UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Análisis de Componentes Principales y su aplicación en dos problemas ecológicos

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Actuario

P R E S E N T A :

Victor Rivera Escobar

DIRECTOR DE TESIS:
Mat. Margarita Elvira Chávez Cano

CIUDAD DE MÉXICO 2022

Índice general

Introducción

El estudio de la ecología, generalmente intenta comprender los factores principales que influyen en la distribución y abundancia de organismos en la Tierra. Los ecologístas están interesados en el estudio de cómo interactúan los organismos entre sí y con su medio ambiente.

Este objetivo, puede irse logrando con detección de ciertos patrones en el ambiente, y luego analizar la respuesta de los organismos a dichos patrones, es decir, se identifica y describe los comportamientos más relevantes en los datos. Pero, una vez teniendo el conjunto de datos, ¿Cómo sabremos qué patrones fueron los más significativos?

Ciertamente, existen diversas técnicas que nos permitirían abordar esta problemática. Una técnica exploratoria que nos podría ser muy útil, es el Análisis de Componentes Principales (ACP). Este análisis extrae en un conjunto más pequeño de variables, la principal información de los datos, es decir, realiza una reducción de dimensión en las variables. El objetivo principal de este trabajo es explicar y entender el análisis de componentes principales, complementándolo con la aplicación de esta técnica a dos problemas ecológicos reales.

En el capítulo uno el objetivo principal es conocer conceptos fundamentales para comprender el análisis de componentes principales como análisis exploratorio.

En el segundo capítulo, se presenta todo el análisis de componentes principales, contiene la construcción del cálculo de esta técnica, algunos criterios para elegir el número de componentes principales a estudiar, así como su interpretación y limitaciones. Para comprender mejor este análisis se acompañó con un problema ecológico, este primer problema presentado contiene información cuantitativa de la morfología cromosómica de algunas especies de leguminosas, dichas especies están clasificadas en un mismo grupo taxonómico. Sin embargo, se realizará un ACP para analizar si existe un agrupamiento entre algunas especies y si efectivamente hay evidencia para pensar que estas leguminosas podrían ser clasificadas en dos grupos diferentes.

En el tercer capítulo se analiza un problema acerca del clima en Kotzbehue, Alaska. Se presentan registros de 1981 hasta el año 2003, la idea principal es verificar

si el ACP detecta un cambio significativo en la información de la base para algunos años y la comprensión de dicho comportamiento. Este problema puede asociarse con la teoría del calentamiento global.

Para aplicar y desarrollar el análisis descriptivo de componentes principales en ambos problemas mencionados, se utilizó el lenguaje de programación **R**, el cual es un entorno de programación que principalmente se utiliza con un enfoque al análisis estadístico.

Capítulo 1

Vectores y álgebra de matrices

En este capítulo se abordarán varios temas básicos para poder comprender el Análisis de Componentes Principales (ACP) y cómo es que se calculan las componentes, algunos temas que encontraremos son: las definiciones de un vector, de una matriz, operaciones entre matrices, proyección de un vector sobre otro, vectores propios y valores propios, vectores y matrices ortogonales, etc. Estos conceptos son fundamentales para entender mejor esta técnica de análisis exploratorio que profundizaremos en el capítulo II.

1.1. Definiciones básicas de vectores

Un vector $\mathbf{x}$ de orden nx1 es un conjunto ordenado de n números reales (llamados escalares), que podemos escribir como:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Las $\mathbf{x}$'s denotan a los números reales y son llamados *componentes, elementos,* o *entradas* de $\mathbf{x}$. La forma en que denotamos a $\mathbf{x}$ es llamado vector columna y consiste de n renglones y 1 columna de elementos (de donde deriva la designación nx1). Por otro lado, podemos escribir $\mathbf{x}$' de orden 1xn como:

$$\mathbf{x}' = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix}$$

y lo llamamos vector fila, el cual consiste de 1 renglón y n columnas de elementos. Usaremos la notación $\mathbf{x}$ para denotar un vector columna y la notación $\mathbf{x}$', el cual

es llamado *la transpuesta* de $\mathbf{x}$, para denotar un vector fila. Por la transpuesta de un vector, generalmente, significa que un vector columna de orden n por 1 se convierte a un vector fila, conservando el mismo número de elementos, pero ahora de orden 1 por n. De manera similar, la transpuesta de un vector fila de orden 1 por n es un vector columna de orden n por 1, conservando el mismo número de elementos.

Por ejemplo, sean $\mathbf{a}$ un vector de orden 4x1 y b de orden 3x1, representados como:

$$\mathbf{a} = \begin{pmatrix} 12 \\ 4.8 \\ 1.2 \\ 9 \end{pmatrix}; \mathbf{b} = \begin{pmatrix} 1 \\ 3 \\ 4 \end{pmatrix}$$

Notemos que $\mathbf{a}$ y $\mathbf{b}$ son vectores columna. Si obtenemos la transpuesta de cada vector respectivamente, tendremos lo siguiente:

$$\mathbf{a}' = \begin{pmatrix} 12 & 4.8 & 1.2 & 9 \end{pmatrix}; \mathbf{b}' = \begin{pmatrix} 1 & 3 & 4 \end{pmatrix}$$

Veamos que $\mathbf{a}$' es de orden 1x4 y $\mathbf{b}$' de orden 1x3, ambos son vectores fila.
Si todos los elementos de un vector son 0's, lo llamaremos *vector nulo* o *vector cero*, denotado como $\bar{0}$, de esa forma no lo podremos confundir con el escalar 0. Si todos los elementos de un vector son 1's, lo llamaremos *vector unitario*, denotado como $\bar{1}$.

1.2. Definiciones básicas y operaciones de matrices

1.2.1. Definición de matriz

Generalmente, *las matrices* son arreglos rectangulares de números reales ordenados en filas y columnas. Si $\mathbf{X}$ tiene n renglones y p columnas, $\mathbf{X}$ es un arreglo de orden nxp, se puede representar como:

$$\mathbf{X}_{nxp} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

Si consideramos la matriz $\mathbf{X}_{nxp}$ como una colección de n vectores fila, cada uno de tamaño p, nosotros podemos interpretar los elementos de $\mathbf{X}$ como las coordenadas de n puntos en p dimensiones. Aquí los renglones de $\mathbf{X}$ corresponden a los puntos, y

sus columnas corresponden a las dimensiones, por ejemplo, si $\mathbf{X}$ fuera una matriz de orden 25x2, entonces esos elementos pueden ser considerados como las coordenadas de 25 puntos en 2 dimensiones, por lo que podrían ser graficados en R^2. Una matriz es rectangular cuando $n > p$ o $p > n$.

La transpuesta de la matriz $\mathbf{X}$ denotada por $\mathbf{X}'$, es obtenida del intercambio de renglones y columnas de la matriz $\mathbf{X}$. Entonces las columnas de $\mathbf{X}$ pasan a ser los renglones de $\mathbf{X}'$, y los renglones de $\mathbf{X}$ son las columnas de $\mathbf{X}'$, por lo que el rango de la matriz $\mathbf{X}'$ es de pxn como se muestra en la siguiente matriz:

$$\mathbf{X}'_{pxn} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{p1} & x_{p2} & \dots & x_{pn} \end{pmatrix}$$

Mostraremos un simple ejemplo, consideremos una matriz $\mathbf{A}$ de 3x2 y obtenemos su transpuesta como se muestra:

$$\mathbf{A} = \begin{pmatrix} 4 & 2.4 \\ 8 & 7.8 \\ 18 & 5.1 \end{pmatrix}; \ \mathbf{A}' = \begin{pmatrix} 4 & 8 & 18 \\ 2.4 & 7.8 & 5.1 \end{pmatrix}$$

La nueva matriz $\mathbf{A}'$ representa la transpuesta de la matriz $\mathbf{A}$.

1.2.2. Formas Particulares de Matrices

Una matriz puede mostrar alguna relación que hay entre n, el número de renglones, y p, el número de columnas. Si $n = p$, quiere decir que la matriz tiene el mismo número de renglones y columnas, la cual es llamada *matriz cuadrada*, por ejemplo $\mathbf{X}_{nxn}$ es una matriz cuadrada de orden n. Cuando una matriz cuadrada $\mathbf{X}$ es tal que $x_{ij} = x_{ji}$, para todos i y j, le llamamos *matriz simétrica*, o bien, si la transpuesta de una matriz $\mathbf{X}$ es la misma que la original entoces se muestra que $\mathbf{X}$ es una matriz simétrica. Por ejemplo, sea $\mathbf{A}$ una matriz de orden 3x3:

$$\mathbf{A} = \begin{pmatrix} 1 & 3 & 8 \\ 3 & 7 & 5 \\ 8 & 5 & 9 \end{pmatrix}$$

Si obtenemos $\mathbf{A}'$ notaremos que obtuvimos la matriz original $\mathbf{A}$, es decir $\mathbf{A}' = \mathbf{A}$. Por lo tanto $\mathbf{A}$ es una matriz simétrica.
Si una matriz cuadrada contiene como elementos al número cero en todas las entradas

que no pertencecen a su diagonal, se dice ser una *matriz diagonal* **D**, una matriz diagonal de orden n puede ser escrita como diag$(d_1, d_2, ..., d_n)$ o diag$(\mathbf{d})$ donde **d** es un vector de orden n. Un ejemplo de una matriz diagonal es el siguiente:

$$\mathbf{D} = diag(1, 7, 9) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 9 \end{pmatrix}$$

Una matriz diagonal de 1's es llamada *matriz identidad* denotada por **I**. Por ejemplo, una matriz identidad de 3x3 está dada por:

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Una matriz cuadrada **X** con $x_{ij} = 0$ para todo $i > j$ es conociada como *matriz triangular superior*, es decir,hay ceros debajo de la diagonal, tal como se ve en la siguiente matriz:

$$\mathbf{T} = \begin{pmatrix} 11 & 12 & 10 \\ 0 & 17 & 20 \\ 0 & 0 & 29 \end{pmatrix}$$

Una *matriz triangular inferior* se define de manera similar, solo cambia que $x_{ij} = 0$ para todo $i < j$, como se muestra en el siguiente ejemplo:

$$\mathbf{T} = \begin{pmatrix} 1 & 0 & 0 \\ 5 & 7 & 0 \\ 18 & 3 & 13 \end{pmatrix}$$

1.2.3. Adición y Diferencia de Matrices

Si dos matrices son del mismo orden, su adición se calcula al sumar los elementos correspondientes, es decir, si **A** es una matriz de orden nxp y la matriz **B** tiene el mismo orden, entonces $\mathbf{C} = \mathbf{A} + \mathbf{B}$ y **C** es de orden nxp donde $(c_{ij}) = (a_{ij} + b_{ij})$. Por ejemplo:

$$\begin{pmatrix} 11 & 8 \\ 15 & 7 \\ 18 & 3 \end{pmatrix} + \begin{pmatrix} 2 & 12 \\ 5 & 17 \\ 8 & 13 \end{pmatrix} = \begin{pmatrix} 13 & 20 \\ 20 & 24 \\ 26 & 16 \end{pmatrix}$$

Análogamente con la diferencia entre dos matrices, donde restamos los elemetos que corresponden y las matrices deben ser del mismo orden. Entonces **C** = **A - B** se calcula como $(c_{ij}) = (a_{ij} - b_{ij})$. Por ejemplo:

$$\begin{pmatrix} 11 & 8 \\ 15 & 7 \\ 18 & 3 \end{pmatrix} - \begin{pmatrix} 2 & 12 \\ 5 & 17 \\ 8 & 13 \end{pmatrix} = \begin{pmatrix} 9 & -4 \\ 10 & -10 \\ 10 & -10 \end{pmatrix}$$

1.2.4. Producto de matrices

Para que se defina el producto **AB**, el número de columnas de la matriz **A** debe ser igual al número de filas de la matriz **B**. Entonces el elemento (ij)-ésimo de **C**= **AB** es:

$$c_{ij} = \sum_k a_{ik} * b_{kj} \tag{1.1}$$

Así, c_{ij} es la suma de los productos de la i-ésima fila de **A** y la j-ésima columna de **B**, por lo tanto, se multiplica cada fila de **A** por cada columna de **B** , y el tamaño de **AB** consiste en el número de filas de **A** y el número de columnas de **B**. Por lo tanto, si **A** es de orden nxm y **B** es mxp, entonces **C** = **AB** es nxp. Por ejemplo, si:

$$\mathbf{A} = \begin{pmatrix} 7 & 4 & 9 \\ 3 & 1 & 6 \\ 0 & 5 & 8 \\ 2 & 3 & 7 \end{pmatrix} \quad y \quad \mathbf{B} = \begin{pmatrix} 1 & 2 \\ 4 & 3 \\ 5 & 7 \end{pmatrix}$$

entonces

$$\mathbf{C} = \mathbf{AB} = \begin{pmatrix} (7)(1)+(4)(4)+(9)(5) & (7)(2)+(4)(3)+(9)(7) \\ (3)(1)+(1)(4)+(6)(5) & (3)(2)+(1)(3)+(6)(7) \\ (0)(1)+(5)(4)+(8)(5) & (0)(2)+(5)(3)+(8)(7) \\ (2)(1)+(3)(4)+(7)(5) & (2)(2)+(3)(3)+(7)(7) \end{pmatrix} = \begin{pmatrix} 68 & 89 \\ 37 & 51 \\ 60 & 71 \\ 49 & 62 \end{pmatrix}$$

En el ejemplo mostrado, **A** es de orden 4x3 y **B** es de 3x2. Como el número de columnas de la matriz **A** es igual al número de filas de la matriz **B** entonces se puede realizar el producto de matrices. Por lo tanto **AB** es de dimensión 4x2.

El producto de un escalar y una matriz es obtenida al multiplicar cada elemento de la matriz por el escalar:

$$c\mathbf{A} = c(a_{ij}) = \begin{pmatrix} ca_{11} & ca_{12} & \dots & ca_{1p} \\ ca_{21} & ca_{22} & \dots & ca_{2p} \\ \vdots & \vdots & & \vdots \\ ca_{n1} & ca_{n2} & \dots & ca_{np} \end{pmatrix} \tag{1.2}$$

La multiplicación de vectores nos permite el uso de combinaciones lineales, tal como:

$$a'x = (a_1, a_2, \dots, a_k) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix} = \sum_{i=1}^{k} a_i x_i = a_1x_1 + a_2x_2 \dots + a_kx_k \tag{1.3}$$

Algunas propiedades del producto de matrices

- Si k es un escalar, entonces tenemos la propiedad de la asociatividad: $k\mathbf{AB} = (k\mathbf{A})\mathbf{B}$
- La transpuesta del producto de dos matrices: $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$

1.3. Determinantes de matrices

El *determinante* de una matriz juega un papel importante para conceptos que presentaremos más adelante, como la inversa y el rango de una matriz. Únicamente las matrices cuadradas tienen determinates. Denotaremos al determinante de una matriz $\mathbf{A}$ por el símobolo $|\mathbf{A}|$. Cabe mencionar que al obtener el determinante de una matriz como resultado nos regresará un escalar.

El determinante de orden 2, se puede ver como un caso especial. Sea $\mathbf{A}$ una matriz 2x2:

$$\mathbf{A} = \begin{pmatrix} \text{a} & \text{b} \\ \text{c} & \text{d} \end{pmatrix}$$

Su determinante se define como:

$$|\mathbf{A}| = \begin{vmatrix} \text{a} & \text{b} \\ \text{c} & \text{d} \end{vmatrix} = ad - bc \tag{1.4}$$

Por ejemplo, sea $\mathbf{A}$ una matriz de orden 2:

$$\mathbf{A} = \begin{pmatrix} \text{-2} & 7 \\ \text{-5} & 14 \end{pmatrix}$$

entonces

$$|\mathbf{A}| = (-2)(14) - (7)(-5) = -28 + 35 = 7$$

Para el caso de una matriz $\mathbf{A}$ de orden 3, el determinante está dado por:

$$\mathbf{A} = (a_{ij}) = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

entonces

$$\begin{aligned} |\mathbf{A}| &= a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} \\ &= a_{11}|\mathbf{A}_{11}| - a_{12}|\mathbf{A}_{12}| + a_{13}|\mathbf{A}_{13}| \end{aligned} \tag{1.5}$$

La expresión para el determinante de una matriz $\mathbf{A}$ de orden n se puede obtener en términos de determinantes de matrices de (n-1)x(n-1). Sea i un número entero, $1 \leq i \leq n$. Se define como:

$$|\mathbf{A}| = (-1)^{i+1} a_{i1} |\mathbf{A}_{i1}| + \ldots + (-1)^{i+n} a_{in} |\mathbf{A}_{in}| \tag{1.6}$$

Donde cada $\mathbf{A}_{ij}$ es una matriz de orden n-1.

El determinante de una matriz diagonal es el producto de los elementos de la diagonal, es decir, si $\mathbf{D} = diag(d_1, d_2, \ldots, d_n)$, entonces:

$$|\mathbf{D}| = \prod_{i=1}^{n} d_i. \tag{1.7}$$

Una matriz cuadrada $\mathbf{A}$ se dice ser *singular* si $|\mathbf{A}| = 0$. Si $|\mathbf{A}| \neq 0$, se dice ser *no singular*.

En general, si un renglón particular (o una columna) de una matriz puede ser perfectamente expresado como una combinación lineal de los otros renglones (o columnas), la matriz se dice ser singular.

1.4. Rango

Antes de definir que es el rango de una matriz, debemos tener una noción de independencia lineal. Un conjunto de vectores $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_n$ son *linealmente dependientes* si existen constantes $c_1, c_2, ..., c_n$ (no todos iguales a cero), tales que:

$$c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + ... + c_n\mathbf{a}_n = \bar{0} \tag{1.8}$$

Si no existen tales constantes que satisfagan la condición 1.8, se dice que el conjunto de vectores son *linealmente independientes.*
El rango de cualquier matriz cuadrada o rectangular $\mathbf{A}$ se define como:

rango($\mathbf{A}$) = número de filas linealmente independientes de $\mathbf{A}$ = número de columnas linealmente independientes de $\mathbf{A}$

Para encontrar el número de filas o columnas independientes haremos uso de los determinantes ya que es un método computacional muy eficaz en vez de resolver el sistema de ecuaciones lineales planteado por 1.8. El rango de una matriz es el orden de la submatriz cuadrada más grande cuyo determinante es distinto de cero. Por ejemplo, sea $\mathbf{A}$ una matriz de orden 3 x 4

$$\mathbf{A} = \begin{pmatrix} 2 & 4 & 3 & 5 \\ 0 & 3 & \text{-1} & 3 \\ 1 & \text{-1} & 2 & 3 \end{pmatrix}$$

Veamos que su rango a lo más es igual a 3, ya que el número de filas linealmente independientes es igual al número de columnas linealmente independietes y tenemos una matriz de 3 filas con 4 columnas. Ahora tomamos una submatriz cuadrada de orden 3 (la más grande posible) y obtenemos su determinante. Sea $\mathbf{B}$ dicha submatriz

$$\mathbf{B} = \begin{pmatrix} 2 & 4 & 5 \\ 0 & 3 & 3 \\ 1 & \text{-1} & 3 \end{pmatrix}$$

entonces

$$\begin{aligned} |\mathbf{B}| &= 2\begin{vmatrix} 3 & 3 \\ \text{-1} & 3 \end{vmatrix} - 4\begin{vmatrix} 0 & 3 \\ 1 & 3 \end{vmatrix} + 5\begin{vmatrix} 0 & 3 \\ 1 & \text{-1} \end{vmatrix} \\ &= 2(9-(-3)) - 4(0-3) + 5(0-3) \\ &= 24 + 12 - 15 = 21 \end{aligned}$$

Como el determinante de la submatriz es distinto de cero, podemos concluir que el rango de $\mathbf{A}$ es 3.
Cuando el determinante de una submatriz es cero, quiere decir que algún vector columna es dependiente, asi que tendríamos que hacer las combinaciones posibles para encontrar los vectores columna independientes. Puede pasar que en todas las combinaciones el determinante nos de cero, entonces disminuiríamos el orden de la submatriz, en este caso sería de 2. Analizaríamos si el determinante es distinto de cero para encontrar cuáles vectores columnas son independientes, que equivalen al rango. Si fuera una matriz más grande, haríamos este proceso, y se termina hasta el análisis de las submatrices de orden 2 si es que aún no hay ninguna submatriz que sea distinto de cero.

1.5. Inversa de una matriz

Una matriz cuadrada $\mathbf{A}$ cuyo determinante es $\neq 0$, o en forma equivalente, que tiene inversa, como vimos anteriormente se conoce como *no singular* y $\mathbf{A}$ tiene una única inversa, denotada por $\mathbf{A}^{-1}$, denotada con la propiedad

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \tag{1.9}$$

Si $\mathbf{A}$ y $\mathbf{B}$ son no singulares y son del mismo orden, entonces la inversa de su producto es el producto de sus inversas en orden contrario, es decir

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1} \tag{1.10}$$

La inversa de la transpuesta de una matriz no singula está dada por la transpuesta de la inversa:

$$(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})' \tag{1.11}$$

Suponiendo que una matriz $\mathbf{A}$ tiene inversa, esta se calcula

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|}\left(Adj(\mathbf{A})\right)' \tag{1.12}$$

donde cada entrada la matriz adjunta $Adj(\mathbf{A})$ se define por el elemento de la fila i y columna j de la matriz $\mathbf{A}$ como se muestra en la siguiente expresión:

$$adj_{ij}(\mathbf{A}) = (-1)^{i+j}|\mathbf{A}_{ij}| \tag{1.13}$$

1.6. Matrices definidas positivas y definidas semipositivas

Se dice que una matriz cuadrada $\mathbf{A}$ es *definida positiva* si $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$ para todos los posibles vectores $\mathbf{x} \neq \bar{0}$. De manera similar una matriz cuadrada es *definida semipositiva* si $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$ para toda $\mathbf{x} \neq \bar{0}$.
Los elementos diagonales a_{ii} de una matriz definida positiva son positivos. Para ver esto, sea $\mathbf{x}' = (0, ..., 0, 1, 0, ..., 0)$ con 1 en la posición i-ésima. Entonces, $\mathbf{x}'\mathbf{A}\mathbf{x} = a_{ii} > 0$. De forma similar, para una matriz $\mathbf{A}$ definida semipositiva para todo i.
Una manera de obtener una matriz definida positiva es como se muestra:

Si $\mathbf{A} = \mathbf{B}'\mathbf{B}$ donde $\mathbf{B}$ es n x p, de rango $p<n$, entonces $\mathbf{B}'\mathbf{B}$ es definida positiva (1.14)

Esto se puede demostrar:

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{B}'\mathbf{B}\mathbf{x} = (\mathbf{B}\mathbf{x})'(\mathbf{B}\mathbf{x}) = \mathbf{z}'\mathbf{z}$$

donde $\mathbf{z} = \mathbf{B}\mathbf{x}$, entonces $\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{n} z_i^2$, el cual es positivo ($\mathbf{B}\mathbf{x}$ no puede ser 0 a menos que el vextor $\mathbf{x} = \bar{0}$)

1.7. Traza

Una función que se aplica a cualquier matriz $\mathbf{A}$ de dimensión n es la *traza*, denotada por $\text{tr}(\mathbf{A})$ y definida como la suma de elementos de la diagonal de $\mathbf{A}$, por consecuencia, resultará un número escalar, es decir:

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}. \tag{1.15}$$

Por ejemplo, sea

$$\mathbf{A} = \begin{pmatrix} 54 & 13 & 24 \\ 46 & -18 & 28 \\ 39 & 12 & 37 \end{pmatrix}$$

entonces

$$tr(\mathbf{A}) = 54 + (-18) + 37 = 73$$

La traza de la suma de dos matrices cuadradas es la suma de las trazas de sus respectivas matrices:

$$tr(\mathbf{A}+\mathbf{B}) = tr(\mathbf{A}) + tr(\mathbf{B}) \tag{1.16}$$

Un resultado importante para el producto de dos matrices es:

$$tr(\mathbf{AB}) = tr(\mathbf{BA}) \tag{1.17}$$

Este resultado es válido para cualesquiera matriz $\mathbf{A}$ y $\mathbf{B}$ definidas, con la condición de que el producto de ambas matrices sea una matriz cuadrada.

1.8. Vectores y matrices ortogonales

Se dice que dos vectores $\mathbf{a}$ y $\mathbf{b}$ de dimensión n x 1 son *ortogonales* entre sí (geométricamentes son perpendiculares) si se cumple:

$$\mathbf{a}'\mathbf{b} = a_1b_1 + a_2b_2 + ... + a_nb_n = 0 \tag{1.18}$$

Si $\mathbf{a}'\mathbf{a} = 1$, se dice que el vector $\mathbf{a}$ está *normalizado*. Se puede normalizar cualquier vector $\mathbf{a}$ de la siguiente forma:

$$\mathbf{c} = \frac{\mathbf{a}}{\sqrt{\mathbf{a}'\mathbf{a}}} \tag{1.19}$$

Ahora tendremos que $\mathbf{c}'\mathbf{c} = 1$

Una matriz $\mathbf{C} = (c_1, c_2, ..., c_n)$ se dice estar *normalizada* si cada vector columna $\mathbf{c}_i$ lo está. Si cuyas columnas también son mutuamente ortogonales, se conoce como *matriz ortogonal*. Dado que los elementos de $\mathbf{C}'\mathbf{C}$ son productos de columnas de $\mathbf{C}$, que tienen las propiedades $\mathbf{c}_i'\mathbf{c}_i = 1$ para todo i y $\mathbf{c}_i'\mathbf{c}_j = 0$ para todo $i \neq j$, tenemos:

$$\mathbf{C}'\mathbf{C} = \mathbf{I} \tag{1.20}$$

Si $\mathbf{C}$ satisface (1.20) necesariamente se sigue que:

$$\mathbf{C}\mathbf{C}' = \mathbf{I} \tag{1.21}$$

La multiplicación por una matriz ortogonal implica la rotación de ejes; es decir, si a un punto $\mathbf{x}$ se le aplica la transformación $\mathbf{z} = \mathbf{C}\mathbf{x}$, donde $\mathbf{C}$ es una matriz ortogonal, entonces:

$$\mathbf{z}'\mathbf{z} = (\mathbf{C}\mathbf{x}')(\mathbf{C}\mathbf{x}) = \mathbf{x}'\mathbf{C}\mathbf{C}\mathbf{x} = \mathbf{x}'\mathbf{I}\mathbf{x} = \mathbf{x}'\mathbf{x} \tag{1.22}$$

Cabe mencionar, que al ser una rotación, la distancia que había del origen a $\mathbf{z}$, es la misma que hay del origen al punto $\mathbf{x}$.

