1*, *ROM2* e *INC1*.

Los demás objetos simbólicos necesitan tres o más componentes principales para poder representarse de buena manera.

La figura 4 13 muestra el círculo de correlaciones simbólico, se puede observar que:

- Las variables BC y AD se encuentran muy correlacionadas positivamente.
- La variable GH se encuentra correlacionada positivamente con las variables BC y EH.
- Las variables EH y AH se encuentran muy correlacionadas negativamente.

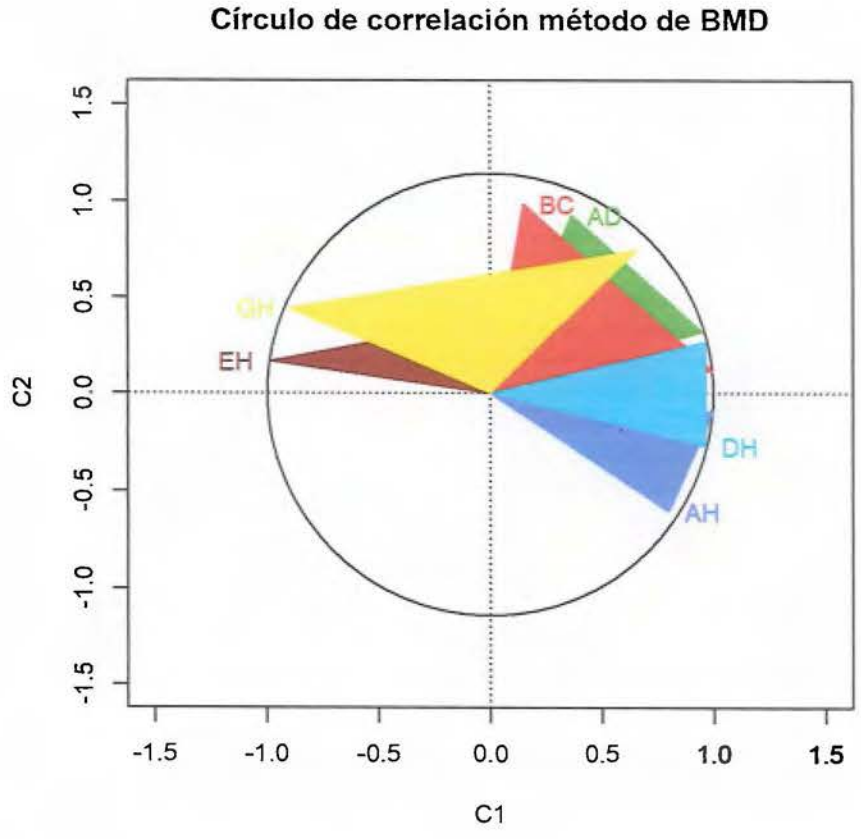


FIGURA 4.13: Círculo de correlaciones del primer y segundo componente principal de Z^{φ}.

La matriz de correlaciones se encuentra en la tabla 4.39 y en la figura 4.14, las correlaciones se puede concluir lo siguiente:

- El primer componente principal tiene mayor correlación con las variables *DH* y *AH*.
- El segundo componente principal tiene mayor correlación con las variables *EH* y *GH*.
- El tercer componente principal posee una correlación negativa con las variables *AD*, *DH* y *GH*.

	AD	BC	AH	DH	EH	GH
comp 1	0.64	0.59	0.80	0.90	-0.37	-0.22
comp 2	0.63	0.59	-0.24	0.04	0.67	0.83
comp 3	-0.28	0.11	0.49	-0.18	0.45	-0.37
comp 4	0.23	-0.42	0.39	0.37	0.12	0.17
comp 5	-0.33	0.06	0.20	0.47	-0.06	0.01
comp 6	-0.03	0.04	0.18	0.04	-0.40	0.36

TABLA 4.39: Matriz de correlaciones suplementarios de Z^{φ}.

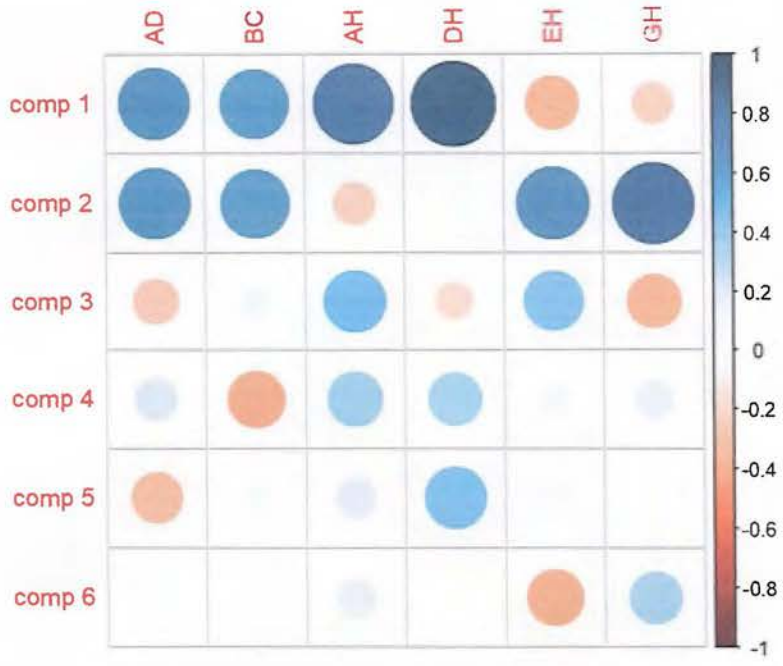


FIGURA 4.14: Matriz de correlaciones suplementarios de Z^{φ}.

En la tabla 4 40, se muestra la contribución de cada varible en el ACP de Z^{φ}. En el primer componente principal las variables que más contribuyen son DH y AH, mientras que en el segundo componente son GH y EH.

	comp 1	comp 2	comp 3	comp 4	comp 5	comp 6
AD	18.07	15.78	8.30	4.73	48.91	4.22
BC	15.24	18.12	4.07	60.04	1.61	0.92
AH	24.79	3.00	40.53	13.15	0.49	18.04
DH	31.03	0.03	10.52	7.62	42.49	8.32
EH	6.65	27.36	30.65	6.85	2.09	26.41
GH	4.23	35.72	5.93	7.62	4.41	42.09

TABLA 4.40: Contribuciones de las variables en el ACP de Z^{φ}.

La figura 4 15 muestra los individuos en el primer plano principal (83.21 % de la varianza), en este plano se pueden obtener 5 grupos *INC*, *KHA*, *LOT*, *ISA* y *ROM*.

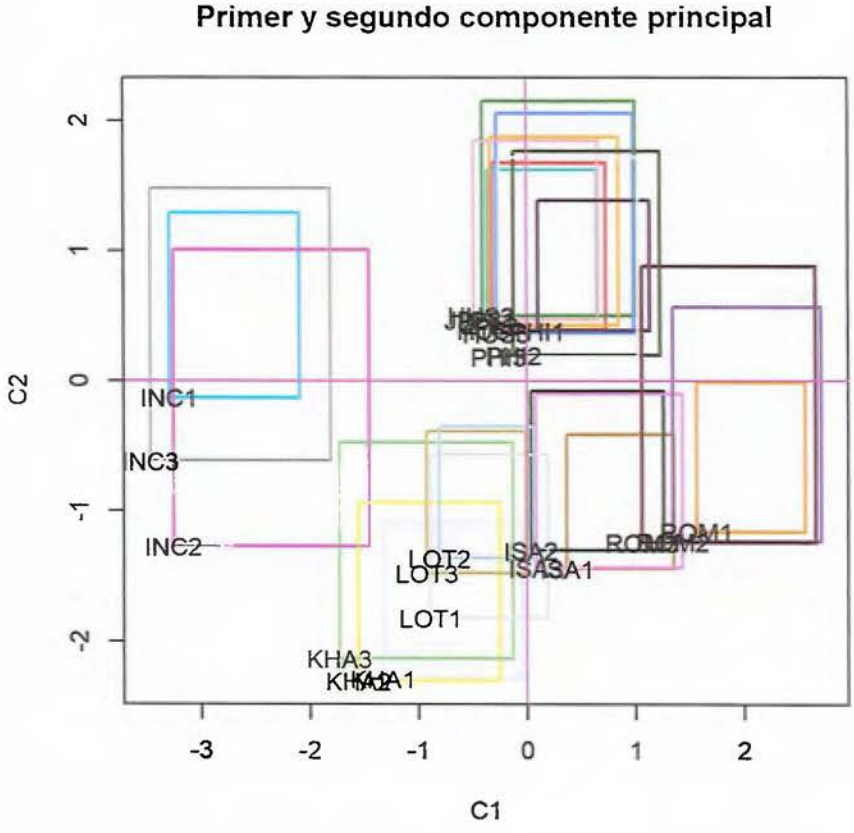


FIGURA 4.15: Primer (C1) y segundo (C2) componente principal de Z^{φ}.

La figura 4.16 muestra los individuos en el plano principal formado por el primer y tercer componente principal (45.29 % de la varianza), en este plano se obtiene una separación (que no se da en el plano principal) para los grupos *JPL*, *HUS* y *PHI*, algunos individuos de estos grupos se intersecan, como por ejemplo *JPL2* y *PHI3*.

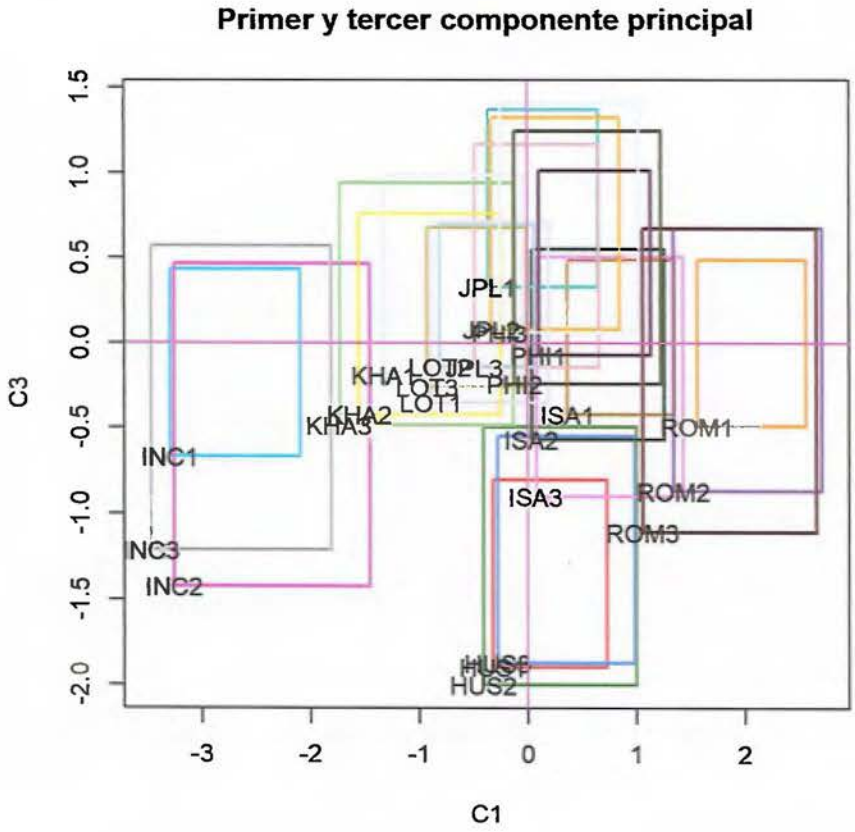


FIGURA 4.16: Primer (C1) y tercer (C3) componente principal de Z^{φ}.

Si se realiza un análisis de dualidad sobre el primer plano principal, se utilizan las figuras 4.15 y 4.13 se puede notar:

- Los grupos *LOT* y *KHA* no tienen medidas muy amplias en las variables *BC* y *AD*.
- El grupo *INC* no tiene medidas muy amplias en la variable *DH*.
- El grupo *ROM* tiene medidas muy amplias en la variable *DH*.

4.1.2.2. Matriz óptima respecto a la varianza (MOV)

La matriz que resuelve el problema 3.28 es

	AD	BC	AH	DH	EH	GH
HUS1	172.48	58.55	106.53	122.38	56.73	64.54
HUS2	175.03	60.21	104.65	120.24	58.99	66.84
HUS3	175.15	61.40	107.22	120.93	59.05	65.75
INC1	156.73	53.15	96.02	93.99	66.22	63.13
INC2	156.26	52.57	98.68	96.83	64.20	61.55
INC3	158.60	55.08	98.98	96.43	65.86	62.59
ISA1	164.00	60.03	123.04	121.02	55.53	53.25
ISA2	163.00	59.00	120.03	120.24	55.69	54.84
ISA3	164.01	59.01	120.44	122.43	54.77	55.07
JPL1	167.11	65.01	118.23	110.63	63.89	57.51
JPL2	169.14	65.07	118.85	113.17	62.63	58.02
JPL3	169.13	65.01	115.88	112.49	63.26	59.46
KHA1	150.21	54.15	115.75	106.41	58.14	50.61
KHA2	150.08	52.04	113.22	105.55	58.14	52.54
KHA3	150.33	52.09	112.44	104.74	58.67	53.00
LOT1	155.57	54.32	117.25	114.62	55.44	53.01
LOT2	157.62	56.32	117.52	114.28	57.63	54.41
LOT3	157.81	55.23	117.59	114.52	56.64	55.23
PHI1	167.07	66.03	119.60	116.10	61.72	57.01
PHI2	168.03	65.09	119.60	115.26	61.58	57.21
PHI3	167.00	64.26	118.79	114.59	61.52	57.16
ROM1	168.01	64.07	126.59	126.37	54.64	53.71
ROM2	172.14	67.24	127.29	126.75	57.14	55.41
ROM3	171.19	65.62	126.57	127.78	55.83	55.32

TABLA 4.41: Matriz Z^{Λ} para datos faciales.

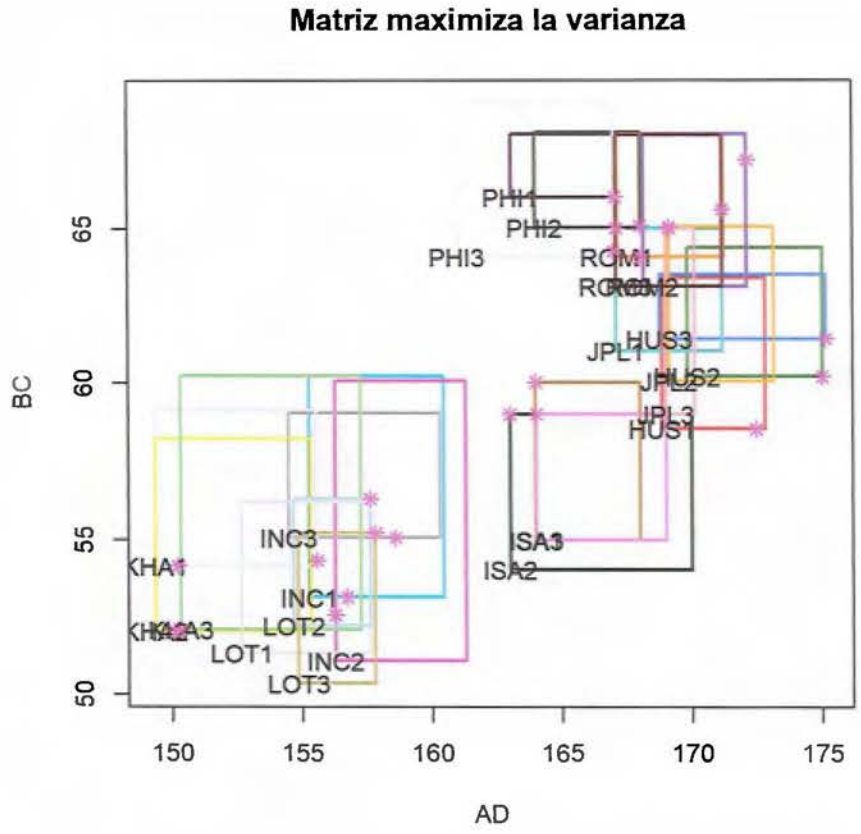


FIGURA 4.17: Matriz con mejor varianza.

La varianza por el primer componente principal del método de maximización de la varianza es de $\frac{\lambda_1}{\sum \lambda_i} = 52.97\,\%$. De igual manera se pueden tomar más componentes principales, si se utilizan 2 o 3 componentes principales, la varianza acumulada será de 87.38 % y 99.83 % respectivamente. En la tabla 4.42, se pueden observar todos los valores propios para este caso.

	Valor propio	Porcentaje de varianza	Porcentaje acumulado de varianza
λ_1	3.18	52.97 %	52.97 %
λ_2	2.06	34.41 %	87.38 %
λ_3	0.75	12.45 %	99.83 %
λ_4	0.01	0.17 %	100.00 %
λ_5	0.00	0.00 %	100.00 %
λ_6	0.00	0.00 %	100.00 %

TABLA 4.42: Valores propios para el ACP de Z^{Λ}.

Los vectores propios se muestran en la tabla 4.43.

	comp 1	comp 2	comp 3	comp 4	comp 5	comp 6
AD	0.35	0.54	-0.11	0.09	-0.73	0.20
BC	0.42	0.36	0.47	-0.53	0.24	-0.37
AH	0.49	-0.25	0.35	0.72	0.02	-0.24
DH	0.54	0.06	-0.33	-0.01	0.51	0.59
EH	-0.37	0.36	0.63	0.24	0.17	0.50
GH	-0.18	0.62	-0.38	0.38	0.36	-0.41

TABLA 4.43: Vectores propios para el ACP de Z^{Λ}.

Las coordenadas en el ACP de Z^{Λ} se encuentran en la tabla 4.44.

	I comp 1	I comp 2	I comp 3	I comp 4	I comp 5	I comp 6
HUS1	[-0.55,1]	[0.76,2.53]	[-2.17,-0.31]	[-1.27,0.24]	[-0.38,0.89]	[-0.41,1.28]
HUS2	[-0.71,1.3]	[0.87,3.2]	[-2.24,0.16]	[-1.35,0.56]	[-0.63,1.1]	[-0.87,1.4]
HUS3	[-0.49,1.25]	[0.64,3.08]	[-2.03,0.1]	[-1.32,0.31]	[-0.71,1.19]	[-0.71,1.44]
INC1	[-4.27,-2.57]	[-0.17,1.77]	[-0.59,1.37]	[-1.4,0.24]	[-1,0.5]	[-1.3,0.51]
INC2	[-4.15,-1.38]	[-1.68,1.41]	[-2.08,1.43]	[-2.28,0.26]	[-1.94,0.31]	[-2.25,0.98]
INC3	[-4.43,-2.05]	[-0.87,2.1]	[-1.41,1.53]	[-1.96,0.28]	[-1.38,0.83]	[-2.07,0.79]
ISA1	[0.51,1.89]	[-1.94,-0.54]	[-1.04,0.47]	[-0.51,0.71]	[-1.21,-0.05]	[-0.35,1.09]
ISA2	[0.04,1.67]	[-1.74,-0.03]	[-1.17,0.5]	[-0.32,1.09]	[-1.31,0.18]	[-0.38,1.22]
ISA3	[0.11,2.04]	[-1.91,-0.05]	[-1.7,0.44]	[-0.63,0.98]	[-1.32,0.2]	[-0.76,1.28]
JPL1	[-0.91,0.55]	[0.57,2.17]	[0.77,2.46]	[0.05,1.39]	[-0.75,0.47]	[-0.53,1.05]
JPL2	[-0.98,0.83]	[0.56,2.55]	[0.32,2.58]	[0.04,1.58]	[-1,0.5]	[-0.67,1.44]
JPL3	[-1.13,0.69]	[0.67,2.51]	[-0.11,2.22]	[-0.09,1.75]	[-0.57,0.55]	[-0.85,1.16]
KHA1	[-1.38,0.38]	[-3.27,-1.6]	[-0.74,1.06]	[-1.5,-0.12]	[-0.97,0.56]	[-1,0.77]
KHA2	[-1.69,0.15]	[-3.24,-1.35]	[-1.25,0.78]	[-1.41,0.07]	[-0.98,0.69]	[-1.14,0.83]
KHA3	[-2.02,0.2]	[-3,-0.73]	[-1.1,1.3]	[-1.68,0.23]	[-1.13,0.82]	[-1.09,1.22]
LOT1	[-1.05,0.5]	[-2.5,-0.78]	[-1.36,0.42]	[-0.28,1.13]	[-0.33,1.02]	[-0.61,1.05]
LOT2	[-1.02,0.22]	[-1.88,-0.49]	[-0.98,0.48]	[0.07,1.22]	[-0.09,1]	[-0.41,0.98]
LOT3	[-1.22,0.17]	[-1.99,-0.54]	[-1.29,0.4]	[0.2,1.45]	[-0.19,0.9]	[-0.46,1.08]
PHI1	[-0.08,1.37]	[0.42,1.84]	[0.4,2]	[-0.8,0.36]	[0.13,1.32]	[-0.62,0.95]

Tabla 4.44 – Continúa en la siguiente página

Tabla 4.44 – Continúa de la página anterior

	I comp 1	I comp 2	I comp 3	I comp 4	I comp 5	I comp 6
PHI2	[-0.49,1.49]	[0.17,2.38]	[0.12,2.52]	[-1,0.76]	[-0.2,1.46]	[-0.81,1.5]
PHI3	[-0.71,1.24]	[0.14,2.24]	[0.42,2.78]	[-0.8,0.76]	[-0.05,1.62]	[-0.76,1.43]
ROM1	[2.19,3.6]	[-1.63,-0.03]	[-0.88,0.74]	[-1.34,-0.06]	[-1.03,0.23]	[-1.1,0.46]
ROM2	[1.75,3.82]	[-1.7,0.79]	[-1.53,1.13]	[-1.49,0.56]	[-1.13,0.57]	[-1.46,0.97]
ROM3	[1.41,3.81]	[-1.7,1.25]	[-2.01,1.1]	[-1.52,0.8]	[-1.07,0.99]	[-2.01,0.93]

TABLA 4.44: Coordenadas de los individuos suplementarios para el ACP de Z^{Λ}.