1.9. Proyección de un vector sobre otro

Usaremos la noción de ortogonalidad entre dos vectores para deducir la idea de una proyección. Sean $\mathbf{a}$ y $\mathbf{b}$ dos vectores y $\mathbf{b} \neq \bar{0}$. Se define el vector $\mathbf{p}$ como la proyección de $\mathbf{a}$ a lo largo de $\mathbf{b}$, como se observa en la siguiente figura:

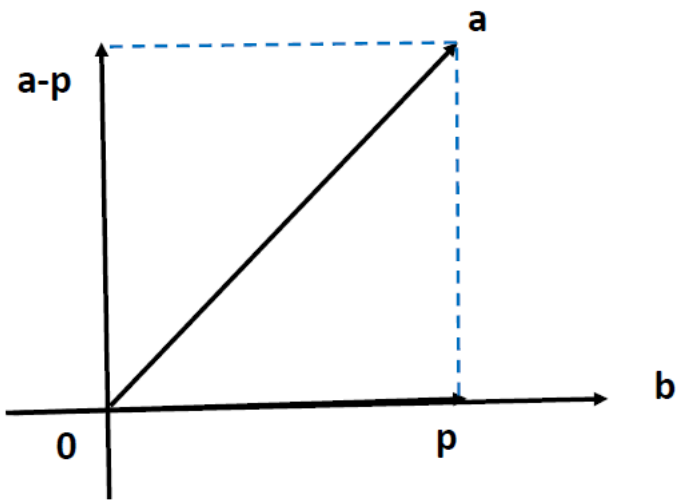

Figura 1.1: Proyección del vector $\mathbf{a}$ a lo largo del vector $\mathbf{b}$

Se quiere encontrar un vector $\mathbf{p}$ tal que $\mathbf{a} - \mathbf{p}$ sea ortogonal a $\mathbf{b}$ y además que se pueda expresar en la forma $\mathbf{p} = c\mathbf{b}$ para algún escalar c. Supongamos que se puede encontrar dicho escalar c si se satisface lo siguiente:

$$(\mathbf{a}' - c\mathbf{b}')\mathbf{b} = 0$$

Entonces se obtiene

$$\mathbf{a}'\mathbf{b} - c\mathbf{b}'\mathbf{b} = 0$$

Por lo tanto

$$c = \frac{\mathbf{a}'\mathbf{b}}{\mathbf{b}'\mathbf{b}}$$

Se puede notar que por la condición de ortogonalidad el escalar c está determinado en forma única.
Ahora definimos al vector $c\mathbf{b}$ como la **proyección** de $\mathbf{a}$ a lo largo de $\mathbf{b}$ denotada como $Proy_{\mathbf{b}}\mathbf{a}$, si c es el escalar obtenido anteriormente:

$$c = \frac{\mathbf{a}'\mathbf{b}}{\mathbf{b}'\mathbf{b}} \tag{1.23}$$

y se define c como la **componente** de $\mathbf{a}$ a lo largo de $\mathbf{b}$. Si $\mathbf{b}$ es un vector que está normalizado, entonces se tiene que:

$$c = \mathbf{a}'\mathbf{b} \tag{1.24}$$

1.10. Vectores y valores propios

Para cada matriz cuadrada $\mathbf{A}$, se obedece la siguiente relación

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \tag{1.25}$$

donde, λ es un número real y $\mathbf{x}$ es un vector distinto del vector cero. Para encontrar λ y $\mathbf{x}$ escribimos (1.25) como:

$$\mathbf{A}\mathbf{x} - \lambda\mathbf{x} = 0$$

equivalentemente,

$$\mathbf{A}\mathbf{x} - \lambda\mathbf{I}\mathbf{x} = 0$$

donde $\mathbf{I}$ es la matriz identidad. En seguida, podemos factorizar $\mathbf{x}$ obteniendo

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0 \tag{1.26}$$

La ecuación $|\mathbf{A} - \lambda\mathbf{I}| = 0$ se conoce como *ecuación característica*. Para obtener soluciones no triviales, se establece que $|\mathbf{A} - \lambda\mathbf{I}| = 0$ para encontrar los valores de λ, conocidos como *valores propios* y así se podrán sustituir en (1.26) para encontrar los valores correspondientes de $\mathbf{x}$, conocidos como *vectores propios*. Entonces, la matriz $\mathbf{A} - \lambda\mathbf{I}$ debe ser singular y nos interesa encontrar un vector solución $\mathbf{x} \neq \bar{0}$. Cabe mencionar que si $|\mathbf{A} - \lambda\mathbf{I}| \neq 0$, entonces $(\mathbf{A} - \lambda\mathbf{I})$ tiene una inversa y $\mathbf{x} = \bar{0}$ es la única solución. Así que no será de nuestro interés este caso.

Supongamos que tenemos una matriz cuadrada $\mathbf{A}$ de orden n, entonces la ecuación característica que le corresponde tendrá n raíces, es decir, tendremos n valores propios $\lambda_1, \lambda_2, ..., \lambda_n$. Si los valores de cada vector columna de $\mathbf{A}$ son números reales y la matriz $\mathbf{A}$ es no singular, tendremos que todas las λ's serán diferentes. Después de encontrar $\lambda_1, \lambda_2, ..., \lambda_n$ se puede encontrar sus vectores propios asociados $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$, esto se hace usando (1.26).

Si multiplicamos al vector $\mathbf{x}$ por un escalar k notamos que k y $\mathbf{A} - \lambda\mathbf{I}$ conmutan, entonces

$$(\mathbf{A} - \lambda\mathbf{I})k\mathbf{x} = k(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = k0 = 0 \tag{1.27}$$

Por lo tanto, si $\mathbf{x}$ es un vector propio de A, $k\mathbf{x}$ también es un vector propio, y los vectores propios son únicos hasta la multiplicación por un escalar. Por lo tanto, podemos ajustar la longitud de $\mathbf{x}$, pero la dirección desde el origen es única, ya que los vectores propios $\mathbf{x} = (x_1, x_2, ..., x_n)'$ son únicos.
Para ilustrar, daremos un ejemplo, sea $\mathbf{A}$ la siguiente matriz:

$$\mathbf{A} = \begin{pmatrix} 1 & 4 \\ 4 & 1 \end{pmatrix}$$

De la ecuación característica tenemos:

$$\begin{aligned} |\mathbf{A} - \lambda\mathbf{I}| &= \left| \begin{pmatrix} 1 & 4 \\ 4 & 1 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = \left| \begin{pmatrix} 1 & 4 \\ 4 & 1 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \right| \\ &= \left| \begin{pmatrix} 1\text{-}\lambda & 4 \\ 4 & 1\text{-}\lambda \end{pmatrix} \right| = (1-\lambda)^2 - 16 = \lambda^2 - 2\lambda - 15 \\ &= (\lambda - 5)(\lambda + 3) = 0 \end{aligned}$$

Por tanto, los únicos valores propios de $\mathbf{A}$ son $\lambda_1 = 5$ y $\lambda_2 = -3$.
Ahora buscaremos el vector propio asociado a λ_1 Este vector debe ser solución al sistema (1.26), es decir,

$$\begin{aligned} (\mathbf{A} - \lambda_1\mathbf{I})x &= \left[\begin{pmatrix} 1 & 4 \\ 4 & 1 \end{pmatrix} - (5) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right] \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= \begin{pmatrix} 1\text{-}5 & 4 \\ 4 & 1\text{-}5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \text{-}4 & 4 \\ 4 & \text{-}4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0 \end{aligned}$$

donde $\mathbf{x} = (x_1 \quad x_2)'$, aplicando la eliminación de Gauss-Jordan:

$$\left(\begin{array}{cc|c} -4 & 4 & 0 \\ 4 & -4 & 0 \end{array} \right) \Rightarrow \left(\begin{array}{cc|c} 1 & -1 & 0 \\ 0 & 0 & 0 \end{array} \right)$$

Finalmente tenemos la ecuación:

$$-x_1 + x_2 = 0 \Rightarrow x_1 = x_2 \Rightarrow \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Lo anterior indica que cualquier vector de la forma:

$$k \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

es un vector propio asociado a $\lambda_1 = 5$. Ahora buscaremos el vector propio asociado a λ_2 Este vector también debe ser solución al sistema (1.26), entonces:

$$\begin{aligned}(\mathbf{A} - \lambda_1\mathbf{I})x &= \left[\begin{pmatrix} 1 & 4 \\ 4 & 1 \end{pmatrix} - (-3)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= \begin{pmatrix} 1{+}3 & 4 \\ 4 & 1{+}3 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 4 & 4 \\ 4 & 4 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0\end{aligned}$$

donde $\mathbf{x} = (x_1 \quad x_2)'$, aplicando la eliminación de Gauss-Jordan:

$$\left(\begin{array}{cc|c} 4 & 4 & 0 \\ 4 & 4 & 0 \end{array}\right) \Rightarrow \left(\begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right)$$

Finalmente tenemos la ecuación:

$$x_1 + x_2 = 0 \Rightarrow x_1 = -x_2 \Rightarrow \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_2\begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

Lo anterior indica que cualquier vector de la forma:

$$k\begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

es un vector propio asociado a $\lambda_2 = -3$.

1.11. Descomposición espectral

Si tenemos una matriz $\mathbf{C} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ que contiene los vectores propios normalizados de una matriz simétrica $\mathbf{A}$ de orden n, entonces $\mathbf{C}$ es ortogonal. Por lo tanto, por (1.21), $\mathbf{I} = \mathbf{C}\mathbf{C}'$, que podemos multiplicar por $\mathbf{A}$ para obtener:

$$\mathbf{A} = \mathbf{A}\mathbf{C}\mathbf{C}'$$

Si sustituimos $\mathbf{C} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ obtenemos

$$\begin{aligned}\mathbf{A} &= \mathbf{A}(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)\mathbf{C}' \\ &= (\mathbf{A}\mathbf{x}_1, \mathbf{A}\mathbf{x}_2, ..., \mathbf{A}\mathbf{x}_n)\mathbf{C}' \qquad \text{por (1.25) tenemos} \\ &= (\lambda_1\mathbf{x}_1, \lambda_2\mathbf{x}_2, ..., \lambda_n\mathbf{x}_n)\mathbf{C}' \\ &= \mathbf{C}\mathbf{D}\mathbf{C}' \qquad (1.28)\end{aligned}$$

donde $\mathbf{D}$ es

$$\mathbf{D} = \begin{pmatrix} \lambda_1 & 0 & \dots 0 \\ 0 & \lambda_2 & \dots 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & \dots \lambda_n \end{pmatrix}$$

La expresión obtenida de una matriz simétrica $\mathbf{A}$ (1.28) en términos de sus valores propios y vectores propios se conoce como la *descomposición espectral* de $\mathbf{A}$.

1.12. Raíz cuadrada de una matriz definida positiva

Si $\mathbf{A}$ es una matriz definida positiva, la descomposición espectral de $\mathbf{A}$ (1.28) puede modificarse tomando las raíces cuadradas de los valores propios para producir una *matriz de raíz cuadrada* $\mathbf{A}$, es decir

$$\mathbf{A}^{1/2} = \mathbf{C}\mathbf{D}^{1/2}\mathbf{C}' \tag{1.29}$$

donde,

$$\mathbf{D}^{1/2} = \begin{pmatrix} \sqrt{\lambda_1} & 0 & \dots 0 \\ 0 & \sqrt{\lambda_2} & \dots 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & \dots \sqrt{\lambda_n} \end{pmatrix}$$

Entonces la matriz de raíz cuadrada $\mathbf{A}^{1/2}$ es simétrica y sirve como la raíz cuadrada de $\mathbf{A}$

$$\mathbf{A}^{1/2}\mathbf{A}^{1/2} = (\mathbf{A}^{1/2})^2 = \mathbf{A} \tag{1.30}$$

Capítulo 2

Análisis de Componentes Principales

En este capítulo se presentarán algunos conceptos básicos para el análisis multivariado, debido a que el ACP es una técnica exploratoria que requiere estos conceptos para la construcción de las componentes principales. Se explicará la medida global de dependencia para saber si hay una alta correlación global entre variables o no, esto nos ayudará a decidir si es conveniente hacer este análisis exploratorio. Sin embargo, en la aplicación, no solo basta que esta medida nos sugiera este análisis, así que explicaremos algunas de las problemáticas más relevantes que podría haber al utilizarlo. También se mostrarán algunos criterios para elegir el número de componentes a estudiar, así como su interpretación y sus limitaciones. Para una mejor comprensión de esta técnica y mostrar con claridad lo bueno que puede resultar este análisis exploratorio en algunos problemas ecológicos, se presenta un estudio taxonómico sobre el análisis de 17 leguminosas, las cuales pertenecen a dos subgéneros diferentes y el objetivo del análisis consiste en averiguar si existe algún patrón que sugiera una constitución de un nuevo género.

2.1. Medidas básicas para el análisis multivariado

2.1.1. Media y varianza muestral de una matriz

Sea $\mathbf{x}$ un vector de p variables medidas en una muestra de datos. Si hay n individuos en la muestra, la *media muestral* para cada variable i se calcula con un promedio de las mediciones que se le hicieron a los n individuos:

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ji} \tag{2.1}$$

Entonces, la media muestral de toda la matriz de datos, se puede calcular como el promedio de cada una de las variables p por separado:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \begin{pmatrix} \bar{x_1} \\ \bar{x_2} \\ \vdots \\ \bar{x_p} \end{pmatrix} \tag{2.2}$$

Adicionalmente, siendo $\mathbf{X}$ la matriz n x p podemos encontrar $\bar{\mathbf{x}}$ utililizando dicha matriz. La idea es sumar las n entradas de cada columna de $\mathbf{X}$ y dividir entre n. Por lo que la expresión a partir de la matriz de datos es:

$$\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\bar{1} = \frac{1}{n}\begin{pmatrix} x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \\ \vdots & \vdots & & \vdots \\ x_{1p} & x_{2p} & \dots & x_{np} \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \bar{x_1} \\ \bar{x_2} \\ \vdots \\ \bar{x_p} \end{pmatrix} \tag{2.3}$$

Para centrar una matriz de datos, nos fijamos en el registro que hay de cada individuo en la i-ésima variable y les restamos su media muestral correspondiente. Entonces si centramos los datos de la matriz $\mathbf{X}$ tendremos la matriz centrada $\widetilde{\mathbf{X}}$ como:

$$\widetilde{\mathbf{X}} = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \dots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \dots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \dots & x_{np} - \bar{x}_p \end{pmatrix} \tag{2.4}$$

La *varianza muestral* se calcula para cada variable con todos los n individuos, entonces para la variable i-ésima tenemos:

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)^2 \tag{2.5}$$

Finalmente, la desviación estandar muestral para la variable i se define como:

$$s_i = +\sqrt{s_i^2} \tag{2.6}$$

2.1.2. Matriz de varianzas y covarianzas muestrales

La covarianza muestral entre dos variables i y k se define como:

$$s_{ik} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \tag{2.7}$$

Observemos que la covarianza muestral entre los datos de la i-ésima variable con ella misma, coincidirá con ser la varianza muestral, la cual se puede denotar como s_{ii}, esto se puede ver de la siguiente manera:

$$s_{ii} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{ji} - \bar{x}_i) = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)^2 = s_i^2 \tag{2.8}$$

Por esa razón, en nuestra matriz resultante, tendremos varianzas en la diagonal y covarianzas en los demás elementos y se denomina como *matriz de varianzas y covarianzas*, la cual será representada por $\mathbf{S}$:

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{21} & \dots & s_{1p} \\ s_{12} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{2p} & \dots & s_{pp} \end{pmatrix} \tag{2.9}$$

Debido a que sólo estamos obteniendo la covarianza entre dos variables para cada elemento de la matriz $\mathbf{S}$ tendremos que $s_{12} = s_{21}$, lo mismo sucederá con los elementos que hayan comparado un mismo par de variables, es decir que $s_{ij} = s_{ji}$, entonces la matriz $\mathbf{S}$ es simétrica. Una forma de calcular $\mathbf{S}$ de manera matricial es haciendo uso de la matriz centrada de datos (2.4), como se muestra:

$$\frac{1}{n-1}\widetilde{\mathbf{X}}'\widehat{\mathbf{X}} = \frac{1}{n-1}\begin{pmatrix} x_{11}-\bar{x}_1 & x_{21}-\bar{x}_1 & \dots & x_{n1}-\bar{x}_1 \\ x_{12}-\bar{x}_2 & x_{22}-\bar{x}_2 & \dots & x_{n2}-\bar{x}_2 \\ \vdots & \vdots & & \vdots \\ x_{1p}-\bar{x}_p & x_{2p}-\bar{x}_p & \dots & x_{np}-\bar{x}_p \end{pmatrix}\begin{pmatrix} x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \dots & x_{1p}-\bar{x}_p \\ x_{21}-\bar{x}_1 & x_{22}-\bar{x}_2 & \dots & x_{2p}-\bar{x}_p \\ \vdots & \vdots & & \vdots \\ x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2 & \dots & x_{np}-\bar{x}_p \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2 & \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2) & \dots & \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{ip}-\bar{x}_p) \\ \frac{1}{n-1}\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)(x_{i1}-\bar{x}_1) & \frac{1}{n-1}\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2 & \dots & \frac{1}{n-1}\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)(x_{ip}-\bar{x}_p) \\ \vdots & \vdots & & \vdots \\ \frac{1}{n-1}\sum_{i=1}^{n}(x_{ip}-\bar{x}_p)(x_{i1}-\bar{x}_1) & \frac{1}{n-1}\sum_{i=1}^{n}(x_{ip}-\bar{x}_p)(x_{i2}-\bar{x}_2) & \dots & \frac{1}{n-1}\sum_{i=1}^{n}(x_{ip}-\bar{x}_p)^2 \end{pmatrix}$$

$$= \begin{pmatrix} s_{11} & s_{21} & \dots & s_{1p} \\ s_{12} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{2p} & \dots & s_{pp} \end{pmatrix}$$

Por lo tanto

$$\frac{1}{n-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} = \mathbf{S} \tag{2.10}$$

2.1.3. Matriz de correlación muestral

La correlación simple entre las variables i y k es definida como:

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}s_{kk}}} = \frac{s_{ik}}{s_i s_k} \tag{2.11}$$

La correlación de dos variables iguales, es 1. Entonces $r_{ii} = 1$. Por lo que si obtenemos la correlación que hay entre cada par de variables, obtendremos la siguiente matriz:

$$\mathbf{R} = \begin{pmatrix} 1 & r_{21} & \dots & r_{1p} \\ r_{12} & 1 & \dots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{2p} & \dots & 1 \end{pmatrix} \tag{2.12}$$

Como estamos obteniendo la correlación entre dos variables para cada elemento de la matriz $\mathbf{R}$ tendremos que $r_{ik} = r_{ki}$. En consecuencia, la matriz $\mathbf{R}$ es simétrica.

La matriz de correlación muestral puede ser obtenida de la matriz de covarianza y viceversa. Sea

$$\begin{aligned} \mathbf{D}_s &= diag(\sqrt{s_{11}}, \sqrt{s_{22}}, \dots, \sqrt{s_{pp}}) \\ &= diag(s_1, s_2, \dots, s_p) \\ &= \begin{pmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & s_p \end{pmatrix} \end{aligned}$$

Entonces

$$\mathbf{R} = \mathbf{D}_s^{-1}\mathbf{S}\mathbf{D}_s^{-1} \tag{2.13}$$

$$\mathbf{S} = \mathbf{D}_s\mathbf{R}\mathbf{D}_s \tag{2.14}$$

2.2. El estudio de los datos en la ecología

Los ecologistas a menudo quieren comprender las causas que pueden afectar en la distribución y abundancia de organismos, identificar las composiciones de los organismos para analizar que puede afectar en su desarrollo, así que intentan asociar algunos patrones que puedan explicar ciertos comportamientos. Por ejemplo, el querer asociar la abundancia de especies animales con la vegetación o el clima. Regularmente,

esto implica un proceso de dos pasos, el primero es identificar y describir patrones de la distribución de organismos y el segundo paso es distinguir factores ecológicos o ambientales que contribuyan en explicar esos patrones. En cualquier caso, el proceso intenta identificar y describir los patrones importantes en los registros de datos.

El análisis de componentes principales es una técnica multivariada que evalúa las relaciones dentro de un único conjunto de variables interdependientes, siendo independientes a las relaciones que puedan tener con variables fuera del conjunto. Generalmente, los datos son un conjunto de observaciones donde tendremos un número de individuos n, el individuo puede referirse a la especie o a un lugar, y un número de variables p donde podrían involucrarse los factores ambientales, como el clima, el tipo de tratamiento, la humedad, la luz, algunas descripciones cuantitativas del organismo en estudio, etc. Por ejemplo, en la tabla 2.1 tenemos 17 especies de leguminosas y sus respectivas variables son datos cuantitativos que describen la morfología cromosómica de las especies en estudio.

Utilizando los datos presentados, obtendremos su matriz de varianzas y covarianzas. Para obtener dicha matriz, omitimos la variable **"2n"** ya que estamos trabajando con puras especies de leguminosas, entonces todas tienen 10 pares de cromosomas (20 cromosomas), entonces es una variable que no afecta en el análisis. También omitimos la variable **"KF"** ya que nos presenta 10 categorías de 17 especies (las cuales son muchas para la poca cantidad de individuos).

	THC	AC	Range	Ratio	TF	CI
THC	9.807481	0.9812172	0.8151533	0.2541198	1.107030	0.4399838
AC	0.981217	0.0981808	0.0816930	0.0255643	0.109576	0.0418897
Range	0.815153	0.0816930	0.1127860	0.0717661	-0.263853	-0.3498018
Ratio	0.254119	0.0255643	0.0717661	0.0673845	-0.405151	-0.473711
TF	1.107030	0.1095764	-0.2638533	-0.405151	7.632181	7.6423349
CI	0.439983	0.0418897	-0.3498018	-0.473711	7.642334	8.0890492

Con las mismas variables, calculamos la matriz de correlación muestral:

	THC	AC	Range	Ratio	TF	CI
THC	1.0000000	0.9999377	0.775055	0.312593	0.127954	0.0493979
AC	0.9999377	1.0000000	0.776325	0.314297	0.126584	0.0470051
Range	0.7750556	0.7763254	1.000000	0.823212	-0.284387	-0.3662230
Ratio	0.3125932	0.3142974	0.823212	1.000000	-0.564953	-0.6416298
TF	0.1279548	0.1265842	-0.284387	-0.564953	1.000000	0.9726419
CI	0.0493979	0.0470051	-0.366223	-0.641629	0.972641	1.0000000

Subgenus Aeschynomene	**2n**	**KF**	**THC** (μm)	**AC** (μm)	**Range** (μm)	**Ratio** (**L**/**S**)	**TF**	**CI**
A. americana var. americana	20	$8m + 1sm + 1st$	12.85	1.28	0.86	1.99	40.75	39.88
A. americana var. flabellata	20	$8m + 1sm + 1st$	13.92	1.39	0.77	1.74	42.02	41.4
A. americana var. glandulosa	20	$8m + 1sm + 1st$	15.86	1.58	0.95	1.98	43.23	42.12
A. sp. aff. americana	20	$8m + 1sm + 1st$	16.54	1.64	0.98	1.86	43.06	42.13
A. villosa var. villosa	20	$4m + 4sm + 2st$	14.16	1.41	0.98	2.16	35.17	34.07
A. villosa var. longifolia	20	$4m + 6sm$	13.68	1.36	0.91	2.01	36.98	36.79
A. sp. aff. villosa	20	$7m + 2sm + 1st$	15.9	1.58	1.09	2.06	40.4	39.79
A. sensitiva I	20	$9m + 1sm$	16.65	1.66	0.9	1.72	42.82	43.11
A. sensitiva II	20	$9m + 1sm$	15.66	1.56	0.77	1.63	43.25	42.97
A. deamii	20	$8m + 2sm$	20.82	2.07	1.01	1.65	41.35	41.48
A. scabra	20	$10m$	15.71	1.56	0.66	1.51	45.54	45.54
A. evenia	20	$7m + 3sm$	14.15	1.41	0.82	1.82	42.04	41.5
A. rudis	20	$8m + 2st$	11.39	1.13	0.6	1.74	39.64	39.13
A. ciliata	20	$7m + 3sm$	15.71	1.56	0.9	1.82	41.13	40.83
Subgenus Ochopodium								
A. paniculata	20	$3m + 7sm$	19.28	1.92	1.82	2.52	36.56	36.63
A. lyonnetii	20	$9m + 1sm$	21.86	2.18	1.67	2.33	43.78	41.88
A. amorphoides	20	$8m + 2st$	22.41	2.24	1.46	1.98	39.66	37.61

Tabla 2.1: Tabla de datos cromosómicos de leguminosas

Descripción de variables
2n: Es el numéro de cromosomas de cada especie de leguminosa
KF:Fórmula cariotípica
THC: Longitud cromosómica total
AC: Talla cromosómica promedio
Range: Es la diferencia entre el cromosoma más grande y más pequeño
Ratio: Cociente del cromosoma más largo entre el más pequeño
TF: Índice de Simetría
CI: Índice Certromedico

Tabla 2.2: breve descripción de las variables de la tabla 2.1

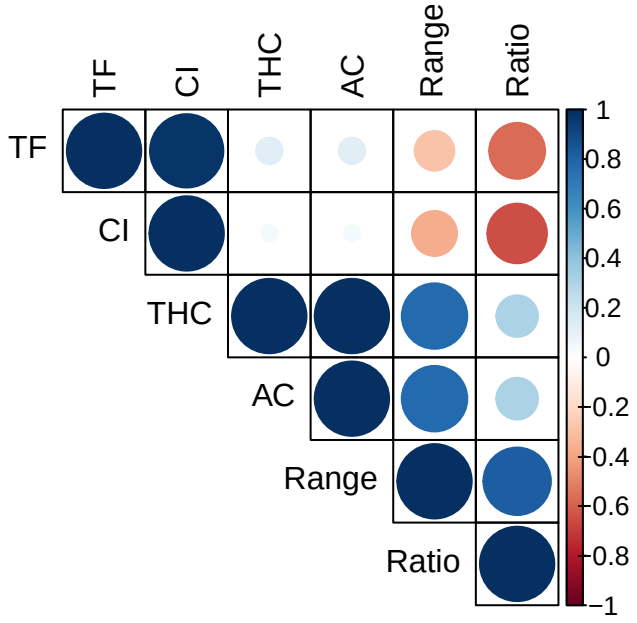

Figura 2.1: Corrplot. Se puede observar que entre más cercano a 1.0 sea el coeficiente de correlación, los círculos se tiñen de azul fuerte, cuando son muy claros quiere decir que la correlación es cercana a cero y si están rojizos tienen correlación negativa. La barra de la derecha muestra la escala que relaciona los colores con los coeficientes de correlación

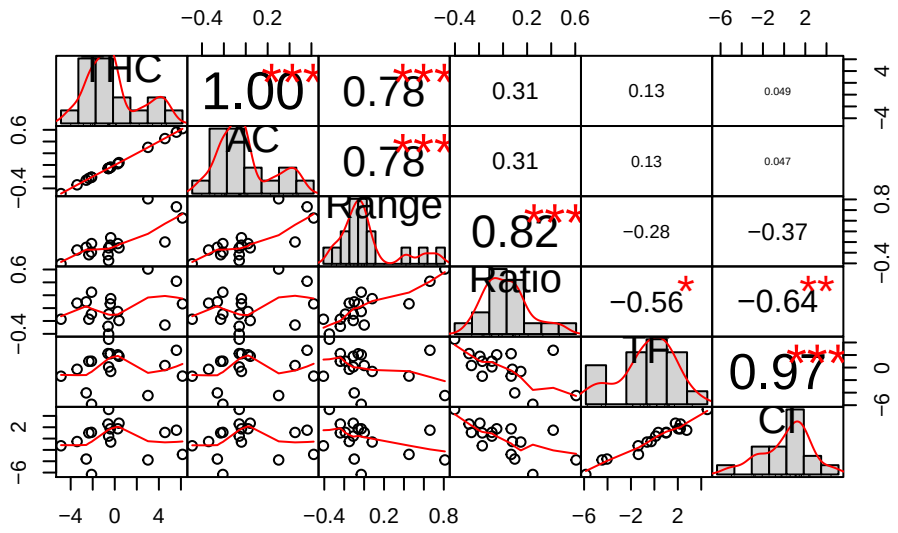

Figura 2.2: chart.Correlation. Se puede observar que las variables que tienen una alta correlación ya sea positiva o negativa, tienen asteriscos y su coeficiente se visualiza de mayor tamaño. Tres asteriscos representa una muy alta correlación, dos una buena correlación y un asterisco representa que hay una correlación mediana y si no tiene quiere decir que hay baja relación entre las variables comparadas

La correlación entre dos variables determina la relación que hay entre ellas. Los coeficientes de correlación puede variar desde -1.0 a 1.0 y son usados para determinar la relación que hay entre variables. Se dice que hay correlación positiva cuando los coeficientes están cercanos a 1.0 y hay corrrelación negativa cuando lo coeficientes están cercanos a -1. En caso de que el coeficiente se encuentre alrededor del cero significa que la relación entre las variables no existe. Se observa que las variables **THC** y **AC** están muy correlacionas. También hay una alta correlación entre las variables **Range** y **Ratio** siendo de 0.823212, la variable **CI** y la variable **TF** tienen una correlación de 0.9726419. Se puede observar una correlación negativa entre las variables **CI** y **Ratio** En los siguientes gráficos se puede visualizar la correlación que hay entre las variables.

2.3. Análisis de Componentes Principales (ACP)

Los primeros desarrollos de esta técnica fueron inicialmente realizados por el científico británico Karl Pearson a finales del siglo XIX. Sus trabajos de ajustes ortogonales constituyeron la base que daría lugar al análisis de componentes principales. Esta técnica se desarrolló de manera formal por Hotelling en 1933. A pesar de haberse desarrollado hace mucho tiempo, está técnica se comenzó a popularizar a finales del siglo XX con el gran avance que dio la tecnología en el desarrollo de las computadoras.

El objetivo principal del análisis de componentes principales (ACP) es la reducción de dimensionalidad, es decir, se condensa la información contenida en las p variables originales (donde cada dimensión está definida por una variable) en un conjunto con un número menor de variables r (menor dimensión) construidas como combinaciones lineales de las originales, estas combinaciones lineales se llaman *componentes principales*. Por lo tanto, si el conjunto de datos original contiene mucha redundancia, ACP extraerá la mayor parte de la variación en las menores dimensiones posible.

De esta manera, el ACP proporciona una interpretación significativa para cada componente, ya que cada uno intenta juntar la información más importante de las variables originales. La interpretación de cada componente en un enfoque ecológico se refleja en la importancia de cada variable que define el componente, es decir, se les da mayor importancia a las variables con mayor peso.