Los cosenos cuadrados para este ACP vienen dados en la tabla 4.45.

	I comp 1	I comp 2	I comp 3
HUS1	[0,0.19]	[0.16,0.86]	[0.03,0.71]
HUS2	[0,0.29]	[0.23,0.96]	[0,0.56]
HUS3	[0,0.28]	[0.14,0.92]	[0,0.43]
INC1	[0.69,0.97]	[0,0.21]	[0,0.13]
INC2	[0.21,0.98]	[0,0.22]	[0,0.31]
INC3	[0.55,0.89]	[0,0.26]	[0,0.16]
ISA1	[0.1,0.73]	[0.12,0.62]	[0,0.22]
ISA2	[0,0.7]	[0,0.67]	[0,0.35]
ISA3	[0.01,0.71]	[0,0.57]	[0,0.5]
JPL1	[0,0.11]	[0.11,0.67]	[0.17,0.76]
JPL2	[0,0.17]	[0.15,0.75]	[0.03,0.66]
JPL3	[0,0.15]	[0.23,0.85]	[0,0.55]
KHA1	[0,0.23]	[0.58,0.94]	[0,0.15]
KHA2	[0,0.36]	[0.55,0.88]	[0,0.21]
KHA3	[0,0.55]	[0.2,0.88]	[0,0.28]
LOT1	[0,0.23]	[0.36,0.93]	[0,0.4]
LOT2	[0,0.28]	[0.17,0.96]	[0,0.3]
LOT3	[0,0.3]	[0.16,0.85]	[0,0.36]
PHI1	[0,0.5]	[0.08,0.62]	[0.05,0.68]
PHI2	[0,0.55]	[0.02,0.83]	[0,0.75]
PHI3	[0,0.37]	[0.01,0.7]	[0.09,0.73]
ROM1	[0.65,0.96]	[0,0.22]	[0,0.07]
ROM2	[0.62,0.98]	[0,0.21]	[0,0.21]
ROM3	[0.47,0.87]	[0,0.25]	[0,0.32]

TABLA 4.45: Cosenos cuadrados de los individuos suplementarios para el ACP de Z^{Λ}.

Los individuos mejor representados (cuyo mínimo es mayor que 0.5) en el primer componente principal de Z^{Λ} son *ROM1*, *ROM2*, *INC1* y *INC3*.

Para los demás se considerean más dimensiones para que puedan ser representados de mejor manera. En la tabla 4.46 se representantan la calidades de los individuos en 2 y 3 dimensiones.

	Componentes 1 y 2	Componentes 1,2 y 3
HUS1	[0.17,0.87]	[0.62,1]
HUS2	[0.26,0.96]	[0.61,0.99]
HUS3	[0.22,0.93]	[0.41,0.99]
INC1	[0.75,1]	[0.77,1]
INC2	[0.26,0.99]	[0.29,1]
INC3	[0.55,0.99]	[0.55,1]
ISA1	[0.46,0.95]	[0.51,0.99]
ISA2	[0.12,0.98]	[0.21,1]
ISA3	[0.14,0.9]	[0.29,0.99]
JPL1	[0.12,0.68]	[0.65,1]
JPL2	[0.15,0.81]	[0.49,0.99]
JPL3	[0.24,0.86]	[0.37,0.97]
KHA1	[0.64,0.97]	[0.64,0.99]
KHA2	[0.61,1]	[0.64,1]
KHA3	[0.41,0.99]	[0.44,0.99]
LOT1	[0.46,0.95]	[0.5,1]
LOT2	[0.23,0.97]	[0.23,0.98]
LOT3	[0.26,0.87]	[0.26,0.97]
PHI1	[0.17,0.78]	[0.58,0.97]
PHI2	[0.1,0.88]	[0.56,0.96]
PHI3	[0.06,0.8]	[0.44,0.99]
ROM1	[0.8,1]	[0.81,1]
ROM2	[0.69,0.98]	[0.78,0.99]
ROM3	[0.48,0.93]	[0.68,0.96]

TABLA 4.46: Calidades de los individuos suplementarios para el ACP de Z^{Λ}.

Con dos componentes principales los individuos que se encuentran bien representados (cuyo mínimo es mayor que 0.69) son *ROM1*, *ROM2*, *INC1* y *INC3*.

Los demás objetos simbólicos necesitan tres o más componentes principales para poder representarse de buena manera.

La figura 4.18 muestra el círculo de correlaciones simbólico, se puede observar que:

- Las variables BC y AD se encuentran muy correlacionadas positivamente.
- Las variables DH y AH se encuentran correlacionadas positivamente.

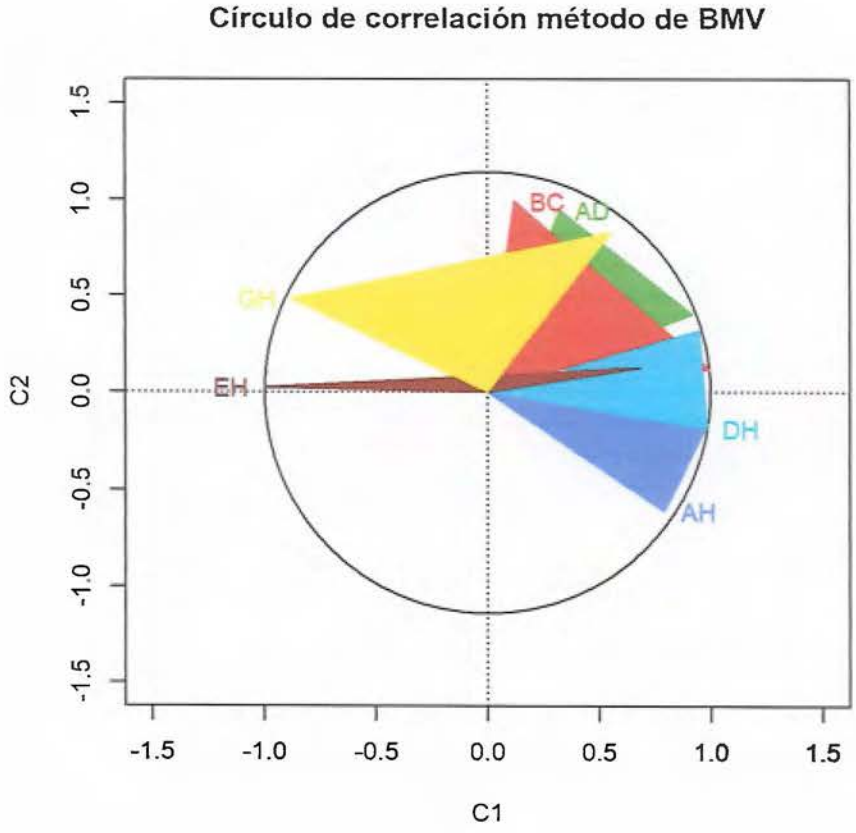


FIGURA 4.18: Círculo de correlaciones del primer y segundo componente principal de Z^{Λ}.

La matriz de correlaciones se encuentra en la tabla 4.47 y en la figura 4.19, de las correlaciones se puede concluir lo siguiente:

- El primer componente principal tiene mayor correlación con las variables DH y AH.
- El segundo componente principal tiene mayor correlación con las variables EH y GH.
- El tercer componente principal posee una correlación negativa con las variables DH y GH.

	AD	BC	AH	DH	EH	GH
comp 1	0.56	0.54	0.78	0.88	-0.51	-0.30
comp 2	0.65	0.56	-0.26	0.06	0.64	0.85
comp 3	0.02	0.42	0.21	-0.22	0.75	-0.06
comp 4	0.16	-0.23	0.49	0.18	0.38	0.25
comp 5	-0.13	0.32	0.06	0.28	0.36	0.43
comp 6	0.19	-0.12	0.17	0.32	0.46	-0.14

TABLA 4.47: Matriz de correlaciones suplementarios de Z^{Λ}.

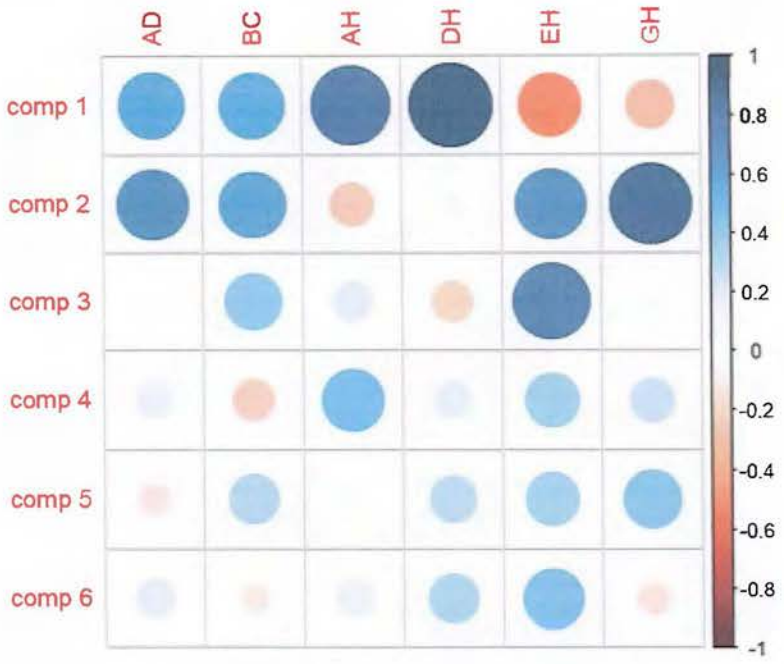


FIGURA 4.19: Matriz de correlaciones suplementarios de Z^{Λ}.

En la tabla 4 48, se muestra la contribución de cada variable en el ACP de Z^{Λ}. En el primer componente principal las variables que más contribuyen son DH y AH, mientras que en el segundo componente son GH y AD.

	comp 1	comp 2	comp 3	comp 4	comp 5	comp 6
AD	12.18	29.20	1.11	0.79	52.88	3.84
BC	17.77	12.89	22.40	27.89	5.62	13.43
AH	24.38	6.29	12.33	51.16	0.05	5.79
DH	28.69	0.35	10.69	0.00	25.64	34.63
EH	13.61	13.31	39.10	5.75	2.84	25.39
GH	3.38	37.97	14.37	14.41	12.96	16.92

TABLA 4.48: Contribuciones de las variables ACP Z^{Λ}.

La figura 4 20 muestra los individuos en el primer plano principal (87.38 % de la varianza), en este plano se pueden obtener 6 grupos *INC*, *KHA*, *LOT*, *ISA*, *ROM* y *PHI*.

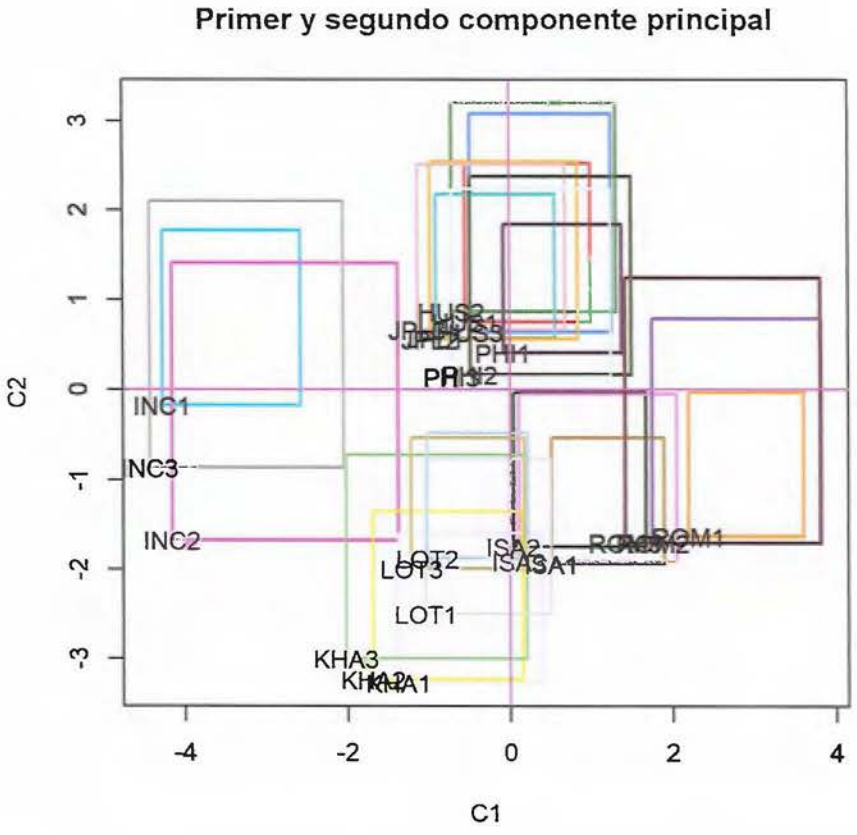


FIGURA 4.20: Primer y segundo componente principal de Z^{Λ}.

La figura 4 21 muestra los individuos en el plano principal formado por el primer y tercer componente principal (65.42 % de la varianza), en este plano se puede obtener una separación (que no se da en el plano principal) para los grupos *JPL* y *HUS*, algunos individuos de estos grupos se intersecan, como por ejemplo *JPL2* y *HUS2*.

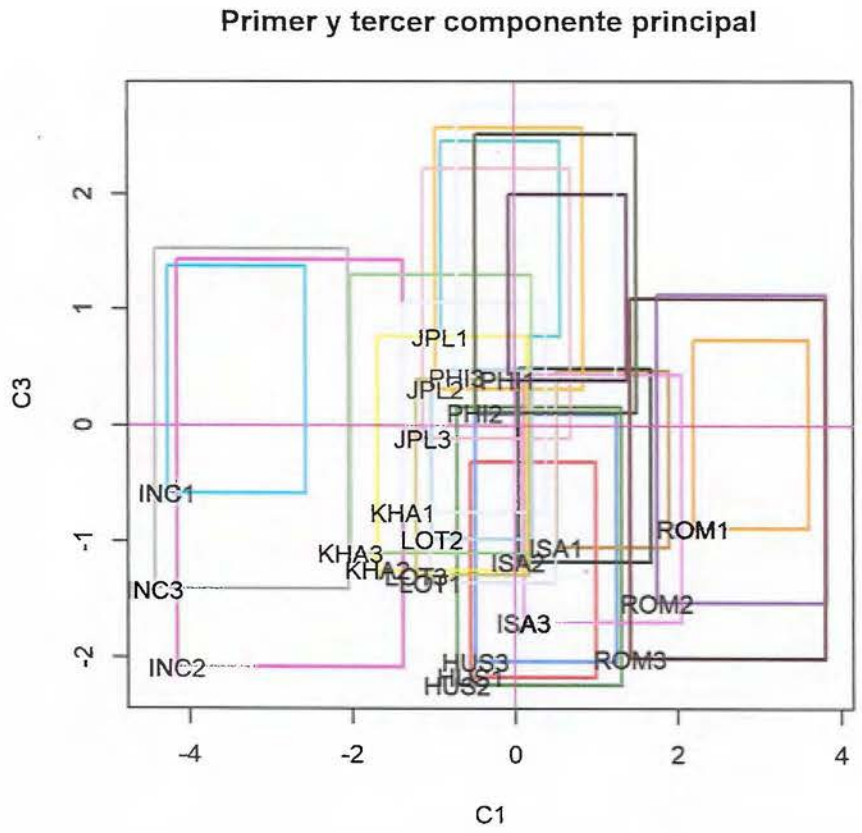


FIGURA 4.21: Primer y tercer componente principal de Z^{Λ}.

Si se realiza un análisis de dualidad sobre el primer plano principal, se utilizan las figuras 4.18 y 4 20 se puede notar:

- Los grupos *LOT* y *KHA* no tienen medidas muy amplias en las variables *BC* y *AD*.
- El grupo *INC* no tiene medidas muy amplias en la variable *AH*.
- El grupo *ROM* tiene medidas muy amplias en la variable *AH*.

4.1.2.3. Método de centros (CM)

La varianza por el primer componente principal del método de maximización de la varianza es de $\frac{\lambda_1}{\sum \lambda_i} = 43.09\,\%$. De igual manera se pueden tomar más componentes principales, si se utilizan 2 o 3 componentes principales, la varianza acumulada será de 80.17 % y 90.50 % respectivamente. En la tabla 4.49 se pueden observar todos los valores propios para este caso.

	Valor propio	Porcentaje de varianza	Porcentaje acumulado de varianza
λ_1	2.59	43.09 %	43.09 %
λ_2	2.23	37.08 %	80.17 %
λ_3	0.62	10.33 %	90.50 %
λ_4	0.35	5.83 %	96.34 %
λ_5	0.17	2.91 %	99.25 %
λ_6	0.05	0.75 %	100.00 %

TABLA 4.49: Valores propios para el ACP de centros.

Los vectores propios se muestran en la tabla 4.50.

	comp 1	comp 2	comp 3	comp 4	comp 5	comp 6
AD	0.49	0.34	-0.26	0.06	-0.72	-0.26
BC	0.46	0.31	0.22	-0.74	0.26	0.15
AH	0.46	-0.30	0.57	0.36	-0.17	0.47
DH	0.56	-0.12	-0.26	0.35	0.58	-0.38
EH	-0.16	0.54	0.64	0.26	0.10	-0.45
GH	-0.03	0.63	-0.30	0.35	0.20	0.60

TABLA 4.50: Vectores propios para el ACP de centros.

Las coordenadas en el ACP de centros se encuentran en la tabla 4.51.

	I comp 1	I comp 2	I comp 3	I comp 4	I comp 5	I comp 6
HUS1	[-0.01,1.28]	[0.51,2.41]	[-2.48,-0.8]	[-0.55,1.15]	[-0.28,0.94]	[-1.08,0.6]
HUS2	[0.04,1.66]	[0.59,3.14]	[-2.64,-0.38]	[-0.77,1.27]	[-0.59,1.03]	[-1.25,1.14]
HUS3	[0.22,1.55]	[0.36,2.98]	[-2.46,-0.42]	[-0.72,1.03]	[-0.57,1.11]	[-1.43,1.04]
INC1	[-3.93,-2.37]	[0.66,2.68]	[-0.89,0.79]	[-1.36,0.65]	[-0.93,0.55]	[-0.76,0.96]
INC2	[-4.01,-1.8]	[-1.21,2.21]	[-2.18,0.83]	[-2.22,0.85]	[-1.6,0.46]	[-1.4,1.58]
INC3	[-3.99,-2.19]	[-0.22,3.06]	[-1.75,0.99]	[-1.79,0.64]	[-1.12,0.87]	[-1.29,1.8]
ISA1	[0.14,1.42]	[-2.32,-0.85]	[-0.75,0.59]	[-0.48,1.02]	[-1.14,0.04]	[-0.95,0.35]
ISA2	[-0.22,1.36]	[-2.03,-0.28]	[-0.92,0.67]	[-0.21,1.42]	[-1.35,0.2]	[-1,0.56]
ISA3	[-0.12,1.53]	[-2.34,-0.29]	[-1.43,0.6]	[-0.45,1.34]	[-1.27,0.26]	[-1.09,0.86]
JPL1	[-0.15,1.09]	[0.69,2.42]	[0.49,2.05]	[-0.51,0.99]	[-1.12,0.07]	[-0.8,0.77]
JPL2	[-0.08,1.35]	[0.58,2.77]	[0.1,2.07]	[-0.53,1.35]	[-1.35,0.04]	[-1.06,0.87]
JPL3	[-0.25,1.13]	[0.7,2.8]	[-0.25,1.79]	[-0.47,1.55]	[-0.93,0.11]	[-0.72,1.14]
KHA1	[-2.19,-0.53]	[-3.16,-1.41]	[-0.43,1.25]	[-1.51,0.19]	[-0.59,1]	[-1.01,0.6]
KHA2	[-2.45,-0.76]	[-3.1,-1.12]	[-0.81,0.96]	[-1.32,0.61]	[-0.59,1.07]	[-0.94,0.81]
KHA3	[-2.61,-0.51]	[-2.83,-0.45]	[-0.79,1.31]	[-1.65,0.77]	[-0.79,1.2]	[-1.37,0.73]
LOT1	[-1.52,-0.16]	[-2.49,-0.66]	[-0.74,0.87]	[0.03,1.67]	[-0.22,1.11]	[-0.63,0.98]
LOT2	[-1.31,-0.25]	[-1.82,-0.32]	[-0.4,0.87]	[0.32,1.71]	[-0.06,0.98]	[-0.47,0.88]
LOT3	[-1.52,-0.34]	[-1.94,-0.35]	[-0.62,0.84]	[0.5,2.02]	[-0.19,0.86]	[-0.44,0.93]
PHI1	[0.47,1.68]	[0.34,1.91]	[0.04,1.6]	[-1.09,0.17]	[0.05,1.23]	[-0.78,0.76]

Tabla 4.51 – **Continúa en la siguiente página**

Tabla 4.51 – Continúa de la página anterior

	I comp 1	I comp 2	I comp 3	I comp 4	I comp 5	I comp 6
PHI2	[0.26,1.76]	[0.06,2.53]	[-0.22,2.03]	[-1.3,0.61]	[-0.25,1.28]	[-1.35,1.02]
PHI3	[-0.02,1.58]	[0.17,2.44]	[0.17,2.28]	[-1.26,0.61]	[-0.14,1.44]	[-1.12,0.87]
ROM1	[1.89,3.1]	[-2.35,-0.64]	[-0.89,0.58]	[-1.71,-0.2]	[-0.84,0.38]	[-0.67,0.89]
ROM2	[1.74,3.29]	[-2.49,0.28]	[-1.48,0.93]	[-1.76,0.45]	[-0.94,0.57]	[-1.08,1.43]
ROM3	[1.51,3.24]	[-2.47,0.84]	[-1.89,0.91]	[-1.81,0.75]	[-0.89,0.92]	[-0.96,2.1]

TABLA 4.51: Coordenadas de los individuos suplementarios para el ACP de centros.