En este trabajo, presentaremos esta técnica como una herramienta exploratoria de datos.

2.4. Medida global de dependencia

Anteriormente mostramos como se analiza la relación entre las variables a partir de la matriz de correlación y también mostramos gráficos que facilitan su interpretación. Sin embargo, eso no muestra de manera formal si las variables se encuentran altamente intercorrelacionadas, ya que si las correlaciones entre todas las variables son bajas, tal vez no sea apropiado realizar el análisis de componentes principales.

Una medida de dependencia global debe ser función de la matriz de correlaciones $\mathbf{R}$. Un coeficiente de dependencia es:

$$\eta^2 = 1 - |\mathbf{R}| \tag{2.15}$$

la cual verifica las siguientes tres propiedades:

1. $0 \leq \eta^2 \leq 1$

2. $\eta^2 = 0$ si y solo si las p variables no están correlacionadas

3. $\eta^2 = 1$ si y solo si hay relaciones lineales entre las variables, es decir, las variables originales están correlacionadas.

Demostraciones:

1. Sean $\lambda_1, \lambda_2, ..., \lambda_p$ los valores propios de $\mathbf{R}$. Si a y b son las medias geométrica y aritmética respectivamente de p números positivos, se verifica que $a \leq b$. Como tenemos p valores propios, por la descomposición espectral definida en (1.28) sabemos que son obtenidos de la matriz R de orden p x p.

   Entonces $\mathbf{tr(R)} = p$ por las definiciones (1.15) y (2.12).
   Como el $|\mathbf{R}| = \lambda_1 \times \lambda_2 \times, ..., \times \lambda_p$
   Entonces usando la media geométrica tenemos que:
   $(|\mathbf{R}|)^{1/p} = (\lambda_1 \times \lambda_2 \times, ..., \times \lambda_p)^{1/p}$
   Y usando la media aritmética tenemos:
   $(\lambda_1 + \lambda_2 +, ..., +\lambda_p)/p$
   entonces:

$$(|\mathbf{R}|)^{1/p} = (\lambda_1 \times \lambda_2 \times, ..., \times \lambda_p)^{1/p} \leq \frac{(\lambda_1 + \lambda_2 +, ..., +\lambda_p)}{p} = \frac{p}{p} = 1$$

   Por lo tanto $0 \leq |\mathbf{R}| \leq 1$ lo cual implica que la propiedad 1 se cumpla.

2. $\mathbf{R} = I$ si y sólo si las variables no están correlacionadas, porque es la matriz identidad, los elementos fuera de la diagonal son ceros y son las correlaciones entre las variables. Entonces si sustituimos en (2.15) tenemos que:

$$\eta^2 = 1 - |\mathbf{I}| = 1 - 1 = 0$$

   Por lo tanto queda demostrada la propiedad 2.

3. Si $|\mathbf{R}| = 0$ se cumple que $\eta^2 = 1$, como el determinante de la matriz de correlaciones es cero, entonces $r(\mathbf{R}) < p$ y por lo tanto hay combinaciones lineales entre las variables, lo cual se debe al siguiente teorema:

   Teorema 2.4.1 *Si $r = r(\boldsymbol{S}) \leq p$ hay r variables linealmente independientes y las otras $p - r$ son combinación lineal de estas r variables.*

Cabe mencionar que por (2.13) y (2.14) se cumple que $r(\mathbf{S}) = r(\mathbf{R})$, es decir, el rango de la matriz de varianzas y covarianzas muestrales es igual al rango de la matriz de correlación muestral.

Por lo tanto si el determinante de la matriz de correlación muestral es cercano a cero, entonces hay variables originales en estudio que están altamente correlacionadas y eso indica que una o más variables podrían ser expresadas como combinación lineal de otras variables, por lo tanto, hacer un análisis de componentes principales es adecuado. Caso contrario sucede cuando el determinante de la matriz de correlación muestral está próximo a 1 y en ese caso tal vez no sea muy apropiado realizar el ACP. El determinante de la matriz de correlación muestral asociada a la (2.1) es muy próximo a cero, siendo de 3.858809×10^{-8}, por lo tanto, a esos datos ecológicos se les puede realizar un ACP.

2.5. Cálculo de las componentes principales

Sea $\mathbf{X}$ una matriz de datos centrada con p variables sobre n individuos. A partir de ellas, se buscará un nuevo conjunto de variables $(\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_p)$ no correlacionadas entre sí, cuyas varianzas vayan decreciendo progresivamente, las cuales son conocidas como *componentes principales*. Cada $\mathbf{z}_i$ $(i = 1, 2, ..., p)$ es una combinación lineal de las variables originales ya centradas, es decir para un sólo individuo se tiene lo siguiente:

$$\mathbf{z}_i = x_1 a_{1i} + x_2 a_{2i} + ... + x_p a_{pi} = \mathbf{x}\mathbf{a}_i \tag{2.16}$$

$$\text{donde} \quad \mathbf{x}' = (x_1, x_2, ..., x_p) \quad \text{y} \quad \mathbf{a}_i = \begin{pmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{pi} \end{pmatrix}$$

Debido a que queremos mantener la ortogonalidad de la transformación se establece que la norma del vector $\mathbf{a}_i$ es igual a 1, es decir

$$||\mathbf{a}_i||^2 = (\sqrt{\mathbf{a}'\mathbf{a}})^2 = a_{1i}^2 + a_{2i}^2 + ... + a_{pi}^2 = 1$$

Para que la idea quede clara, primero vamos a proyectar cada individuo a lo largo del vector $\mathbf{a}_1$ como se muestra a continuación:

Sea

$$\begin{cases} \mathbf{x}_1' = & (x_{11} & x_{12} & \dots & x_{1p}) \\ \mathbf{x}_2' = & (x_{21} & x_{22} & \dots & x_{2p}) \\ \vdots & & & & \\ \mathbf{x}_n' = & (x_{n1} & x_{n2} & \dots & x_{np}) \end{cases} \quad \text{y} \quad \mathbf{a}_1 = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{pmatrix}$$

Entonces

$$\mathbf{X}\mathbf{a}_1 = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1'\mathbf{a}_1 \\ \mathbf{x}_2'\mathbf{a}_1 \\ \vdots \\ \mathbf{x}_n'\mathbf{a}_1 \end{pmatrix} = \begin{pmatrix} Proy_{\mathbf{a}_1}\mathbf{x}_1 \\ Proy_{\mathbf{a}_1}\mathbf{x}_2 \\ \vdots \\ Proy_{\mathbf{a}_1}\mathbf{x}_n \end{pmatrix}$$

Como queremos maximizar la varianza que hay en las proyecciones de los n individuos a lo largo de $\mathbf{a}_1$ tenemos que: Sea $\mathbf{b}_1 = \mathbf{X}\mathbf{a}_1$, entonces

$$\mathbf{Var}(\mathbf{b}_1) = \frac{\mathbf{b}_1'\mathbf{b}_1}{n-1} = \frac{(\mathbf{a}_1'\mathbf{X}')(\mathbf{X}\mathbf{a}_1)}{n-1} = \frac{\mathbf{a}_1'(\mathbf{X}'\mathbf{X})\mathbf{a}_1}{n-1} = \mathbf{a}_1'\mathbf{S}\mathbf{a}_1$$

ya se mostró que $\mathbf{S} = \frac{\mathbf{X}'\mathbf{X}}{n-1}$ donde $\mathbf{X}$ es la matriz de datos centrada.
Entonces buscamos maximizar la $\mathbf{Var}(\mathbf{b}_1) = \mathbf{a}_1'\mathbf{S}\mathbf{a}_1$ sujeto a $||\mathbf{a}_1||^2 = \mathbf{a}_1'\mathbf{a}_1 = 1$

Un método para maximizar una función de varias variables sujeta a restricciones es el método de los multiplicadores de Lagrange. Notemos que la variable $\mathbf{a}_1$ es un vector desconocido y deseamos encontrarlo, de tal manera que nos de la combinación lineal óptima. Por lo cual, utilizando dicho método tenemos lo siguiente:

$$\text{Sea} \quad L(\mathbf{a}_1) = \mathbf{a}_1'\mathbf{S}\mathbf{a}_1 - \lambda_1(\mathbf{a}_1'\mathbf{a}_1 - 1)$$

Derivamos para optimizar la función

$$\begin{aligned} \frac{\partial L(\mathbf{a}_1)}{\partial \mathbf{a}_1} &= 2\mathbf{S}\mathbf{a}_1 - 2\lambda_1\mathbf{a}_1 = 0 \\ &\Rightarrow 2\mathbf{S}\mathbf{a}_1 = 2\lambda_1\mathbf{a}_1 \Rightarrow \mathbf{S}\mathbf{a}_1 = \lambda_1\mathbf{a}_1 \end{aligned}$$

Notemos que se cumple (1.25), entonces $\mathbf{a}_1$ es un vector propio de $\mathbf{S}$ y λ_1 es el valor propio asociado. Sustituyendo $\mathbf{S}\mathbf{a}_1 = \lambda_1\mathbf{a}_1$ en la función a maximizar obtenemos:

$$\mathbf{a}_1'\mathbf{S}\mathbf{a}_1 = \mathbf{a}_1'\lambda_1\mathbf{a}_1 = \lambda_1\mathbf{a}_1'\mathbf{a}_1 = \lambda_1(1) = \lambda_1$$

Por lo tanto λ_1 es el valor propio más grande.
Como $\mathbf{S}$ es la matriz de varianzas y covarianzas entonces es una matriz simétrica y

positiva definida. Por lo tanto, los valores propios son no negativos.

Ahora haremos la proyección de cada individuo en dos direcciones, teniendo lo siguiente:

Sea

$$\begin{cases} \mathbf{x}_1' = & (x_{11} & x_{12} & \dots & x_{1p}) \\ \mathbf{x}_2' = & (x_{21} & x_{22} & \dots & x_{2p}) \\ \vdots & & & & \\ \mathbf{x}_n' = & (x_{n1} & x_{n2} & \dots & x_{np}) \end{cases} \quad \mathbf{a}_1 = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{pmatrix} \quad \text{y} \quad \mathbf{a}_2 = \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{p2} \end{pmatrix}$$

Entonces

$$\mathbf{X}\mathbf{a}_1 = \begin{pmatrix} \mathbf{x}_1'\mathbf{a}_1 \\ \mathbf{x}_2'\mathbf{a}_1 \\ \vdots \\ \mathbf{x}_n'\mathbf{a}_1 \end{pmatrix} = \begin{pmatrix} Proy_{\mathbf{a}_1}\mathbf{x}_1 \\ Proy_{\mathbf{a}_1}\mathbf{x}_2 \\ \vdots \\ Proy_{\mathbf{a}_1}\mathbf{x}_n \end{pmatrix} \quad \text{y}$$

$$\mathbf{X}\mathbf{a}_2 = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{p2} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1'\mathbf{a}_2 \\ \mathbf{x}_2'\mathbf{a}_2 \\ \vdots \\ \mathbf{x}_n'\mathbf{a}_2 \end{pmatrix} = \begin{pmatrix} Proy_{\mathbf{a}_2}\mathbf{x}_1 \\ Proy_{\mathbf{a}_2}\mathbf{x}_2 \\ \vdots \\ Proy_{\mathbf{a}_2}\mathbf{x}_n \end{pmatrix}$$

Sea $\mathbf{b}_1 = \mathbf{X}\mathbf{a}_1$ y $\mathbf{b}_2 = \mathbf{X}\mathbf{a}_2$

Entonces

$$\mathbf{Var}(\mathbf{b}_1) = \frac{\mathbf{b}_1'\mathbf{b}_1}{n-1} = \frac{(\mathbf{a}_1'\mathbf{X}')(\mathbf{X}\mathbf{a}_1)}{n-1} = \frac{\mathbf{a}_1'(\mathbf{X}'\mathbf{X})\mathbf{a}_1}{n-1} = \mathbf{a}_1'\mathbf{S}\mathbf{a}_1 \tag{2.17}$$

$$\mathbf{Var}(\mathbf{b}_2) = \frac{\mathbf{b}_2'\mathbf{b}_2}{n-1} = \frac{(\mathbf{a}_2'\mathbf{X}')(\mathbf{X}\mathbf{a}_2)}{n-1} = \frac{\mathbf{a}_2'(\mathbf{X}'\mathbf{X})\mathbf{a}_2}{n-1} = \mathbf{a}_2'\mathbf{S}\mathbf{a}_2 \tag{2.18}$$

Al tener dos proyecciones, queremos maximizar lo siguiente:
$\mathbf{Var}(\mathbf{b}_1) + \mathbf{Var}(\mathbf{b}_2)$ sujeto a $||\mathbf{a}_1||^2 = \mathbf{a}_1'\mathbf{a}_1 = 1$ y $||\mathbf{a}_2||^2 = \mathbf{a}_2'\mathbf{a}_2 = 1$
Utilizando los multiplicadores de Lagrange tenemos que:

$$\text{Sea} \quad L(\mathbf{a}_1, \mathbf{a}_2) = \mathbf{a}_1'\mathbf{S}\mathbf{a}_1 - \lambda_1(\mathbf{a}_1'\mathbf{a}_1 - 1) + \mathbf{a}_2'\mathbf{S}\mathbf{a}_2 - \lambda_2(\mathbf{a}_2'\mathbf{2}_1 - 1)$$

Obtenemos las derivadas parciales para maximizar la función:

$$\frac{\partial L(\mathbf{a}_1, \mathbf{a}_2)}{\partial \mathbf{a}_1} = 2\mathbf{S}\mathbf{a}_1 - 2\lambda_1\mathbf{a}_1 = 0 \Rightarrow 2\mathbf{S}\mathbf{a}_1 = 2\lambda_1\mathbf{a}_1 \Rightarrow \mathbf{S}\mathbf{a}_1 = \lambda_1\mathbf{a}_1$$

$$\frac{\partial L(\mathbf{a}_2, \mathbf{a}_2)}{\partial \mathbf{a}_2} = 2\mathbf{S}\mathbf{a}_2 - 2\lambda_2\mathbf{a}_2 = 0 \Rightarrow 2\mathbf{S}\mathbf{a}_2 = 2\lambda_2\mathbf{a}_2 \Rightarrow \mathbf{S}\mathbf{a}_2 = \lambda_2\mathbf{a}_2$$

Observemos que $\mathbf{a}_1$ y $\mathbf{a}_2$ son vectores propios de $\mathbf{S}$ mientras λ_1 y λ_2 son los valores propios. Si sustituimos $\mathbf{S}\mathbf{a}_1 = \lambda_1\mathbf{a}_1$ y $\mathbf{S}\mathbf{a}_2 = \lambda_2\mathbf{a}_2$ en (2.17) y (2.18) respectivamente se obtiene:

$$\mathbf{a}_1'\mathbf{S}\mathbf{a}_1 = \mathbf{a}_1'\lambda_1\mathbf{a}_1 = \lambda_1\mathbf{a}_1'\mathbf{a}_1 = \lambda_1(1) = \lambda_1$$
$$\mathbf{a}_2'\mathbf{S}\mathbf{a}_2 = \mathbf{a}_2'\lambda_1\mathbf{a}_2 = \lambda_2\mathbf{a}_2'\mathbf{a}_2 = \lambda_2(1) = \lambda_2$$

Por lo tanto, λ_1 es el valor propio más grande y λ_2 es el segundo valor propio más grande. Como $\mathbf{S}$ es simétrica y sus vectores propios están normalizados, se cumple (1.28) entonces los vectores propios son ortogonales.

Generalizando para obtener todos los componentes, vamos a proyectar cada individuo a lo largo de cada vector $\mathbf{a}_i$. Sea $\mathbf{X}$ una matriz centrada (como en los dos casos anteriores), además:
Sea

$$\begin{cases} \mathbf{x}_1' = & (x_{11} & x_{12} & \dots & x_{1p}) \\ \mathbf{x}_2' = & (x_{21} & x_{22} & \dots & x_{2p}) \\ \vdots & & & & \\ \mathbf{x}_n' = & (x_{n1} & x_{n2} & \dots & x_{np}) \end{cases} \quad \text{y} \quad \mathbf{a}_1 = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{pmatrix} \mathbf{a}_2 = \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{p2} \end{pmatrix} \dots \mathbf{a}_p = \begin{pmatrix} a_{1p} \\ a_{2p} \\ \vdots \\ a_{pp} \end{pmatrix}$$

donde $||\mathbf{a}_i||^2 = 1$ para toda $i = (1, 2, \dots, p)$
Entonces

$$\mathbf{X}\mathbf{a}_1 = \begin{pmatrix} Proy_{\mathbf{a}_1}\mathbf{x}_1 \\ Proy_{\mathbf{a}_1}\mathbf{x}_2 \\ \vdots \\ Proy_{\mathbf{a}_1}\mathbf{x}_n \end{pmatrix} \mathbf{X}\mathbf{a}_2 = \begin{pmatrix} Proy_{\mathbf{a}_2}\mathbf{x}_1 \\ Proy_{\mathbf{a}_2}\mathbf{x}_2 \\ \vdots \\ Proy_{\mathbf{a}_2}\mathbf{x}_n \end{pmatrix} \dots \mathbf{X}\mathbf{a}_p = \begin{pmatrix} Proy_{\mathbf{a}_p}\mathbf{x}_1 \\ Proy_{\mathbf{a}_p}\mathbf{x}_2 \\ \vdots \\ Proy_{\mathbf{a}_p}\mathbf{x}_n \end{pmatrix}$$

Sea $\mathbf{b}_i = \mathbf{X}\mathbf{a}_i$ para toda $i = (1, 2, \dots, p$, entonces

$$\mathbf{Var}(\mathbf{b}_i) = \frac{\mathbf{b}_i'\mathbf{b}_i}{n-1} = \frac{(\mathbf{a}_i'\mathbf{X}')(\mathbf{X}\mathbf{a}_i)}{n-1} = \frac{\mathbf{a}_i'(\mathbf{X}'\mathbf{X})\mathbf{a}_i}{n-1} = \mathbf{a}_i'\mathbf{S}\mathbf{a}_i$$

Por lo que nuestro objetivo es maximizar:
$\mathbf{Var}(\mathbf{b}_1) + \mathbf{Var}(\mathbf{b}_2) + \dots + \mathbf{Var}(\mathbf{b}_p)$ sujeto a $||\mathbf{a}_i||^2 = \mathbf{a}_i'\mathbf{a}_i = 1 \ \forall \ i = (1, 2, \dots, p)$
Nuevamnte, utilizando los multiplicadores de Lagrange:
Sea $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p)$ y :

$$L(\mathbf{A}) = (\mathbf{a}_1'\mathbf{S}\mathbf{a}_1 - \lambda_1(\mathbf{a}_1'\mathbf{a}_1 - 1)) + (\mathbf{a}_2'\mathbf{S}\mathbf{a}_2 - \lambda_2(\mathbf{a}_2'\mathbf{2}_1 - 1)) + \dots + (\mathbf{a}_p'\mathbf{S}\mathbf{a}_p - \lambda_p(\mathbf{a}_p'\mathbf{a}_p - 1))$$

Calculamos las derivadas parciales

$$\begin{aligned}
&\frac{\partial L(\mathbf{A})}{\partial \mathbf{a}_1} = 2\mathbf{S}\mathbf{a}_1 - 2\lambda_1\mathbf{a}_1 = 0 \Rightarrow 2\mathbf{S}\mathbf{a}_1 = 2\lambda_1\mathbf{a}_1 \Rightarrow \mathbf{S}\mathbf{a}_1 = \lambda_1\mathbf{a}_1 \\
&\frac{\partial L(\mathbf{A})}{\partial \mathbf{a}_2} = 2\mathbf{S}\mathbf{a}_2 - 2\lambda_2\mathbf{a}_2 = 0 \Rightarrow 2\mathbf{S}\mathbf{a}_2 = 2\lambda_2\mathbf{a}_2 \Rightarrow \mathbf{S}\mathbf{a}_2 = \lambda_2\mathbf{a}_2 \\
&\vdots \\
&\frac{\partial L(\mathbf{A})}{\partial \mathbf{a}_p} = 2\mathbf{S}\mathbf{a}_p - 2\lambda_p\mathbf{a}_p = 0 \Rightarrow 2\mathbf{S}\mathbf{a}_p = 2\lambda_p\mathbf{a}_p \Rightarrow \mathbf{S}\mathbf{a}_p = \lambda_p\mathbf{a}_p
\end{aligned}$$

Por lo que

$$\frac{\partial L(\mathbf{A})}{\partial \mathbf{a}_i} = 2\mathbf{S}\mathbf{a}_i - 2\lambda_i\mathbf{a}_i = 0 \Rightarrow 2\mathbf{S}\mathbf{a}_i = 2\lambda_i\mathbf{a}_i \Rightarrow \mathbf{S}\mathbf{a}_i = \lambda_i\mathbf{a}_i \quad \forall\, i = (1, 2, \ldots, p)$$

Así que $\lambda_1 > \lambda_2 > \ldots > \lambda_p$ son los valores propios y la matriz $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p)$ contiene a los vectores propios asociados que están normalizados. Como $\mathbf{S}$ es simétrica entonces sus vectores propios son mutuamente ortogonales, por lo tanto las componentes principales contruidas no están correlacionadas.

2.6. Proporciones de varianza

Los valores propios representan las varianzas de las componentes principales, por lo que cada valor propio λ_i corresponde a la varianza de la componente $\mathbf{z}_i$, esto se debe a que en la sección anterior se mostró que al maximizar la varianza de las proyecciones obtuvimos que correspondían a los valores propios, es decir, $\mathbf{Var}(\mathbf{z}_i) = \mathbf{Var}(\mathbf{X}\mathbf{a}_i) = \lambda_i$. Todos los valores propios son positivos o cero, se toman los mayores ya que estos explican una mayor variación de la muestra en esa componente principal. Como los vectores propios son ortogonales, eso implica que las componentes principales también lo sean, por lo cual si tenemos p variables entonces tendremos p componentes principales que forman p dimensiones. De esta manera, la componente con el mayor valor propio asociado (siempre es la primer componente) es la que contiene mayor varianza (la varianza más grande entre entidades) y por lo tanto es la que mejor explica la estructura de la muestra (ya que contiene mayor información), la segunda componente corresponde al segundo valor propio más grande que mide la varianza a lo largo de la segunda componente principal, el cual es ortogonal a la primer componente, el tercer valor propio corresponde a la componente con la siguiente variación mayor, que es ortogonal a la primera y segunda componente, esta proporciona la mayor explicación adicional de la varianza de la muestra después de que la primera y la segunda hayan hecho su mejor posible y así sucesivamente hasta asociar los p valores propios. Por tal motivo, cuando los valores propios que se asocian a las componentes principales son cercanos a cero o cero, quiere decir que el poder explicativo de la estructura de la muestra es insignificante para dichas componentes. Para obtener la varianza total de las componentes, basta con sumar todos los valores propios que se asocian, es decir:

$$\sum_{i=1}^{p} \mathbf{Var}(\mathbf{z}_i) = \sum_{i=1}^{p} \lambda_i \tag{2.19}$$

Si utilizamos una matriz $\mathbf{A}$ que contiene los vectores propios ($\mathbf{A}$=($\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_p$)) y la matriz diagonal $\mathbf{D}$ que contiene los valores propios, ambos obtenidos de la matriz de varianzas y covarianzas $\mathbf{S}$ podremos hacer una descomposición espectral (1.28) entonces:

$$\mathbf{tr}(\mathbf{S}) = \mathbf{tr}(\mathbf{A}\mathbf{D}\mathbf{A}') = \mathbf{tr}(\mathbf{D}\mathbf{A}\mathbf{A}') = \mathbf{tr}(\mathbf{D}) = \sum_{i=1}^{p} \lambda_i \tag{2.20}$$

Por lo tanto, la suma de varianzas de las variables originales y la suma de varianzas de las componentes principales son iguales.

Para saber cuál es la proporción de varianza que contiene cada componente principal se tiene:

$$\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} = \frac{\mathbf{Var}(\mathbf{z}_i)}{\sum_{i=1}^{p} \mathbf{Var}(\mathbf{z}_i)} \tag{2.21}$$

También se puede expresar la proporción de varianza acumulada por las primeros k componentes (donde $k < p$) :

$$\frac{Varianza\ acumulada\ por\ k\ componentes}{Varianza\ total} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \tag{2.22}$$

En general, se pretende que las primeras dos o tres componentes principales expliquen una buena proporción de la varianza total de la muestra, con la finalidad de poderlos graficar e intentar darles una interpretación adecuada. Sin embargo, el identificar la información que contienen las componenes, a menudo resulta ser un problema complicado.

2.7. Componentes principales utilizando la matriz de correlaciones

Anteriormente se mostró el cálculo para encontrar las componentes principales a partir de una matriz centrada de datos y al obtener la varianza de cada proyección obteníamos una expresión que dependía de la matriz de varianzas y covarianzas $\mathbf{S}$. Por otro lado, si utilizamos una matriz donde las variables están estandarizadas, es decir con media 0 y varianza 1, equivaldría a tomar las componentes principales de la matriz de correlaciones $\mathbf{R}$ en vez del uso de la matriz $\mathbf{S}$, ya que si las variables están estandarizadas entonces las covarianzas coinciden con las correlaciones. Por tal motivo, al maximizar la proyección de un individuo a lo largo del vector $\mathbf{a}_i$ tendremos que los valores propios más grandes son los que maximizan cada componente al igual que con una matriz centrada de datos. Sin embargo los vectores propios y valores propios de la matriz de correlaciones si difieren a los que se obtienen con la matriz de varianzas y covarianzas.

Si se decide estandarizar cada variable, le estaríamos dando la misma importancia a cada una y al obtener la varianza de cada componente obtendremos una expresión que depende de la matriz de correlaciones $\mathbf{R}$. Se sabe que todos los elementos de la diagonal de dicha matriz son 1. Por lo que

$$\mathbf{tr}(\mathbf{R}) = \sum_{i=1}^{p} 1 = p \tag{2.23}$$

Si nuevamente $\mathbf{A}$ es una matriz que contiene los vectores propios ($\mathbf{A}$=($\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_p$)) y la matriz diagonal $\mathbf{D}$ que contiene los valores propios, podremos hacer una descomposición espectral entonces:

$$\mathbf{tr}(\mathbf{R}) = \mathbf{tr}(\mathbf{ADA}') = \mathbf{tr}(\mathbf{DAA}') = \mathbf{tr}(\mathbf{D}) = \sum_{i=1}^{p} \lambda_i = p$$

Por lo tanto

$$\sum_{i=1}^{p} \mathbf{Var}(\mathbf{z}_i) = \sum_{i=1}^{p} \lambda_i = p \tag{2.24}$$

Con este resultado podemos concluir que la varianza total de los componentes principales es igual al número de variables que hay en la muestra.
Para obtener la proporción de varianza de cada i-ésima compontente se tiene que:

$$\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} = \frac{\lambda_i}{p} \tag{2.25}$$

De forma análoga la proporción de varianza acumulada por las primeras k componentes ($k < p$) se expresa como:

$$\frac{\sum_{i=1}^{k} \lambda_i}{p} \tag{2.26}$$

En la mayoría de análisis, se prefiere utilizar la matriz de correlación ya que le da el mismo peso a cada variable; en muchos casos es lo más adecuado, y siempre es más apropiada si la escala o unidad de medida de las variables difiere mucho entre ellas porque generalmente tenemos pocos argumentos comprobados que permitan decidir si una variable es más importante que otra en el punto de vista ecológico.
Sin embargo, en algunas circunstancias puede que la matriz de varianzas y covarianzas sea más deseable, en particular cuando todas las variables tienen una escala de

medición común. Como lo hemos mencionado, la interpretación de las componentes principales pueden ser complicada, pues el hecho de obtener buenas proporciones de varianza en las primeras dos o tres componentes, no garantizan que las conclusiones sean buenas o sencillas, así que un gran número de correlaciones altas puede indicar que varias variables tendrán un gran peso en cada componente principal, esto complicaría mucho la interpretación. Lo más sencillo son lo casos donde hay correlaciones altas entre pocas variables.

En seguida mostraremos algunos resultados que obtuvimos utilizando la matriz de correlaciones correspondiente a los datos de la tabla 2.1 sin considerar las variables **"2n"** y **"KF"**, su exclusión fue explicada con anterioridad.

	Valores propios	**Proporción de varianza**	**Proporción acumulada**
λ_1	3.26077	0.5434628	0.5434628
λ_2	2.28092	0.3801533	0.9236161
λ_3	0.42025	0.0700429	0.9936591
λ_4	0.02791	0.0046528	0.9983119
λ_5	0.01008	0.0016807	0.9999927
λ_6	0.000043	0.0000073	1.0000000

Tabla 2.3: Tabla de los valores propios de la matriz **R** y sus respectivas proporciones de varianza

Notamos que las primeras dos componentes principales acumulan el 92.36 % de la varianza total y si decidiéramos tomar tres componentes se obtendría un acumulado de 99.33 % de la varianza total. Así que en la siguiente tabla mostraremos los vectores propios correspondientes a las primeras tres componentes principales.

	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$
THC	0.402	0.434	-0.318
AC	0.403	0.433	-0.317
Range	0.532	0.120	0.289
Ratio	0.464	-0.211	0.675
TF	-0.275	0.545	0.391
CI	-0.319	0.518	0.327

Tabla 2.4: Tabla de los vectores propios de la matriz **R**

Entonces las primeras tres componentes principales son:

$$\mathbf{z}_1 = \mathbf{X}\mathbf{a}_1 = 0.402\mathbf{x}_1 + 0.403\mathbf{x}_2 + 0.532\mathbf{x}_3 + 0.464\mathbf{x}_4 - 0.275\mathbf{x}_5 - 0.319\mathbf{x}_6$$
$$\mathbf{z}_2 = \mathbf{X}\mathbf{a}_2 = 0.434\mathbf{x}_1 + 0.433\mathbf{x}_2 + 0.120\mathbf{x}_3 - 0.211\mathbf{x}_4 + 0.545\mathbf{x}_5 + 0.518\mathbf{x}_6$$
$$\mathbf{z}_3 = \mathbf{X}\mathbf{a}_3 = 0.318\mathbf{x}_1 + 0.317\mathbf{x}_2 - 0.289\mathbf{x}_3 - 0.675\mathbf{x}_4 - 0.391\mathbf{x}_5 - 0.327\mathbf{x}_6$$

2.8. Elección e importancia del número de componentes

Una parte fundamental en ACP es determinar cuántas componentes principales se deben seleccionar e interpretar. Recuerde que entre más grande sea el valor propio, mayor será el poder explicativo de esa componente. Nuestro objetivo es que por medio de estas componentes se pueda explicar la mayor variabilidad de la muestra original de datos. Se han desarrollado distintos enfoques para determinar la importancia de cada componente principal, seleccionar el número adecuado de componentes e interpretar. En esta sección se muestran algunos criterios que nos ayudarán a determinar el número de componentes que se interpretarán más adelante.

2.8.1. Criterio de Kaiser-Guttman

Este enfoque heurístico es muy común y también es conocido como el criterio de la raíz latente, el cual consiste en tomar las componentes que tengan valores propios mayores a la unidad ($\lambda_i \geq 1$) para realizar un análisis más detallado, esta regla sólo se puede utilizar cuando estandarizamos cada variable que equivale a utilizar la matriz de correlaciones. Inicialmente fue propuesta por Guttman (1954). Claramente si los valores propios son menores a la unidad, sus respectivos componentes son descartados para el estudio, según el criterio se debe a que cada componente debe tener en cuenta al menos la varianza de una variable, sin embargo todas las variables tienen varianza 1 ya que se utilizó la matriz de correlaciones. El criterio de Kaiser-Guttman es poco confiable cuando hay demasiadas variables, también cuando hay muy pocas, y suele ser un poco más confiable cuando el número de variables está entre 20 y 50. Generalmente, los investigadores utilizan este criterio para determinar el número máximo de componentes a utilizar y a pesar de que esta regla no tiene un buen comportamiento en distintos casos, su uso es muy común debido a la simplicidad del cálculo. Una de las críticas más fuertes es porque si generamos datos aleatorios no correlacionados y se realiza un ACP se producirán valores propios mayores a uno. Este es un motivo por el cual recalcamos la importancia de estudiar si hay correlación entre las variables antes de realizar el ACP.

Utilizando este criterio se puede ver en la Tabla 2.3 que sólo utilizaríamos dos componentes principales ya que $\lambda_1 = 3.26077$ y $\lambda_2 = 2.280902$ son mayores a 1.

2.8.2. Criterio de la gráfica de codo

Una gráfica de codo consiste en trazar los valores propios (λ_i) contra el número de componentes en el orden de extracción, es decir, del índice i, esta representción fue propuesta por Cattell (1966). Se conoce como de "codo" porque al seleccionar las componentes i se va formando una curva que se asemeja a un codo, es decir, hay un cambio de pendiente que desciente abruptamente en el comienzo y que al final se aproxima a cero. Esto se debe a que los valores propios disminuyen secuencialmente desde el primero hasta la última componente.
El número i donde la curva empieza a formarse una línea que se asemeja a una recta horizontal, indica el número máximo de componentes a extraer, es decir, tomamos la primera componente hasta la componente i donde se presenta el comportamiendo mencionado. El problema de este gráfico es que en ocasiones (generalmente cuando hay muchas variables) no es obvio a simple vista el punto donde ocurren los cambios de pendiente que se aproximan a la pendiente cero.
En la siguiente gráfica representaremos los valores propios obtenidos en la Tabla 2.3

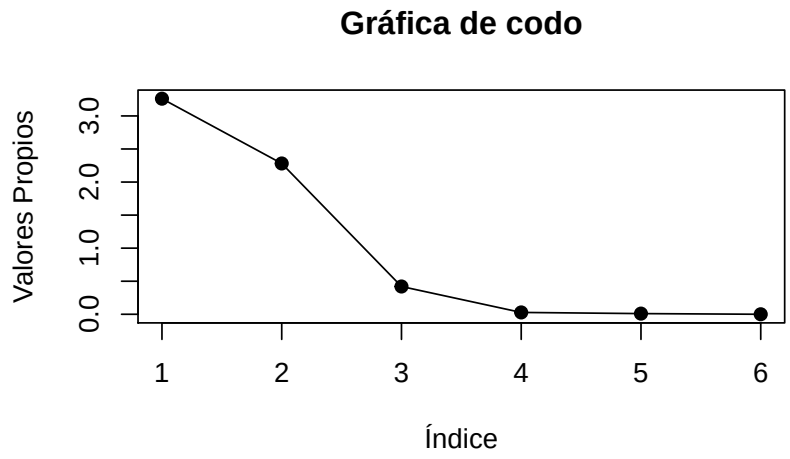

Figura 2.3: Gráfica de codo para seleccionar las componentes principales a partir de la matriz de correlación $\mathbf{R}$

También se puede hacer un diagrama de barras para representar cada valor propio.

Y la idea sería similar, solo que en vez de fijarnos en el comportamiento de la curva, visualizaremos que valores propios explican una mayor varianza y cuales tienen poco poder explicativo.

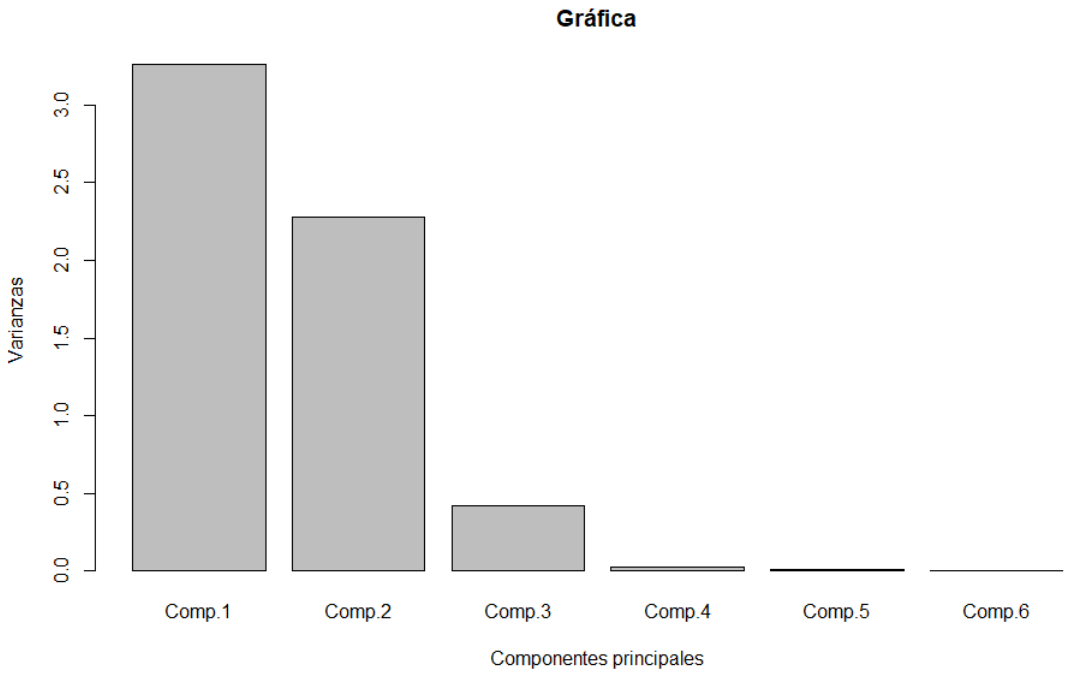

Figura 2.4: Diagrama de barras que representa el poder explicativo de los componentes principales

2.8.3. Criterio del palo roto

Frontier (1976) propuso el método de palo roto (broken stick method) que consiste en obtener valores propios de datos aleatorios. La idea que dio origen al modelo consiste en que si un palo se rompe aleatoriamente en p piezas, λ_1^* sería la pieza más grande en el conjunto de palos rotos, λ_2^* sería el tamaño del segundo palo más grande, y así sucesivamente. Así que el modelo de palo roto asume que si la variación total (la suma de todos los valores propios) se distribuye aleatoriamente entre las componentes y se acomodan de mayor a menor, entonces la gráfica de codo mostrará una distribución de palo roto (broken stick distribution). El criterio para seleccionar el número de componentes usando este modelo consiste en comparar los valores propios observados contra los valores propios esperados bajo la distribución de palo

roto y se conservan los valores propios observados que hayan excedido a los esperados. Según estudios (Jackson 1993) que comparan diversos métodos, el modelo de palo roto funcionó igual o mejor que cualquier técnica estadística y además cuenta con la ventaja de ser un cálculo sencillo de realizar. Los valores propios esperados bajo el modelo del palo roto (broken stick model) se calculan cómo:

$$\lambda_i^* = \sum_{k=i}^{p} \frac{1}{k} \tag{2.27}$$

donde k es el número de la i-ésima componente principal y λ_i^* es el valor propio estimado de la i-ésima componente principal bajo el modelo del palo roto(Jackson 1993). Utilizando dicho modelo para 6 variables tenemos lo siguiente:

$$\begin{aligned}
\lambda_1^* &= \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} = 2.45 \\
\lambda_2^* &= \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} = 1.45 \\
\lambda_3^* &= \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} = 0.95 \\
\lambda_4^* &= \frac{1}{4} + \frac{1}{5} + \frac{1}{6} = 0.616667 \\
\lambda_5^* &= \frac{1}{5} + \frac{1}{6} = 0.366667 \\
\lambda_6^* &= \frac{1}{6} = 0.166667
\end{aligned}$$

Claramente se puede observar en la figura 2.5 que los primeros dos valores propios observados son mayores a los esperados por el modelo del palo roto, por lo que este modelo sugiere conservar las primeras dos componentes. Otros autores representan el modelo del palo roto en una gráfica de barras comparativa como se muestra en la figura 2.6, en la cual se puede observar hasta la tercera componente el valor propio esperado por el modelo del palo roto es mayor que el observado, por lo que el modelo sugiere tomar las primeras dos componentes.

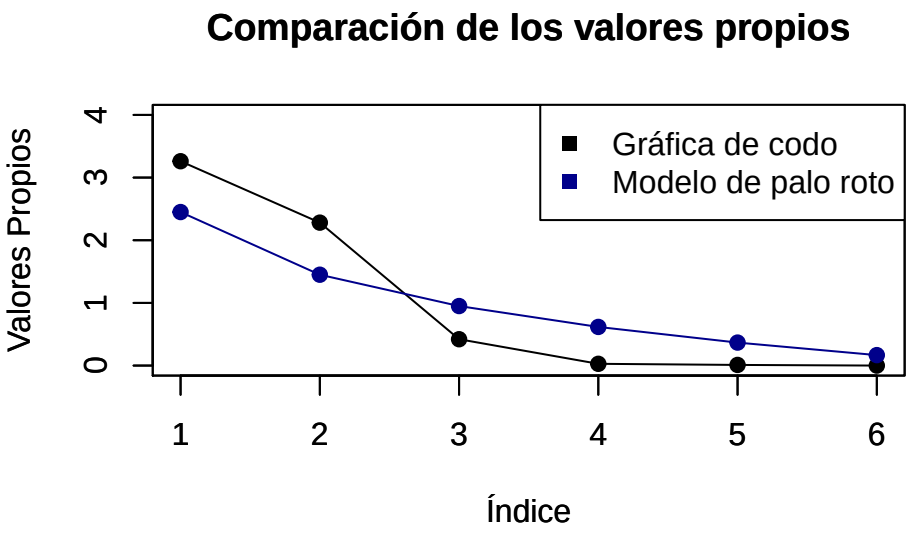

Figura 2.5: Gráfica comparativa para seleccionar las componentes principales utilizando el modelo del palo roto

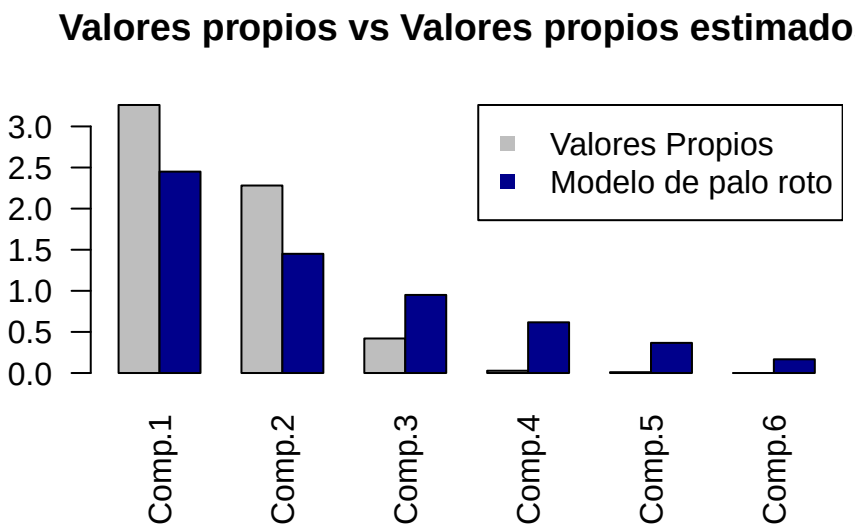

Figura 2.6: Gráfica de barras comparativa

2.8.4. Criterio de la variación porcentual

Anteriormente hemos explicado como encontrar la proporción de varianza que contiene cada componente principal y también la varianza acumulada por los primeros k componentes (expresión 2.22). Estas proporciones se pueden multiplicar por 100 para obtener los porcentajes de variación acumulada de cada componente principal, de esta forma es más sencillo explicar la variabilidad que hay en cada componente. Lo ideal es que la variación porcentual acumulada de las primeras tres componentes sea alta (más del 70 por ciento), esto hablando de un enfoque ecológico, esto significaría que la estructura de los datos fue resumida en pocas dimensiones. En resumen, el criterio de variación porcentual puede ser útil de las siguientes maneras:

- Puede evaluar la importancia de cada componente principal
- Puede determinar cuántas componentes principales retener

Sin embargo, el criterio de variación porcentual ha sido criticado por ser muy difundido y en algunos casos no es muy confiable. Esto se debe a que repercute mucho sobre el número de variables que haya en el conjunto de datos y también los porcentajes pueden ser influenciados por el tamaño de la muestra (Karr y Martin 1981), en ocasiones nos puede llevar a conclusiones erróneas. Por ejemplo, supongamos que solo estudiamos tres variables en un conjunto de datos, entonces la variación porcentual acumulada de las tres componentes obtenidas es del 100 por ciento, sin embargo la segunda y tercera componente pueden carecer de significado. En cambio, suponiendo que tenemos 25 variables en un conjunto de datos, la variación porcentual acumulada de las primeras tres componentes tendría que ser mucho menor (a menos que las últimas 22 variables seas muy redundantes), pero seguramente en este caso las primeras tres componentes puede que sean muy significativas. En cuanto a la influencia del número de observaciones, la variación porcentual disminuye con el aumento de tamaño de la muestra. Por lo tanto, el uso de este criterio para determinar cuántas componentes principales se retendrán es bueno, pero eso no significa que las interpretaciones que se hagan sean las adecuadas, pues si hay muchas variables (más de 20) con muchos individuos (más de 100), la explicación de las componentes significativas puede ser poco clara.

2.8.5. Pruebas de significancia

Existen distintas técnicas para evaluar la importancia estadística de las componentes principales que son obtenidas de una muestra y éstas están basadas en usar pruebas no paramétricas en procedimientos de *remuestreo*. Por otro lado, también

hay pruebas paramétricas que determinan la importancia de los valores propios, sin embargo estas pruebas rara vez son usadas debido a las suposiciones involucradas (Tatsuoka, 1971).
El remuestreo como su nombre lo indica, esta técnica implica "remuestrear" la muestra original, es decir, a partir de los datos observados se generan nuevas muestras simuladas ya sea del mismo tamaño o un subconjunto de la muestra original, eso depende de la técnica a usar. La función del remuestreo es generar una distribución para el estadístico que permita el cálculo de las estimaciones a partir de las muestras obtenidas por el remuestreo. La ventaja principal de estos procedimientos sobre los métodos paramétricos es que en el remuestreo las inferencias estadísticas se basan en las propiedades de distribución de una "pseudo muestra", la cual se genera al remuestrear la muestra original, y no se requiere conocer la distribución de la muestra para poder realizar inferencia estadística sobre la población. Además, no se requieren de cálculos complejos o supuestos que restrinjan a la población muestral, ya que las propiedades estadíscas son fáciles de calcular, las cuales usan los resultados de las pseudomuestras que son calculadas por medio de simulaciones. El gran avance en la informática ha permitido una mayor aplicabilidad pues ha facilidado la realización de simulaciones. El único supuesto requerido es que las observaciones son independientes en la muestra.
Los procedimientos de remuestreo más populares son el jackknife y el bootstrap (Efron 1979, 1982; Efron y Gong 1983). Y es precisamente que mostraremos estos dos procedimientos que pueden utilizarse para probar la importancia de cada componente principal. Solo queda aclarar que usaremos la variación porcentual estimada y observada de cada componente.

Procedimiento Jackknife

La idea del procedimiento de jackknife es determinar el efecto de cada entidad de muestreo en una estadística mediante la eliminación iterativa de entidades de muestreo y el recálculo de la estadística deseada.
En el método jackknife se define de la siguiente manera: Sean $X_1, X_2, \ldots, X_n$ una muestra aleatoria y Φ un estimador del parámetro θ, basado en la muestra de tamaño n. Denotemos Φ_i al estimador Φ evaluado para los $n-1$ elementos que quedan después de quitar el i-ésimo elemento de la muestra:

$$\Phi_i = \Phi(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$$

Y definimos un pseudovalor como:

$$s_i = n\Phi - (n-1)\Phi_i, \qquad i = 1, 2, \ldots, n$$

Finalmente el estimador jackknife de θ asociado al estimador inicial Φ y a la muestra $X_1, X_2, \ldots, X_n$, es el promedio de dichos pseudovalores:

$$\Phi_J = \frac{\sum_{i=1}^{n} s_i}{n} \tag{2.28}$$

Con base en lo anterior, cabe destacar que para obtener estimaciones "jackknife" se requiere un procedimiento de estimación inicial Φ. Sabiendo esto, en el contexto actual, el procedimiento de jackknife comienza con el cálculo de la variación porcentual para cada valor propio de la muestra original (conjunto de datos nxp). En seguida se elimina la primera entidad de los datos (un individuo) y se recalcula Φ_l, utilizando las $(n-1)$ entidades restantes. Luego se elimina la segunda entidad de la muestra y de nuevo se calcula Φ_i utilizando las $(n-1)$ entidades restantes. Este procedimiento debe repetirse n veces, eliminando sucesivamente cada entidad con remplazo hasta que se hayan obtenido n "pseudoestimaciones" Φ_l. Sea $\Phi_{i(\tau)}$ el estimador de la i-ésima componente de la muestra original y sea $\Phi_{i(j)}$ el estimador de la i-ésima componente con la j-ésima entidad eliminada. Se define un pseudovalor $\Phi^*_{i(j)}$ como:

$$\Phi^*_{i(j)} = n\Phi_{i(\tau)} - (n-1)\Phi_{i(j)}$$

donde n es el tamaño de la muestra. El estimador jackknife de Φ^*_i es obtenido como la media de los pseudovalores

$$\Phi^*_i = \frac{\sum_{j=1}^{n} \Phi^*_{i(j)}}{n} \tag{2.29}$$

El error estándar de la estimación jackknife está dada por:

$$SE(\Phi^*_i) = \sqrt{\frac{\sum_{j=1}^{n}(\Phi^*_{i(j)} - \Phi^*_i)^2}{n(n-1)}}$$

Para determinar si el estimador observado difiere significativamente de los esperados (Φ^*_i) sin ninguna estructura "real" (es decir, no aleatoria), se divide la diferencia del estimador bootstrap esperado y el observado ($\Phi^*_i - \Phi_i$) sobre su error estándar estimado ($SE(\Phi^*_i)$). Esta relación se puede tratar como una estadística de distribución t con $(n-1)$ grados de libertad (Mosteller y Tukey 1977).
Las tablas siguientes muestran las pseudoestimaciones y pseudovalores que obtuvimos al utilizar el procedimiento jackknife y que corresponden a las 6 variables de la tabla 2.1 que hemos trabajado.

	1	2	3	4	5	6
$\Phi_{i(1)}$	55.18135	37.56283	6.625657	0.4602278	0.16915301	0.00077841
$\Phi_{i(2)}$	53.43422	38.75597	7.183383	0.4724794	0.15329379	0.00065133
$\Phi_{i(3)}$	54.49858	38.42523	6.513105	0.4215291	0.14087892	0.00068086
$\Phi_{i(4)}$	54.42839	37.94791	6.985438	0.4694354	0.16841472	0.00041706
$\Phi_{i(5)}$	57.77168	33.99803	7.541222	0.5411204	0.14720078	0.00074452
$\Phi_{i(6)}$	56.71753	35.36503	7.23896	0.5082373	0.16947278	0.00076549
$\Phi_{i(7)}$	54.50717	37.93178	6.927635	0.4673729	0.16533391	0.00070667
$\Phi_{i(8)}$	54.46209	37.74266	7.194037	0.4327203	0.16792157	0.00056907
$\Phi_{i(9)}$	53.65213	38.35939	7.333291	0.4821291	0.17234109	0.00071618
$\Phi_{i(10)}$	57.7131	36.88427	4.836863	0.4607928	0.10414158	0.00082813
$\Phi_{i(11)}$	52.68354	38.52125	8.058475	0.5444952	0.19154075	0.00070764
$\Phi_{i(12)}$	53.70357	38.65018	7.005617	0.4710782	0.16882415	0.00072623
$\Phi_{i(13)}$	55.23339	36.72409	7.399804	0.4842247	0.1577228	0.00076946
$\Phi_{i(14)}$	54.27231	38.08157	7.013496	0.4633488	0.16858904	0.0006839
$\Phi_{i(15)}$	49.88246	40.99644	8.745067	0.2326712	0.14258595	0.00077185
$\Phi_{i(16)}$	57.11222	36.30535	5.95711	0.423272	0.20111464	0.00092748
$\Phi_{i(17)}$	51.94115	42.03225	5.525422	0.4392347	0.06104742	0.00089781

Tabla 2.5: Tabla de las pseudoestimaciones

	1	2	3	4	5	6
$\Phi^*_{i(1)}$	40.9850738	45.255324	13.0625462	0.5462135	0.15087274	-0.00003028
$\Phi^*_{i(2)}$	68.9392408	26.165012	4.1389369	0.3501874	0.40462026	0.002003
$\Phi^*_{i(3)}$	51.9095008	31.456936	14.8633809	1.1653936	0.60325818	0.00153052
$\Phi^*_{i(4)}$	53.032534	39.094081	7.306057	0.3988912	0.16268538	0.00575132
$\Phi^*_{i(5)}$	-0.4601686	102.292098	-1.5864809	-0.7480688	0.50210842	0.00051196
$\Phi^*_{i(6)}$	16.4062687	80.420028	3.2497097	-0.2219389	0.14575642	0.00017644
$\Phi^*_{i(7)}$	51.7720525	39.352061	8.2308993	0.4318914	0.21197834	0.00111756
$\Phi^*_{i(8)}$	52.4932815	42.378011	3.9684783	0.9863339	0.17057578	0.00331916
$\Phi^*_{i(9)}$	65.452646	32.510319	1.7404137	0.1957922	0.09986346	0.0009654
$\Phi^*_{i(10)}$	0.4771135	56.112226	41.683257	0.5371741	1.19105562	-0.0008258
$\Phi^*_{i(11)}$	80.9501832	29.920644	-9.8625327	-0.8020651	-0.2073311	0.00110204
$\Phi^*_{i(12)}$	64.6296026	27.857663	6.9831877	0.3726077	0.1561345	0.0008046
$\Phi^*_{i(13)}$	40.1525695	58.675088	0.6762094	0.1622639	0.3337561	0.00011292
$\Phi^*_{i(14)}$	55.5297391	36.95547	6.8571348	0.4962778	0.15989626	0.00148188
$\Phi^*_{i(15)}$	125.7673216	-9.682462	-20.8479985	4.1871187	0.5759457	0.00007468
$\Phi^*_{i(16)}$	10.091193	65.374915	23.7593135	1.1375069	-0.36051334	-0.0024154
$\Phi^*_{i(17)}$	92.8284146	-26.255456	30.6663158	0.8821039	1.88056218	-0.00194068

Tabla 2.6: Tabla de los pseudovalores

Los estimadores jackknife Φ_i^* para $i = 1, 2, \ldots, 6$ y sus respectivos errores estándar son los siguientes:

Φ_1^*	Φ_2^*	Φ_3^*	Φ_4^*	Φ_5^*	Φ_6^*
51.2327392	39.8754093	7.9346369	0.5928049	0.3636015	0.0008082
$SE(\Phi_1^*)$	$SE(\Phi_2^*)$	$SE(\Phi_3^*)$	$SE(\Phi_4^*)$	$SE(\Phi_5^*)$	$SE(\Phi_6^*)$
7.96812557	7.23489825	3.51569044	0.26232862	0.12584802	0.00045447

Tabla 2.7: Tabla de estimadores jackknife y sus errores estándar

Como se dijo anteriormente, la relación $\frac{\Phi_i^* - \Phi_i}{SE(\Phi_i^*)}$ puede tratarse como una estadística para una distribución t con $(n-1)$ grados de libertad, con la finalidad de decidir si el estimador observado difiere significativamente de los esperados. En seguida mostraremos dicha relación con los resultados obtenidos:
Tenemos $n = 17$, por lo que los grados de libertad de las distribución t son $n-1 = 16$. Supongamos que tenemos un nivel de significancia $\alpha = 0.05$, entonces los cuantiles 0.025 y 0.975 de una distribución t_{n-1} son $\omega_{\alpha/2} = \omega_{0.025} = -2.119905$ y $\omega_{1-\alpha/2} = \omega_{0.975} = 2.119905$ respectivamente.
Por otro lado, sea $T_i = \frac{\Phi_i^* - \Phi_i}{SE(\Phi_i^*)}$ entonces las estadísticas y las decisiones son las siguientes:

T_1	T_2	T_3	T_4	T_5	T_6
-0.3907494	0.2570985	0.2646248	0.4861045	1.5536499	0.1702203
No se rechaza	No se rechaza	No se rechaza	No se rechaza	No se rechaza	No se rechaza

Tabla 2.8: Estadísticas T_i y las decisiones de rechazar o no rechazar la hipótesis nula

Estos resultados nos muestran que los estimadores jackknife obtenidos no difieren significativamente a los estimadores calculados de los datos originales, es decir que los porcentajes de varianza de cada componente obtenidos por el procedimiento jackknife no difieren de forma significativa a los que fueron calculados con los datos reales. En la siguiente tabla se muestra la comparación:

Φ_1^*	Φ_2^*	Φ_3^*	Φ_4^*	Φ_5^*	Φ_6^*
51.2327392	39.8754093	7.9346369	0.5928049	0.3636015	0.0008082
Φ_1	Φ_2	Φ_3	Φ_4	Φ_5	Φ_6
54.34627956	38.01532812	7.00429799	0.46528579	0.16807770	0.00073084

Tabla 2.9: Tabla de estimadores jackknife y estimaciones de los datos originales

Finalmente si utilizamos el criterio de la variación porcentual tenemos que las primeras tres estimaciones por el procedimiento jackknife acumulan el 99.04279 % de la varianza total de la muestra. A pesar de que explican casi el 100 % de la varianza de la muestra, se puede observar que podríamos tener problemas con la interpretación de la tercer componente ya que explica el 7.934 % de varianza. Por otro lado, las primeras dos estimaciones acumulan 91.042 % de la varianza total de la muestra, así que lo ideal sería elegir solo las primeras dos componentes, pues acumulan un gran porcentaje de varianza de la muestra y la interpretación debe ser más clara.