Los cosenos cuadrados para este ACP vienen dados en la tabla 4.52.

	I comp 1	I comp 2	I comp 3
HUS1	[0,0.29]	[0.06,0.66]	[0.17,0.84]
HUS2	[0,0.41]	[0.09,0.78]	[0.04,0.79]
HUS3	[0.01,0.4]	[0.04,0.73]	[0.06,0.69]
INC1	[0.48,0.96]	[0.03,0.49]	[0,0.05]
INC2	[0.37,0.91]	[0,0.43]	[0,0.32]
INC3	[0.42,0.92]	[0,0.54]	[0,0.19]
ISA1	[0.01,0.56]	[0.33,0.85]	[0,0.11]
ISA2	[0,0.58]	[0.06,0.89]	[0,0.19]
ISA3	[0,0.63]	[0.08,0.81]	[0,0.37]
JPL1	[0,0.26]	[0.2,0.7]	[0.06,0.6]
JPL2	[0,0.41]	[0.18,0.71]	[0,0.51]
JPL3	[0,0.36]	[0.23,0.83]	[0,0.41]
KHA1	[0.04,0.5]	[0.37,0.86]	[0,0.16]
KHA2	[0.08,0.62]	[0.29,0.78]	[0,0.1]
KHA3	[0.05,0.75]	[0.07,0.72]	[0,0.2]
LOT1	[0.01,0.37]	[0.22,0.95]	[0,0.17]
LOT2	[0.02,0.35]	[0.06,0.78]	[0,0.25]
LOT3	[0.04,0.36]	[0.06,0.72]	[0,0.21]
PHI1	[0.05,0.72]	[0.04,0.67]	[0,0.54]
PHI2	[0.01,0.76]	[0,0.8]	[0,0.65]
PHI3	[0,0.65]	[0.02,0.65]	[0.01,0.65]
ROM1	[0.4,0.86]	[0.05,0.47]	[0,0.08]
ROM2	[0.38,0.97]	[0,0.48]	[0,0.22]
ROM3	[0.34,0.87]	[0,0.49]	[0,0.31]

TABLA 4.52: Cosenos cuadrados de los individuos suplementarios para el ACP de centros.

El método de centros no presenta individuos bien representados (cuyo mínimo es mayor que 0.5) en el primer componente principal.

Se consideran más dimensiones para que puedan ser representados de mejor manera. En la tabla 4 53 se representantan la calidades de los individuos en 2 y 3 dimensiones.

	Componentes 1 y 2	Componentes 1,2 y 3
HUS1	[0.1,0.78]	[0.67,0.98]
HUS2	[0.18,0.88]	[0.71,0.99]
HUS3	[0.16,0.82]	[0.55,0.98]
INC1	[0.83,0.99]	[0.83,1]
INC2	[0.41,0.97]	[0.5,0.98]
INC3	[0.66,0.98]	[0.69,0.99]
ISA1	[0.5,0.98]	[0.5,0.99]
ISA2	[0.15,0.99]	[0.18,0.99]
ISA3	[0.17,0.96]	[0.25,0.99]
JPL1	[0.22,0.84]	[0.62,1]
JPL2	[0.23,0.89]	[0.42,0.99]
JPL3	[0.28,0.89]	[0.38,0.99]
KHA1	[0.66,0.99]	[0.72,1]
KHA2	[0.74,0.99]	[0.76,0.99]
KHA3	[0.44,0.99]	[0.52,1]
LOT1	[0.39,0.98]	[0.41,0.99]
LOT2	[0.16,0.91]	[0.24,0.92]
LOT3	[0.2,0.84]	[0.28,0.85]
PHI1	[0.35,0.86]	[0.5,0.98]
PHI2	[0.23,0.9]	[0.44,0.98]
PHI3	[0.13,0.92]	[0.47,0.96]
ROM1	[0.75,0.98]	[0.76,0.98]
ROM2	[0.65,0.97]	[0.75,0.98]
ROM3	[0.42,0.92]	[0.61,0.96]

TABLA 4.53: Calidades de los individuos suplementarios para el ACP de centros.

Con dos componentes principales los individuos que se encuentran bien representados (cuyo mínimo es mayor que 0.69) son: *ROM1* e *INC1*.

Los demás objetos simbólicos necesitan tres o más componentes principales para poder representarse de buena manera.

La figura 4.22 muestra el círculo de correlaciones simbólico, se puede observar que:

- Las variables *BC* y *DH* se encuentran muy correlacionadas positivamente.
- Las variables *EH* y *GH* se encuentran correlacionadas positivamente.

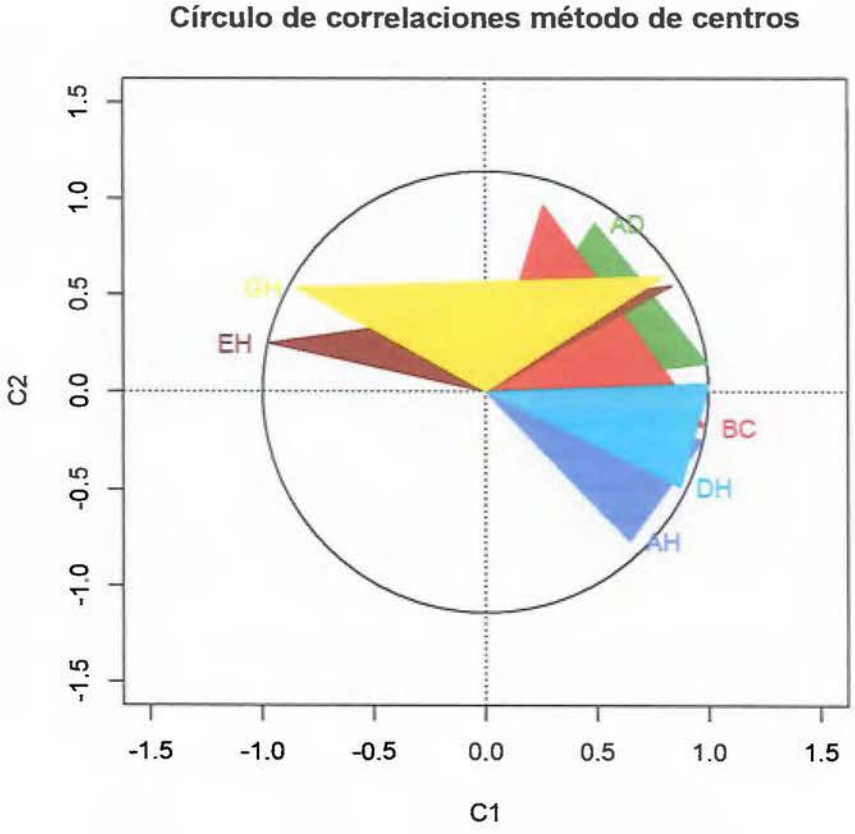


FIGURA 4.22: Círculo de correlaciones del primer y segundo componente principal de centros.

La matriz de correlaciones se encuentra en la tabla 4.54 y en la figura 4.23, las correlaciones se puede concluir lo siguiente:

- El primer componente principal tiene mayor correlación con las variables *DH* y *AD*.
- El segundo componente principal tiene mayor correlación con las variables *EH* y *GH*.
- El tercer componente principal posee una correlación negativa con las variables *DH*, *GH* y *AD*.

	AD	BC	AH	DH	EH	GH
comp 1	0.75	0.71	0.71	0.88	-0.24	-0.04
comp 2	0.47	0.43	-0.41	-0.17	0.75	0.87
comp 3	-0.19	0.19	0.40	-0.18	0.60	-0.27
comp 4	0.04	-0.55	0.18	0.18	0.22	0.29
comp 5	-0.37	0.20	-0.07	0.23	0.09	0.18
comp 6	-0.08	0.08	0.08	-0.06	-0.32	0.41

TABLA 4.54: Matriz de correlaciones suplementarios del ACP de centros.

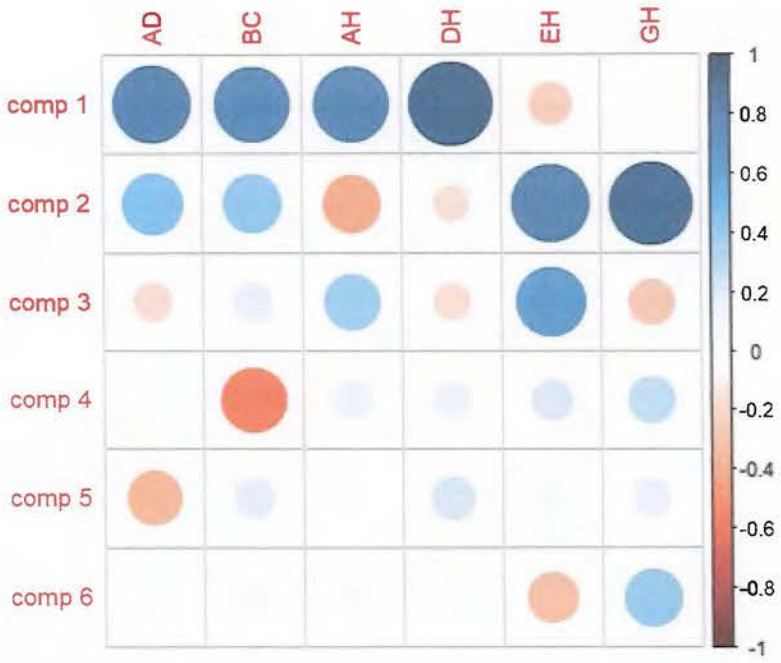


FIGURA 4.23: Matriz de correlaciones suplementarios del ACP de centros.

En la tabla 4 55 se muestra la contribución de cada varible en ACP de centros. En el primer componente principal las variables que más contribuyen son DH y AD, mientras que en el segundo componente son GH y EH.

	comp 1	comp 2	comp 3	comp 4	comp 5	comp 6
AD	23.62	11.42	6.69	0.38	51.26	6.63
BC	21.31	9.57	4.82	55.17	6.92	2.21
AH	20.90	8.88	32.53	12.99	3.00	21.71
DH	31.61	1.49	6.51	12.52	33.80	14.07
EH	2.50	29.16	40.67	6.83	1.00	19.84
GH	0.07	39.47	8.77	12.11	4.03	35.54

TABLA 4.55: Contribuciones de las variables ACP de centros.

La figura 4.24 muestra los individuos en el primer plano principal (80.17 % de la varianza), en este plano se pueden obtener 5 grupos *INC*, *KHA*, *LOT*, *ISA* y *ROM*.

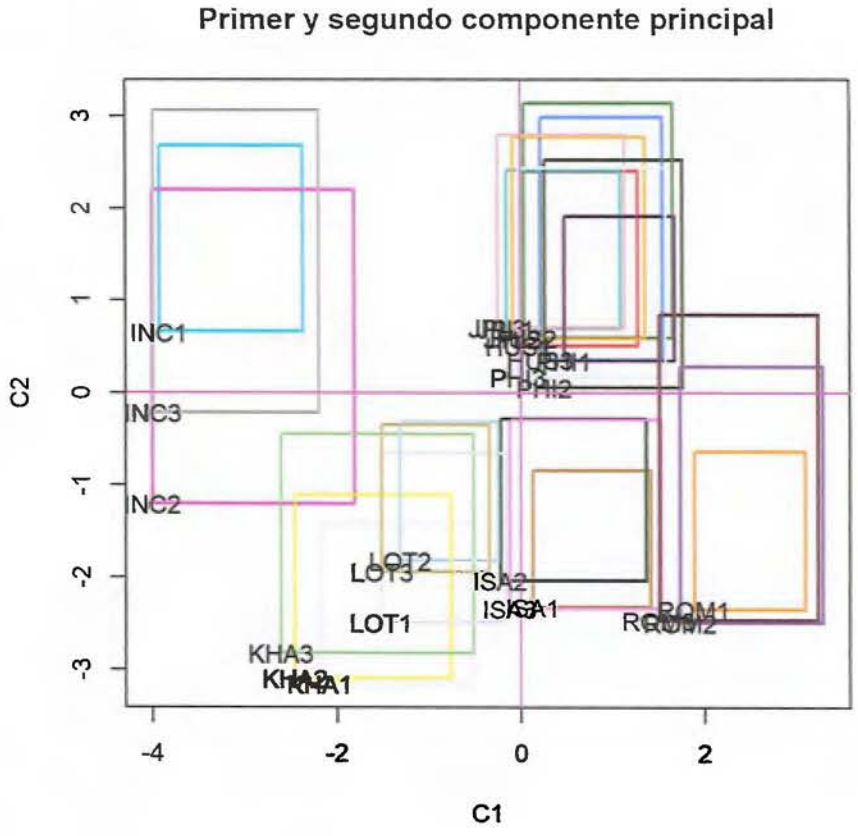


FIGURA 4.24: Primer y segundo componente principal de centros.

La figura 4.25 muestra los individuos en el plano principal formado por el primer y tercer componente principal (53.42 % de la varianza), en este plano se puede obtener una separación (que no se da en el plano principal) para los grupos *JPL*, *HUS* y *PHI*, algunos individuos de estos grupos se intersecan, como por ejemplo *PHI1* y *JPL2*.

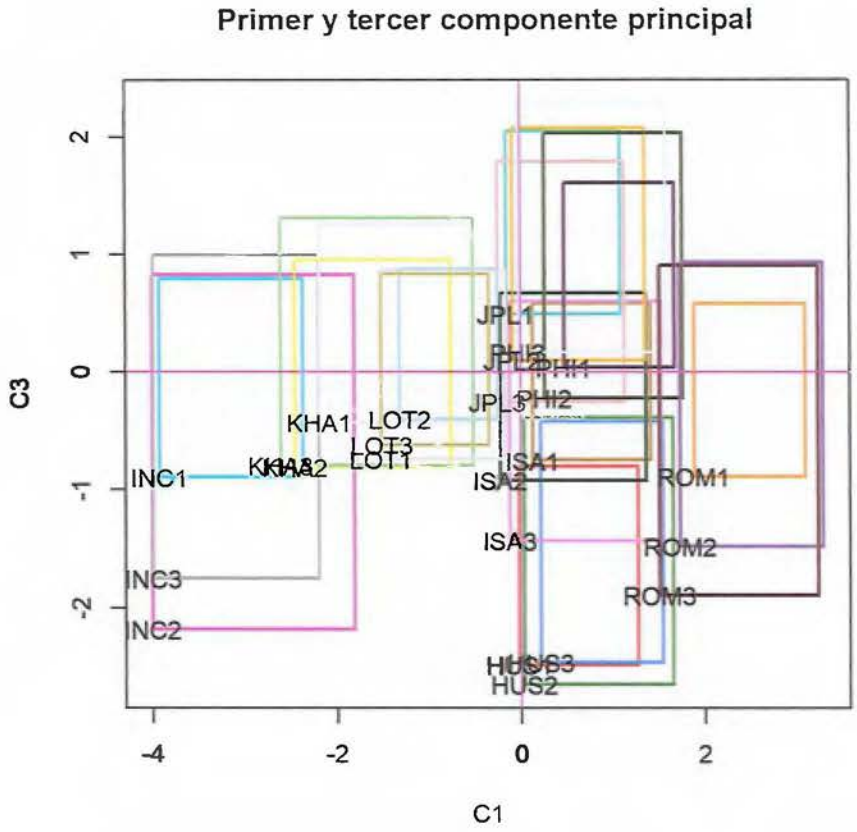


FIGURA 4.25: Primer y tercer componente principal de centros.

Si se realiza un análisis de dualidad sobre el primer plano principal, utilizando las figuras 4.24 y 4.22 se puede notar:

- Los grupos *LOT* y *KHA* no tienen medidas muy amplias en las variables *BC* y *AD*.
- El grupo *PHI*, *HUS* y *JPL* tienen medidas muy amplias en las variables *BC* y *AD*.
- El grupo *ROM* tiene medidas muy amplias en las variables *AH* y *DH*.

4.1.2.4. Método de vértices (VM)

La varianza por el primer componente principal del método de vértices es de $\frac{\lambda_1}{\sum \lambda_i} = 40.07\,\%$. De igual manera se pueden tomar más componentes principales, si se utilizan 2 o 3 componentes principales, la varianza acumulada será de 72.01 % y 83.71 % respectivamente. En la tabla 4.56, se pueden observar todos los valores propios para este caso.

	Valor propio	Porcentaje de varianza	Porcentaje acumulado de varianza
λ_1	2.40	40.07 %	40.07 %
λ_2	1.92	31.94 %	72.01 %
λ_3	0.70	11.71 %	83.71 %
λ_4	0.49	8.09 %	91.80 %
λ_5	0.29	4.78 %	96.58 %
λ_6	0.20	3.42 %	100.00 %

TABLA 4.56: Valores propios para el ACP de vértices.

Los vectores propios se muestran en la tabla 4.57.

	comp 1	comp 2	comp 3	comp 4	comp 5	comp 6
AD	0.48	0.36	-0.24	0.09	-0.76	0.11
BC	0.44	0.32	0.22	-0.76	0.28	0.04
AH	0.47	-0.30	0.47	0.36	0.13	0.56
DH	0.57	-0.11	-0.19	0.31	0.31	-0.65
EH	-0.16	0.52	0.70	0.29	-0.09	-0.34
GH	-0.04	0.63	-0.37	0.33	0.48	0.36

TABLA 4.57: Vectores propios para el ACP de vértices.

Las coordenadas en el ACP de vértices se encuentran en la tabla 4.58.

	I comp 1	I comp 2	I comp 3	I comp 4	I comp 5	I comp 6
HUS1	[-1.14,0.04]	[0.51,2.16]	[-2.2,-0.67]	[-0.55,0.95]	[-0.67,0.62]	[-1.21,-0.12]
HUS2	[-1.48,0.02]	[0.58,2.79]	[-2.37,-0.31]	[-0.75,1.06]	[-0.88,0.85]	[-1.27,0.37]
HUS3	[-1.39,-0.15]	[0.39,2.65]	[-2.22,-0.31]	[-0.68,0.84]	[-0.97,0.98]	[-1.36,0.22]
INC1	[2.34,3.74]	[0.51,2.28]	[-0.85,0.69]	[-1.3,0.48]	[-0.6,0.94]	[-0.35,0.71]
INC2	[1.81,3.8]	[-1.03,1.9]	[-1.99,0.77]	[-2.04,0.68]	[-0.56,1.57]	[-0.71,1.22]
INC3	[2.16,3.84]	[-0.21,2.6]	[-1.63,0.89]	[-1.68,0.45]	[-1.02,1.19]	[-0.83,1.26]
ISA1	[-1.38,-0.22]	[-1.99,-0.69]	[-0.64,0.56]	[-0.36,0.98]	[-0.05,1.09]	[-0.43,0.47]
ISA2	[-1.33,0.12]	[-1.76,-0.21]	[-0.81,0.62]	[-0.14,1.33]	[-0.26,1.26]	[-0.46,0.63]
ISA3	[-1.49,0.04]	[-1.99,-0.22]	[-1.26,0.55]	[-0.36,1.25]	[-0.3,1.14]	[-0.67,0.82]
JPL1	[-0.99,0.14]	[0.56,2.06]	[0.38,1.79]	[-0.42,0.92]	[-0.23,0.98]	[-0.07,1]
JPL2	[-1.23,0.06]	[0.49,2.37]	[0.02,1.83]	[-0.43,1.23]	[-0.17,1.22]	[-0.21,1.07]
JPL3	[-1.02,0.22]	[0.58,2.38]	[-0.3,1.55]	[-0.41,1.39]	[-0.32,0.73]	[-0.15,1.13]
KHA1	[0.47,2]	[-2.77,-1.23]	[-0.33,1.16]	[-1.33,0.19]	[-0.87,0.6]	[-0.76,0.45]
KHA2	[0.69,2.23]	[-2.72,-0.99]	[-0.69,0.89]	[-1.18,0.54]	[-1.01,0.59]	[-0.73,0.48]
KHA3	[0.48,2.38]	[-2.49,-0.41]	[-0.67,1.22]	[-1.47,0.69]	[-1.03,0.9]	[-1.01,0.45]
LOT1	[0.08,1.31]	[-2.23,-0.64]	[-0.68,0.78]	[0.02,1.48]	[-1.2,0.15]	[-0.59,0.45]
LOT2	[0.15,1.12]	[-1.65,-0.36]	[-0.39,0.77]	[0.28,1.51]	[-1.09,-0.01]	[-0.48,0.42]
LOT3	[0.24,1.29]	[-1.76,-0.39]	[-0.59,0.73]	[0.43,1.78]	[-0.99,0.04]	[-0.41,0.5]
PHI1	[-1.53,-0.4]	[0.28,1.64]	[0.03,1.43]	[-0.97,0.16]	[-1.06,0.05]	[-0.78,0.42]

Tabla 4.58 – Continúa en la siguiente página

Tabla 4.58 – Continúa de la página anterior

	I comp 1	I comp 2	I comp 3	I comp 4	I comp 5	I comp 6
PHI2	[-1.61,-0.21]	[0.05,2.16]	[-0.22,1.83]	[-1.14,0.54]	[-1.13,0.45]	[-1.08,0.62]
PHI3	[-1.42,0.03]	[0.11,2.07]	[0.13,2.04]	[-1.12,0.55]	[-1.27,0.28]	[-0.86,0.46]
ROM1	[-2.89,-1.78]	[-1.94,-0.45]	[-0.77,0.57]	[-1.43,-0.1]	[-0.48,0.79]	[-0.35,0.7]
ROM2	[-3.07,-1.65]	[-2.05,0.32]	[-1.33,0.89]	[-1.48,0.47]	[-0.75,0.93]	[-0.64,0.99]
ROM3	[-3.02,-1.41]	[-2.03,0.78]	[-1.74,0.85]	[-1.54,0.7]	[-1.23,0.79]	[-0.67,1.35]

TABLA 4.58: Coordenadas para el ACP de vértices.