Procedimiento Bootstrap

El método bootstrap introducido por Efron (1979) es un procedimiento de re-muestreo de los datos originales que requiere de observaciones independientes. De manera general, se forman muchas muestras, digamos B muestras de dimensión nxp y a cada muestra se le calcula una estadística deseada.

El procedimiento bootstrap se define de la siguiente forma: Supongase que se observa una muestra $X = x_1, x_2, \ldots, x_n$ con n entidades independientes, sobre la que se calcula una estadística Φ, es decir $\Phi(X)$. Se define como una muestra bootstrap $X^* = x_1^*, x_2^*, \ldots, x_n^*$ que se obtiene muestreando n veces con remplazamiento a partir de los datos originales $x_1, x_2, \ldots, x_n$. Por ejemplo si $n = 8$ se podría obtener como una posible muestra bootstrap:

$$X^* = (x_3^*, x_7^*, x_1^*, x_5^*, x_7^*, x_2^*, x_1^*, x_4^*)$$

Como podemos observar, una consecuencia de obtener muestras bootstrap, es que algunas entidades puedan representarse más de una vez o no remprensentarse en absoluto en algunas de las muestras bootstrap. El procedimieno bootstrap comienza con generar un gran número B de muestras bootstrap $x^{*1}, x^{*2}, \ldots, x^{*B}$, cada una de tamaño n. Para cada muestra bootstrap x^{*b} donde $b = 1, 2, \ldots, B$, se calcula la estadística Φ, es decir Φ^{*b}. Al tener muchas muestras bootstrap de la muestra original, se puede calcular una estimación bootstrap del error estándar y un intervalo de confianza.

Hablando en el contexto de nuestro interés, la estadística que usaremos consiste en el cálculo de la variación porcentual para cada valor propio de la muestra, por lo que definimos a Φ_i como la estimación bootstrap donde i hace referencia al estimador de la i-ésima componente. En este caso $\Phi^*_{i(j)}$ representa el cálculo Φ_i para la j-ésima muestra bootstrap. La estimación bootstrap de Φ_i^* está defina como la media de las B estimaciones bootstrap:

$$\Phi_i^* = \frac{\sum_{j=1}^{B} \Phi^*_{i(j)}}{B} \tag{2.30}$$

El estimador bootstrap del error estándar es la desviación estándar de las B muestras bootstrap y está dado por:

$$SE(\Phi_i^*) = \sqrt{\frac{\sum_{j=1}^{B}(\Phi_{i(j)}^* - \Phi_i^*)^2}{B-1}} \tag{2.31}$$

De forma parecida al procedimiento de jackknife, para determinar si el estimador Φ_i observado difiere significativamente del estimador esperado Φ_i^* pero sin ninguna estructura "real" (es decir, no aleatoria), se divide la diferencia del estimador bootstrap esperado y el observado ($\Phi_i^* - \Phi_i$) entre su error estándar estimado ($SE(\Phi_i^*)$). Esta relación se puede tratar como una estadística de ditribución t con n-1 grados de libertad.

Como se puede ver, los procedimientos de jackknife y bootstrap son parecidos, aunque se ha demostrado que el bootstrap es superior al jackknife (Efron y Gong 1983).
En la siguiente tabla se tienen las estimaciones bootstrap de Φ_i^* para $i = 1, 2, \ldots, 6$ con sus respectivos errores estándar, dónde se generaron 10000 muestras bootstrap.

Φ_1^*	Φ_2^*	Φ_3^*	Φ_4^*	Φ_5^*	Φ_6^*
54.3624724	38.0098179	6.9944427	0.4645034	0.1680380	0.0007256
$SE(\Phi_1^*)$	$SE(\Phi_2^*)$	$SE(\Phi_3^*)$	$SE(\Phi_4^*)$	$SE(\Phi_5^*)$	$SE(\Phi_6^*)$
0.14567297	0.14742856	0.10340860	0.00637800	0.00230025	0.00001662

Tabla 2.10: Tabla de estimaciones bootstrap y sus errores estándar

De manera similar a lo visto en el procedimiento jackknife, la relación $T_i = \frac{\Phi_i^* - \Phi_i}{SE(\Phi_i^*)}$ puede tratarse como una estadística para una distribución t con $(n-1)$ grados de libertad. Suponiendo que tenemos un nivel de significancia $\alpha = 0.05$, las estadísticas T_i y las decisiones son las siguientes:

T_1	T_2	T_3	T_4	T_5	T_6
0.11115858	-0.03737566	-0.09530465	-0.12266228	-0.01724595	-0.31528279
No se rechaza	No se rechaza	No se rechaza	No se rechaza	No se rechaza	No se rechaza

Tabla 2.11: Estadísticas T_i y las decisiones de rechazar o no rechazar la hipótesis nula

Estos resultados nos muestran que los porcentajes de varianza de cada componente obtenidos por el procedimiento bootstrap no difieren de forma significativa a los que fueron calculados con los datos originales. En la siguiente tabla se muestra la comparación:

Φ_1^*	Φ_2^*	Φ_3^*	Φ_4^*	Φ_5^*	Φ_6^*
54.3624724	38.0098179	6.9944427	0.4645034	0.1680380	0.0007256
Φ_1	Φ_2	Φ_3	Φ_4	Φ_5	Φ_6
54.34627956	38.01532812	7.00429799	0.46528579	0.16807770	0.00073084

Tabla 2.12: Tabla de estimadores jackknife y estimaciones de los datos originales

Utilizando el criterio de la variación porcentual tenemos que las primeras dos estimaciones por el procedimiento bootstrap acumulan el 92.37229 % de la varianza total de la muestra, y del mismo modo, elegiríamos solo dos componentes ya que la tercera estimación no explica un porcentaje de varianza tan grande como las primeras dos, además su interpretación sería poco clara.

En resumen tanto el procedimiento jackknife y bootstrap se enfocan principalmente por la confiabilidad de los resultados en submuestras extraídas de la población original. Las diferencias entre ambos procedimientos se sintetizan en la manera de extraer las muestras del conjunto de datos original y cómo se otienen las estimaciones de cada procedimiento. Ya hemos explicado como se extraen muestras de los datos originales para los dos procedimientos, y cabe mencionar que se en las distribuciones bootstrap y jackknife se pueden usar para estimar intervalos de confianza y para probar hipótesis nulas sobre el valor de la estadística de prueba en la población.

2.9. Interpretación de las componentes principales

Una vez que ya hemos obtenido las componentes principales y analizado su importancia de cada una para saber cuántas se van a retener, se puede proseguir con la interpretación de las componentes retenidas. Como ya se ha mencionado, los resultados dependen de la decisión de utilizar la matriz de varianzas y covarianzas o la matriz de correlaciones, es decir, de la escala de medida de las variables originales. Cabe recordar que si todas las variables están medidas en la misma escala (siempre que sean de la misma naturaleza) lo ideal es utilizar la matriz de varianzas y covarianzas, por el contrario, si las escalas son distintas, se verán incrementadas con una mayor aportación a la varianza total las de mayor escala, y dichas variables

serán más importantes en las componentes principales, para resolver este problema, se estandarizan las variables originales, lo cual equivale a trabajar con la matriz de correlaciones.

En esta sección, revisaremos algunos de los métodos más comunes para interpretar los componentes principales.

2.9.1. Estructura de las componentes principales

Para interpretar las componentes principales $\mathbf{z}_i$, es necesario relacionarlas con las variables orignales por lo que nos interesa conocer la correlación que hay entre ellas, es decir:

$$r_{jk} = Cor(\mathbf{x}_j, \mathbf{z}_k) = \frac{Cov(\mathbf{x}_j, \mathbf{z}_k)}{\sqrt{V(\mathbf{x}_j)V(\mathbf{z}_k)}} \tag{2.32}$$

Así que necesitamos necesitamos conocer qué es la covarianza entre una variable y una componente. Primero consideraremos las p variables, es decir, a la matriz $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ y a la componente $\mathbf{z}_k$ dónde $k = 1, 2, \ldots, p$, entonces:

$$\begin{aligned} Cov(\mathbf{X}, \mathbf{z}_k) &= Cov(\mathbf{X}, \mathbf{X}\mathbf{a}_k) = Cov(\mathbf{X}, \mathbf{X})\mathbf{a}_k = Var(\mathbf{X})\mathbf{a}_k \\ &= \mathbf{S}\mathbf{a}_k = \lambda_k \mathbf{a}_k = \lambda_k(a_{1k}, a_{2k}, \ldots, a_{pk}) \end{aligned}$$

Por lo que la covarianza entre la j-ésima variable y la k-ésima componente es:

$$Cov(\mathbf{x}_j, \mathbf{z}_k) = \lambda_k \mathbf{a}_{jk}$$

Como $Var(\mathbf{x}_{jj}) = \mathbf{s}_j^2$ y la $Var(\mathbf{z}_k) = \lambda_k$, ya podemos escribir la correlación existente entre la variable $\mathbf{x}_j$ y la componente $\mathbf{z}_k$ sustituyendo en la expresión (2.32) tenemos lo siguiente:

$$r_{jk} = Cor(\mathbf{x}_j, \mathbf{z}_k) = \frac{\lambda_k \mathbf{a}_{jk}}{\sqrt{\mathbf{s}_j^2 \lambda_k}} = \frac{\lambda_k \mathbf{a}_{jk}}{\mathbf{s}_j \sqrt{\lambda_k}} = \frac{\sqrt{\lambda_k}\mathbf{a}_{jk}}{\mathbf{s}_j} \tag{2.33}$$

De este modo, si las variables originales de la matrix $\mathbf{X}$ están estandarizadas, la correlación entre las variables $\mathbf{x}_j$ y la componente $\mathbf{z}_k$ es la siguiente:

$$r_{jk} = Cor(\mathbf{x}_j, \mathbf{z}_k) = \frac{\lambda_k \mathbf{a}_{jk}}{\sqrt{\mathbf{s}_j^2 \lambda_k}} = \frac{\lambda_k \mathbf{a}_{jk}}{\sqrt{\lambda_k}} = \sqrt{\lambda_k}\mathbf{a}_{jk} \tag{2.34}$$

Donde r_{jk} es la correlación entre la j-ésima variable y la componente k, a_{jk} es el peso de la j-ésima variable en la componente principal k, es decir, el vector propio y λ_k es el valor propio asociado a la componente principal k.

Estas correlaciones se conocen como cargas de componentes principales, y la matriz de cargas se denomina *estructura de componentes principales* (o matriz de cargas de factores). Entonces, las cargas de las componentes principales nos dicen qué tan cerca están relacionadas una variable y una componente. Cuando la carga es grande (es decir, se aproxima a 1), significa que la componente lleva casi la misma información que la variable. En contraste, cuando la carga es cercana a cero, entonces la información que hay de la variable en la componente es poca. Esto es muy importante, porque nos puede ayudar a definir la interpretación de las componentes principales en base al análisis de las cargas, prestando atención a las variables con las cargas más grandes.

La estructura de la componente principal proporciona una manera de establecer una interpretación ecológica de cada componente principal y quizá sea lo más importante del análisis de las componentes principales en la ecología.

2.9.2. Importancia de las cargas e interpretación

Hay algunas reglas que se pueden utilizar para determinar que cargas de las componentes principales seleccionadas son importantes para la interpretación (Hair, Anderson y Tatham 1987). En seguida, mostraremos las reglas más usadas:

1. Se consieran cargas significativas cuando son mayores a 0.30 y menores a -0.30. Las cargas mayores que 0.40 y menores que -0.40 se consideran más importantes, y cuando las cargas son mayores de 0.50 y menores de -0.50 se consideran muy significativas. Esta regla se considera útil cuando el tamaño de la muestra es mayor e igual a 100.

2. Una regla similar discutida por Tabachnik y Fidell (1989) sugiere que las cargas mayores que 0.45 o menores que -0.45 son justas, mayores que 0.55 o menores que -0.55 son buenas, mayores de 0.63 o menores de -0.63 son muy buenas y más de 0.71 o menos de -0.71 son excelentes. Si las cargas están entre -0.45 y 0.45 se consideran cargas deficientes.

3. Un método empírico, sugiere que si el tamaño de la muestra es de 100 (Hair, Anderson y Tatham 1987) las cargas mayores o iguales que ± 0.19 y ± 0.26 con niveles de significancia del 5 % y 1 % respectivamente. Cuando el tamaño de la muestra es 200 se recomiendan ± 0.14 y ± 0.18 con niveles de significancia del 5 % y 1 % respectivamente, y cuando el tamaño de la muestra es mayor o igual a 300, se recomiendan cargas de ± 0.11 y ± 0.15 con los niveles de significancia ya mencionados.

Una gran desventaja de utilizar estas reglas es que no analizan con precisión el número de entidades de muestreo ni la cantidad de variables en el estudio, por lo que la interpretación queda a criterio del que esté analizando la muestra de datos. Quizá, estas reglas podrían marcar una pauta para determinar qué variables tienen más peso en cada componente principal. Se pueden establecer comportamientos generales (Hair, Anderson y Tatham 1987):

- Cuanto mayor sea el número de entidades (n individuos), más pequeña será la carga para que se considere significativa;
- Cuanto mayor sea el número de variables (p variables) en estudio, menor será la carga que se se considere significativa;
- Cuanto mayor sea el número de componentes ($\mathbf{z}_k$), más grande será la carga en factores posteriores para que se consideren significativos en la interpretación.

Hemos mencionado que a menudo, la interpretación es una tarea que resulta ser complicada. Sin embargo, hay algunos procedimientos que nos pueden ayudar, en los cuales haremos uso de la estructura de las componentes principales (en esencia, son las cargas). Primero se comienza con la primera variable, nos fijamos en la carga que hay entre la primera variable con la primera componente, luego con la segunda componente, y así sucesivamente hasta observar la carga con la componente u (es la última componente que seleccionamos para el estudio) y se toma la carga absoluta más alta que se observó. Una vez tomada la mayor carga absoluta entre la primera variable y las componentes, nos fijamos si esa carga es significativa o no lo es. Se repite este proceso para cada variable, hasta haberlo hecho para las p variables. Como dijimos antes, las cargas son equivalentes a la correlación que hay entre la j-ésima variable y la k-ésima componente. Supongamos que seleccionamos u componentes para el estudio donde claramente $u < p$ y $1 < k < u$, entonces el proceso que acabamos de mencionar, se puede describir de la siguiente manera:

	$CP1$	$CP2$	$\ldots$	CPu	Máximos
$variable_1$	r_{11}	r_{12}	$\ldots$	r_{1u}	$max(\lvert r_{11}\rvert, \lvert r_{12}\rvert, \ldots, \lvert r_{1u}\rvert)$
$variable_2$	r_{21}	r_{22}	$\ldots$	r_{2u}	$max(\lvert r_{21}\rvert, \lvert r_{22}\rvert, \ldots, \lvert r_{2u}\rvert)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
$variable_p$	r_{p1}	r_{p2}	$\ldots$	r_{pu}	$max(\lvert r_{p1}\rvert, \lvert r_{p2}\rvert, \ldots, \lvert r_{pu}\rvert)$

Y después se analiza si la carga absoluta máxima de cada variable es significativa o no. Lo ideal, es que cada variable tenga sólo una carga en una componente que se

considere significativa, y en ese caso se tendría una "solución de estructura simple", de esta manera la interpretación se simplificaría considerablemente. Sin embargo, en la práctica, es poco común que solo haya una carga significativa, normalmente hay varias en cada variable que pueden ser significativas, y es por esa razón que tomamos la carga absoluta máxima. La idea es minimizar el número de cargas significativas en cada fila (es decir, las cargas asociadas con cada variable) y maximizar el número de cargas con valores no significativos. Una vez que se ha identificado qué cargas resultaron ser significativas de cada fila, se debe examinar de la estructura de componentes principales qué variables no tienen una carga significativa con alguna componente, en tal caso tenemos dos opciones:

- Interpretar la solución con las variables a las que se les pudo asociar una carga significativa e ignorar las variables que no tuvieron una carga significativa con alguna componente; o
- Evaluar críticamente cada una de las variables que no tiene carga significativa en ninguna componente. Si la(s) variable(s) no son de gran importancia en el objetivo del estudio, se puede decidir eliminar la(s) variable(s) y repetir todo el análisis para la obtención de una nueva solución de componentes.

Cuando se haya obtenido una solución final, se puede intentar asignar algún significado al patrón de las cargas. Las variables con cargas más altas se consideran más importantes en la interpretación de cada componente. Lo ideal es etiquetar dichas variables para darles cierto énfasis.

Finalmente, es importante tener en cuenta el signo (+ o -) de cada carga significativa. Ya dijimos anteriormente que las cargas son correlaciones entre variable-componente, así que las correlaciones (o cargas) positivas indican una relación directa entre una variable y una componente. Por el contrario, las correlaciones negativas indican una relación inversa entre una variable y una componente. En conclusión, la estructura de las componentes principales tienen cargas tanto positivas como negativas; las cargas positivas más grandes y las cargas negativas más grandes que se asocian a cada variable son las que jugarán el papel más importante para la interpretación. Cabe señalar, que los programas informáticos no derivan alguna conclusión, sino que el investigador lo desarrolla de manera intuitiva.

Hemos trabajado con los datos de la tabla 2.1 sin considerar las variables "**2n**" y "**KF**" (ya se ha explicado el motivo). De esos datos, calculamos sus componentes principales y se seleccionaron las primeras dos para el estudio del problema. En la siguiente tabla 2.13 se muestra la estructura de las componentes principales que fueron seleccionadas:

	CP_1	CP_2
THC	0.7263275	0.6551600
AC	0.7277712	0.6537082
Range	0.9600907	0.1817856
Ratio	0.8387488	0-.3179565
TF	-0.4963731	0.8237657
CI	-0.57612	0.7820604

Tabla 2.13: Tabla de las cargas que hay entre las variables y las primeras dos componentes

Una vez ya obtenida la estructura de las componentes principales en estudio, se debe identificar la carga absoluta máxima entre una variable y una componentes, en la siguiente tabla 2.14, se observan las cargas absolutas máximas resaltadas en amarrillo:

	CP_1	CP_2
THC	0.7263275	0.6551600
AC	0.7277712	0.6537082
Range	0.9600907	0.1817856
Ratio	0.8387488	-0.3179565
TF	-0.4963731	0.8237657
CI	-0.57612	0.7820604

Tabla 2.14: Tabla de las cargas absolutas máximas

Como la muestra es de 17 entidades (es una muestra pequeña) se hace uso de la segunda regla que presentamos al comienzo de la sección, la cual indica que si las cargas son mayores que 0.71 o menores que -0.71 entonces dichas cargas son excelentes para el estudio. Note que todas las cargas que resaltamos son excelentes.

Así que la primera componente explica principalmente la variabilidad de las variables **THC**, **AC**, **Range** y **Ratio**. Este resultado de datos cromosómicos de dos subgéneros de leguminosas sugiere que la primera componente refleja a las especies que tienen una mayor longitud cromosómica y también que tengan una mayor diferencia entre la longitud del cromosoma más grande y el más pequeño. Con esto, podemos observar que las especies: *A. paniculata*, *A. lyonnetii* y *A. amorphoides*

tienen una mayor longitud cromosómica total (**THC**) y una mayor talla cromosómica promedio (**AC**) en comparación a las demás especies, además también hay una mayor diferencia entre el cromosoma más grande y el más pequeño (**Range**), y el cociente del cromosoma más largo entre el más pequeño (**Ratio**) es mayor. Sólo esas especies cumplen esas condiciones, por otro lado, hay especies que pueden tener una longitud cromosómica total bastante grande en comparación de las demás pero no hay mucha diferencia entre su cromosoma más grande y el más pequeño como en el caso de la especie *A. deamii*.

La segunda componente explica principalmente la variabilidad de las variables **TF** y **CI**.

2.9.3. Biplot y círculos de correlación

Hemos visto que las cargas son los coeficientes de correlación que hay entre variables y componentes, anteriormente nos enfocamos en analizar los pesos de cada variable en cada componente. En ese sentido, dichas cargas se pueden representar en un plano, el cual se denomina **círculo de correlación**. Este gráfico está formado por vectores que representan a cada variable por medio de dos coordenadas y su inicio es en el origen, es decir, la coordenada (0,0). La primera entrada correspone al coeficiente de correlación que hay entre una variable y la primera componente, y la segunda entrada es la correlación que hay entre una variable y la segunda componente, así que todas las coordenadas estarán contenidas dentro del círculo unitario. En este caso sólo estamos considerando las primeras dos componentes, sin embargo, también podríamos utilizar tres componentes, entonces las coordenadas estarían en $\mathbf{R}^3$ y dentro de una esfera de radio 1.

Por otro lado, también nos interesa calcular los valores de las componentes principales, también conocidos como "**scores**", las cuales son calculadas con la multiplicación de variables centradas o estandarizadas por los vectores propios, dependiendo el caso. Si trabajamos con la matriz $\mathbf{S}$ entonces debemos usar las variables centradas y si decidimos trabajar con la matriz $\mathbf{R}$ entonces las variables deben estar estandarizadas para calcular sus scores. En seguida, se muestra el cálculo de los "scores" si se decide trabajar con la matriz $\mathbf{S}$:

La j-ésima componente principal está dada por:

$$\mathbf{z}_j = \widetilde{\mathbf{x}}_1\mathbf{a}_{1j} + \widetilde{\mathbf{x}}_2\mathbf{a}_{2j} + ... + \widetilde{\mathbf{x}}_p\mathbf{a}_{pj} = \widetilde{\mathbf{X}}\mathbf{a}_j$$

donde $\widetilde{\mathbf{X}}$ es la matriz centrada de los datos originales y $\mathbf{a}_j$ es el j-ésimo vector propio.

Explícitamente, la observación i-ésima de la j-ésima variable se expresa como función lineal de las p variables originales de la i-ésima observación:

$$\mathbf{z}_{ij} = \mathbf{a}_{i1}(x_{i1} - \overline{x}_1) + \mathbf{a}_{i2}(x_{i2} - \overline{x}_2) + \ldots + \mathbf{a}_{ip}(x_{ip} - \overline{x}_p) \tag{2.35}$$

donde $\overline{x}_j$ es la media muestral de la j-ésima variable Claramente, si decide utilizar la matriz $\mathbf{R}$ entonces los datos de cada variable deben estar estandarizados, es decir, $\frac{x_{ij}-\overline{x}}{\mathbf{s}_j}$ donde $\mathbf{s}$ es la varianza muestral de la j-ésima variable.

Como ya vimos, la j-ésima componente principal se expresa como:

$$\mathbf{z}_j = \mathbf{X}\mathbf{a}_j$$

Si se calculan los "scores" para el conjunto de las n observaciones muestrales, la ecuación puede expresarse de forma matricial:

$$\mathbf{z}_j = \mathbf{X}\mathbf{a}_j = \begin{pmatrix} z_{1j} \\ z_{2j} \\ \vdots \\ z_{nj} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{pj} \end{pmatrix} \tag{2.36}$$

El biplot es una representación gráfica simultánea de los individuos (mediante puntos) y las variables (mediante vectores), en un plano cartesiano construido en base a las dos primeras componentes principales. La primera componente es el eje horizontal y la segunda es el eje vertical, se puede intentar dar una interpretación al observar las direcciones de los vectores graficados. Este gráfico fue introducido por Gabriel (1971).

Los puntos que estén próximos en el biplot, se puede decir que los individuos correspondientes son parecidos. También se dirá que hay una alta correlación entre variables si la dirección de sus respectivos vectores graficados son parecidas.

Por ejemplo, si graficamos el círculo de correlación de los datos que hemos trabajado en este capítulo, se puede verificar que coincide con las cargas que mostramos en la tabla 2.13:

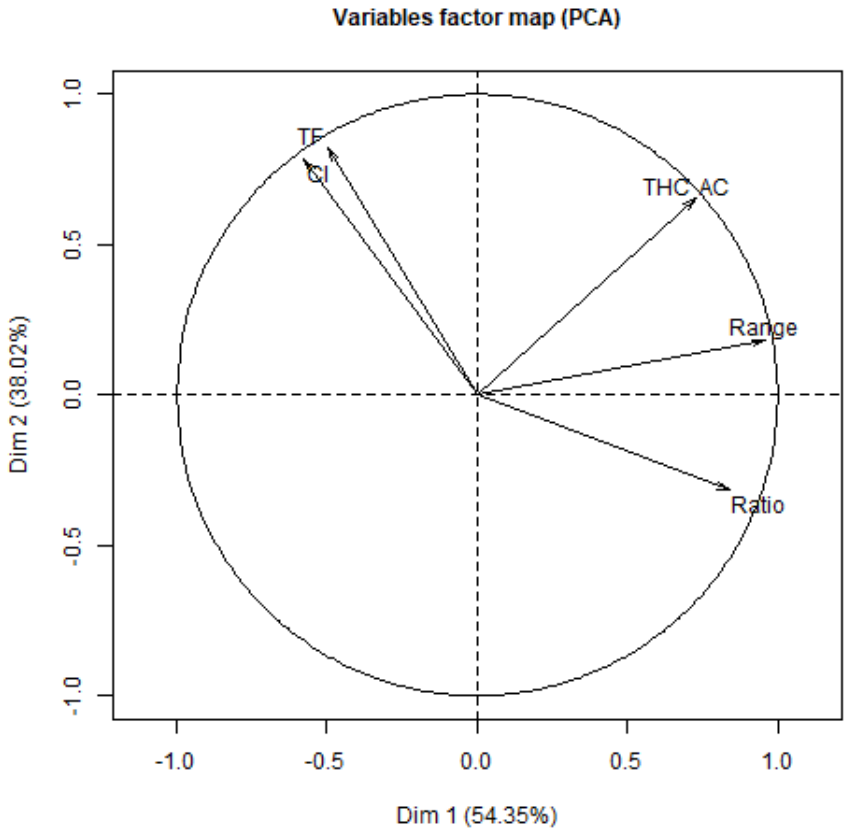

Figura 2.7: Círculo de correlación

Se puede observar que las variables **THC** y **AC** tienen una correlación muy alta, y que las variables **TF** y **CI** también.
Por otro lado si deseamos calcular el valor de las componentes (los scores), lo ponemos hacer al multiplicar la matriz de datos estandarizada para cada variable por los valores propios asociados a la matriz **R**, esto es porque se decidió darle el mismo peso a cada variable, es decir, se eligió trabajar con la matriz de correlaciones. En la siguiente tabla, se muestra la matriz de datos ya estandarizada:

	THC	AC	Range	Ratio	TF	CI
A. americana var. americana	-1.09131082	-1.0832119	-0.443143054	0.2968535	-0.09858429	-0.184074
A. americana var. flabellata	-0.74964225	-0.7321536	-0.71113075	-0.6662208	0.36112086	0.350361
A. americana var. glandulosa	-0.1301684	-0.1257802	-0.175155357	0.2583305	0.79910766	0.6035144
A. sp. aff. americana	0.08696677	0.0657061	-0.085826125	-0.2039452	0.73757233	0.6070304
A. villosa var. villosa	-0.67300631	-0.6683249	-0.085826125	0.9517441	-2.11839119	-2.2268812
A. villosa var. longifolia	-0.82627819	-0.8278968	-0.294261	0.3738995	-1.46322085	-1.2705239
A. sp. aff. villosa	-0.11739574	-0.1257802	0.241714393	0.5665143	-0.22527469	-0.2157181
A. sensitiva I	0.12209157	0.1295349	-0.324037411	-0.7432668	0.65069891	0.9516003
A. sensitiva II	-0.19403168	-0.189609	-0.71113075	-1.0899736	0.80634711	0.902376
A. deamii	1.45364103	1.4380248	0.003503107	-1.0129276	0.11859925	0.3784891
A. scabra	-0.17806586	-0.189609	-1.038671268	-1.5522492	1.63526428	1.805993
A. evenia	-0.67619947	-0.6683249	-0.562248696	-0.3580371	0.36836031	0.3855212
A. rudis	-1.55751279	-1.5619277	-1.217329732	-0.6662208	-0.50037384	-0.4477754
A. ciliata	-0.17806586	-0.189609	-0.324037411	-0.3580371	0.03896528	0.1499479
A. paniculata	0.96189375	0.959309	2.415392375	2.3385711	-1.61524933	-1.3267803
A. lyonnetii	1.78573011	1.7890831	1.968746214	1.6066346	0.99819257	0.5191299
A. amorphoides	1.96135414	1.9805694	1.343441589	0.2583305	-0.49313438	-0.9822104

Tabla 2.15: Matriz de datos ya estandarizada por cada variable

Anteriormente, con distintos criterios se concluyó que nuestros datos solo requieren de ser analizados con dos componentes. Así que sólo nos interesan los scores de las 17 entidades, correspondientes a las primeras dos componentes principales, así que solo utilizaremos los primeros dos vectores propios de la matriz $\mathbf{R}$:

	$\mathbf{a}_1$	$\mathbf{a}_2$
THC	0.402	0.434
AC	0.403	0.433
Range	0.532	0.120
Ratio	0.464	-0.211
TF	-0.275	0.545
CI	-0.319	0.518

Tabla 2.16: Tabla de los primeros dos vectores propios de la matriz $\mathbf{R}$

Los scores se calculan utilizando la expresión (2.35), si lo hacemos para las primeras dos entidades con la primera componente tendremos lo siguiente:

$$z_{11} = 0.402(-1.09131082) + 0.403(-1.0832119) + 0.532(-0.443143054) + 0.464(0.2968535) + (-0.275)(-0.09858429) + (-0.319)(-0.184074) = -0.887$$

$$z_{21} = 0.402(-0.74964225) + 0.403(-0.7321536) + 0.532(-0.71113075) + 0.464(-0.6662208) + (-0.275)(0.36112086) + (-0.319)(0.350361) = -1.494935491$$

Y con la segunda componente se tiene:

$$z_{12} = 0.434(-1.09131082) + 0.433(-1.0832119) + 0.120(-0.443143054) + (-0.211)(0.2968535) + 0.545(-0.09858429) + 0.518(-0.184074) = -1.207$$
$$z_{22} = 0.434(-0.74964225) + 0.433(-0.7321536) + 0.120(-0.71113075) + (-0.211)(-0.6662208) + 0.545(0.36112086) + 0.518(0.350361) = -0.209$$

Lo conveniente es hacerlo de forma matricial. En la siguiente tabla, se muestran los scores obtenidos para las 17 entidades y las primeras dos componentes principales:

	Comp.1	Comp.2
A. americana var. americana	-0.887	-1.207
A. americana var. flabellata	-1.495	-0.209
A. americana var. glandulosa	-0.488	0.562
A. sp. aff. americana	-0.475	0.815
A. villosa var. villosa	1.149	-3.101
A. villosa var. longifolia	0.159	-2.287
A. sp. aff. villosa	0.424	-0.430
A. sensitiva I	-0.899	1.074
A. sensitiva II	-1.548	0.885
A. deamii	0.542	1.727
A. scabra	-2.447	1.870
A. evenia	-1.231	-0.174
A. rudis	-1.932	-1.863
A. ciliata	-0.545	-0.024
A. paniculata	4.011	-0.937
A. lyonnetii	2.792	2.261
A. amorphoides	2.870	1.038

Tabla 2.17: Scores de las 17 entidades que se observan en la matriz de datos

En la figura 2.8 se puede ver la representación de los individuos en las primeras dos componentes que fueron observadas en la matriz de datos original. Finalmente se puede hacer un biplot para representar de forma simultánea las entidades en las primeras dos componentes principales y las variables como se muestra en la figura 2.9.

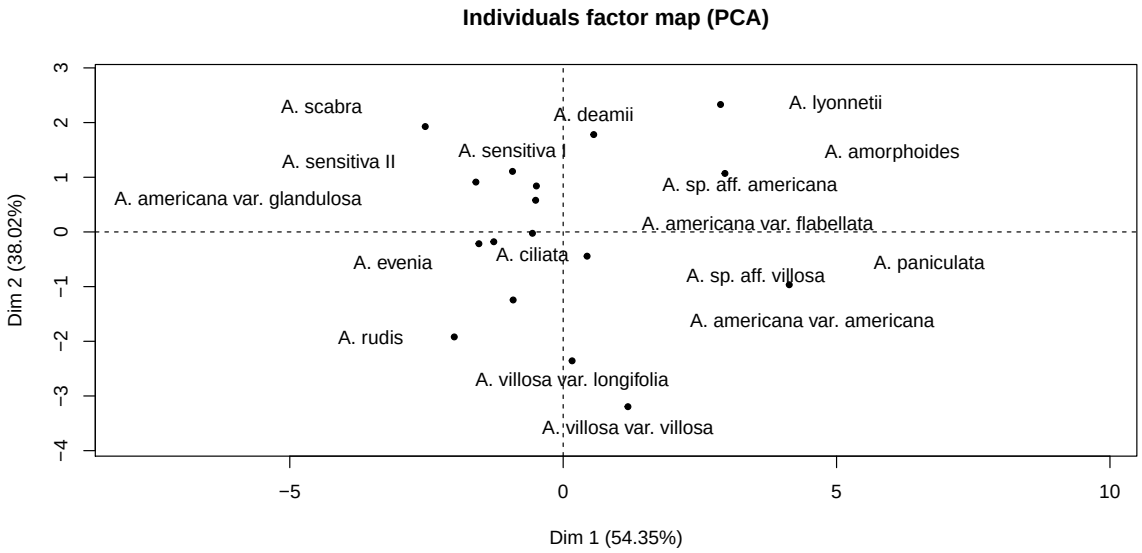

Figura 2.8: Gráfica de los Scores de las 17 entidades

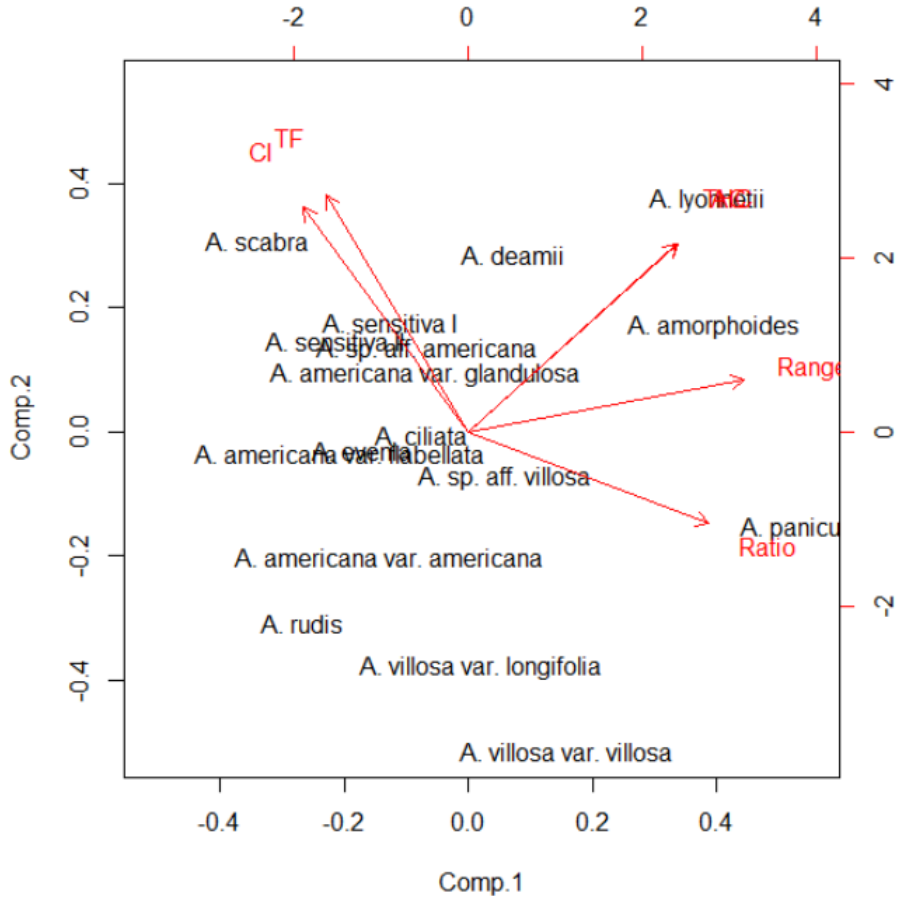

Figura 2.9: Gráfica Biplot

2.9.4. Limitaciones de la interpretación de Componentes Principales

Ya se han mencionado algunos problemas que se pueden encontrar para la interpretación de las componentes principales elegidas. Sin embargo, deseamos hacer énfasis en las limitaciones que puede haber en la interpretación, a pesar de que la variación porcentual acumulada de las primeras tres componentes sea alta:

- Al interpretar solo las primeras componentes principales, uno puede pasar por alto una componente posterior que explique la mayor parte de la variación de alguna variable; esto implica que la información de cierta variable se pierda. Por lo tanto las componentes principales a interpretar, habrán perdido mucha información de dicha variable. En ocasiones, este comportamiento se puede ver reflejado en las cargas que hay entre componentes y variables.

- En ocasiones se cuestiona la selección de las combinaciones lineales de variables (es decir, de las componentes principales) ya que la justificación es maximizar la varianza dentro de una muestra original de datos, es decir, se hace un "mejor" resumen (Johnson 1981). Sin embargo, uno puede aumentar el porcentaje de la variación explicada de las primeras componentes si se agregan variables redundantes al conjunto de datos. Así que a medida de que se incluyan más variables redundantes, el análisis de componentes principales mejorará, pero en realidad este hecho puede que complique la interpretación ya que aumentará el número de variables significativas en las componentes. Por lo tanto, en algunos casos, un gran porcentaje de la variación explicada puede reflejar solo ignorancia en la selección de variables.

2.9.5. Valores atípicos de la muestra

La mayoría de los procedimientos estadísticos asumen independencia entre las muestras, y el análisis de componentes principales no es una excepción. En el trabajo de campo, un ecólogo puede tratar de recolectar una muestra aleatoria de una serie de entidades en un área determinada, pero esto no garantiza que lo sea en el análisis. Sin embargo, en este trabajo, hemos utilizado el ACP con fines exploratorios o descriptivos, así que no nos interesa si se cumple o no dicho supuesto. Además es común que no se puedan recolectar muestras indepndientes en los estudios que se hacen de la vida silvestre, generalmente por buenas razones.

Es probable que una muestra aleatoria incluya valores atípicos (es decir, valores que se distinguen claramente de todos los demás) como resultado de condiciones ambientales, comportamientos históricos, respuestas de los organismos, etc. Aunque en teoría los valores atípicos deben ser eliminados, eso no es del todo cierto, ya que todo depende de lo que se esté estudiando. Hablando de datos ecológicos, es necesario consultar con un especialista, que observaciones pueden ser extremas y cuales en verdad son valores atípicos ecológicos. Cuando históricamente no se han presentado ciertos comportamientos de algunas entidades y con el conocimiento que hay en la actualidad, si la observación es irracional, entonces es un verdadero dato atípico y puede se eliminado. Por otro lado, las observaciones extrema se pueden desviar considerablemente de la media del grupo, pero aún pueden representar condiciones ecológicas significativas. Por lo tanto, se debe tener precaución al eliminar observaciones sospechosas, ya que la eliminación negligente de los datos atípicos puede resultar una pérdida de información significativa.

Para identificar los datos atípicos (no necesariemente se eliminan por lo mencionado con anterioridad) se puede realizar cualquiera de los siguientes diagnóstcos (el que vea más conveniente):

1. Desviaciones estándar. En ocasiones, la forma más sencilla de detectar los valores atípicos es estandarizar los datos originales, reste la media muestral y divida sobre la desviación estándar muestral para cada variable, es decir:

$$\frac{x - \bar{x}}{s}$$

   Luego inspeccione los datos datos de las entidades que tengan cualquier valor mayor que o menor que, por ejemplo, 2.5 desviaciones estándar de la media en cualquier variable.

2. Diagramas de caja de variables centradas. Realice un diagrama de caja para cada variable ya centrada para visualizar los valores atípicos sospechosos.

3. Gráficos de puntajes de componentes principales. Primero construya los diagramas de caja de los puntajes obtenidos para cada componente principal y verifique si hay valores atípicos. Adicionalmente, construya diagramas de dispersión para cada par de componentes principales y verifique si hay datos atípicos sospechosos.

   Ahora, si se utilizan los datos que hemos estado trabajando en este capítulo y se hace con el uso del diagnóstico número 2 anterior, tendremos los siguientes Diagramas de Caja para cada variable:

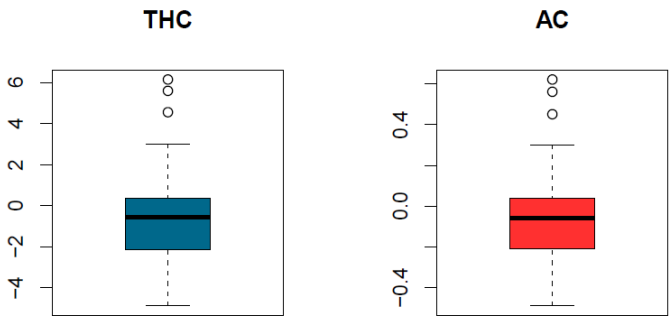

Figura 2.10: Diagrama de Caja de las variables THC y AC

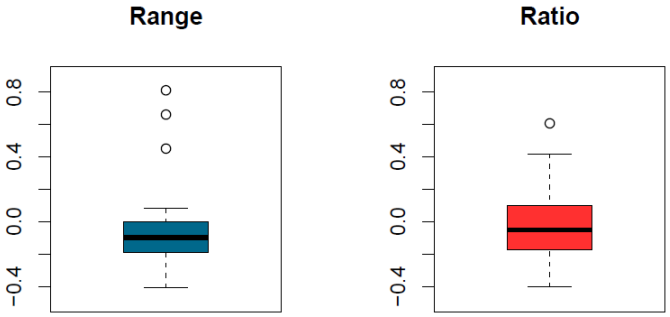

Figura 2.11: Diagrama de Caja de las variables Range y Ratio

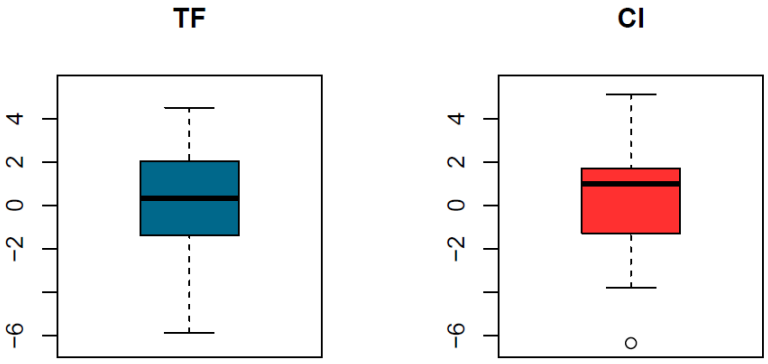

Figura 2.12: Diagrama de Caja de las variables TF y CI

Claramente, para las variables THC, AC y Range tienen 3 valores, sin embargo, sus correspondientes observaciones son valores extremos que requerimos en el estudio, ya que nuestros datos son de medidas cromosómicas y además nuestro objetivo es saber si las las componentes principales apoyan la idea de clasificar ciertas especies de leguminosas en dos categorías que se han hecho con el estudio macro (es decir, a base de la observación de la vista) o no la apoyan.

Si se construye un Diagrama de Dispersión como lo sugiere el diagnóstico número 3 se tiene:

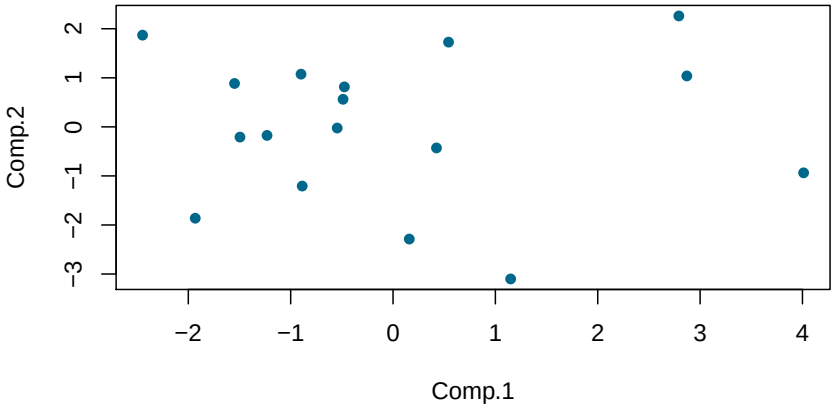

Figura 2.13: Diagrama de Dispersión de los puntajes de las primeras dos componentes de los datos originales trabajados en este capítulo

Se puede observar que 3 observaciones se separan del resto. Consideremos ahora el biplot de la figura 2.9, examinándolo se concluye que esta separación se debe a que las 3 especies de leguminosas tienen una mayor longitud cromosómica total, un rango, un radio y una talla cromosómica promedio mayor a las otras leguminosas. Deseo subrayar que estas especies tienen las mayores diferencias entre el cromosoma más grande y el más pequeño.

Capítulo 3

Estudio de información climática

En este tercer y último capítulo, se mostrará un Análisis de Componentes Principales con una base de datos que proviene de un proyecto de investigación marina en Kotzbehue, Alaska. La muestra contiene datos anuales de 23 años, desde 1981 hasta 2003 de 17 variables climáticas. El objetivo principal de este estudio exploratorio consiste en averiguar si existe algún patrón que pueda mostrar un cambio abrupto e identificar las variables que lo expliquen. Se aplicarán todos los criterios mencionados en el capítulo anterior con sus conclusiones correspondientes.

3.1. Planteamiento del problema climático y cálculo de las componentes

Diversos estudios demuestran que la Tierra se ha calentado y enfriado de forma natural, sin embargo, ese proceso fue muy lento, tardando millones de años en efectuarse, aunque en la actualidad, las actividades que practica la humanidad ha traído cambios muy acelerados, alcanzando niveles que podrían traer muchas más consecuencias de las que ya hay. La extinción de algunas especies, y el peligro de sufrir lo mismo para otras, son algunas consecuencias que ha provocado el ser humano.

Algunos estudios, indican que la causa principal del cambio climático se debe al calentamiento global. El Centro Internacional para la Investigación del Fenómeno de El Niño, denomina al efecto invernadero como un fenómeno por el cual determinados gases, retienen parte de la energía que el suelo emite por haber sido calentado por la radiación solar,dichos gases son componentes de la atmósfera planetaria. Este fenómeno afecta a todos los cuerpos planetarios dotados de atmósfera. Un problema en la

actualidad, es que muchas actividades que el humano realiza diariamente, aumenta la emisión de gases a la atmósfera de nuestro planeta y ésta retiene más calor del necesario, entonces provoca un aumento de temperatura a la Tierra, y se produce lo que se conoce como calentamiento global.

En este capítulo, nuestro objetivo será analizar una base de datos que contiene información de los cambios en el clima en Kotzbehue, Alaska, donde los registros fueron realizados cada año desde 1981 hasta el año 2003. Esta información se muestra en la tabla 3.2.
Los conjuntos de datos del clima, son importantes para los proyectos de investigación ecológica, ya que algunos indicadores ambientales pueden ayudar a explicar patrones biológicos o encontrar relaciones entre el clima y la biodiversidad.

La descripción del conjunto de datos es la siguiente:

Descripción de variables
AO: Oscilación Ártica (índice climático)
AO_wi: Oscilación Ártica durante el invierno
AO_su: Oscilación Ártica durante el verano
NPI: Índice del Pacífico Norte
NPI_sp: Índice del Pacífico Norte durante la primavera
NPI_wi: Índice del Pacífico Norte durante el invierno
Temp: Temperatura
Temp_su: Temperatura promedio durante el verano
Temp_wi: Temperatura promedio durante el invierno
Rain: Precipitaciones
Rain_su: Precipitaciones durante el verano
Rain_wi: Precipitaciones durante el invierno
Ice: Hielo
Ice_JanJul: Hielo entre los meses de enero a julio
Ice_OctDec: Hielo entre los meses de octubre a diciembre
IceCover: Cubierta de Hielo
IceFreeDays: Días libres de hielo

Tabla 3.1: Breve descripción del conjunto de variables de la tabla 3.2

	AO	AO_ wi	AO_ su	NPI	NPI_ sp	NPI_ wi	Temp	Temp_ su	Temp_ wi	Rain	Rain_ su	Rain_ wi	Ice	Ice_ JanJul	Ice_ OctDec	IceCover	IceFreeDays
1981	-0.4346	-0.1683	-0.2410	-2.09	-0.15	-4.46	-3.9	8.8	-16.66	23.62	4.25	1.46	12.309	14.647	11.367	-0.64	140
1982	0.2977	-0.3750	0.3083	0.75	0.13	1.70	-4.7	10.5	-17.17	27.03	3.71	1.92	12.673	14.983	11.907	-1.65	144
1983	0.0319	0.1733	0.4653	-2.54	0.30	-5.44	-4.4	8.2	-17.75	28.75	5.65	0.68	12.493	14.735	11.573	-0.34	116
1984	-0.1917	0.2627	0.0240	-1.20	-0.23	-2.62	-7.0	8.7	-20.03	26.04	5.75	0.74	12.089	14.428	11.103	0.15	134
1985	-0.5192	-1.2667	0.2678	0.52	-0.43	1.11	-5.9	9.6	-15.15	27.28	3.62	1.24	12.208	14.758	11.153	-0.21	120
1986	0.0848	-1.8067	-1.8067	-1.84	-0.38	-4.11	-5.7	9.0	-14.65	25.32	6.01	0.91	12.404	14.627	11.690	-0.32	154
1987	-0.5442	-0.8537	-0.8537	-1.25	-0.30	-2.93	-5.2	9.6	-16.41	20.45	3.22	1.35	12.266	14.828	10.405	-1.43	152
1988	0.0402	-0.4450	-0.4450	-0.24	0.35	-0.77	-5.1	9.5	-16.63	22.45	3.60	1.70	12.094	14.532	11.647	0.19	142
1989	0.9500	2.6880	2.6880	1.35	-0.18	2.84	-4.8	10.7	-17.47	34.75	6.25	1.91	12.147	14.322	11.497	-1.16	159
1990	1.0241	1.2530	1.2530	0.91	0.33	2.00	-6.4	11.0	-22.81	37.49	5.27	1.40	11.888	14.342	11.310	0.16	149
1991	0.1970	0.3747	0.3747	0.99	0.20	2.34	-5.2	10.4	-20.46	22.00	2.66	2.31	11.914	14.315	11.150	0.53	154
1992	0.4366	1.0950	1.0950	-1.10	1.33	-2.45	-7.3	8.0	-20.02	24.94	3.35	1.42	12.226	14.295	11.640	0.65	137
1993	0.0792	1.7687	1.7687	-0.16	1.53	-0.15	-4.2	9.4	-19.06	34.06	4.17	1.52	12.108	14.567	11.480	-0.63	173
1994	0.5324	-0.4180	-0.4180	0.67	0.23	1.46	-6.3	9.3	-15.46	34.52	7.31	1.38	12.176	14.465	11.437	0.00	148
1995	-0.2746	0.7230	0.7230	-0.02	0.53	-0.34	-4.9	10.8	-17.96	23.80	2.28	1.06	11.601	14.060	10.963	0.74	159
1996	-0.4565	-1.0547	-1.0547	-0.57	0.30	-1.13	-5.6	8.4	-19.15	26.09	3.89	1.58	11.948	13.980	11.030	-1.89	157
1997	-0.0398	-0.0963	-0.0963	-0.22	1.05	-0.23	-4.5	10.9	-18.06	34.87	6.07	1.20	11.831	14.230	10.987	-0.38	183
1998	-0.2709	-0.7783	-0.7783	-1.57	-1.73	-3.42	-3.8	9.6	-19.66	36.04	6.83	0.91	11.943	14.463	10.953	0.89	162
1999	0.1126	0.6483	0.6483	0.26	-0.73	0.55	-6.2	10.0	-19.84	21.34	4.07	1.33	11.876	14.388	10.990	0.43	133
2000	-0.0465	1.1297	1.1297	-0.29	-0.40	-0.45	-5.0	7.4	-19.70	30.28	5.93	1.56	11.661	14.063	10.757	0.67	156
2001	-0.1615	-1.3117	-1.3117	-0.95	-0.63	-1.82	-5.3	9.7	-11.32	25.73	4.40	1.59	11.771	14.243	10.783	-2.24	137
2002	0.0717	0.4543	0.0187	0.13	-0.18	0.30	-3.3	9.4	-20.00	26.31	3.68	1.40	11.568	14.072	10.803	0.78	203
2003	0.1521	-0.6453	0.0399	-1.67	-0.40	-3.84	-3.8	8.9	-16.89	31.98	4.52	1.38	11.563	14.090	10.587	-1.60	179

Tabla 3.2: Tabla de datos de las variables del clima en Kotzbehue, Alaska.

Notemos que algunas variables fueron medidas bajo una escala de medición distinta, por ejemplo, los índices no tienen alguna escala, por otro lado la temperatura está medida por grados Celcius, la precipitación pluvial se mide en milímetros, que equivale a una lámina de agua de dicho espesor sobre una superficie plana y es equivalente a los litros de agua por metro cuadrado ($\frac{1litro}{m^2}$) y en otra variable mide el conteo de los días sin hielo que hubo en el año. Por lo tanto usaremos la matriz de correlaciones para darle el mismo peso a cada variable, ya que no tienen una medida de escala en común, la cual se muestra en la figura (3.3).

Se observa que algunas variables tienen una alta correlación, por ejemplo hay una alta correlación positiva entre las variables **NPI** y **NPI_wi** siendo de 0.9978961, la variable **Rain_wi** y la variable **NPI_wi** tienen una correlación de 0.631, las variables **NPI_su** y **NPI** con una correlación de 0.610. En la matriz de correlaciones se pueden observar y también resalta la correlación positiva que hay entre **Ice** y**Ice_OctDec**, también entre las variables **Ice** y**Ice_JanJul**.
Por otro lado, hay correlaciones negativas entre las variables **Temp_**wi y **AO_wi** siendo de -0.591, la variable **Temp_wi** y la variable **IceCover** con una correlación de -0.563, las variables **Ice** y **IceFreeDays** tienen una correlación de -0.565, etc.

	AO	AO_ wi	AO_ su	NPI	NPI_ sp	NPI_ wi	Temp	Temp_ su	Temp_ wi	Rain	Rain_ su	Rain_ wi	Ice	Ice_ JanJul	Ice_ OctDec	IceCover	IceFreeDays
AO	1.000	0.589	0.562	0.503	0.258	0.498	-0.222	0.314	-0.305	0.494	0.312	0.319	0.045	-0.097	0.451	0.115	0.102
AO_ wi	0.589	1.000	0.932	0.435	0.404	0.436	-0.016	0.158	-0.591	0.307	0.021	0.257	-0.165	-0.252	0.169	0.323	0.170
AO_ su	0.562	0.932	1.000	0.495	0.383	0.494	-0.024	0.186	-0.499	0.346	-0.035	0.279	-0.071	-0.114	0.216	0.237	0.039
NPI	0.503	0.435	0.495	1.000	0.186	0.998	-0.192	0.610	-0.238	0.189	-0.133	0.623	-0.166	-0.134	0.130	0.127	0.159
NPI_ sp	0.258	0.404	0.383	0.186	1.000	0.194	-0.173	0.046	-0.206	0.061	-0.276	0.180	0.132	-0.061	0.404	0.007	0.115
NPI_ wi	0.498	0.436	0.494	0.998	0.194	1.000	-0.185	0.600	-0.241	0.205	-0.118	0.632	-0.165	-0.138	0.131	0.118	0.159
Temp	-0.222	-0.016	-0.024	-0.192	-0.173	-0.185	1.000	0.095	0.097	0.177	-0.016	0.081	-0.207	-0.038	-0.221	-0.141	0.566
Temp_ su	0.314	0.158	0.186	0.610	0.046	0.600	0.095	1.000	-0.030	0.173	-0.142	0.278	-0.100	0.081	0.035	-0.042	0.211
Temp_ wi	-0.305	-0.591	-0.499	-0.238	-0.206	-0.241	0.097	-0.030	1.000	-0.198	0.046	-0.027	0.241	0.298	0.011	-0.564	-0.233
Rain	0.494	0.307	0.346	0.189	0.061	0.205	0.177	0.173	-0.198	1.000	0.703	-0.122	-0.128	-0.147	0.091	0.007	0.327
Rain_ su	0.312	0.021	-0.035	-0.133	-0.276	-0.118	-0.016	-0.142	0.046	0.703	1.000	-0.378	0.107	0.018	0.124	0.052	0.032
Rain_ wi	0.319	0.257	0.279	0.623	0.180	0.632	0.081	0.278	-0.027	-0.122	-0.378	1.000	-0.054	-0.099	0.106	-0.248	0.166
Ice	0.045	-0.165	-0.071	-0.166	0.132	-0.165	-0.207	-0.100	0.241	-0.128	0.107	-0.054	1.000	0.867	0.710	-0.243	-0.566
Ice_ JanJul	-0.097	-0.252	-0.114	-0.134	-0.061	-0.138	-0.038	0.081	0.298	-0.147	0.018	-0.099	0.867	1.000	0.472	-0.179	-0.514
Ice_ OctDec	0.451	0.169	0.216	0.130	0.404	0.131	-0.221	0.035	0.011	0.091	0.124	0.106	0.710	0.472	1.000	0.064	-0.360
IceCover	0.115	0.323	0.237	0.127	0.007	0.118	-0.141	-0.042	-0.564	0.007	0.052	-0.248	-0.243	-0.179	0.064	1.000	0.048
IceFreeDays	0.102	0.170	0.039	0.159	0.115	0.159	0.566	0.211	-0.233	0.327	0.032	0.166	-0.566	-0.514	-0.360	0.048	1.000

Tabla 3.3: Matriz de Correlación de los datos del clima en Kotzbehue, Alaska

En el siguiente gráfico se puede visualizar la correlación positiva o negativa que hay entre las variables.