Los cosenos cuadrados para este ACP vienen dados en la tabla 4.59.

	I comp 1	I comp 2	I comp 3
HUS1	[0,0.26]	[0.07,0.65]	[0.14,0.75]
HUS2	[0,0.39]	[0.11,0.78]	[0.03,0.7]
HUS3	[0,0.37]	[0.05,0.73]	[0.04,0.56]
INC1	[0.55,0.97]	[0.02,0.41]	[0,0.06]
INC2	[0.43,0.92]	[0,0.36]	[0,0.32]
INC3	[0.49,0.92]	[0,0.46]	[0,0.2]
ISA1	[0.02,0.62]	[0.27,0.84]	[0,0.11]
ISA2	[0,0.63]	[0.03,0.87]	[0,0.2]
ISA3	[0,0.68]	[0.05,0.79]	[0,0.37]
JPL1	[0,0.27]	[0.16,0.68]	[0.05,0.61]
JPL2	[0,0.41]	[0.15,0.69]	[0,0.53]
JPL3	[0,0.36]	[0.19,0.83]	[0,0.41]
KHA1	[0.04,0.5]	[0.36,0.87]	[0,0.17]
KHA2	[0.09,0.62]	[0.28,0.78]	[0,0.11]
KHA3	[0.06,0.75]	[0.06,0.72]	[0,0.23]
LOT1	[0,0.33]	[0.25,0.94]	[0,0.17]
LOT2	[0.01,0.32]	[0.09,0.77]	[0,0.24]
LOT3	[0.02,0.32]	[0.09,0.72]	[0,0.19]
PHI1	[0.05,0.75]	[0.04,0.67]	[0,0.57]
PHI2	[0.01,0.79]	[0,0.8]	[0,0.73]
PHI3	[0,0.68]	[0.01,0.64]	[0.01,0.73]
ROM1	[0.47,0.9]	[0.03,0.42]	[0,0.08]
ROM2	[0.45,0.97]	[0,0.42]	[0,0.22]
ROM3	[0.4,0.88]	[0,0.43]	[0,0.35]

TABLA 4.59: Cosenos cuadrados de los individuos ACP de vértices.

El método de vértices solamente presenta un individuo bien representado (cuyo mínimo es mayor que 0.5), *INC1*.

Se consideran más dimensiones para que puedan ser representados de mejor manera. En la tabla 4 60 se representantan la calidades de los individuos en 2 y 3 dimensiones.

	Componentes 1 y 2	Componentes 1,2 y 3
HUS1	[0.11,0.76]	[0.59,0.99]
HUS2	[0.19,0.88]	[0.62,0.99]
HUS3	[0.16,0.81]	[0.46,1]
INC1	[0.82,1]	[0.83,1]
INC2	[0.45,0.98]	[0.53,0.99]
INC3	[0.68,0.98]	[0.7,1]
ISA1	[0.49,0.99]	[0.49,1]
ISA2	[0.15,0.99]	[0.18,0.99]
ISA3	[0.17,0.97]	[0.24,1]
JPL1	[0.18,0.82]	[0.56,0.98]
JPL2	[0.21,0.87]	[0.39,0.98]
JPL3	[0.25,0.88]	[0.32,0.99]
KHA1	[0.67,1]	[0.75,1]
KHA2	[0.72,0.99]	[0.76,0.99]
KHA3	[0.46,0.99]	[0.57,1]
LOT1	[0.38,0.98]	[0.4,0.98]
LOT2	[0.17,0.89]	[0.23,0.89]
LOT3	[0.2,0.83]	[0.26,0.84]
PHI1	[0.34,0.87]	[0.49,0.99]
PHI2	[0.23,0.91]	[0.45,0.98]
PHI3	[0.11,0.92]	[0.44,0.98]
ROM1	[0.77,0.99]	[0.78,0.99]
ROM2	[0.69,0.97]	[0.78,0.98]
ROM3	[0.46,0.94]	[0.67,0.98]

TABLA 4.60: Calidades de los individuos para el ACP de vértices.

Con dos componentes principales los individuos que se encuentran bien representados (cuyo mínimo es mayor que 0.69) son *KKA2*, *ROM1* y *INC1*.

Los demás objetos simbólicos necesitan tres o más componentes principales para poder representarse de buena manera.

La matriz de correlaciones se encuentra en la tabla 4.61 y en la figura 4.26, de las correlaciones se puede concluir lo siguiente:

- El primer componente principal tiene mayor correlación con las variables *DH* y *AD*.
- El segundo componente principal tiene mayor correlación con las variables *EH* y *GH*.
- El tercer componente principal posee una correlación negativa con las variables *DH*, *GH* y *AD*.

	AD	BC	AH	DH	EH	GH
comp 1	0.74	0.68	0.74	0.89	-0.25	-0.06
comp 2	0.50	0.45	-0.41	-0.15	0.72	0.87
comp 3	-0.20	0.19	0.40	-0.16	0.59	-0.31
comp 4	0.06	-0.53	0.25	0.22	0.20	0.23
comp 5	-0.40	0.15	0.07	0.17	-0.05	0.26
comp 6	0.05	0.02	0.25	-0.30	-0.16	0.16

TABLA 4.61: Matriz de correlaciones método de vértices.

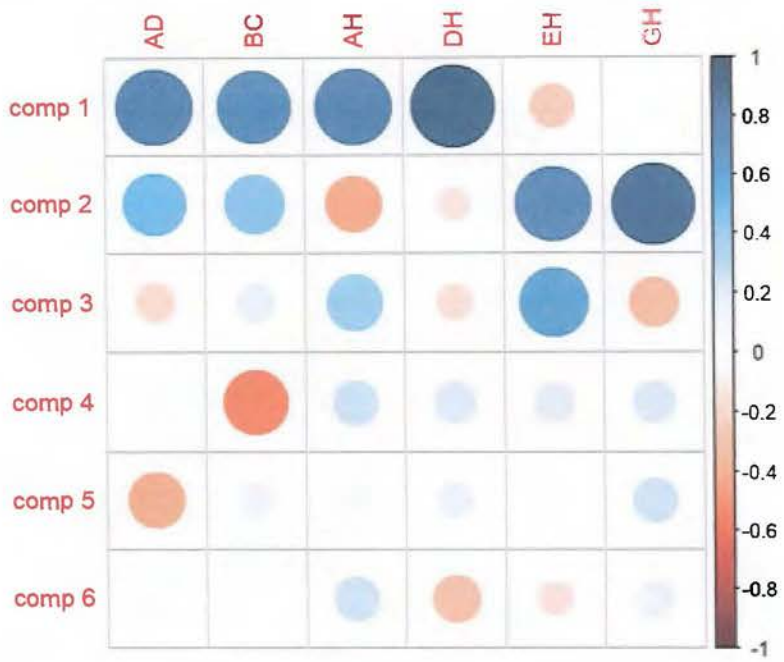


FIGURA 4.26: Representación gráfica de la matriz de correlaciones método de vértices.

En la tabla 4.62, se muestra la contribución de cada variable en el ACP de vértices. En el primer componente principal las variables que más contribuyen son DH y AD, mientras que en el segundo componente son AH y EH.

	comp 1	comp 2	comp 3	comp 4	comp 5	comp 6
AD	22.62	12.91	5.59	0.73	57.01	1.13
BC	19.25	10.52	5.00	57.28	7.83	0.13
AH	22.51	8.77	22.46	13.14	1.68	31.44
DH	32.78	1.14	3.80	9.87	9.90	42.51
EH	2.66	27.36	49.15	8.25	0.80	11.78
GH	0.17	39.31	14.00	10.72	22.77	13.02

TABLA 4.62: Contribuciones de las variables ACP de vértices.

La figura 4.27 muestra los individuos en el primer plano principal (72.01 % de la varianza), en este plano se pueden obtener 5 grupos *INC*, *KHA*, *LOT*, *ISA* y *ROM*.

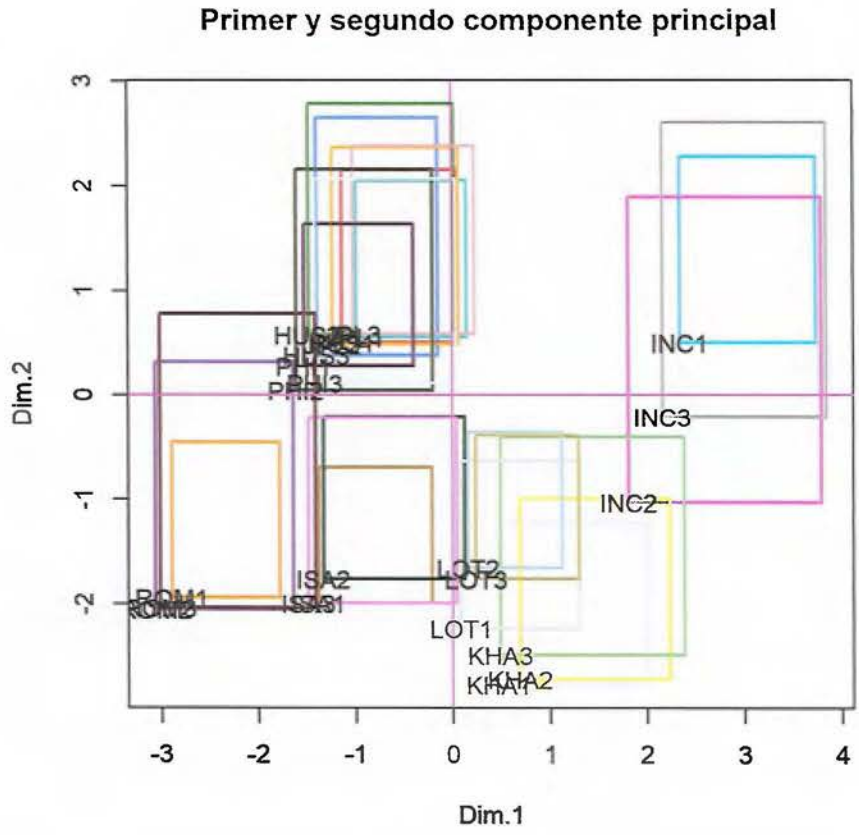


FIGURA 4.27: Primer y segundo componente principal del ACP de vértices.

La figura 4.28 muestra los individuos en el plano principal formado por el primer y tercer componente principal (51.78 % de la varianza), en este plano se pueden obtener una separación (que no se da en el plano principal) para los grupos *JPL*, *HUS* y *PHI*, algunos individuos de estos grupos se intersecan, como por ejemplo *PHI1* y *JPL3*.

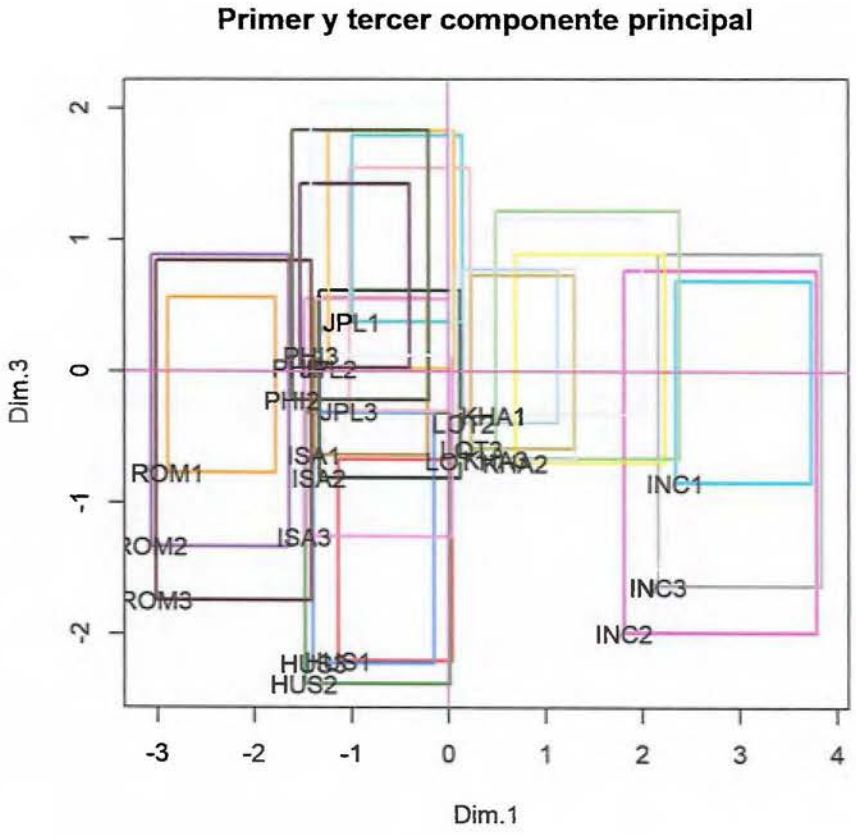


FIGURA 4.28: Primer y tercer componente principal del ACP de vértices.

4.1.2.5. Comparación de métodos

En esta sección se realizará una comparación de los métodos antes expuestos, utilizando los datos de la tabla 4 32, datos de reconocimiento facial.

4.1.2.5.1 Varianza

En la tabla 4 63, se muestra la varianza acumulada en cada uno de los componentes principales, el método de vértices es el que presenta una menor varianza en los primeros tres componentes principales 83.71 %, mientras que el ACP de maximización de varianza obtiene un 99.83 % de la varianza en los primeros tres componentes principales.

	Vértices	Centros	Z^{φ}	Z^{Λ}
comp 1	40.07 %	43.09 %	45.16 %	52.97 %
comp 2	72.01 %	80.17 %	83.21 %	87.38 %
comp 3	83.71 %	90.50 %	92.34 %	99.83 %
comp 4	91.80 %	96.34 %	96.45 %	100.00 %
comp 5	96.58 %	99.25 %	99.59 %	100.00 %
comp 6	100.00 %	100.00 %	100.00 %	100.00 %

TABLA 4.63: Comparación de la varianza para diferentes ACP, datos de reconocimiento facial.

Independiente del componente principal la matriz Z^{Λ} es mejor (respecto a la varianza acumulada) en todos los componentes.

4.1.2.5.2 Distancias

En la tabla 4 64, se muestran las distancias de los vértices a cada uno de los componentes principales, para este caso la mayor distancia (11392.81) se alcanza en la matriz centros y la menor distancia (5929.60) se obtiene con Z^{φ}.

Vértices	Centros	Z^{φ}	Z^{Λ}
9216.00	11392.81	5929.60	11063.17

TABLA 4.64: Comparación de la distancia para diferentes ACP, datos de reconocimiento facial.

4.1.2.5.3 Cosenos cuadrados

En las tablas 4.65 y 4.66 se muestran los valores de los cosenos cuadrados para cada individuo, se puede notar que entre los diversos tipos de ACP existe consistencia en estos resultados, esto quiere decir que si un individuo se encuentra bien representado (mayor a 50 %) en un tipo de ACP, en los demás también se encuentra muy bien representado.

1. **Primer componente principal.**

	Vértices	**Centros**	Z^{φ}	Z^{Λ}
HUS1	[0,0.26]	[0,0.29]	[0,0.17]	[0,0.19]
HUS2	[0,0.39]	[0,0.41]	[0,0.27]	[0,0.29]
HUS3	[0,0.37]	[0.01,0.4]	[0,0.28]	[0,0.28]
INC1	[0.55,0.97]	[0.48,0.96]	[0.69,0.96]	[0.69,0.97]
INC2	[0.43,0.92]	[0.37,0.91]	[0.38,0.98]	[0.21,0.98]
INC3	[0.49,0.92]	[0.42,0.92]	[0.6,0.9]	[0.55,0.89]
ISA1	[0.02,0.62]	[0.01,0.56]	[0.1,0.79]	[0.1,0.73]
ISA2	[0,0.63]	[0,0.58]	[0,0.72]	[0,0.7]
ISA3	[0,0.68]	[0,0.63]	[0.01,0.77]	[0.01,0.71]
JPL1	[0,0.27]	[0,0.26]	[0,0.18]	[0,0.11]
JPL2	[0,0.41]	[0,0.41]	[0,0.31]	[0,0.17]
JPL3	[0,0.36]	[0,0.36]	[0,0.24]	[0,0.15]
KHA1	[0.04,0.5]	[0.04,0.5]	[0,0.34]	[0,0.23]
KHA2	[0.09,0.62]	[0.08,0.62]	[0.02,0.47]	[0,0.36]
KHA3	[0.06,0.75]	[0.05,0.75]	[0.01,0.62]	[0,0.55]

Tabla 4.65 – Continúa en la siguiente página

Tabla 4.65 – Continúa de la página anterior

	Vértices	Centros	Z^{φ}	Z^{Λ}
LOT1	[0,0.33]	[0.01,0.37]	[0,0.26]	[0,0.23]
LOT2	[0.01,0.32]	[0.02,0.35]	[0,0.28]	[0,0.28]
LOT3	[0.02,0.32]	[0.04,0.36]	[0,0.29]	[0,0.3]
PHI1	[0.05,0.75]	[0.05,0.72]	[0.01,0.65]	[0,0.5]
PHI2	[0.01,0.79]	[0.01,0.76]	[0,0.74]	[0,0.55]
PHI3	[0,0.68]	[0,0.65]	[0,0.6]	[0,0.37]
ROM1	[0.47,0.9]	[0.4,0.86]	[0.68,0.97]	[0.65,0.96]
ROM2	[0.45,0.97]	[0.38,0.97]	[0.66,0.96]	[0.62,0.98]
ROM3	[0.4,0.88]	[0.34,0.87]	[0.46,0.92]	[0.47,0.87]

TABLA 4.65: Comparación del coseno cuadrado (primer componente principal) para diferentes ACP, datos de reconocimiento facial.

Para los datos de reconocimiento facial el ACP que cuenta con mayor número de individuos bien representados en el primer componente principal es el ACP de Z^{φ}.

2. Segundo componente principal.

	Vértices	Centros	Z^{φ}	Z^{Λ}
HUS1	[0.07,0.65]	[0.06,0.66]	[0.07,0.66]	[0.16,0.86]
HUS2	[0.11,0.78]	[0.09,0.78]	[0.12,0.8]	[0.23,0.96]
HUS3	[0.05,0.73]	[0.04,0.73]	[0.07,0.74]	[0.14,0.92]
INC1	[0.02,0.41]	[0.03,0.49]	[0,0.2]	[0,0.21]
INC2	[0,0.36]	[0,0.43]	[0,0.21]	[0,0.22]
INC3	[0,0.46]	[0,0.54]	[0,0.24]	[0,0.26]
ISA1	[0.27,0.84]	[0.33,0.85]	[0.13,0.78]	[0.12,0.62]
ISA2	[0.03,0.87]	[0.06,0.89]	[0.01,0.83]	[0,0.67]
ISA3	[0.05,0.79]	[0.08,0.81]	[0.01,0.79]	[0,0.57]
JPL1	[0.16,0.68]	[0.2,0.7]	[0.16,0.76]	[0.11,0.67]
JPL2	[0.15,0.69]	[0.18,0.71]	[0.15,0.78]	[0.15,0.75]

Tabla 4.66 – Continúa en la siguiente página

Tabla 4.66 – Continúa de la página anterior

	Vértices	**Centros**	Z^{φ}	Z^{Λ}
JPL3	[0.19,0.83]	[0.23,0.83]	[0.2,0.9]	[0.23,0.85]
KHA1	[0.36,0.87]	[0.37,0.86]	[0.42,0.94]	[0.58,0.94]
KHA2	[0.28,0.78]	[0.29,0.78]	[0.37,0.9]	[0.55,0.88]
KHA3	[0.06,0.72]	[0.07,0.72]	[0.12,0.85]	[0.2,0.88]
LOT1	[0.25,0.94]	[0.22,0.95]	[0.32,0.97]	[0.36,0.93]
LOT2	[0.09,0.77]	[0.06,0.78]	[0.14,0.84]	[0.17,0.96]
LOT3	[0.09,0.72]	[0.06,0.72]	[0.15,0.82]	[0.16,0.85]
PHI1	[0.04,0.67]	[0.04,0.67]	[0.1,0.79]	[0.08,0.62]
PHI2	[0,0.8]	[0,0.8]	[0.03,0.89]	[0.02,0.83]
PHI3	[0.01,0.64]	[0.02,0.65]	[0.04,0.77]	[0.01,0.7]
ROM1	[0.03,0.42]	[0.05,0.47]	[0,0.25]	[0,0.22]
ROM2	[0,0.42]	[0,0.48]	[0,0.26]	[0,0.21]
ROM3	[0,0.43]	[0,0.49]	[0,0.28]	[0,0.25]

TABLA 4.66: Comparación del coseno cuadrado (segundo componente principal) para diferentes ACP, datos de reconocimiento facial.

Para el segundo componente los individuos que mejor se representan son *KHA1* y *KHA2*, el ACP que cuenta con mayor número de individuos bien representados en el primer componente principal es el ACP de Z^{Λ}.

4.1.2.5.4 Calidades

La calidad de un individuo en s componentes principales (no necesariamente consecutivos) es la suma de sus cosenos cuadrados en dichos s componentes principales.