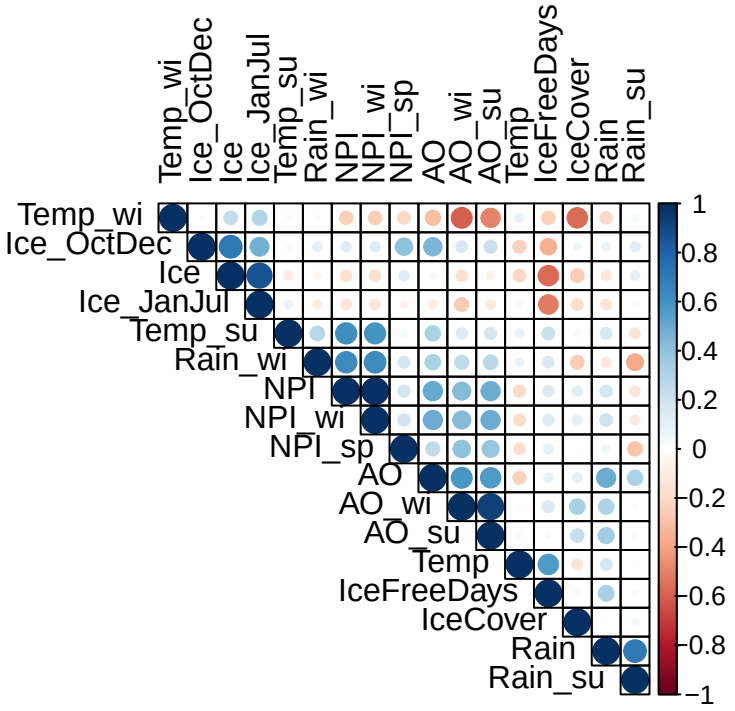

Figura 3.1: Imagen de la matriz de correlaciones

Todas nuestras variables son continuas, para saber si las variables se encuentran altamente relacionadas entre ellas, analizaremos la medida global de dependencia. Como el determinante de la matriz de correlación muestral obtenida es de 9.273148×10^{-10}, es un valor cercano a cero, este resultado indica que hay una buena correlación entre algunas variables en el estudio y que una o más variables podrían ser expresadas como combinación lineal de otras variables, por lo tanto, hacer un análisis de componentes principales es adecuado para la exploración de los datos.

En las siguientes figuras se muestran los valores propios y sus vectores propios asociados que fueron obtenidos al calcular las componentes.

	Valores propios	**Proporción de varianza**(%)	**Proporción acumulada**(%)
λ_1	4.730501804	27.8264812	27.82648
λ_2	3.01621566	17.74244506	45.56893
λ_3	2.258355521	13.28444424	58.85337
λ_4	1.832477713	10.77928066	69.63265
λ_5	1.319726209	7.76309535	77.39575
λ_6	0.994386302	5.84933119	83.24508
λ_7	0.779684543	4.58637967	87.83146
λ_8	0.623597193	3.66821878	91.49968
λ_9	0.419578838	2.46811081	93.96779
λ_{10}	0.402257786	2.36622227	96.33401
λ_{11}	0.223923964	1.31719979	97.65121
λ_{12}	0.158764955	0.9339115	98.58512
λ_{13}	0.132422463	0.77895567	99.36408
λ_{14}	0.063694701	0.37467471	99.73875
λ_{15}	0.026094811	0.15349889	99.89225
λ_{16}	0.01723628	0.10138988	99.99364
λ_{17}	0.001081256	0.00636033	100

Tabla 3.4: Tabla de los valores propios de la matriz **R** y sus respectivas proporciones de varianza en términos porcetuales

Notamos que las primeras dos componentes principales acumulan el 45.56 % de la varianza total, lo cual es un porcentaje bajo, seguramente necesitaremos elegir más de dos componentes, por otro lado si decidiéramos tomar hasta ocho componentes se obtendría un porcentaje acumulado de 91.49 % de la varianza total. En la siguiente tabla mostraremos los vectores propios asociados a las primeras 5 componentes:

	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$	$\mathbf{a}_4$	$\mathbf{a}_5$
AO	0.33509993	0.16777339	0.179538088	0.16941291	-0.047665394
AO_wi	0.37287477	0.01714593	0.184320875	-0.17110901	0.204802048
AO_su	0.36583517	0.08487849	0.13847401	-0.12831681	0.191270566
NPI	0.37651718	0.0682142	-0.257474947	0.0990942	-0.26579787
NPI_sp	0.18551215	0.15891357	-0.007158265	-0.21989463	0.467161051
NPI_wi	0.37703247	0.06742711	-0.253147991	0.10606353	-0.258000927
Temp	-0.035691	-0.26145276	-0.0411471	0.2991994	0.506939813
Temp_su	0.22267722	0.02740679	-0.264384066	0.28543183	-0.158793727
Temp_wi	-0.25818752	0.10642022	-0.23256549	0.3371859	0.001902442
Rain	0.19041188	-0.06035678	0.387459973	0.43870448	0.021106423
Rain_su	-0.01042119	0.01730847	0.488109199	0.40764924	-0.217912412
Rain_wi	0.23611124	0.05625147	-0.424430981	0.07555968	0.135769944
Ice	-0.12808621	0.51509352	0.045379924	0.07464844	0.144014171
Ice_JanJul	-0.15121315	0.43533391	-0.036773945	0.13270723	0.06299382
Ice_OctDec	0.08363692	0.4715141	0.125745841	0.02458535	0.167371382
IceCover	0.13228157	-0.08291703	0.265574314	-0.37428484	-0.269551472
IceFreeDays	0.15200331	-0.39789724	-0.015451433	0.20789683	0.299659759

Tabla 3.5: Tabla de los vectores propios de la matriz $\mathbf{R}$

3.2. Elección del número de componentes principales

Para decidir cuántas componentes elegiremos interpretar, hemos mostrado distintos criterios, los cuales serán utilizados para la elección del número de componentes principales.

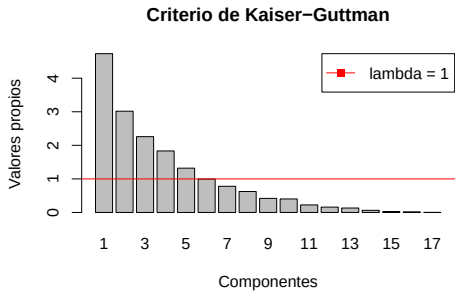

Figura 3.2: Diagrama de barras que muestra los valores propios que son mayores a 1

Recordemos que para aplicar el criterio de Kaiser-Guttman se deben estandarizar

las variables, como estamos ocupando la matriz de correlaciones el criterio es válido. Observemos que sólo cinco componentes tienen valores propios mayores a 1. Usualmente este criterio nos puede servir para conocer el número máximo de componentes a interpretar (figura 3.2).

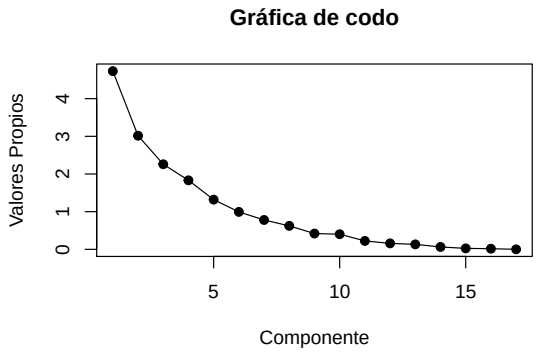

Figura 3.3: Gráfica de codo

El criterio de la gráfica de codo no muestra con claridad en que componente hay un cambio abrupto en la pendiente. Se puede observar que entre la novena y la décima componente la pendiente es casi cero, por lo que en teoría, este criterio sugiere seleccionar ocho componentes (tabla 3.3), sin embargo, no es muy claro un cambio antes de dicha componente, este criterio está abierto a lo que decida el analista.

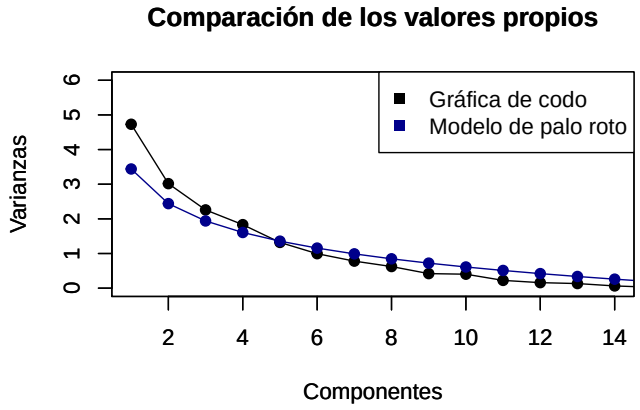

Figura 3.4: Modelo del palo roto

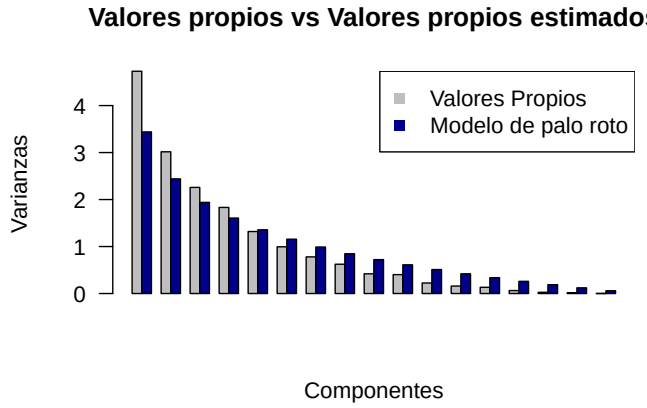

Figura 3.5: Gráfica de barras comparativa utilizando el modelo del palo roto

Ser puede observar en las figuras (3.4) y (3.5) que los primeros cuatro valores propios observados son mayores a los esperados por el modelo del palo roto, por lo que este modelo sugiere conservar las primeras cuatro componentes.

El criterio de la variación porcentual se basa en fijarnos en el porcentaje de variación de cada componente y en la variación acumulada de las primeras componentes, así que en la figura (3.4) muestra que el porcentaje de varianza acumulada por la primeras 4 componentes es de 69.63 % lo cual ya es un porcentaje aceptable con un enfoque ecológico como lo es este estudio. Mientras que si se decide tomar cinco componentes, se tendría un porcentaje de variaza acumulada del 77.39 %, sin embargo el porcentaje de variación de la quinta componente explica un porcentaje menor al 10 % siendo del 7.76 %, así que lo más adecuado es tomar 4 componentes principales para realizar una interpretación, ya que la quinta componente explica poca variación porcentual y además, su inclusión podría complicar la interpretación.

Ahora haremos las pruebas de significancia que fueron presentadas en el capítulo anterior.
Las tablas siguientes muestran las pseudoestimaciones y pseudovalores que obtuvimos al utilizar el procedimiento jackknife y que corresponden a las 23 variables de la tabla (3.2) que hemos trabajado.

Los estimadores jackknife Φ_i^* para $i = 1, 2, \ldots, 17$ y sus respectivos errores estándar son los siguientes:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
$\Phi_{i(1)}$	27.27719	18.18834	13.65444	11.031797	7.35666	5.842596	4.413148	3.653145	2.509897	2.479579	1.328577	0.9527202	0.6661664	0.3786337	0.1575659	0.10242992	0.00712101
$\Phi_{i(2)}$	29.07199	17.75144	12.47699	9.887793	7.734328	5.945241	4.746824	3.805182	2.500365	2.342699	1.366556	1.0023605	0.690722	0.3832759	0.1856772	0.10202354	0.00652613
$\Phi_{i(3)}$	28.23637	18.54845	11.8991	10.980188	7.517586	6.003067	4.669322	3.700762	2.498966	2.443072	1.313578	0.981163	0.8416427	0.2054384	0.1043488	0.04997506	0.00696794
$\Phi_{i(4)}$	28.05819	18.08449	13.14253	10.337822	7.493361	6.104528	4.719238	3.694345	2.41824	2.398237	1.315807	0.8305573	0.7824544	0.3915834	0.1263163	0.09578354	0.00651409
$\Phi_{i(5)}$	28.63309	18.03341	13.03984	11.089486	7.436564	5.868065	4.704754	3.585029	2.577898	1.778225	1.306768	0.9192259	0.4901479	0.3182547	0.1411931	0.07190171	0.00615672
$\Phi_{i(6)}$	27.16199	17.97959	13.82426	10.845848	8.017738	5.978837	4.407384	3.548095	2.586261	2.163014	1.220421	0.9343418	0.7104491	0.3459456	0.1651875	0.10416534	0.00646827
$\Phi_{i(7)}$	26.73105	19.34008	12.83383	11.312121	7.94385	6.018145	4.646166	3.561927	2.528771	2.20932	1.034229	0.8240029	0.4121226	0.3899155	0.1456048	0.06666088	0.00221072
$\Phi_{i(8)}$	28.08935	17.76271	13.1829	10.89141	7.828243	5.886253	4.576675	3.462485	2.381877	2.285083	1.327766	0.9457137	0.7724448	0.3770739	0.1269087	0.10066147	0.00244368
$\Phi_{i(9)}$	26.8029	17.78623	14.67528	10.430716	8.035907	6.084129	4.026713	3.815869	2.659716	2.287157	1.339008	0.8677192	0.6396274	0.3304857	0.129398	0.08373332	0.00541744
$\Phi_{i(10)}$	25.8274	18.09328	13.75451	11.813952	7.535547	6.242132	4.649521	3.806982	2.839343	2.025342	1.272271	0.8101602	0.703912	0.386951	0.1434427	0.08839976	0.00685087
$\Phi_{i(11)}$	27.72167	17.88083	13.21036	10.608966	7.935445	5.992052	4.738927	3.421774	2.543912	2.391204	1.334197	0.9522968	0.6386346	0.3859423	0.1402759	0.09747087	0.00604416
$\Phi_{i(12)}$	28.77367	17.91848	13.51944	9.254817	8.140323	5.253302	4.912015	3.759112	2.576803	2.354883	1.356472	0.8845791	0.8022988	0.288897	0.1147185	0.08452598	0.00565672
$\Phi_{i(13)}$	27.76426	18.20376	13.49934	11.242862	6.408895	5.979113	5.059287	3.64725	2.508454	2.007215	1.361694	0.9581968	0.8040939	0.368463	0.1130451	0.0681414	0.00593333
$\Phi_{i(14)}$	28.58612	17.6418	13.96651	10.745125	7.375844	5.593338	4.355	3.548385	2.55374	2.098701	1.238889	0.9099545	0.7913335	0.3460021	0.1516722	0.0919958	0.0055848
$\Phi_{i(15)}$	28.1972	17.94258	13.77497	10.596609	7.965305	5.98036	4.517122	3.344549	2.453393	1.686827	1.266546	0.896045	0.7529576	0.3738881	0.1509696	0.09566369	0.00500812
$\Phi_{i(16)}$	28.45634	17.96895	13.33126	11.131259	7.921042	5.373474	4.659677	3.779672	2.478578	2.038098	1.165944	0.8094445	0.3968008	0.2802346	0.1374607	0.06738882	0.00437307
$\Phi_{i(17)}$	28.02251	17.85973	13.73164	10.783205	8.016404	6.058414	4.099691	3.216154	2.478025	2.266029	1.388892	0.9545812	0.5368119	0.3815133	0.15124	0.05030301	0.00486294
$\Phi_{i(18)}$	28.56483	17.46116	13.00922	10.969924	8.392189	5.169734	4.511653	3.581586	2.499594	2.420989	1.168401	0.9790736	0.8141195	0.2344548	0.1520985	0.06439994	0.00657979
$\Phi_{i(19)}$	28.21349	18.08755	13.52628	10.721003	7.099607	5.791465	4.449339	3.677596	2.533642	2.262159	1.342121	0.9017613	0.790738	0.3644055	0.1436779	0.08894444	0.00621767
$\Phi_{i(20)}$	28.639	17.74562	13.06319	10.933305	7.9545	6.074138	4.232294	3.40387	2.561579	1.690858	1.313437	0.965229	0.7760966	0.3752073	0.1631627	0.10215164	0.00635996
$\Phi_{i(21)}$	27.87267	18.42571	12.63611	10.910426	8.007318	5.223068	4.575828	3.728748	2.940963	2.073107	1.317127	0.9541662	0.7359196	0.386227	0.1458163	0.06540897	0.00138276
$\Phi_{i(22)}$	27.9128	16.58215	13.44056	11.447487	8.203686	5.852899	4.724771	3.351483	2.602865	2.389489	1.368165	0.9648254	0.6337428	0.2787621	0.1671972	0.07433373	0.00479179
$\Phi_{i(23)}$	28.60431	16.95223	13.62378	10.664022	7.861263	5.724271	4.569645	3.796546	2.578641	2.181286	1.399107	0.8978676	0.546729	0.390519	0.1685297	0.03697131	0.00428337

Tabla 3.6: Tabla de las pseudoestimaciones

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
$\Phi^*_{i(1)}$	39.9109413	7.93274604	5.144623	5.2239216	16.704668	5.9974948	8.397487	3.9998463	1.54880561	-0.1276154	1.0669047	0.52012054	3.26032049	0.28757737	0.06402533	0.078509	-0.01037463
$\Phi^*_{i(2)}$	0.4251992	17.54458212	31.048359	30.3920153	8.395972	3.7393143	1.056596	0.6550216	1.75852567	2.8837294	0.231368	-0.57196562	2.72009575	0.18544875	-0.55442327	0.08744936	0.00271273
$\Phi^*_{i(3)}$	18.8089793	0.01034496	43.761915	6.3593209	13.164291	2.4671374	2.761648	2.9522789	1.78928563	0.6755223	1.3968752	-0.10562084	-0.60015899	4.09787265	1.23479999	1.23251592	-0.00700709
$\Phi^*_{i(4)}$	22.7288273	10.2174676	16.406519	20.4913677	13.697257	0.2350122	1.663505	3.0934325	3.56526841	1.6619075	1.3478399	3.20770434	0.70198427	0.00268397	0.75151609	0.22472936	0.00297761
$\Phi^*_{i(5)}$	10.0811897	11.34127886	18.665776	3.9547691	14.946781	5.4371775	1.98215	5.4983836	0.05279857	15.3021594	1.546704	1.25699492	7.13272639	1.61591515	0.42422649	0.75012962	0.01083975
$\Phi^*_{i(6)}$	42.4452443	12.52519016	1.408531	9.3147895	2.160949	3.0002062	8.524283	6.3109474	-0.13120217	6.8368003	3.4463349	0.92444424	2.28610109	1.00671513	-0.10364965	0.04032976	0.00398565
$\Phi^*_{i(7)}$	51.9260063	-17.40541538	23.19797	-0.9432044	3.786492	2.1354192	3.271072	6.0066296	1.13357761	5.818083	7.5425596	3.3519007	8.84928233	0.03937821	0.32716865	0.86542788	0.09765175
$\Phi^*_{i(8)}$	22.0433669	17.29663706	15.518377	8.31244	6.329841	5.0370619	4.799876	8.1943531	4.36525155	4.1512858	1.0847326	0.67426376	0.92219525	0.32189231	0.73848307	0.1174149	0.09252663
$\Phi^*_{i(9)}$	50.3453409	16.77919442	-17.313913	18.4476946	1.761236	0.6837706	16.899055	0.4199243	-1.74719897	4.1056573	0.8374286	2.39014188	3.84417673	1.34683403	0.68371869	0.4898342	0.02710391
$\Phi^*_{i(10)}$	71.8062128	10.02408144	2.94296	-11.9834796	12.769167	-2.7922862	3.197264	0.6154262	-5.69900419	9.8655968	2.3056312	3.6564401	2.42991663	0.10459743	0.37473551	0.38717252	-0.00443155
$\Phi^*_{i(11)}$	30.1323256	14.69791632	14.914388	14.5261979	3.971402	2.7094679	1.230349	9.0900103	0.80048705	1.8166284	0.9432574	0.52943578	3.86601921	0.12678707	0.44440401	0.1876081	0.01331607
$\Phi^*_{i(12)}$	6.9882966	13.86957606	8.11451	44.3174792	-0.535911	18.9619839	-2.577607	1.6685737	0.07687889	2.6156825	0.4532176	2.01922518	0.26540703	2.26178367	1.00666791	0.47239568	0.02183975
$\Phi^*_{i(13)}$	29.1952389	7.59361582	8.556821	0.5805015	37.555507	2.9941375	-5.817581	4.1295223	1.58056019	10.2643829	0.3383199	0.3996349	0.22591373	0.51133321	1.04348117	0.83285644	0.01575433
$\Phi^*_{i(14)}$	11.1143827	19.95654662	-1.720953	11.5306977	16.282617	11.4811739	9.676724	6.3045571	0.58426159	8.2516966	3.0400473	1.46096484	0.50664253	1.00547213	0.19368695	0.30805964	0.02342199
$\Phi^*_{i(15)}$	19.6706973	13.33937078	2.492814	14.7980482	3.314478	2.966691	6.110058	10.788959	2.79190021	17.3129198	2.4315779	1.76697538	1.35091299	0.39197969	0.20914327	0.22736606	0.03610895
$\Phi^*_{i(16)}$	13.9695251	12.75940656	12.254413	3.0357475	4.288277	16.3181977	2.973827	1.2162517	2.23784099	9.5849549	4.6448228	3.67218462	9.18636237	2.45235779	0.50633797	0.8494132	0.05008005
$\Phi^*_{i(17)}$	23.5139292	15.16215438	3.446211	10.6929496	2.190315	1.249518	15.29353	13.6136398	2.25000545	4.5704663	-0.2600394	0.47917876	6.10611927	0.22422463	0.20319447	1.22530102	0.03930291
$\Phi^*_{i(18)}$	11.5829165	23.93072078	19.339373	6.5851303	-6.076965	20.8004727	6.230366	5.5741481	1.77548767	1.1613628	4.5907677	-0.05965536	0.00535119	3.45951361	0.18430835	0.91516856	0.00153221
$\Phi^*_{i(19)}$	19.3123004	10.15007302	7.964024	12.0613863	22.359843	7.1223801	7.601281	3.4619298	1.02643013	4.6556206	0.7689244	1.64121612	0.51974397	0.60059777	0.36956133	0.37518956	0.00949885
$\Phi^*_{i(20)}$	9.9509913	17.67264522	18.152036	7.3907551	3.552187	0.9035759	12.376262	9.4838908	0.41181019	17.2242415	1.3999706	0.24492738	0.84185565	0.36295795	-0.05910603	0.08463116	0.00636847
$\Phi^*_{i(21)}$	26.810321	2.71054642	27.54779	7.894088	2.390193	19.6271282	4.818508	2.3365678	-7.93462681	8.8147554	1.3187924	0.4883092	1.72574899	0.12052477	0.32251697	0.8929699	0.11586687
$\Phi^*_{i(22)}$	25.9275554	43.26896872	9.849987	-3.9212617	-1.929891	5.7708317	1.541763	10.6363967	-0.49648973	1.8543434	0.1959722	0.25380482	3.97363969	2.48475103	-0.14786459	0.69662518	0.04086821
$\Phi^*_{i(23)}$	10.7142885	35.12710158	5.819122	13.3149624	5.603404	8.6006582	4.954544	0.8450213	0.03644135	6.4348149	-0.48475	1.72687796	5.88794175	0.02609945	-0.17717937	1.51859842	0.05205345

Tabla 3.7: Tabla de los pseudovalores

Φ_1^*	Φ_2^*	Φ_3^*	Φ_4^*	Φ_5^*	Φ_6^*
24.75669898	13.76107607	12.06572409	10.10331812	8.11661343	6.32376194
$SE(\Phi_1^*)$	$SE(\Phi_2^*)$	$SE(\Phi_3^*)$	$SE(\Phi_4^*)$	$SE(\Phi_5^*)$	$SE(\Phi_6^*)$
3.53490865	2.40461715	2.61252161	2.38123765	1.96811359	1.38748966
Φ_7^*	Φ_8^*	Φ_9^*	Φ_{10}^*	Φ_{11}^*	Φ_{12}^*
5.08543306	5.08242225	0.5120476	6.33630417	1.74727215	1.30119581
$SE(\Phi_7^*)$	$SE(\Phi_8^*)$	$SE(\Phi_9^*)$	$SE(\Phi_{10}^*)$	$SE(\Phi_{11}^*)$	$SE(\Phi_{12}^*)$
1.09703444	0.78545806	0.56266983	1.06380246	0.38943617	0.26184979
Φ_{13}^*	Φ_{14}^*	Φ_{15}^*	Φ_{16}^*	Φ_{17}^*	
2.86992601	1.00162164	0.34955449	0.55911719	0.02791291	
$SE(\Phi_{13}^*)$	$SE(\Phi_{14}^*)$	$SE(\Phi_{15}^*)$	$SE(\Phi_{16}^*)$	$SE(\Phi_{17}^*)$	
0.59414979	0.24616334	0.08948279	0.08897748	0.00716593	

Tabla 3.8: Tabla de estimadores jackknife y sus errores estándar

Como se dijo en el capítulo anterior, la relación $\frac{\Phi_i^* - \Phi_i}{SE(\Phi_i^*)}$ puede tratarse como una estadística para una distribución t con $(n-1)$ grados de libertad, con la finalidad de decidir si el estimador observado difiere significativamente de los esperados. Tenemos $n = 23$, por lo que los grados de libertad de las distribución t son $n - 1 = 22$. Supongamos que tenemos un nivel de significancia $\alpha = 0.05$, entonces los cuantiles 0.025 y 0.975 de una distribución t_{n-1} son $\omega_{\alpha/2} = \omega_{0.025} = -2.073873$ y $\omega_{1-\alpha/2} = \omega_{0.975} = 2.073873$ respectivamente. Por otro lado las estadísticas y las decisiones se muestran en la tabla (3.9)

T_1	T_2	T_3	T_4	T_5	T_6
-0.8684191	-1.6557185	-0.4664919	-0.2838703	0.1796228	0.3419346
No se rechaza	No se rechaza	No se rechaza	No se rechaza	No se rechaza	No se rechaza
T_7	T_8	T_9	T_{10}	T_{11}	T_{12}
0.4549113	1.8004825	-3.4763961	3.7319729	1.1043462	1.4026527
No se rechaza	No se rechaza	Se rechaza	Se rechaza	No se rechaza	No se rechaza
T_{13}	T_{14}	T_{15}	T_{16}	T_{17}	
3.5192646	2.5468737	2.1909867	5.1443052	3.0076459	
Se rechaza	Se rechaza	Se rechaza	Se rechaza	Se rechaza	

Tabla 3.9: Estadísticas T_i y las decisiones de rechazar o no rechazar la hipótesis nula

Estos resultados nos muestran que los primeros 8 estimadores jackknife obtenidos no difieren significativamente de los primeros estimadores calculados de los datos originales, y como elegiremos interpretar 4 componentes entonces los porcentajes de varianza por el procedimiento jackknife no difieren en las primeras componentes de forma significativa a los que fueron calculados con los datos reales. Sin embargo, el criterio de la variación porcentual indica que las primeras 4 componentes acumulan un 60.67 % de la varianza total de la muestra. Dicho porcentaje no es tan bueno, por otro lado, las primeras 5 componentes acumulan un 68.78 % de la varianza total. Por lo tanto, el criterio de la variación porcentual para los estimadores jackknife sugiere considerar las primeras 5 componentes.