1. Componentes principales 1 y 2.

	Vértices	**Centros**	Z^{φ}	Z^{Λ}
HUS1	[0.11,0.76]	[0.1,0.78]	[0.07,0.69]	[0.17,0.87]
HUS2	[0.19,0.88]	[0.18,0.88]	[0.13,0.83]	[0.26,0.96]
HUS3	[0.16,0.81]	[0.16,0.82]	[0.11,0.76]	[0.22,0.93]
INC1	[0.82,1]	[0.83,0.99]	[0.74,1]	[0.75,1]
INC2	[0.45,0.98]	[0.41,0.97]	[0.42,0.98]	[0.26,0.99]
INC3	[0.68,0.98]	[0.66,0.98]	[0.6,0.98]	[0.55,0.99]
ISA1	[0.49,0.99]	[0.5,0.98]	[0.56,0.99]	[0.46,0.95]
ISA2	[0.15,0.99]	[0.15,0.99]	[0.2,0.97]	[0.12,0.98]
ISA3	[0.17,0.97]	[0.17,0.96]	[0.23,0.98]	[0.14,0.9]
JPL1	[0.18,0.82]	[0.22,0.84]	[0.16,0.8]	[0.12,0.68]
JPL2	[0.21,0.87]	[0.23,0.89]	[0.18,0.85]	[0.15,0.81]
JPL3	[0.25,0.88]	[0.28,0.89]	[0.22,0.9]	[0.24,0.86]
KHA1	[0.67,1]	[0.66,0.99]	[0.6,0.99]	[0.64,0.97]
KHA2	[0.72,0.99]	[0.74,0.99]	[0.67,0.98]	[0.61,1]
KHA3	[0.46,0.99]	[0.44,0.99]	[0.4,0.98]	[0.41,0.99]
LOT1	[0.38,0.98]	[0.39,0.98]	[0.38,0.97]	[0.46,0.95]
LOT2	[0.17,0.89]	[0.16,0.91]	[0.18,0.88]	[0.23,0.97]
LOT3	[0.2,0.83]	[0.2,0.84]	[0.22,0.87]	[0.26,0.87]
PHI1	[0.34,0.87]	[0.35,0.86]	[0.37,0.88]	[0.17,0.78]
PHI2	[0.23,0.91]	[0.23,0.9]	[0.26,0.9]	[0.1,0.88]
PHI3	[0.11,0.92]	[0.13,0.92]	[0.11,0.94]	[0.06,0.8]
ROM1	[0.77,0.99]	[0.75,0.98]	[0.8,0.99]	[0.8,1]
ROM2	[0.69,0.97]	[0.65,0.97]	[0.71,0.98]	[0.69,0.98]
ROM3	[0.46,0.94]	[0.42,0.92]	[0.48,0.96]	[0.48,0.93]

TABLA 4.67: Comparación de las coordenadas en el primer plano principal, para diferentes ACP, datos de reconocimiento facial.

Los únicos individuos muy bien representados en el primer plano principal (para todos los ACP) son *ROM1* y *INC1*.

2. **Componentes principales 1, 2 y 3.**

	Vértices	**Centros**	Z^{φ}	Z^{Λ}
HUS1	[0.59,0.99]	[0.67,0.98]	[0.74,0.99]	[0.62,1]
HUS2	[0.62,0.99]	[0.71,0.99]	[0.75,1]	[0.61,0.99]
HUS3	[0.46,1]	[0.55,0.98]	[0.62,1]	[0.41,0.99]
INC1	[0.83,1]	[0.83,1]	[0.74,1]	[0.77,1]
INC2	[0.53,0.99]	[0.5,0.98]	[0.47,0.99]	[0.29,1]
INC3	[0.7,1]	[0.69,0.99]	[0.62,0.99]	[0.55,1]
ISA1	[0.49,1]	[0.5,0.99]	[0.57,1]	[0.51,0.99]
ISA2	[0.18,0.99]	[0.18,0.99]	[0.23,0.99]	[0.21,1]
ISA3	[0.24,1]	[0.25,0.99]	[0.3,0.99]	[0.29,0.99]
JPL1	[0.56,0.98]	[0.62,1]	[0.52,1]	[0.65,1]
JPL2	[0.39,0.98]	[0.42,0.99]	[0.36,0.98]	[0.49,0.99]
JPL3	[0.32,0.99]	[0.38,0.99]	[0.37,1]	[0.37,0.97]
KHA1	[0.75,1]	[0.72,1]	[0.66,0.99]	[0.64,0.99]
KHA2	[0.76,0.99]	[0.76,0.99]	[0.7,1]	[0.64,1]
KHA3	[0.57,1]	[0.52,1]	[0.43,0.99]	[0.44,0.99]
LOT1	[0.4,0.98]	[0.41,0.99]	[0.46,0.97]	[0.5,1]
LOT2	[0.23,0.89]	[0.24,0.92]	[0.3,0.92]	[0.23,0.98]
LOT3	[0.26,0.84]	[0.28,0.85]	[0.35,0.88]	[0.26,0.97]
PHI1	[0.49,0.99]	[0.5,0.98]	[0.45,0.99]	[0.58,0.97]
PHI2	[0.45,0.98]	[0.44,0.98]	[0.38,0.99]	[0.56,0.96]
PHI3	[0.44,0.98]	[0.47,0.96]	[0.42,0.95]	[0.44,0.99]
ROM1	[0.78,0.99]	[0.76,0.98]	[0.81,1]	[0.81,1]
ROM2	[0.78,0.98]	[0.75,0.98]	[0.76,0.99]	[0.78,0.99]
ROM3	[0.67,0.98]	[0.61,0.96]	[0.6,0.98]	[0.68,0.96]

TABLA 4.68: Comparación de las calidades para los primeros tres componente principal, para diferentes ACP, datos de reconocimiento facial.

Los individuos muy bien representados en los componentes principales 1, 2 y 3 (para

todos los ACP) son *ROM1*, *ROM2* y *INC1*, para los demás individuos es necesario tener al menos estos tres componentes principales, esto con el fin de obtener una mejor interpretación.

4.1.2.5.5 Coordenadas

En la figura 4 29 se muestra el primer plano principal, de acuerdo con el análisis de la varianza acumulada (subsección 4 1.2.5.1) la varianza oscila entre 72.01 % y 87.38 %, se puede observar lo siguiente:

1. En todos los análisis existe una segmentación clara de 5 grupos.
2. El ACP de Z^{φ} es el que más se parece al método de vértices.
3. El único individuo bien representado (análisis de calidades 4 1.2.5.4) es *ROM1*, el cual es mejor representado por el método de vértices y después por el método Z^{φ}.

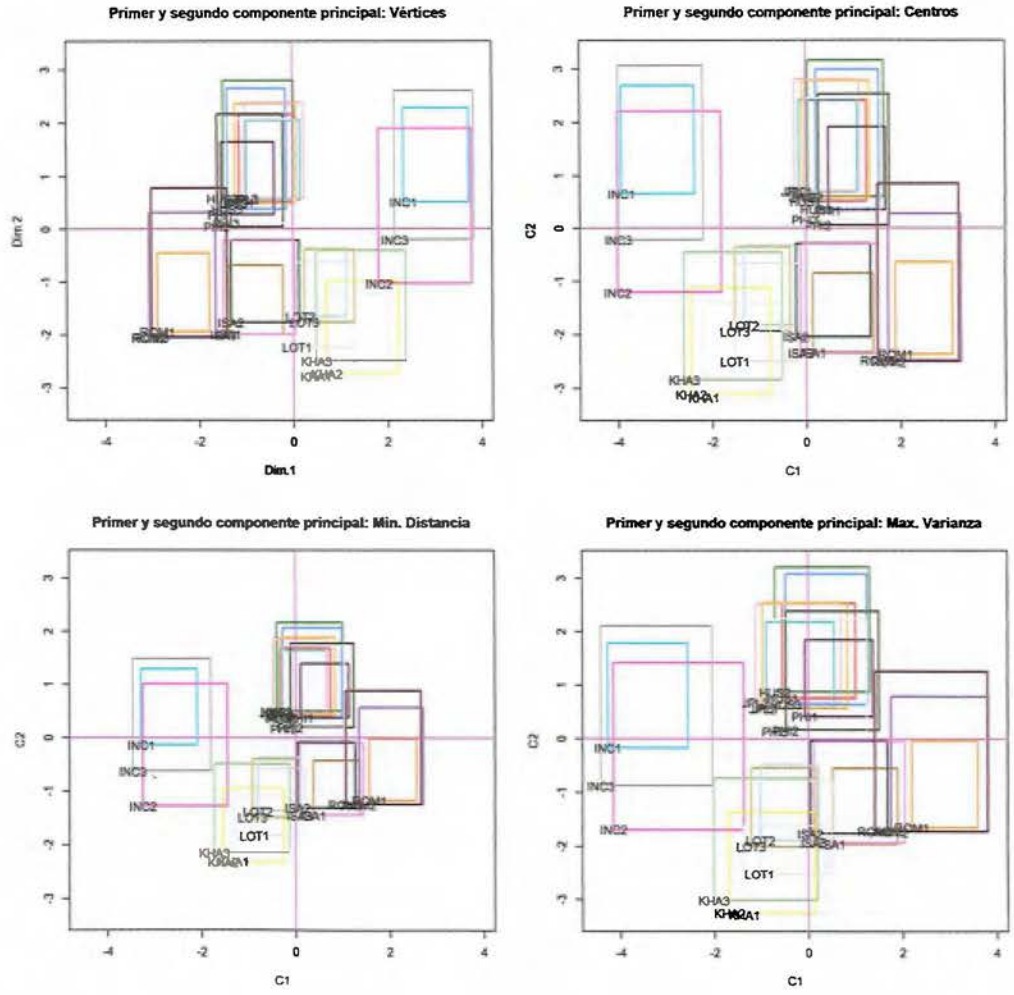


FIGURA 4.29: Comparación de ACP: datos de reconocimiento facial.

4.2. Superficies principales simbólicas vs ACP simbólico

4.2.1. Datos de Aceite

Para esta comparación se utilizan los datos de aceite de Ichino, presentados en el cuadro 4.1

Al realizar el análisis de curvas principales en los datos de Ichino los resultados se muestran en la tabla 4 69.

	I superficie 1	I superficie 2	I superficie 3	I superficie 4	I lambda
B	[1.42,1.87]	[-1.2,-1]	[-1.91,-1.54]	[0.36,0.4]	[9.3,9.92]
Ca	[-0.7,-0.51]	[-0.17,-0.02]	[0.43,0.51]	[0.19,0.22]	[6.1,6.36]
Co	[-0.52,-0.19]	[-0.35,-0.17]	[0.28,0.43]	[0.18,0.19]	[6.35,6.76]
H	[1.24,1.69]	[-1.11,-0.92]	[-1.76,-1.35]	[0.33,0.38]	[9.02,9.66]
L	[-1.26,-0.65]	[0.99,1.52]	[0.81,0.91]	[-3.88,0.35]	[0,5.03]
O	[-0.14,0.12]	[-0.49,-0.38]	[0.06,0.25]	[0.18,0.19]	[6.82,7.17]
P	[-0.6,-0.5]	[1.59,1.75]	[0.93,0.95]	[0.01,0.24]	[3.99,4.34]
S	[-0.64,-0.37]	[-0.26,-0.08]	[0.37,0.48]	[0.18,0.21]	[6.19,6.54]

TABLA 4.69: Curvas principales en datos de aceite Ichino.

La figura 4 69 representa la curva principal simbólica ($\hat{f}$) de los datos de aceite de Ichino, en la columna lambda se encuentra el intervalo de los parámetros para estimar las 4 curvas principales.

En la figura 4 30 se puede observar una clara separación entre los aceites animales B y H, además de la separación de los aceites vegetales L y P.

En la figura 4 31 se puede observar una clara separación entre los objetos simbólicos B y H, además se forman tres grupos de aceites vegetales, cuyos elementos son:

- L y P.
- Ca, Co y S.
- O.

La figura 4 32 y en la tabla 4 70, muestran la matriz de correlación entre las superficies principales y las variables.

Se puede concluir a partir de la matriz de correlaciones:

- La superficie principal 1 se encuentra muy correlacionada con las variables GRA y FRE.

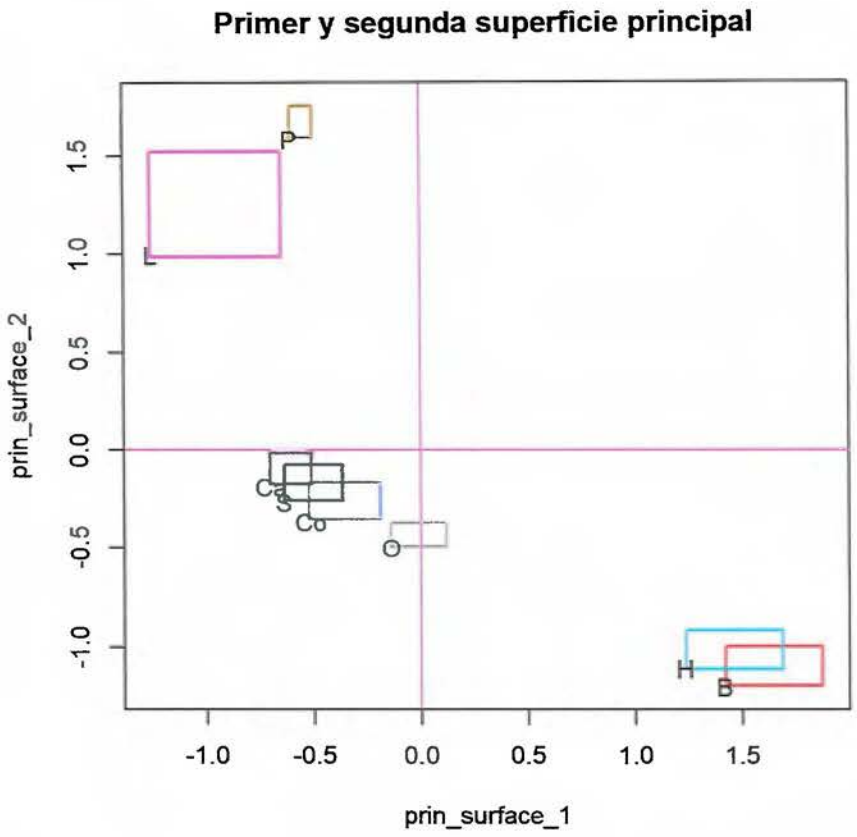


FIGURA 4.30: Primer y segunda curva principal.

	GRA	FRE	IOD	SAP
superficie 1	-0.96	0.96	-0.71	0.39
superficie 2	0.78	-0.69	0.97	-0.40
superficie 3	0.99	-0.92	0.77	-0.29
superficie 4	-0.31	0.37	-0.42	0.98

TABLA 4.70: Correlación de las variables con las superficies principales.

- La superficie principal 2 se encuentra muy correlacionada con la variable *IOD*.
- La superficie principal 4 se encuentra muy correlacionada con la variable *SAP*.

En la tabla 4.71, se muestra una comparación de la distancia de los diferentes ACP y las superficies principales, se puede observar que las superficies principales poseen la menor distacia (27.03), que es menor que la distancia de Z^{φ} (368.51).

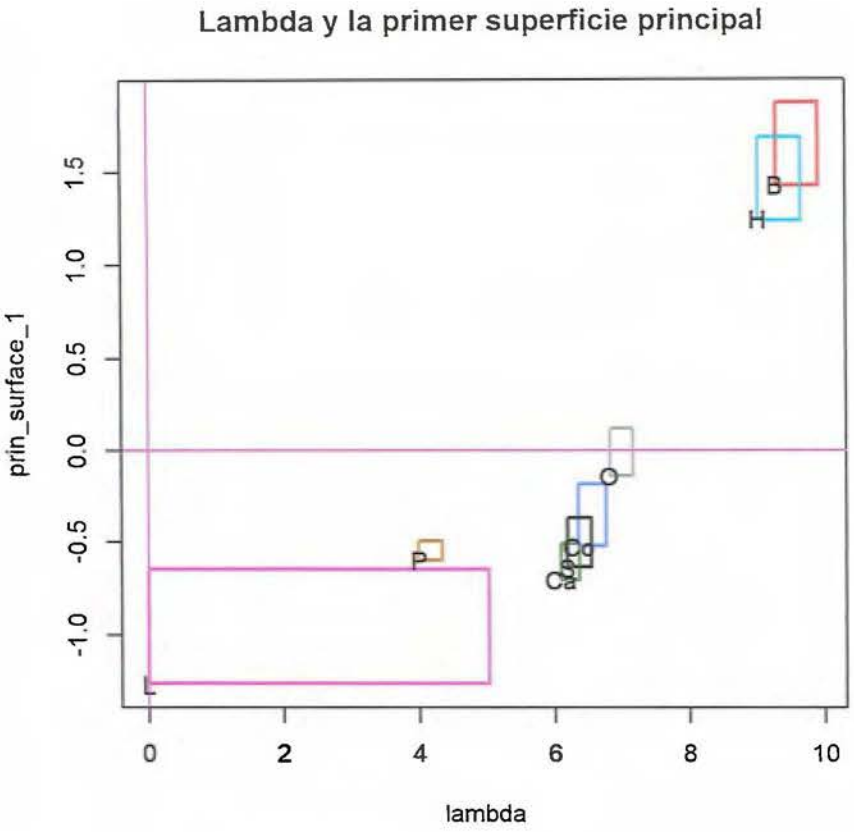


FIGURA 4.31: Primer curva principal.

vértices	centros	Z^{φ}	Z^{Λ}	superficies
512.00	706.80	368.51	982.12	27.03

TABLA 4.71: Distancia para los distintos ACP vs superficies principales.

En la figura 4.33, se muestra una comparación del ACP de vértices y las superficies principales de vértices.

De la comparación 4 33 se puede concluir:

- La variable *SAP* correlaciona mejor con la superficie principal que con el ACP, ya que correlaciona 0.84 con el segundo componente principal, mientras que con la cuarta superficie principal correlaciona un 0.98.

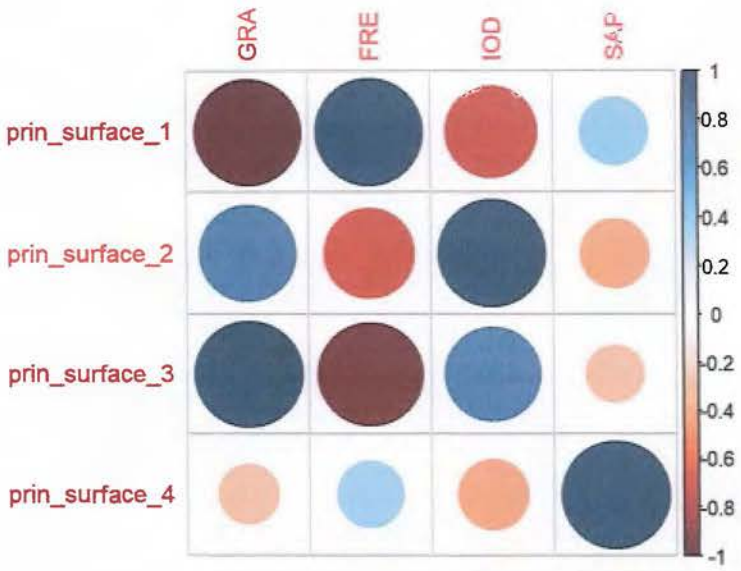


FIGURA 4.32: Correlación de las variables con las superficies principales.

- En las superficies principales se forman 4 grupos de aceites, mientras que en el ACP se forman 3 grupos, la diferencia radica en la separación del aceite vegetal *O* en el análisis de superficies principales.

4.2.2. Datos de reconocimiento facial

Para esta comparación se utilizan los datos faciales de la tabla 4.32.

Al realizar el análisis de curvas principales en los datos de Ichino los resultados se muestran en la tabla 4.72.

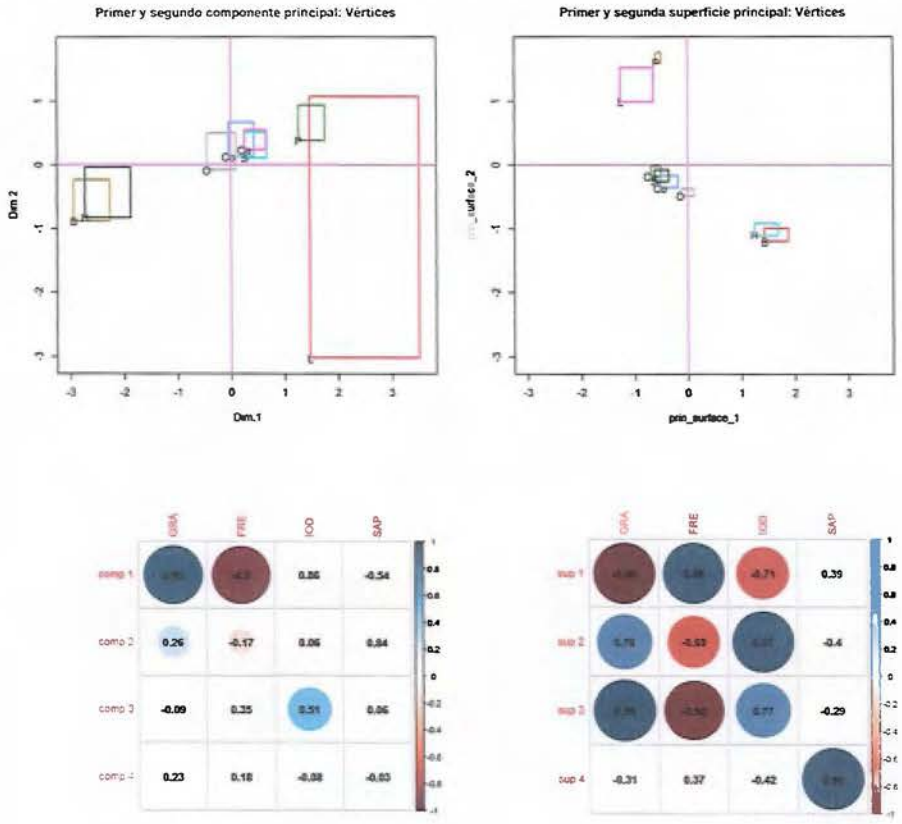


FIGURA 4.33: Comparación de las curvas principales vs los componentes principales.

	I superficie 1	I superficie 2	I superficie 3	I superficie 4	I superficie 5	I superficie 6	I lambda
HUS1	[-0.02,0.1]	[0.81,0.88]	[0.73,0.85]	[0.49,0.71]	[-0.02,0.47]	[0.53,0.74]	[3.15,3.77]
HUS2	[-0.02,0.13]	[0.78,0.88]	[0.69,0.86]	[0.44,0.73]	[-0.08,0.54]	[0.48,0.74]	[3.07,3.88]
HUS3	[-0.02,0.16]	[0.81,0.88]	[0.72,0.87]	[0.49,0.76]	[-0.15,0.47]	[0.43,0.74]	[2.97,3.78]
INC1	[-2.34,-1.85]	[-0.96,-0.68]	[-0.71,-0.53]	[-2.44,-2.02]	[0.41,1.1]	[0.4,1.51]	[10.93,12.43]
INC2	[-2.21,-1.26]	[-1.23,-0.76]	[-0.87,-0.58]	[-2.33,-1.51]	[-0.17,0.91]	[-0.45,1.21]	[9.6,12.02]
INC3	[-2.49,-1.52]	[-1.13,-0.6]	[-0.81,-0.46]	[-2.57,-1.75]	[0.05,1.32]	[-0.13,1.86]	[10.14,12.9]
ISA1	[0.38,1.02]	[-0.68,0.85]	[-0.83,0.97]	[0.03,1.09]	[-1.37,0.2]	[-0.82,-0.14]	[0.99,6.8]
ISA2	[0.39,0.84]	[-0.68,0.86]	[-0.83,0.94]	[0.04,1.03]	[-1.13,0.36]	[-0.58,0.03]	[1.4,6.79]
ISA3	[0.36,1.02]	[-0.74,0.85]	[-0.86,0.97]	[0.01,1.09]	[-1.37,0.33]	[-0.82,0]	[0.99,6.89]
JPL1	[0.11,0.37]	[0.35,0.69]	[0.1,0.56]	[0.22,0.34]	[0.59,0.71]	[0.29,0.66]	[4.14,4.91]
JPL2	[0.04,0.38]	[0.34,0.75]	[0.08,0.65]	[0.22,0.4]	[0.57,0.71]	[0.27,0.72]	[3.97,4.94]
JPL3	[0.02,0.38]	[0.34,0.77]	[0.09,0.68]	[0.22,0.43]	[0.56,0.71]	[0.28,0.73]	[3.91,4.93]
KHA1	[-0.47,0.07]	[-1.36,-1.12]	[-0.99,-0.97]	[-0.81,-0.29]	[-0.53,-0.48]	[-0.97,-0.88]	[7.56,8.37]
KHA2	[-0.55,0.02]	[-1.36,-1.16]	[-0.99,-0.97]	[-0.88,-0.33]	[-0.53,-0.49]	[-0.97,-0.9]	[7.63,8.47]
KHA3	[-0.88,0.36]	[-1.36,-0.74]	[-0.99,-0.86]	[-1.17,0.01]	[-0.53,-0.29]	[-0.97,-0.62]	[6.89,8.96]
LOT1	[0.07,0.46]	[-1.12,-0.44]	[-0.97,-0.68]	[-0.29,0.13]	[-0.48,-0.09]	[-0.88,-0.42]	[6.41,7.55]
LOT2	[0.21,0.49]	[-0.97,-0.3]	[-0.94,-0.58]	[-0.14,0.16]	[-0.41,0.01]	[-0.78,-0.31]	[6.17,7.27]
LOT3	[0.2,0.48]	[-0.99,-0.33]	[-0.95,-0.61]	[-0.16,0.15]	[-0.42,-0.01]	[-0.79,-0.34]	[6.23,7.3]
PHI1	[-0.01,0.33]	[0.43,0.88]	[0.21,0.88]	[0.23,0.79]	[-0.23,0.71]	[0.37,0.73]	[2.85,4.73]

Tabla 4.72 – Continúa en la siguiente página

Tabla 4.72 – Continúa de la página anterior

	I superficie 1	I superficie 2	I superficie 3	I superficie 4	I superficie 5	I superficie 6	I lambda
PHI2	[-0.02,0.39]	[0.3,0.88]	[0.04,0.89]	[0.22,0.84]	[-0.39,0.71]	[0.23,0.74]	[2.62,5.01]
PHI3	[-0.02,0.41]	[0.25,0.88]	[-0.02,0.87]	[0.21,0.75]	[-0.14,0.71]	[0.19,0.74]	[2.99,5.12]
ROM1	[0.86,1.39]	[0.77,0.83]	[0.94,1.03]	[1.04,1.21]	[-1.87,-1.15]	[-1.41,-0.58]	[0.12,1.36]
ROM2	[0.51,1.44]	[0.77,0.86]	[0.91,1.04]	[0.92,1.23]	[-1.94,-0.67]	[-1.49,-0.04]	[0,2.17]
ROM3	[0.28,1.41]	[0.77,0.88]	[0.88,1.03]	[0.82,1.22]	[-1.9,-0.34]	[-1.45,0.27]	[0.07,2.69]

TABLA 4.72: Curvas principales en datos faciales.

La tabla 4.72 representa la curva principal simbólica ($\hat{f}$) de los datos de aceite de reconocimiento facial, en la columna lambda se encuentra el intervalo de los parámetros para estimar las 6 curvas principales.

En la figura 4.34 se pueden observar 4 grupos *KHA*, *INC*, *LOT* e *ISA*.

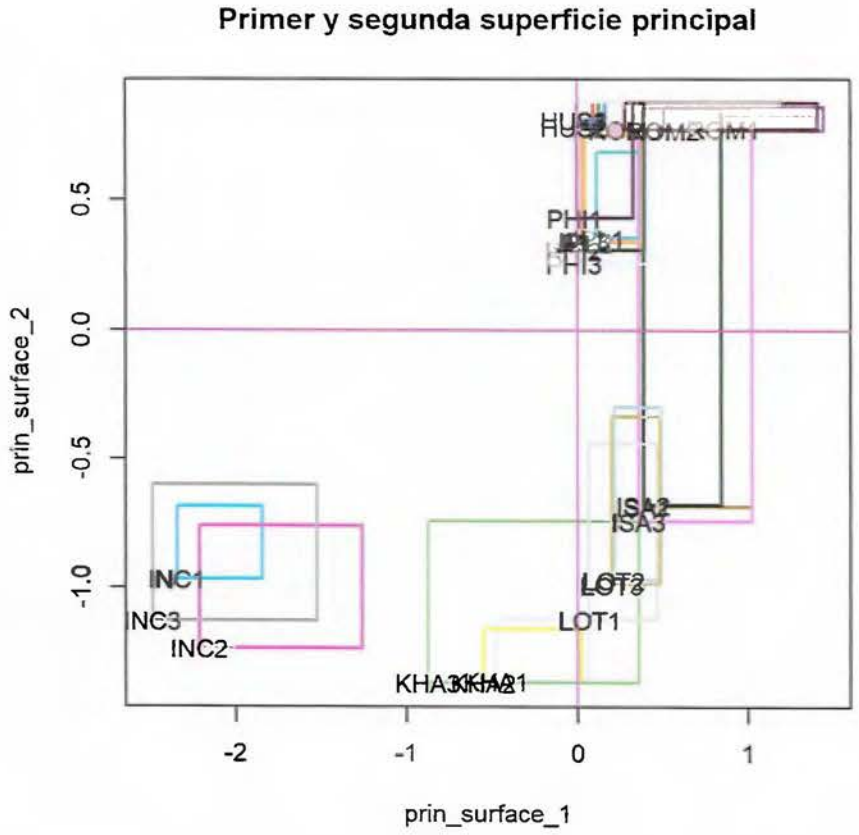


FIGURA 4.34: Primer y segunda curva principal.

En la figura 4.35 se puede observar una clara separación entre los objetos simbólicos *INC* y *KHA*, además se forman tres grupos de razas, cuyos elementos son:

- *ROM* e *ISA*.
- *PHI* y *JPL*.
- *LOT* y *KHA*.

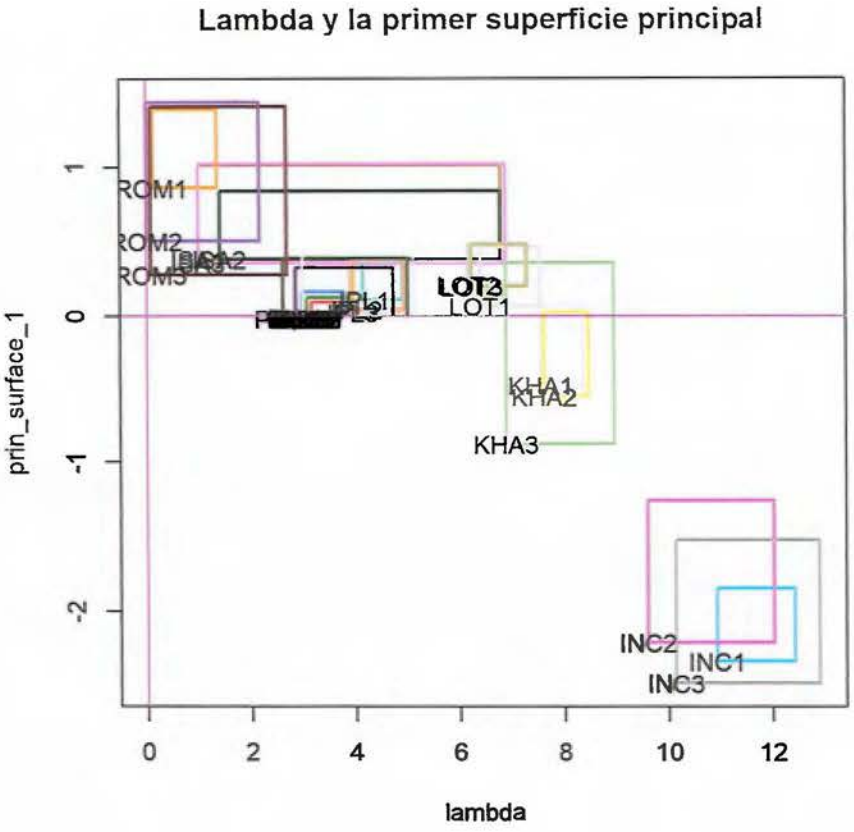


FIGURA 4.35: Primer curva principal.

La tabla 4 73, muestran la matriz de correlación entre las superficies principales y las variables.

	AD	BC	AH	DH	EH	GH
superficie 1	0.43	0.32	0.88	0.87	-0.36	-0.34
superficie 2	0.88	0.77	0.35	0.65	0.12	0.39
superficie 3	0.83	0.80	0.28	0.60	0.04	0.34
superficie 4	0.63	0.52	0.74	0.93	-0.26	-0.10
superficie 5	0.07	0.00	-0.47	-0.45	0.81	0.68
superficie 6	0.38	0.30	-0.50	-0.26	0.71	0.84

TABLA 4.73: Correlación de las variables con las superficies principales.

Se puede concluir a partir de la matriz de correlaciones:

- La superficie principal 1 se encuentra muy correlacionada con las variables *DH* y *AH*.

- La superficie principal 2 se encuentra muy correlacionada con la variable *AD*.
- La superficie principal 6 se encuentra muy correlacionada con la variable *GH*.

En la tabla 4 74, se muestra una comparación de la distancia de los diferentes ACP y las superficies principales, se puede observar que las superficies principales poseen la menor distacia (2512.21), que es menor que la distancia de Z^{φ} (5929.60).

vertices	centros	Z^{φ}	Z^{Λ}	superficies
9216.00	11392.81	5929.60	11063.17	2512.21

TABLA 4.74: Distancia distintos ACP vs superficies principales.

En la figura 4 36, se muestra una comparación del ACP de vértices y las superficies principales de vértices.

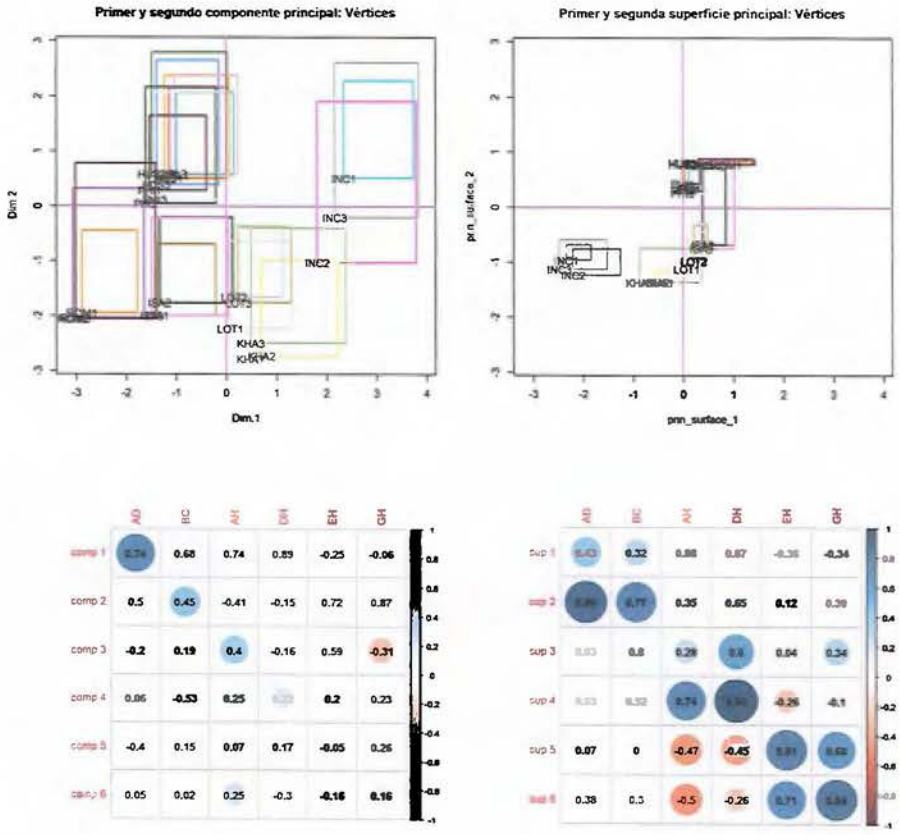


FIGURA 4.36: Comparación de las curvas principales vs los componentes principales.

De la comparación mostrada en la figura 4 33 se puede concluir:

- La variable BC no se encuentra correlacionada con la quinta superficie principal.
- Las superficies principales separan mejor que los componentes principales a la raza *INC*.
- La variable DH correlaciona mejor con la sexta curva principal, mientras que en el ACP correlaciona mejor con el primer componente principal.

Capítulo 5

Conclusiones y Recomendaciones

En este trabajo se ha generalizado el método de centros (expuesto en la sección 2.5.2.2) a un método de componentes principales que se aplica a cualquier matriz que pertenezca a una matriz de intervalos (sección 3.1) Las coordenadas de los vértices en el método de centros vienen dadas por (2.42) y (2.43), estas fórmulas son generalizadas en el teorema 3.1. Respecto a la dualidad para el método de centros propuesto por [Rodríguez,O. (2012)] (sección 2.5.2.3), las coordenadas en el círculo de correlación vienen dadas en el teorema 2.11, se generaliza en el teorema 3.2. Los teoremas 3.1 y 3.2, permiten encontrar las coordenadas de los vértices de manera explícita (sección 3.1), para la selección de los puntos se realiza por medio de dos criterios de optimización:

1. Minimizar la distancia al cuadrado de los vértices a los ejes principales (sección 3.1.1)
2. Maximizar la varianza en los primeros componentes (sección 3.1.2)

Además se propone el algoritmo (6), con el cual se construyen las curvas principales para variables simbólicas de tipo intervalo, para futuros trabajos se deben generalizar los conceptos de esperanza condicionada y autoconsistencia a variables simbólicas de tipo intervalo, esto con el fin de construir de manera formal las curvas principales para variables de tipo intervalo. Los resultados expuestos en el capítulo 4.1, cumplen lo esperado para los casos analizados y esto es:

1. Z^{φ} posee menor distancia al cuadrado de los vértices a los ejes principales.

2. Z^{Λ} maximizar la varianza en los primeros s componentes.

Por su parte las curvas principales simbólicas segmentan mejor los individuos, con lo cual es más sencillo interpretar los resultados. En futuros trabajos se pueden definir nuevas funciones para la selección de la matriz óptima (Z^*), de manera similar se pueden utilzar los conceptos expuestos en esta tesis, con el fin de mejorar otros métodos de reducción de la dimensionalidad (análisis de correspondencias simples, análisis de correspondencias múltiples, entre otros). Por último la implementación de los métodos expuestos en este trabajo fue realizada en **R**, esta implementación se encuentra en el paquete **RSDA** [Rodríguez, O. with contributions from Olger Calderón, Roberto Zúñiga and Jorge Arce (2015)].

Apéndice A

Código en R

A.1. Código

Para el desarrollo de los ejemplos se programaron las siguientes funciones en **R**. Para los objetos simbólicos se utiliza el paquete **RSDA** [Rodríguez, O. with contributions from Olger Calderón, Roberto Zúñiga and Jorge Arce (2015)], para la construcción de las curvas principales se utiliza el paquete **princurve** [Hastie, T.; Weingessel, A. (2014)].

A.1.1. Código para generar LaTeX

Estas funciones tienen como objetivo, escribir objetos **RSDA** en LaTeX.

- **generate.columns.set:** Genera una visualización para las variables simbólicas de tipo conjunto.

```
generate.columns.set<-function(data)
{
  data.colnames<-colnames(data)
  sal<-apply(data,1,function(x) {
    paste0(paste0("{",paste(data.colnames[x==1], collapse = ",")),"}")
```

```
  })
  return(sal)
}
```

- **generate.columns.interval:** Genera una visualización para las variables simbólicas de tipo intervalo.

```
generate.columns.interval<-function(data)
{
  min.char<-as.character(round(data[,1],2))
  max.char<-as.character(round(data[,2],2))
  return(paste0(paste0("[",paste(min.char,max.char,sep = ",")),"]"))
}
```

- **generate.columns.multivalued:** Genera una visualización para las variables simbólicas de tipo multievaluadas.

```
generate.columns.multivalued<-function(data)
{
  data.colnames<-colnames(data)
  sal<-apply(data,1,function(x) {
    paste0(paste0("{",paste( paste(data.colnames[x>0], x[x>0] , sep =
    ↪ ";") , collapse = ",")),"}")
  })
  return(sal)
}
```

- **generate.sym.table:** Genera una visualización para una tabla de datos simbólica.

```
generate.sym.table<-function(sym.data)
{
  sym.var.names<-sym.data$sym.var.names
  sym.var.starts<-sym.data$sym.var.starts
```

```
sym.var.length<-sym.data$sym.var.length
sym.var.types<-sym.data$sym.var.types
sym.obj.names<-sym.data$sym.obj.names
sym.var.names.length<-length(sym.var.names)
sym.obj.names.length<-length(sym.obj.names)
sym.tbl.columns<-rep("X",sym.var.names.length)
sym.tbl<-matrix(rep("X",sym.obj.names.length*sym.var.names.length),
 ↪ nrow=sym.obj.names.length,ncol=sym.var.names.length, byrow =
 ↪ TRUE)
for (i in 1:sym.var.names.length) {
  data<-sym.data$meta[,sym.var.starts[i]:(sym.var.starts[i]+sym.var.
 ↪ length[i]-1)]
  switch(sym.var.types[i], '$C' = {
    sym.tbl[,i]<-as.character(data)
    sym.tbl.columns[i]<-paste0('C ', sym.var.names[i])
  }, '$I' = {
    sym.tbl[,i]<-generate.columns.interval(data)
    sym.tbl.columns[i]<-paste0('I ', sym.var.names[i])
  }, '$H' = {
    sym.tbl[,i]<-generate.columns.multivalued(data)
    sym.tbl.columns[i]<-paste0('H ', sym.var.names[i])
  }, '$S' = {
    sym.tbl[,i]<-generate.columns.set(data)
    sym.tbl.columns[i]<-paste0('S ', sym.var.names[i])
  }, stop("Invalid variable type"))
}
sym.tbl<-as.data.frame(sym.tbl)
colnames(sym.tbl)<- sym.tbl.columns
row.names(sym.tbl)<-sym.obj.names
return(sym.tbl)
```

```
}
```

- **RSDA.to.latex:** Genera el código LaTeXpara una tabla de datos simbólica (formato **RSDA**).

```
RSDA.to.latex<-function(sym.data)
{
  return(xtable(generate.sym.table(sym.data)))
}
```

A.1.2. Código para calcular matrices de vértices y centros

- **vertex.interval.new.j:** Genera la matriz de vértices de una matriz simbólica de tipo intervalo definida en (35).

```
vertex.interval.new.j<-function (sym.data)
{
  if ((sym.data$sym.var.types[1] != "$I")) {
    stop("Variables have to be Interval")
  }
  else {
    nn <- sym.data$N
    mm <- sym.data$M
    num.vertex <- rep(-1, nn)
    vertex <- matrix(0, 1, mm)
    vertex <- as.data.frame(vertex)
    colnames(vertex) <- sym.data$sym.var.names
    sym.text <- "as.matrix(sym.data$data["
    for (i in 1:nn) {
      current.row <- as.character(i)
      previous <- "1:2"
```

```
      command <- paste0(sym.text, current.row, ",", previous,
                        "])")
      for (j in 2:mm) {
        col.current.min <- 2 * j - 1
        col.current.max <- 2 * j
        nxt.grid <- paste0(as.character(col.current.min),
                           ":", as.character(col.current.max))
        command <- paste0(command, ",", sym.text, current.row,
                          ",", nxt.grid, "])")
      }
      command <- paste0("expand.grid(", command, ")")
      aux <- eval(parse(text = command))
      aux <- sqldf("select distinct * from aux")
      num.vertex[i] <- dim(aux)[1]
      colnames(aux) <- sym.data$sym.var.names
      vertex <- rbind(vertex, aux)
    }
    num.vertexf <- dim(vertex)[1]
    return(list(vertex = vertex[2:num.vertexf, ], num.vertex = num.
    ↪ vertex))
  }
}
```

- **centers.interval.j:** Genera la matriz de centros de una matriz simbólica de tipo intervalo definida en 2.39.

```
centers.interval.j<-function (sym.data)
{
  idn <- all(sym.data$sym.var.types == sym.data$sym.var.types[1])
  if (idn == FALSE)
    stop("All variables have to be of the same type")
  if ((sym.data$sym.var.types[1] != "$I"))
```

```
    stop("Variables have to be continuos or Interval")
  else nn <- sym.data$N
  mm <- sym.data$M
  centers <- matrix(0, nn, mm)
  ratios <- matrix(0, nn, mm)
  centers <- as.data.frame(centers)
  ratios <- as.data.frame(ratios)
  rownames(centers) <- sym.data$sym.obj.names
  colnames(centers) <- sym.data$sym.var.names
  rownames(ratios) <- sym.data$sym.obj.names
  colnames(ratios) <- sym.data$sym.var.names
  for (i in 1:nn) {
    for (j in 1:mm) {
      sym.var.act <- sym.var(sym.data, j)
      min.val <- sym.var.act$var.data.vector[i, 1]
      max.val <- sym.var.act$var.data.vector[i, 2]
      centers[i, j] <- (min.val + max.val)/2
      ratios[i, j] <- (-min.val + max.val)/2
    }
  }
  return(list(centers = centers, ratios = ratios))
}
```

A.1.3. Código para calcular los límites del ACP general

- **get.limits.PCA:** Encuentra las coordenadas de los individuos en el ACP, aplicando el teorema 3.1.

```
get.limits.PCA<-function(sym.data,matrix.stan,min.stan,max.stan,svd,nn
    ↪ ,mm){
```

```
sym.comp <- sym.data
for (i in 1:nn) {
  posd <- 1
  for (j in 1:mm) {
    smin <- 0
    smax <- 0
    for (k in 1:mm) {
      if (svd$vectors[k, j] < 0) {
        smin <- smin + max.stan[i, k] * svd$vectors[k,j]
        smax <- smax + min.stan[i, k] * svd$vectors[k,j]
      }
      else{
        smin <- smin + min.stan[i, k] * svd$vectors[k,j]
        smax <- smax + max.stan[i, k] * svd$vectors[k,j]
      }
    }
    sym.comp$meta[i, sym.comp$sym.var.starts[j]] <- smin
    sym.comp$meta[i, sym.comp$sym.var.starts[j] + 1] <- smax
    sym.comp$data[i, posd] <- smin
    sym.comp$data[i, posd + 1] <- smax
    posd <- posd + 2
  }
}
pos <- 1
for (j in 1:mm) {
  comp.name <- paste("C", j, sep = "")
  sym.comp$sym.var.names[j] <- comp.name
  comp.name <- paste("Min.C", j, sep = "")
  colnames(sym.comp$data)[pos] <- comp.name
  comp.name <- paste("Max.C", j, sep = "")
```

```
  colnames(sym.comp$data)[pos + 1] <- comp.name
  pos <- pos + 2
  comp.name <- paste("Min.C", j, sep = "")
  colnames(sym.comp$meta)[sym.comp$sym.var.starts[j]] <- comp.name
  comp.name <- paste("Max.C", j, sep = "")
  colnames(sym.comp$meta)[sym.comp$sym.var.starts[j] + 1] <- comp.
↪ name
}

svdV <- matrix(0, nn, nn)

for (i in 1:nn) {
  for (j in 1:mm) {
    ss <- 0
    for (k in 1:mm) {
      ss <- ss + matrix.stan[i, k] * svd$vectors[k,j]
    }
    svdV[i, j] <- (1/sqrt(svd$values[j])) * ss
  }
}

IPrinCorre <- matrix(0, mm, 2 * mm)
for (i in 1:mm) {
  pcol <- 1
  for (j in 1:mm) {
    smin <- 0
    smax <- 0
    for (k in 1:nn) {
      if (svdV[k, j] < 0) {
        smin <- smin + (1/sqrt(nn)) * max.stan[k,i] * svdV[k, j]
```

```
            smax <- smax + (1/sqrt(nn)) * min.stan[k,i] * svdV[k, j]
          }
          else{
            smin <- smin + (1/sqrt(nn)) * min.stan[k,i] * svdV[k, j]
            smax <- smax + (1/sqrt(nn)) * max.stan[k,i] * svdV[k, j]
          }
        }
        IPrinCorre[i, pcol] <- smin
        IPrinCorre[i, pcol + 1] <- smax
        pcol <- pcol + 2
      }
    }
    IPrinCorre <- as.data.frame(IPrinCorre)
    rownames(IPrinCorre) <- sym.data$sym.var.names
    class(sym.comp) <- "sym.data.table"
    return(list(Sym.Components = sym.comp, Sym.Prin.Correlations =
      ↪ IPrinCorre))
}
```

- **sym.circle.plot.new:** Encuentra las coordenadas de las variables en el círculo de correlaciones, aplicando el teorema 3.2.

```
sym.circle.plot.new<-function (prin.corre,msg = paste("Correlation
   ↪ Circle"))
{
  v <- c("green", "red", "blue", "cyan", "brown", "yellow",
         "pink", "purple", "orange", "gray")

  plot(-1.5:1.5, -1.5:1.5, type = "n", xlab = "C1", ylab = "C2",
       main = msg)
  abline(h = 0, lty = 3)
  abline(v = 0, lty = 3)
```

```
symbols(0, 0, circles = 1, inches = FALSE, add = TRUE)
c1 = 1
c2 = 2
n <- dim(prin.corre)[1]
f <- dim(prin.corre)[2]
CRTI <- matrix(nrow = n, ncol = f)
CRTI <- prin.corre
vars <- rownames(prin.corre)
for (k in 1:n) {
  x1 <- min(CRTI[k, c1], CRTI[k, c2])
  x2 <- max(CRTI[k, c1], CRTI[k, c2])
  y1 <- min(CRTI[k, c2 + 1], CRTI[k, c2 + 2])
  y2 <- max(CRTI[k, c2 + 1], CRTI[k, c2 + 2])
  if (((x1 > 0) && (x2 > 0) && (y1 > 0) && (y2 > 0)) ||
      ((x1 < 0) && (x2 < 0) && (y1 < 0) && (y2 < 0))) {
    RSDA:::plotX.slice(x1, y2, x2, y1, v, vars, k)
  }
  if (((x1 < 0) && (x2 < 0) && (y1 > 0) && (y2 > 0)) ||
      ((x1 > 0) && (x2 > 0) && (y1 < 0) && (y2 < 0))) {
    RSDA:::plotX.slice(x1, y1, x2, y2, v, vars, k)
  }
  if ((y1 > 0) && (y2 > 0) && (x1 < 0) && (x2 > 0)) {
    RSDA:::plotX.slice(x1, y1, x2, y1, v, vars, k)
  }
  if ((y1 < 0) && (y2 < 0) && (x1 < 0) && (x2 > 0)) {
    RSDA:::plotX.slice(x1, y2, x2, y2, v, vars, k)
  }
  if ((x1 > 0) && (x2 > 0) && (y1 < 0) && (y2 > 0)) {
    RSDA:::plotX.slice(x1, y1, x1, y2, v, vars, k)
  }
```

```
    if ((x1 < 0) && (x2 < 0) && (y1 < 0) && (y2 > 0)) {
      RSDA:::plotX.slice(x2, y1, x2, y2, v, vars, k)
    }
    if ((x1 < 0) && (x2 > 0) && (y1 < 0) && (y2 > 0)) {
      RSDA:::plotX.slice(x2, y1, x2, y2, v, vars, k)
    }
  }
}
```

A.1.4. Código para calcular el ACP de vértices

- **sym.interval.vertex.pca.j.new:** Realiza el ACP de vértices, utilizando las ecuaciones (2.36) y (2.37) para encontrar las coordenadas de los individuos.

```
sym.interval.vertex.pca.j.new<-function (data.sym)
{
  vertex.sym <- vertex.interval.new.j(data.sym)
  data.vertex <- as.matrix(vertex.sym$vertex)
  dim.sym <- dim(data.vertex)
  indx.cols <- data.frame(i = 1:dim.sym[2])
  medias <- apply(indx.cols, 1, function(i) {
    mean(data.vertex[, i])
  })
  data.vertex.centrada <- t(t(data.vertex) - medias)
  desviaciones <- apply(indx.cols, 1, function(i) {
    sd(data.vertex[, i])
  })
  desviaciones <- desviaciones * sqrt((dim.sym[1] - 1)/dim.sym[1])
  data.vertex.centrada <- t(t(data.vertex.centrada)/desviaciones)
  matrix.data <- as.matrix(data.sym$data)
  matrix.data.centrada <- as.matrix(data.sym$data)
```

```
m <- data.sym$M
n <- data.sym$N
indx <- data.frame(pos = 1:m)
list.stand <- apply(indx, 1, function(i) {
  pos.ini <- 2 * (i - 1) + 1
  pos.fin <- 2 * i
  matrix.data.centrada[, pos.ini:pos.fin] <<- (matrix.data[,
                                                          pos.ini:
 ↪ pos.fin] - medias[i])/desviaciones[i]
})
cor.matrix <- t(data.vertex.centrada) %*% data.vertex.centrada/dim.
 ↪ sym[1]
cor.matrix.eigen <- eigen(cor.matrix)
vector.propios <- cor.matrix.eigen$vectors
vector.propios.pos <- vector.propios
vector.propios.pos <- apply(indx, 1, function(i) {
  apply(indx, 1, function(j) {
    if (vector.propios[j, i] > 0) {
      vector.propios[j, i]
    }
    else {
      0
    }
  })
})
vector.propios.neg <- vector.propios - vector.propios.pos
indx.max <- seq(2, to = 2 * m, by = 2)
indx.min <- seq(1, to = 2 * m, by = 2)
max.neg <- matrix.data.centrada[, indx.max] %*% vector.propios.neg
min.neg <- matrix.data.centrada[, indx.min] %*% vector.propios.neg
```

```
  max.pos <- matrix.data.centrada[, indx.max] %*% vector.propios.pos
  min.pos <- matrix.data.centrada[, indx.min] %*% vector.propios.pos
  maximos <- max.pos + min.neg
  minimos <- min.pos + max.neg
  names.sal <- paste0("Dim.", t(indx))
  sal.sym <- as.data.frame(matrix(rep(0, n * 2 * m), nrow = n))
  colnames(sal.sym)[indx.min] <- names.sal
  colnames(sal.sym)[indx.max] <- paste0(names.sal, ".1")
  sal.sym[, indx.max] <- maximos
  sal.sym[, indx.min] <- minimos
  row.names(sal.sym) <- data.sym$sym.obj.names
  return(list(Sym.Components = data.frame.to.RSDA.inteval.table.j(sal.
   ↪ sym),
              pos.coord.eigen = vector.propios.pos, neg.coord.eigen =
   ↪ vector.propios.neg,
              mean.vertex = medias, sd.vertex = desviaciones))
}
```

A.1.5. Código para calcular el ACP de centros

- **centers.pca.j.new:** Realiza el ACP de centros, utilizando las ecuaciones (2.42) y (2.43) para encontrar las coordenadas de los individuos.

```
centers.pca.j.new<- function(sym.data)
{
  nn <- sym.data$N
  mm <- sym.data$M
  seq.min<-seq(from = 1, to = 2*mm,by = 2)
  seq.max<-seq(from = 2, to = 2*mm,by = 2)

  centers<- RSDA:::centers.interval(sym.data)
```

```
  sym.data.vertex <- vertex.interval.new.ja(sym.data)
  sym.data.vertex.matrix <- sym.data.vertex$vertex
  dim.vertex <- dim(sym.data.vertex.matrix)[1]
  tot.individuals <- N + dim.vertex
  centers <- rbind(centers, sym.data.vertex.matrix)

  PCA.centers<-PCA(X = centers, scale.unit = TRUE,
                   ind.sup = (N + 1):tot.individuals,
                   ncp = mm, graph = FALSE)

  centers.stan.mean<-PCA.centers$call$centre
  centers.stan.stand<-PCA.centers$call$ecart.type
  centers.stan<-scale.matrix.j(centers,centers.stan.mean,centers.stan.
   ↪ stand,nn,mm)
  data<-stand.data(sym.data,centers.stan.mean,centers.stan.stand,nn,mm
   ↪ )

  svd<-list(values = PCA.centers$eig[,1],
            vectors = PCA.centers$svd$V)

  sym.PCA.res<-get.limits.PCA(sym.data,centers.stan,data[,seq.min],
                              data[,seq.max],svd,nn,mm)

  return(list(classic.PCA = PCA.centers,
              symbolic.PCA = sym.PCA.res))
}
```

A.1.6. Código para calcular las funciones $\varphi(Z)$ y $\Lambda(Z, s)$

- **pca.supplementary.vertex.fun.j.new:** Calcula la función $\varphi(Z)$ definida en (3.23), utilizando el algoritmo 2.

```
pca.supplementary.vertex.fun.j.new<-function (x, N, M, sym.var.names,
                                              sym.data.vertex.matrix,
                                              tot.individuals)
{
  M.x <- matrix(x, nrow = N)
  colnames(M.x) <- sym.var.names
  M.x <- rbind(M.x, sym.data.vertex.matrix)
  pca.min <- PCA(X = M.x, scale.unit = TRUE,
                 ind.sup = (N + 1):tot.individuals,
                 ncp = M, graph = FALSE)
  min.dist.pca <- pca.min$ind.sup$dist * pca.min$ind.sup$dist
  return(sum(min.dist.pca))
}
```

- **pca.supplementary.vertex.lambda.fun.j.new:** Calcula la función $\Lambda(Z, s)$ definida en 3.26, utilizando el algoritmo 4.

```
pca.supplementary.vertex.lambda.fun.j.new<-function (x, M, N, sym.var.
    ↪ names,
                                                     sym.data.vertex.
    ↪ matrix,
                                                     tot.individuals,
    ↪ num.dimen.aux)
{
  M.x <- matrix(x, nrow = N)
  colnames(M.x) <- sym.var.names
  M.x<-scale(M.x)
```

```
  pca.max <- PCA(X = M.x, scale.unit = FALSE, ncp = M, graph = FALSE)
  return(-sum(pca.max$eig$eigenvalue[(1:num.dimen.aux)]))
}
```

A.1.7. Código para optimizar las funciones $\varphi(Z)$ y $\Lambda(Z)$

- **optim.pca.distance.j.new:** Resuelve el problema (3.24), utilizando el algoritmo 3.

```
optim.pca.distance.j.new<-function (sym.data)
{
  N <- sym.data$N
  M <- sym.data$M
  seq.min <- seq(from = 1, by = 2, length.out = M)
  seq.max <- seq(from = 2, by = 2, length.out = M)
  sym.var.names <- sym.data$sym.var.names
  sym.data.vertex <- RSDA:::vertex.interval.new.j(sym.data)
  sym.data.vertex.matrix <- sym.data.vertex$vertex
  dim.vertex <- dim(sym.data.vertex.matrix)[1]
  tot.individuals <- N + dim.vertex
  min.interval <- as.vector(as.matrix(sym.data$data[, seq.min]))
  max.interval <- as.vector(as.matrix(sym.data$data[, seq.max]))
  init.point <- as.vector(as.matrix(RSDA:::centers.interval.j(sym.data
   ↪ )$centers))
  res.min <- lbfgs(init.point, pca.supplementary.vertex.fun.j.new,
                   lower = min.interval, upper = max.interval, nl.info
   ↪  = FALSE,
                   control = list(xtol_rel = 1e-08, maxeval = 20000),
   ↪ N = N,
                   M = M, sym.var.names = sym.var.names,
                   sym.data.vertex.matrix = sym.data.vertex.matrix,
                   tot.individuals = tot.individuals)
```

```
  M.x <- matrix(res.min$par, nrow = N)
  colnames(M.x) <- sym.var.names

  M.x <- rbind(M.x, sym.data.vertex.matrix)

  pca.min <- PCA(X = M.x, scale.unit = TRUE,
                 ind.sup = (N + 1):tot.individuals,
                 ncp = M, graph = FALSE)

  svd<-list(values = pca.min$eig$eigenvalue,
            vectors = pca.min$svd$V)

  best.stan.mean<-pca.min$call$centre
  best.stan.stand<-pca.min$call$ecart.type
  best.stan<-scale.matrix.j(M.x,best.stan.mean,best.stan.stand,N,M)

  data<-stand.data(sym.data,best.stan.mean,best.stan.stand,N,M)

  sym.PCA.res<-get.limits.PCA(sym.data,best.stan,data[,seq.min],
                              data[,seq.max],svd,N,M)

  return(list(symbolic.PCA = sym.PCA.res,
              classic.PCA = pca.min,
              res.best = res.min))
}
```

- **optim.pca.variance.j.new:** Resuelve el problema (3.27), utilizando el algoritmo 5.

```
optim.pca.variance.j.new<-function (sym.data, num.dimension)
{
  N <- sym.data$N
  M <- sym.data$M
```

```
num.dimen.aux <- num.dimension
seq.min <- seq(from = 1, by = 2, length.out = M)
seq.max <- seq(from = 2, by = 2, length.out = M)
sym.var.names <- sym.data$sym.var.names
sym.data.vertex <- RSDA:::vertex.interval.new.j(sym.data)
sym.data.vertex.matrix <-  sym.data.vertex$vertex
dim.vertex <- dim(sym.data.vertex.matrix)[1]
tot.individuals <- N + dim.vertex
min.interval <- as.vector(as.matrix(sym.data$data[, seq.min]))
max.interval <- as.vector(as.matrix(sym.data$data[, seq.max]))
init.point <- as.vector(as.matrix(RSDA:::centers.interval.j(sym.data
 ↪ )$centers))
res.min <- lbfgs(init.point, pca.supplementary.vertex.lambda.fun.j.
 ↪ new,
                 lower = min.interval, upper = max.interval, nl.info
 ↪  = FALSE,
                 control = list(xtol_rel = 1e-10, maxeval = 20000),
 ↪ N = N,
                 M = M, sym.var.names = sym.var.names,
                 sym.data.vertex.matrix = sym.data.vertex.matrix,
                 tot.individuals = tot.individuals,
                 num.dimen.aux = num.dimen.aux)
M.x <- matrix(res.min$par, nrow = N)
colnames(M.x) <- sym.var.names

M.x <- rbind(M.x, sym.data.vertex.matrix)
pca.max <- PCA(X = M.x, scale.unit = TRUE,
               ind.sup = (N + 1):tot.individuals,
               ncp = M, graph = FALSE)
```

```
    svd<-list(values = pca.max$eig$eigenvalue,
              vectors = pca.max$svd$V)

    best.stan.mean<-pca.max$call$centre
    best.stan.stand<-pca.max$call$ecart.type
    best.stan<-scale.matrix.j(M.x,best.stan.mean,best.stan.stand,N,M)

    data<-stand.data(sym.data,best.stan.mean,best.stan.stand,N,M)

    sym.PCA.res<-get.limits.PCA(sym.data,best.stan,data[,seq.min],
                                data[,seq.max],svd,N,M)

    return(list(symbolic.PCA = sym.PCA.res,
                classic.PCA = pca.max,
                res.best = res.min))
}
```

A.1.8. Código para calcular las curvas principales

- **sym.interval.pc.limits:** Genera las coordenadas de los individuos en las curvas principales, utilizando el algoritmo 6.

```
sym.interval.pc.limits<-function (sym.data, prin.curve, num.vertex,
    ↪ lambda, var.ord)
{
  num.vars <- sym.data$M
  num.ind <- sym.data$N
  res <- as.data.frame(prin.curve)
  res$lambda <- lambda
  sym.indiv <- rep("X", sum(num.vertex))
  start <- 1
```

```
  finish <- num.vertex[1]
  sym.indiv[start:finish] <- sym.data$sym.obj.names[1]
  for (i in 2:num.ind) {
    previous <- num.vertex[i - 1]
    start <- start + previous
    finish <- num.vertex[i] + finish
    sym.indiv[start:finish] <- sym.data$sym.obj.names[i]
  }
  res$symindiv <- sym.indiv
  var.type <- rep("$I", num.vars + 1)
  variables <- rep("X", num.vars)
  for (i in 1:num.vars) {
    variables[var.ord[i]] <- paste0("prin_surface_", as.character(i))
  }
  colnames(res)[1:num.vars] <- variables
  variables <- c(variables[var.ord], "lambda")
  sym.res <- classic.to.sym(dataTable = res, concept = c("symindiv"),
                            variables = variables, variables.types =
   ↪ var.type)
  return(sym.res)
}
```

- **variance.princ.curve:** Calcula la inercia en cada una de las curvas principales, según [8].

```
variance.princ.curve<-function (data, curve)
{
  var.data <- diag(var(data))
  var.curve <- diag(var(curve))
  dist <- sum((data - curve)^2)/dim(data)[1]
  ord <- order(x = var.data, decreasing = TRUE)
  var.data.cum <- cumsum(var.data[ord])
```

```
  var.curve.cum <- cumsum(var.curve[ord])
  return(list(var.data = var.data, var.data.cum = var.data.cum,
              var.curve = var.curve, var.curve.cum = var.curve.cum,
              dist = dist, var.order = ord))
}
```

- **sym.interval.pc:** Calcula la curva principal para una matriz de intervalos.

```
sym.interval.pc<-function (sym.data, method = c("vertex", "centers"),
  ↪ maxit, plot,
                           scale, center)
{
  idn <- all(sym.data$sym.var.types == sym.data$sym.var.types[1])
  if (idn == FALSE)
    stop("All variables have to be of the same type")
  method <- match.arg(method)
  if ((sym.data$sym.var.types[1] != "$C") && (sym.data$sym.var.types
   ↪ [1] !=
                                                "$I"))
    stop("Variables have to be continuos or Interval")
  else if (sym.data$sym.var.types[1] == "$C")
    res <- principal.curve(sym.data$data, plot.true = plot,
                           maxit = maxit)
  else if (sym.data$sym.var.types[1] == "$I") {
    vertex <- vertex.interval(sym.data)
    individuals <- scale(as.matrix(vertex$vertex), scale = scale,
                         center = center)
    if (method == "centers") {
      centers <- centers.interval(sym.data)
      res <- principal.curve(as.matrix(centers), plot.true = plot,
                             maxit = maxit)
      n <- dim(individuals)
```

```
projection.matrix <- matrix(data = NA, nrow = n[1],
                            ncol = n[2])
distance.vector <- rep(NA, n[1])
lambda <- rep(NA, n[1])
orthogonal.projection <- rep(NA, n[1])
for (i in 1:n[1]) {
  neig <- neighbors.vertex(as.matrix(individuals[i,
                                                 ]), res$s, 2)
  v <- -neig$neighbors[1, ] + neig$neighbors[2,
                                             ]
  vp <- -neig$neighbors[1, ] + individuals[i, ]
  proy <- sum(v * vp)/(norm.vect(v)^2) * v
  proy.point <- neig$neighbors[1, ] + proy
  projection.matrix[i, ] <- proy.point
  orthogonal.projection[i] <- sum((vp - proy) *
                                  v)
  distance.vector[i] <- norm.vect(vp - proy)
  lambda1 <- res$lambda[neig$order[1:2]]
  if (lambda1[1] <= lambda1[2]) {
    lambda[i] <- -lambda1[1] + norm.vect(proy)
  }
  else {
    lambda[i] <- lambda1[1] - norm.vect(proy)
  }
}
res.var.ind <- variance.princ.curve(data = individuals,
                                    curve = projection.matrix)
res.var.mid <- variance.princ.curve(data = as.matrix(centers),
                                    curve = res$s)
```

```
  res.var <- list(res.var.ind = res.var.ind, res.var.mid = res.var
↪ .mid)
  colnames(projection.matrix) <- sym.data$sym.var.names
  res.limits <- sym.interval.pc.limits(sym.data = sym.data,
                                       prin.curve = projection.
↪ matrix, num.vertex = vertex$num.vertex,
                                       lambda = lambda, res.var$
↪ res.var.mid$var.order)
  num.vars <- sym.data$M
  variables <- rep("X", num.vars)
  for (i in 1:num.vars) {
    variables[res.var.ind$var.order[i]] <- paste0("prin_surface_",
                                                  as.character(i))
  }
  colnames(projection.matrix) <- variables
  projection.matrix <- projection.matrix[, res.var.ind$var.order]
  correl <- cor(x = projection.matrix, y = vertex$vertex)
}
else if (method == "vertex") {
  res <- principal.curve(individuals, plot.true = plot,
                         maxit = maxit)
  res.var <- variance.princ.curve(data = individuals,
                                  curve = res$s)
  res.limits <- sym.interval.pc.limits(sym.data = sym.data,
                                       prin.curve = res$s, num.
↪ vertex = vertex$num.vertex,
                                       lambda = res$lambda, res.
↪ var$var.order)
  num.vars <- sym.data$M
  variables <- rep("X", num.vars)
```

```
      for (i in 1:num.vars) {
        variables[res.var$var.order[i]] <- paste0("prin_surface_",
                                                  as.character(i))
      }
      colnames(res$s) <- variables
      res$s <- res$s[, res.var$var.order]
      correl <- cor(x = res$s, y = vertex$vertex)
    }
    return(list(prin.curve = res, sym.prin.curve = res.limits,
                var.curve = res.var, cor.ps = correl))
  }
  return(TRUE)
}
```

A.1.9. Función para invocar los métodos de reducción de la dimensionalidad

- **sym.interval.pca.new:** Ejecuta el análisis de reducción de la dimensionalidad, según el método que se desee.

```
sym.interval.pca.new<-function (sym.data, method = c("classic", "tops"
    ↪ , "centers",
                                "principal.curves", "optimized.distance
    ↪ ", "optimized.variance"))
{
  idn <- all(sym.data$sym.var.types == sym.data$sym.var.types[1])
  if (idn == FALSE)
    stop("All variables have to be of the same type")
  method <- match.arg(method)
  if (method == "classic") {
```

```
  if ((sym.data$sym.var.types[1] != "$C") && (sym.data$sym.var.types
↪ [1] !=
                                                "$I"))
    stop("Variables have to be continuos or Interval")
  if (sym.data$sym.var.types[1] == "$C")
    res <- PCA(sym.data$data, scale.unit = TRUE, ncp = sym.data$M,
               graph = FALSE)
  else if (sym.data$sym.var.types[1] == "$I") {
    nn <- sym.data$N
    mm <- sym.data$M
    centers <- matrix(0, nn, mm)
    centers <- as.data.frame(centers)
    rownames(centers) <- sym.data$sym.obj.names
    colnames(centers) <- sym.data$sym.var.names
    for (i in 1:nn) for (j in 1:mm) centers[i, j] <- (sym.var(sym.
↪ data,
                                                            j)$var
↪ .data.vector[i, 1] + sym.var(sym.data,

↪                                  j)$var.data.vector[i, 2])/2
    res <- FactoMineR::PCA(centers, scale.unit = TRUE,
                           ncp = sym.data$M, graph = FALSE)
  }
  return(res)
}
if (method == "centers") {
  res<-centers.pca.j.new(sym.data)
  return(res)
}
if (method == "tops") {
```

```
    res <- vertex.pca.j(sym.data)
    return(res)
  }
  if (method == "principal.curves") {
    res <- sym.interval.pc(sym.data, "vertex", 150, FALSE,
                           FALSE, TRUE)
    return(res)
  }
  if (method == "optimized.distance") {
    res <- optim.pca.distance.j.new(sym.data)
    return(res)
  }
  if (method == "optimized.variance") {
    if(sym.data$M > 3){
      num.dimension<-3
    }
    else if(sym.data$M > 1){
      num.dimension<- sym.data$M -1

    }
    else{
      num.dimension<-1
    }
    res <- optim.pca.variance.j.new(sym.data, num.dimension = num.
    ↪ dimension)
    return(res)
  }
  return(TRUE)
}
```

Apéndice B

Congresos en los que se han presentado resultados preliminares de este trabajo

B.1. SDA 2015

MAPMO MATHÉMATIQUES - ANALYSE, PROBABILITÉS, MODÉLISATION - ORLÉANS

Unité Mixte de Recherche 6628
Richard EMILION
Professor
richard.emilion@univ-orleans.fr
http://www.univ-orleans.fr/mapmo/membres/emilion

Orléans, on July 13th. 2015

Acceptance letter

The paper 'Principal Curves and Surfaces to Interval Valued Variables' of Jorge Arce G. and Oldemar Rodriguez R., and the paper 'Latest developments of the RSDA: An R package for Symbolic Data Analysis' of Oldemar Rodriguez R. have heen accepted to be presented in the 5th Workshop on Symbolic Data Analysis, SDA2015, that will take place in Orléans, France, from November 17th to 19th, 2015.

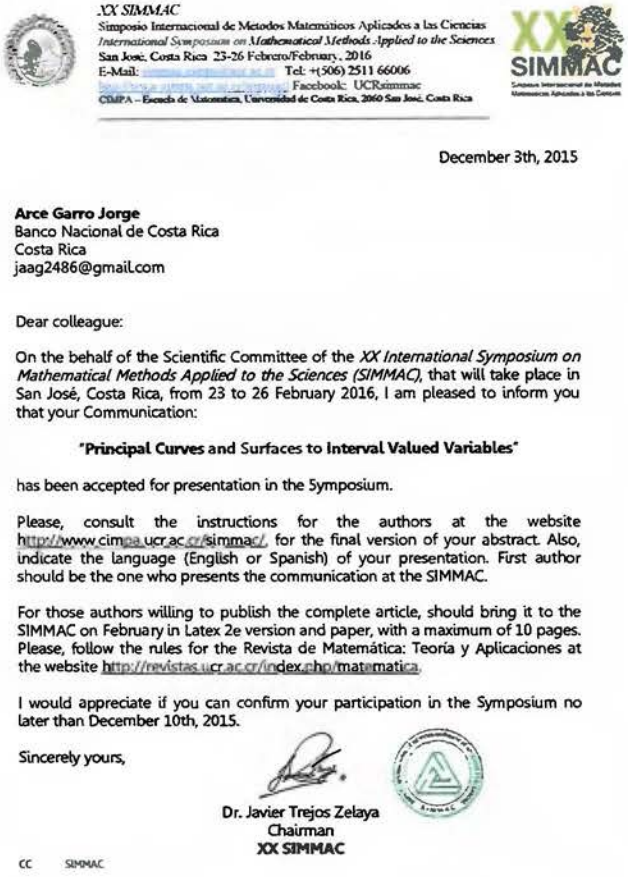
For the steering committee and the scientific committee,

R. EMILION

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
Université d'Orléans - MAPMO - B.P. 6759 - 45067 Orléans Cedex 2
Tél. : (33) 02 38 49 48 95. - Télécopie : (33) 02 38 41 72 05 - email : @labomath.univ-orleans.fr

FIGURA B.1: SDA 2015.

B.2. SIMMAC 2016



XX SIMMAC
Simposio Internacional de Metodos Matemáticos Aplicados a las Ciencias
International Symposium on Mathematical Methods Applied to the Sciences
San José, Costa Rica 23-26 Febrero/February, 2016
E-Mail: [illegible] Tel: +(506) 2511 66006
[illegible] Facebook: UCRsimmac
CIMPA – Escuela de Matemática, Universidad de Costa Rica, 2060 San José, Costa Rica

December 3th, 2015

Arce Garro Jorge
Banco Nacional de Costa Rica
Costa Rica
jaag2486@gmail.com

Dear colleague:

On the behalf of the Scientific Committee of the *XX International Symposium on Mathematical Methods Applied to the Sciences (SIMMAC)*, that will take place in San José, Costa Rica, from 23 to 26 February 2016, I am pleased to inform you that your Communication:

"Principal Curves and Surfaces to **Interval Valued Variables"**

has been accepted for presentation in the Symposium.

Please, consult the instructions for the authors at the website http://www.cimpa.ucr.ac.cr/simmac/, for the final version of your abstract. Also, indicate the language (English or Spanish) of your presentation. First author should be the one who presents the communication at the SIMMAC.

For those authors willing to publish the complete article, should bring it to the SIMMAC on February in Latex 2e version and paper, with a maximum of 10 pages. Please, follow the rules for the Revista de Matemática: Teoría y Aplicaciones at the website http://revistas.ucr.ac.cr/index.php/matematica.

I would appreciate if you can confirm your participation in the Symposium no later than December 10th, 2015.

Sincerely yours,

Dr. Javier Trejos Zelaya
Chairman
XX SIMMAC

cc SIMMAC

FIGURA B.2: SIMMAC 2016.

B.3. IBERAMIA 2016

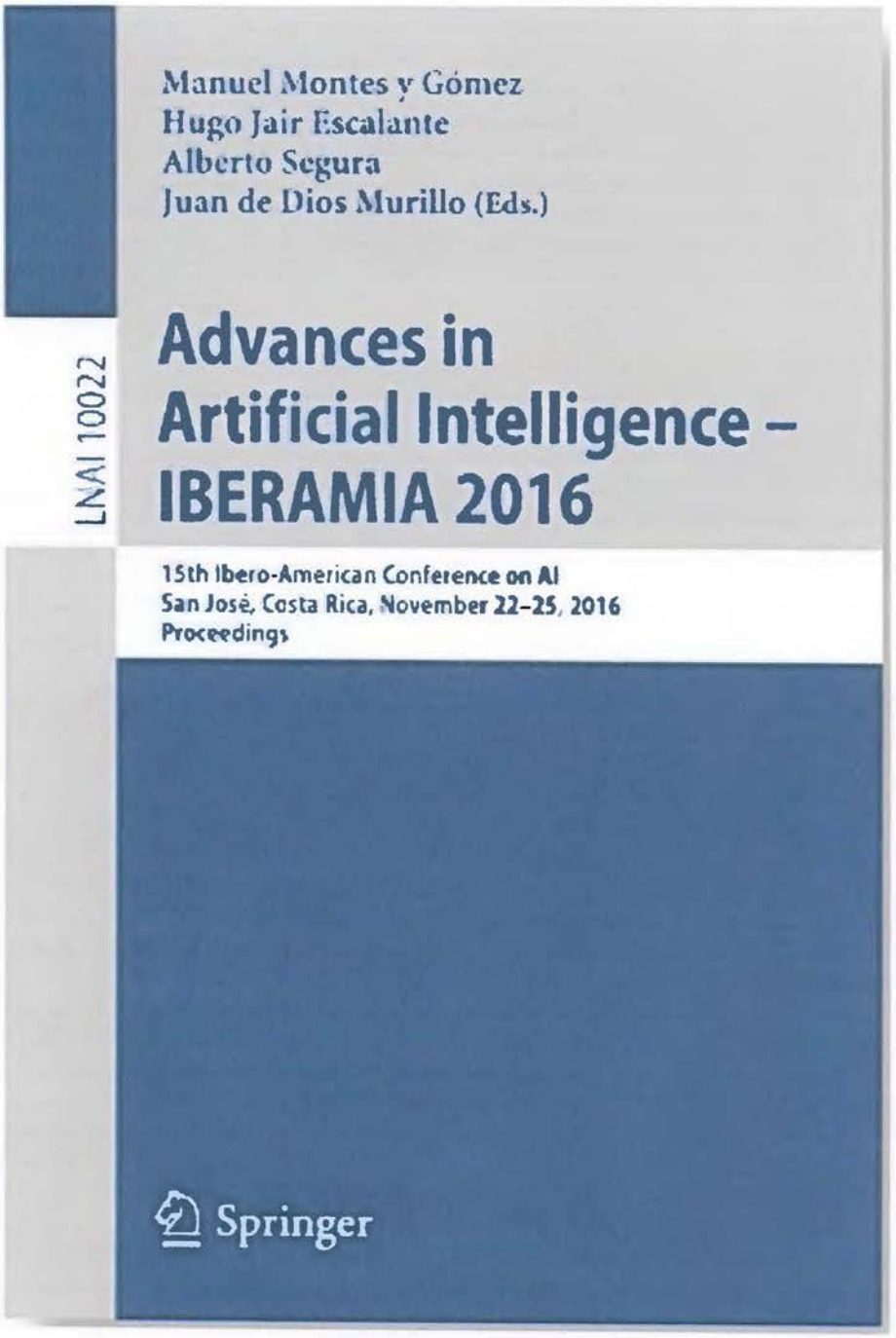


FIGURA B.3: Portada Iberamia 2016.

Principal Curves and Surfaces to Interval Valued Variables

Jorge Arce C.[1,2(✉)] and Oldemar Rodríguez R.[1]

[1] University of Costa Rica, San José, Costa Rica
oldemar.rodriguez@ucr.ac.cr
[2] National Bank of Costa Rica, San José, Costa Rica
jarce@bncr.fi.cr
http://www.cimpa.ucr.ac.cr/

Abstract. In this paper we propose a generalization to symbolic interval valued variables, of the Principal Curves and Surfaces method proposed by Hastie in [6]. Given a data set X with n observations and m continuous variables, the main idea of Principal Curves and Surfaces method is to generalize the principal component line, providing a smooth one-dimensional curved approximation to a set of data points in $\mathbb{R}^m$. A principal surface is more general, providing a curved manifold approximation of dimension 2 or more. In our case we are interested in finding the main principal curve that approximates better symbolic interval data variables. In [3,4], authors proposed the Centers Method and the Vertices Method to extend the well-known principal components analysis method to a particular kind of symbolic objects characterized by multi-valued variables of interval type. In this paper we generalize both, the Centers Method and the Vertices Method, finding a smooth curve that passes through the middle of the data X in an orthogonal sense. Some comparisons of the proposed method regarding the Centers and the Vertices Methods are made, this was done with the RSDA package using Ichino data set, see [1, 10]. To make these comparisons we have used the correlation index.

Keywords: Interval-valued variables · Principal curves and surfaces · Symbolic data analysis

1 Symbolic Data

Symbolic data was first introduced by [12]. In the classical data analysis a variable takes a single value, in the symbolic analysis the variable may take a finite or an infinite set of values. A type of symbolic variable may take an infinite set of numerical values ranging from a low to a high value (interval).

As the Principal Component Analysis (PCA) is one of the most popular multivariate methods, it is tempting to extend the PCA analysis to symbolic data. Some methods can be found in the literature, among them the vertex

M. Montes-y-Gómez et al. (Eds.): IBERAMIA 2016, LNAI 10022, pp. 297–309, 2016.
DOI: 10.1007/978-3-319-47954-2_25

FIGURA B.4: Iberamia 2016.

B.4. SDA 2017

9.00–10.30 **PCA. FA** (EF) *Chair: Richard Emilion*

1. **Symbolic covariance for interval data: Properties and interpretation**
   M. Rosário Oliveira. Margarida Vilela and António Pacheco
2. **Maximum likelihood factor analysis of interval data**
   Paula Cheira. Paula Brito and A. Pedro Duarte Silva
3. **Optimized symbolic principal component for interval-valued variables**
   Jorge Arce and Oldemar Rodríguez

FIGURA B.5: SDA 2017.

B.5. SIMMAC 2018

XXI SIMMAC
Simposio Internacional de Métodos Matemáticos Aplicados a las Ciencias
International Symposium on Mathematical Methods Applied to the Sciences
III Congreso Internacional en Estadística Computacional
III Latin American Conference on Statistical Computing
San José, Costa Rica 27 Febrero-2 Marzo/February 27th-March 2nd, 2018
E-Mail: **Tel:** +(506) 2511 66006
Facebook: UCRsimmac
CIMPA - Escuela de Matemática, Universidad de Costa Rica, 2060 San José, Costa Rica

October 24th, 2017

Jorge Arce / Oldemar Rodríguez
Universidad de Costa Rica
Banco Nacional de Costa Rica
jarceg@bncr.fi.cr
oldemar.rodriguez@ucr.ac.cr

Dear colleagues:

On behalf of the University of Costa Rica and the Organizing Committee of *International Symposium on Mathematical Methods Applied to the Sciences, SIMMAC XXI*, which will take place in San José, Costa Rica in February 27th-March 02nd, 2018, I am pleased to inform you that your communication:

"Best point principal component for interval-valued variables"

has been accepted for presentation in the Symposium.

Please, consult the instructions for the authors at the website http://www.cimpa.ucr.ac.cr/simmac/, for the final version of your abstract. Also, indicate the language (English or Spanish) of your presentation. First author should be the one who presents the communication at the SIMMAC.

For those authors willing to publish the complete article, should bring it to the SIMMAC on February in Latex 2e version and paper, with a maximum of 10 pages. Please, follow the rules for the Revista de Matemática: Teoría y Aplicaciones at the website http://revistas.ucr.ac.cr/index.php/matematica.

I would appreciate if you can confirm your participation in the Symposium no later than January 5th, 2018.

Sincerely yours,

Dr. Javier Trejos Zelaya
Chairman
XXI SIMMAC

FIGURA B.6: SIMMAC 2018.

Bibliografía

[1] Bickel, P. J.; Doksum, K. A. (1977). *Mathematical Statistics.* Prentice Hall, United States of America.

[2] Billard, L.; Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining.* John Wiley & Sons Ltd, United Kingdom.

[3] Cazes, P.; Chouakria, A.; Diday, E.; Schektman, Y. (1997). "Extension de l'analyse en composantes principales à des données de type intervalle", *Rev. Statistique Appliquée, Vol. 45(3)*, 5–24.

[4] Chouakria, A.; Billard, L.; Diday, E. (2011). "Principal component analysis for interval-valued observations", *Statistical Analysis and Data Mining, Vol. 4(2)*, 229–246.

[5] Diday, E. (1987). "Introduction à l'approache symbolique en analyse des Données", *Premières Journées Symbolique-Numérique*, CEREMADE, Université Paris, 21–56.

[6] Gallier, J.; Quaintance, J. (2018). *Fundamentals of Linear Algebra and Optimization,* Department of Computer and Information Science, Philadelphia, USA.

[7] Hastie, T. (1984). *Principal Curves and Surface.* Ph.D Thesis, Stanford University.

[8] Hastie, T.; Stuetzle, W. (1989). "Principal Curves", *Journal of the American Statistical Association, Vol. 84(406)*, 502–516.

[9] Hastie, T.; Tibshirani, R.; Friedman, J. (2008). *The Elements of Statistical Learning; Data Mining, Inference and Prediction.* Springer, New York.

[10] Hastie, T.; Weingessel, A. (2014). princurve - Fits a Principal Curve in Arbitrary Dimension. R package version 1.1–12 http://cran.r-project.org/web/packages/princurve/index.html.

[11] Ichino, M. (1994). "Generalized Minkowski metrics for mixed featuretype data analysis", *IEEE , Transactions on Systems, Man and Cybermetrics, 24(4).*

[12] Nocedal, J.; Wright, S. (1999). *Numerical optimization,* Springer, New York, USA.

[13] Rodríguez, O. (2000). *Classification et Modèles Linèaires en Analyse des Données Symboliques.* Ph.D Thesis, Paris IX-Dauphine University.

[14] Rodríguez,O. (2012). "The Duality Problem in Interval Principal Components Analysis", *The 3rd Workshop in Symbolic Data Analysis, Madrid.*

[15] Rodríguez, O. with contributions from Olger Calderón, Roberto Zúñiga and Jorge Arce (2015). RSDA - R to Symbolic Data Analysis. R package version 1.3. http://CRAN.R-project.org/package=RSDA.

[16] Ugalde, W. (2009). *"MA 0450 Cálculo en varias variables",* Universidad de Costa Rica, San José, Costa Rica.

[17] Trejos, J.; Castillo, W.; González, J. (2014). *Análisis Multivariado de Datos: Métodos y Aplicaciones,* Editorial UCR, San José, Costa Rica.