Hemos mencionado que el procedimiento bootstrap es superior al jackknife, por lo que también hicimos una prueba de significancia utilizando dicho método. Para ello, se generaron 10000 muestras bootstrap con el apoyo del lenguaje de programación R. Se tienen 17 variables, así que obtendremos el mismo número de estimadores bootstrap de los porcentajes de la varianza explicada por cada componente y sus respectivos errores estándar, lo cual se muestra en la tabla (3.10)

Φ_1^*	Φ_2^*	Φ_3^*	Φ_4^*	Φ_5^*	Φ_6^*
27.84909851	17.73030775	13.29206302	10.79766428	7.75172646	5.85766830
$SE(\Phi_1^*)$	$SE(\Phi_2^*)$	$SE(\Phi_3^*)$	$SE(\Phi_4^*)$	$SE(\Phi_5^*)$	$SE(\Phi_6^*)$
0.10590716	0.08590043	0.03515463	0.06061098	0.07268738	0.03596755
Φ_7^*	Φ_8^*	Φ_9^*	Φ_{10}^*	Φ_{11}^*	Φ_{12}^*
4.59143755	3.66099931	2.46291769	2.34746615	1.31301239	0.93137634
$SE(\Phi_7^*)$	$SE(\Phi_8^*)$	$SE(\Phi_9^*)$	$SE(\Phi_{10}^*)$	$SE(\Phi_{11}^*)$	$SE(\Phi_{12}^*)$
0.02917842	0.03253093	0.02639006	0.02910369	0.01085501	0.01016228
Φ_{13}^*	Φ_{14}^*	Φ_{15}^*	Φ_{16}^*	Φ_{17}^*	
0.77716656	0.37633594	0.15305235	0.10137712	0.00633028	
$SE(\Phi_{13}^*)$	$SE(\Phi_{14}^*)$	$SE(\Phi_{15}^*)$	$SE(\Phi_{16}^*)$	$SE(\Phi_{17}^*)$	
0.00984938	0.00466768	0.00196393	0.00147953	0.00008360	

Tabla 3.10: Tabla de estimadores bootstrap y sus errores estándar

En el capítulo 2 se mostró que la relación $T_i = \frac{\Phi_i^* - \Phi_i}{SE(\Phi_i^*)}$ puede tratarse como una estadística con una distribución t con $(n-1)$ grados de libertad. Suponiendo que tenemos un nivel de significancia $\alpha = 0.05$, entonces los cuantiles 0.025 y 0.975 de una distribución t_{22} son $\omega_{\alpha/2} = \omega_{0.025} = -2.073873$ y $\omega_{1-\alpha/2} = \omega_{0.975} = 2.073873$ respectivamente. Por lo tanto las estadísticas T_i y las decisiones como resultado son mostradas en la tabla (3.11)

T_1	T_2	T_3	T_4	T_5	T_6
0.21355789	-0.14129510	0.21672195	0.30330511	-0.15640803	0.23179533
No se rechaza	No se rechaza	No se rechaza	No se rechaza	No se rechaza	No se rechaza
T_7	T_8	T_9	T_{10}	T_{11}	T_{12}
0.17334318	-0.22192633	-0.19678318	-0.64445849	-0.38575736	-0.24946764
No se rechaza	No se rechaza	No se rechaza	No se rechaza	No se rechaza	No se rechaza
T_{13}	T_{14}	T_{15}	T_{16}	T_{17}	
-0.18164697	0.35590058	-0.22737063	-0.00862436	-0.35944976	
No se rechaza	No se rechaza	No se rechaza	No se rechaza	No se rechaza	

Tabla 3.11: Estadísticas T_i y las decisiones de rechazar o no rechazar la hipótesis nula

Esta prueba de significancia muestra que los estimadores bootstrap no difieren significativamente de los estimadores obtenidos de los datos originales (a los porcentajes de la varianza total de cada componente de los datos originales).
El criterio de la variación porcentual indica que las primeras 4 componentes principales acumulan un 69.66 % de la varianza total de la muestra. Este porcentaje es cercano a la varianza acumulada de las primeras 4 componentes con las estimaciones originales. Así que este criterio sugiere considerar las primeras 4 componentes.

Por lo tanto, la mayoría de los criterios sugieren que se consideren cuatro componentes principales para la interpretación de los resultados.

3.3. Interpretación de las componentes principales

Se decidió interpretar las primeras cuatro componentes principales. Para su interpretación, analizamos la estructura de componentes principales, es decir, la correlación entre la j-ésima variable y la componente k. Una vez obtenidas cada una de las correlaciones se debe identificar la carga absoluta máxima que hay para cada variable con alguna de las componentes. La tabla (3.12) muestra las correlaciones que hay entre las variables y componentes.

En el capítulo anterior, mostramos algunas reglas para determinar que cargas son importantes para la interpretación, como nuestra muestra es de $n = 23$ (es pequeña) entonces haremos uso de la regla discutida por Tabachnik y Fidell (1989). En la tabla

(3.12) se señala la importancia de cada correlación según la regla mencionada.
Con estos resultados, se puede deducir que la primera componente explica principalmente la variabilidad de las variables **AO**, **AO_wi**, **AO_su**, **NPI**, **NPI_wi**, **Temp_su** y **Temp_wi**. Así que la primera componente se enfoca en las variables que son Índices de la Oscilación Ártica, los Índices del Pacífico Norte y las temperaturas dadas durante el verano y el invierno. La segunda componente explica principalmente la variabilidad de **Temp**, **Ice**, **Ice_JanJul**, **Ice_OctDec** y **IceFreeDays**, es decir, esta componente explica la temperatura, el hielo, el hielo entre los meses de enero a julio y de octubre a diciembre y los días libres de hielo. La tercera componente explica principalmente la variabilidad de las variables **Rain_su** y **Rain_wi**, es decir, de las lluvias que hay en verano e invierno. Finalmente, la cuarta componente explica principalmente la variabilidad de la variable **Rain** y **IceCover**, es decir, de la lluvia y la cubierta de hielo.

	CP_1	CP_2	CP_3	CP_4	Carga
AO	0.72883285	0.29137634	0.26980671	0.22933263	Excelente
AO_wi	0.81099206	0.02977777	0.27699420	-0.23162862	Excelente
AO_su	0.79568113	0.14741065	0.20809633	-0.17370124	Excelente
NPI	0.81891421	0.11846935	-0.38692887	0.13414287	Excelente
NPI_sp	0.40348367	0.27598925	-0.01075732	-0.29766926	Deficiente
NPI_wi	0.82003495	0.11710239	-0.38042639	0.14357718	Excelente
Temp	-0.07762691	-0.45407169	-0.06183515	0.40502336	Justa
Temp_su	0.48431665	0.04759808	-0.39731177	0.38638633	Justa
Temp_wi	-0.56155054	0.18482271	-0.34949537	0.45644531	Buena
Rain	0.41414045	-0.10482316	0.58226810	0.59387004	Buena
Rain_su	-0.02266579	0.03006006	0.73352201	0.55183086	Excelente
Rain_wi	0.51353527	0.09769336	-0.63782749	0.10228441	Muy buena
Ice	-0.27858388	0.89457607	0.06819616	0.10105088	Excelente
Ice_JanJul	-0.32888431	0.75605553	-0.05526324	0.17964451	Excelente
Ice_OctDec	0.18190793	0.81889059	0.18896866	0.03328095	Excelente
IceCover	0.28770868	-0.14400412	0.39910046	-0.50666579	Justa
IceFreeDays	0.33060289	-0.69103830	-0.02322014	0.28142794	Muy buena

Tabla 3.12: Cargas que hay entre las variables y las primeras cuatro componentes

En la tabla (3.13) se presentan los 23 individuos (registros realizados desde 1981 hasta el año 2003) evaluados en las primeras cuatro componentes, los resultados son conocidos como "scores". En las siguientes figuras, se muestran algunas gráficas de los scores obtenidos y sus respectivos biplot, que dependen de las componentes que se estén eligiendo. Las gráficas que se muestran son los casos donde claramente se puede distinguir una separación de los scores para los años que transcurrieron de 1981 hasta 1994 y el periodo de 1995 al año 2003.

	Com.1	Com.2	Com.3	Com.4
1981	-2.6451051	0.66308562	-0.08145502	-0.30886226
1982	0.7419438	3.56690049	-2.04077821	1.59942994
1983	-2.6249103	2.05815712	2.77303517	-0.52913363
1984	-1.2980478	0.38209882	1.89681388	-1.57570972
1985	-1.2177972	1.54198463	-1.43185238	-0.06467871
1986	-3.4890247	1.39005055	0.97846365	0.93160061
1987	-2.9852849	-0.15456506	-1.98043790	-0.21919175
1988	-0.3542059	1.21525095	-1.24337693	-0.65428749
1989	4.6010392	1.17229506	0.35587847	2.18084005
1990	4.1397675	0.60866336	1.35225822	0.37837031
1991	2.5095328	-0.10203901	-2.57732204	-1.21776928
1992	0.8775581	1.84778908	1.23945764	-3.21975183
1993	2.3028453	0.56871895	0.57903762	0.08764496
1994	0.7965139	1.39619963	1.04764694	1.75349747
1995	1.0618381	-1.82464886	-1.10193777	-1.74552518
1996	-1.2425973	-1.21935026	-1.20389948	-0.51801616
1997	1.0866369	-1.53733427	0.49450959	1.54393645
1998	-1.8209544	-2.01060416	2.42875205	1.53063421
1999	0.8105551	0.01596589	-0.55397658	-1.50320970
2000	0.9990413	-2.10675181	1.51374037	-1.01764424
2001	-2.5762239	-0.81053048	-2.19347718	1.61857755
2002	1.4203040	-3.51922276	-0.43562249	-0.28154934
2003	-1.0934246	-3.14211349	0.18454238	1.23079775

Tabla 3.13: Scores de los 23 años de registros

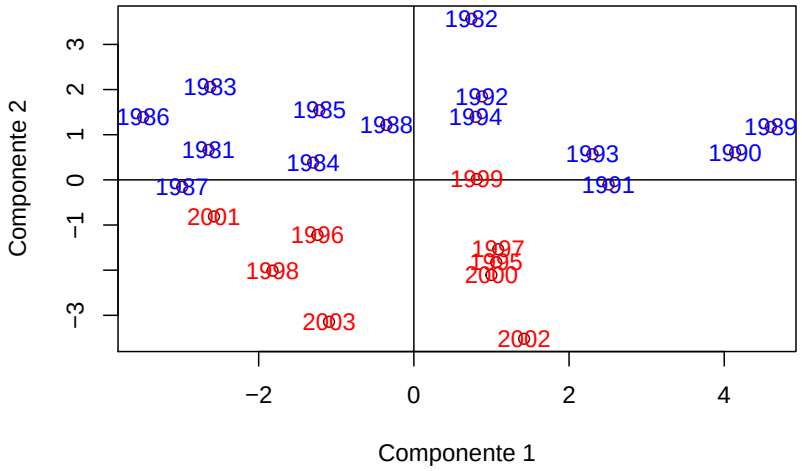

Figura 3.6: Gráfica de scores de los 23 años de registros en las primeras dos componentes

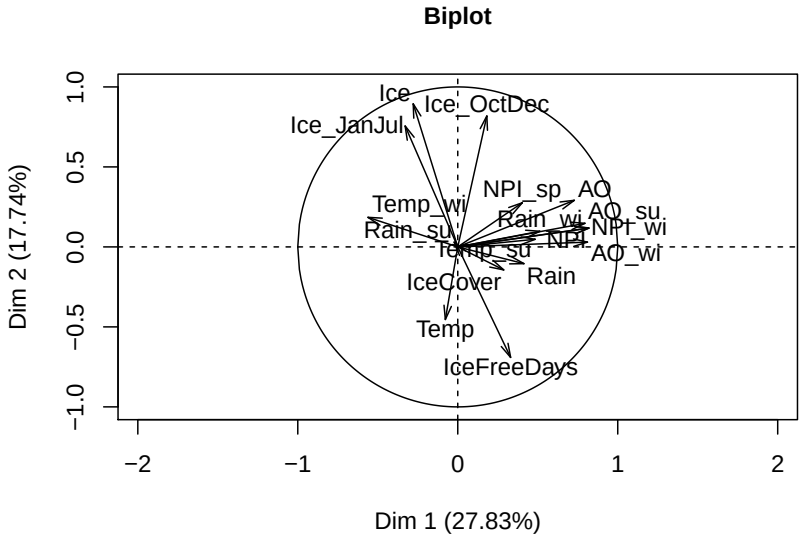

Figura 3.7: Círculo de correlación

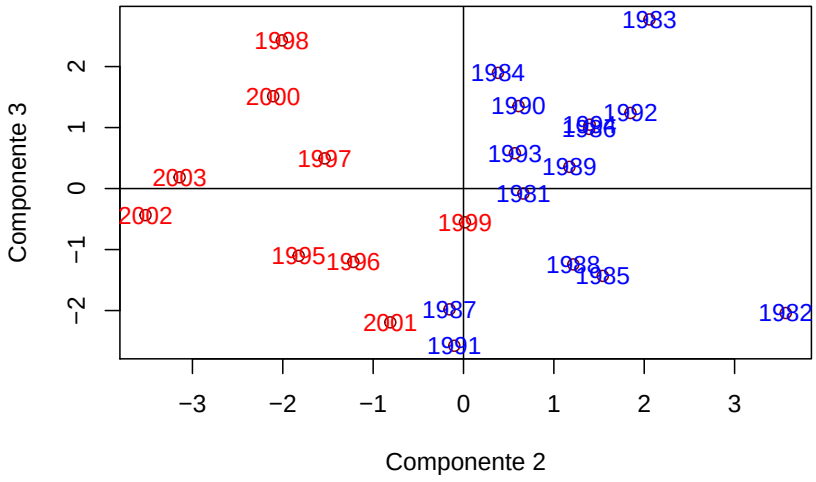

Figura 3.8: Gráfica de scores de los 23 individuos en la segunda y tercera componente

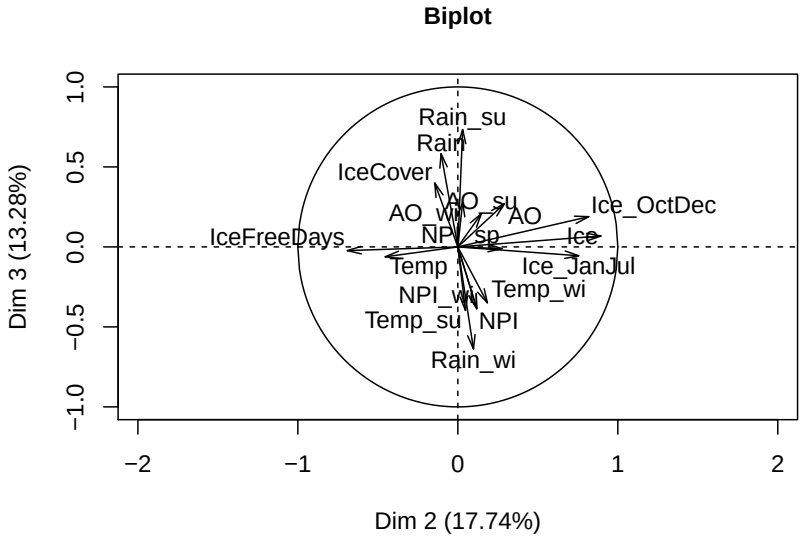

Figura 3.9: Círculo de correlación

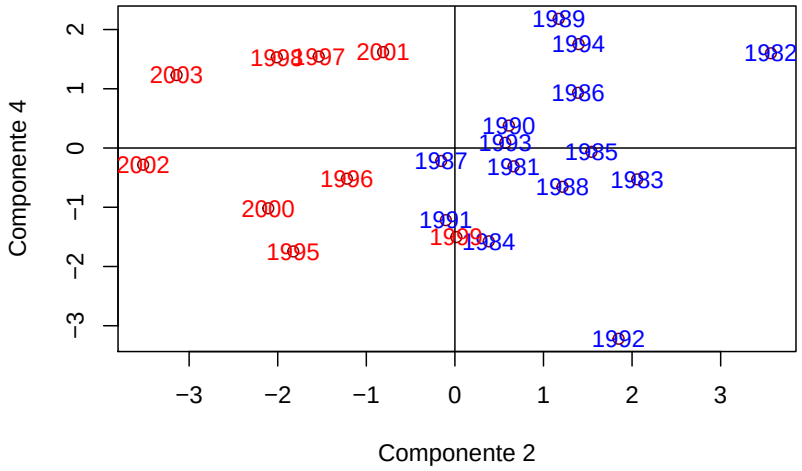

Figura 3.10: Gráfica de scores de los 23 individuos en la segunda y cuarta componente

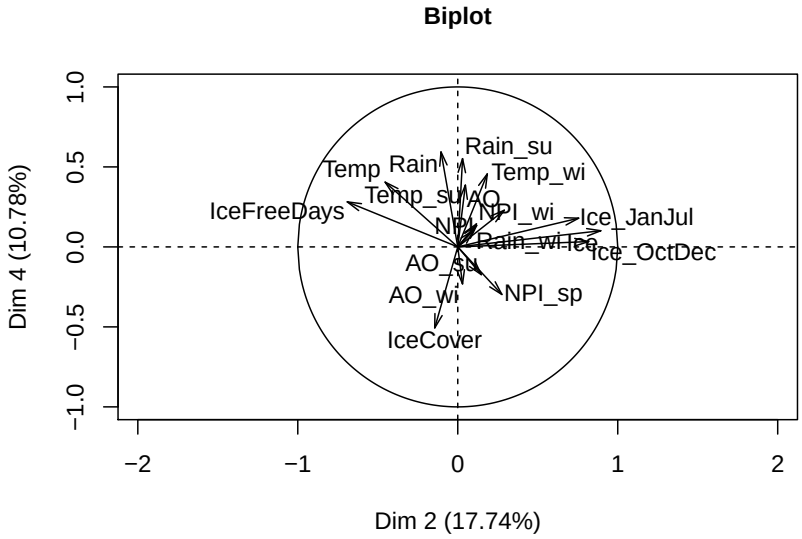

Figura 3.11: Círculo de correlación

Las primeras dos componentes explican una varianza del 45.56 %. Anteriormente ya se ha mencionado que variables se asocian a la segunda y primera componente. Por lo que en la figura (3.7) se puede observar una separación. La segunda componente es la que categoriza los registros en dos grupos. El primer grupo está compuesto por los scores de 1981 hasta 1994 que están cercanos al eje de la primera componente o están por encima de dicho eje, ya que dichos scores son cercanos a cero o positivos, mientras que el segundo grupo está compuesto por los scores que hay entre 1995 al año 2003, estos están por debajo y alejado de dicho eje (ya que sus scores son negativos), excepto por el año 1999. El círculo de correlación de la figura (3.7) representa las correlaciones entre variables y componentes y además nos indica su dirección, la cual nos ayuda en la interpretación. Relacionando ambas figuras y enfocándonos solo en la segunda componente podemos decir que en el primer grupo los registros de temperaturas y presencia de hielo en los plazos mencionados son mas bajos en comparación con el segundo grupo, entonces entre más positivos sean los scores respecto a la segunda componente menor es la temperatura y presencia de hielo, a su vez se observa que la variable **IceFreeDays** tiene una dirección opuesta comparada con el resto de las variables significativas, por lo que podemos decir que entre más negativos sean los scores en dicha componente tienen más días libres de hielo. Por lo tanto, el segundo grupo tiene más días libres de hielo respecto al primero. En relación con la primera componente, los scores de los años 1989 y 1990 se encuentran alejados al restante de los años en estudio a causa de elevadas Oscilaciones Árticas, altos índices del Pacífico Norte y una alta temperatura promedio durante el verano.

Las figuras (3.8) y (3.9) muestran el comportamiento entre la tercera y la segunda componente, la varianza acumulada que explican es del 31.02 %. Igualmente, la segunda componente puede dividir los registros en los mismos grupos que se obtuvieron al comparar las primeras dos componentes. Habría que decir también que el año 1999 es el único individuo del segundo grupo que podría agruparse en el primer grupo. Dado que la segunda componente hace distinción entre dichos grupos, se reafirma que en el primer grupo las temperaturas y la presencia de hielo son menores que el segundo grupo, además el primer grupo tiene menos días libres de hielo con respecto al segundo. Mientras que la tercera componente que es explicada por las precipitaciones durante el verano e invierno, no nos muestra una tendencia clara para su interpretación.

Las figuras (3.10) y (3.11) muestran el comportamiento entre la cuarta y la segunda componente, las cuales explican una varianza del 28.52 %. Nuevamente la segunda componente puede dividir los registros como ya lo hemos explicado.

Por lo tanto, si existe un aumento de temperatura promedio, una disminución de presencia de hielo entre enero a julio, también entre octubre a diciembre y una disminución en el hielo promedio para los registros de 1995 al 2003 con respecto al primer grupo que va de 1981 al año 1994. Más aún los días libres de hielo aumentaron para el periodo mas reciente. Lo cual nos muestra que, si hay un cambio abrupto entre dichos periodos, y el comportamiento principal es la disminución de presencia de hielo entre dichos periodos.

Capítulo 4

Comentarios generales

El primer capítulo presenta conceptos básicos de álgebra lineal con el propósito de disponer y comprender de mejor manera la teoría que se desarrolla para las componentes principales. Posteriormente en el segundo capítulo se explican algunas medidas básicas del análisis multivariado, luego se muestra el cálculo de las componentes principales, algunos criterios para elegir el número de componentes a analizar y su importancia, también su interpretación y sus limitaciones. Simultáneamente, los anteriores temas se esclarecieron con la aplicación de un problema con información cromosómica de algunas especies de leguminosas, de dicho problema se pueden hacer los siguientes comentarios:

- Se presenta un estudio taxonómico sobre el análisis de 17 leguminosas, las cuales pertenecen a dos subgéneros diferentes. El Maestro en Ciencias Luis Fernando Tapia Pastrana obtuvo la información que contiene la tabla (2.1), quien propone que dichas especies deben clasificarse en dos géneros diferentes. Así que se aplicó la técnica de ACP como análisis exploratorio, debido a que al ser 6 variables de nuestro interés, la aplicación de esta técnica podría reducir la dimensión de manera importante, lo cual facilitaría la interpretación, ya que las nuevas variables no correlacionadas recogen la mayor parte de la variabilidad.

- Las primeras dos componentes explican el 92.36 % de la varianza total de los datos, razón por la cual dichas componentes son muy significativas.

- En acorde a los criterios utilizados para el ACP, se eligieron interpretar dos componentes, la primera componente explica principalmente la variabilidad de las variables longitud de cromosoma total (THC), la talla cromosómica

promedio (AC), la diferencia entre el cromosoma más grande y el pequeño (Range) y el cociente del cromosoma más grande entre el más pequeño (Ratio). La segunda componente explica principalmente la variabilidad de las variables de Índice de asimetría(TF) e Índice Centromérico (CI).

- La gráfica de los puntajes de cada componente (scores) muestra claramente que las leguminas de subgénero Ochopodium tienen una mayor longitud cromosómica total (THC) y una mayor talla cromosómica promedio (AC) en comparación a las demás especies de subgénero Aeschynomene, además también hay una mayor diferencia entre el cromosoma más grande y el más pequeño (Range), y el cociente del cromosoma más largo entre el más pequeño (Ratio) es mayor. Sólo esas especies cumplen esas condiciones, por otro lado, hay especies que pueden tener una longitud cromosómica total bastante grande en comparación de las demás pero no hay mucha diferencia entre su cromosoma más grande y el más pequeño como en el caso de la especie A. deamii.

- Todo indica que Ochopodium debería separarse de Aeschynomene y constituir un nuevo género, aunque esto último debe ser corroborado por estudios que incluyen un mayor número de especies, ya que debido al muestreo limitado en este estudio, no permite sustentar completamente esta propuesta en este momento desde la perspectiva cromosómica.

Finalmente en el tercer capítulo se realizó un ACP a una base de datos que tiene medidas de hielo, temperaturas promedio en distintos periodos de tiempo e Índices del clima en Kotzbehue, los registros corresponden del año 1981 hasta el 2003, de dicho análisis exploratorio se puede realizar la siguiente síntesis:

- Se analizaron 17 variables diferentes de 23 años del clima en Kotzbehue con el propósito de realizar un análisis exploratorio de ACP e identificar si existía un cambio abrupto en el comportamiento del clima en el periodo mencionado con anterioridad. Este trabajo mostró que si existe un cambio increíble entre dos periodos de tiempo y el motivo principal es la disminución de presencia de hielo.

- Los criterios sugirieron analizar cuatro componentes, las cuales explican el 69.6 % de la varianza total, lo que es un porcentaje significativo.

- La primera componente explica principalmente los Índices de Oscilación Ártica y del Pacífico Norte, y las temperaturas dadas durante el verano e invierno. La

segunda componente explica principalmente la variabilidad de la temperatura promedio anual, los días libres de hielo, el hielo promedio anual, el hielo entre los meses de enero a julio y de octubre a diciembre. La tercera componente explica principalmente las lluvias que hay en verano e invierno. La cuarta componente explica con mayor variabilidad la cubierta de hielo y la lluvia promedio anual.

- Un patrón interesante que se puede detectar al observar las gráficas de los registros scores de la segunda componente comparada con las demás, es que en todos los casos se observa una separación entre los periodos de 1981 a 1994 con respecto a los registros que van de 1995 a 2003, divididos por el propio eje. Se puede interpretar que el primer periodo mencionado tiene una temperatura promedio anual mas baja, también hay mayor presencia de hielo entre enero a julio, entre octubre a diciembre y hay menos días libres de hielo con respecto al segundo periodo señalado. Por lo tanto si se puede detectar un cambio abrupto entre tales años, y el comportamiento principal es la disminución de presencia de hielo.

- Otro patrón interesante es que al graficar los scores de las primeras dos componentes, se puede observar que en los años 1990 y 1989 hay un alejamiento positivo en el eje de la componente 1 comparada contra la segunda. Esto podría interpretarse en que en esos años aumentaron de manera importante los Índices de Oscilación Ártica y del Pacífico Norte, además las temperaturas durante el verano e invierno alcanzaron los niveles más altos de la muestra.

- Como resultado, el ACP sugirió que si existe un cambio abrupto en la década de los 90's, principalmente a causa de la disminución de presencia de hielo. A pesar de que diversos estudios demuestran que la tierra se ha enfriado y calentado de manera natural a lo largo de millones de años, esto nos habla de un proceso lento, en contraste, este trabajo muestra que ese cambio se puede distinguir en un par de décadas, por lo que es un cambio muy acelerado, al menos en el lugar de estudio Kotzbehue, Alaska.

Apéndice A

Códigos en R

A.1. Gráficos

El documento "matriz1.csv" corresponde a la información de la tabla 2.1

Figura 2.1

```
library(corrplot)
datosS<-read.csv("matriz1.csv",header=TRUE)
rownames(datosS)<-datosS[,1]
datosS<-scale(datosS,T,F)
corrplot(cor(datosS), type = "upper", order = "hclust", tl.col =
    "black", tl.srt = 90,addgrid.col = "black")
```

Figura 2.2

```
library(corpcor)
pairs.panels(datosS,hist.col = "blue2")
```

Figura 2.3

```
library(bpca)
p.comp <- princomp(datosS,cor=TRUE)
plot(p.comp$sdev^2, type = "o", pch = 19, main = "Grafica de codo
    ",xlab = "Indice",ylab = "Valores Propios")
```

Figura 2.4

```
library(bpca)
p.comp <- princomp(datosS,cor=TRUE)
valores.propios <- p.comp$sdev^2
plot(p.comp, main="Varianzas de los componentes")
abline(h=mean(valores.propios),col="red")
legend("topright", c("Varianza promedio"),
```

```
       pch=15, col=c("red"),lty=1,box.lty = 0)
```

Figura 2.5

```
library(bpca)
p.comp <- princomp(datosS,cor=TRUE)
lambdaEst<-function(n)
{
    lambdaest<-0
    for(i in 1:n){
    lambdaest<-lambdaest+(1/(n-i+1))
    }
    return(lambdaest)
}

estimaciones<-c(lambdaEst(6),lambdaEst(6)-lambdaEst(1),lambdaEst
   (6)-lambdaEst(2),lambdaEst(6)-lambdaEst(3),lambdaEst(6)-
   lambdaEst(4),lambdaEst(6)-lambdaEst(5))

plot(p.comp$sdev^2, type = "o", pch = 19, main = "Comparacion de
   los valores propios", ylab = "Valores Propios", xlab = "Indice
   ", ylim=c(0,4))
par(new=TRUE)
plot(estimaciones, type = "o", pch = 19, main = "Comparacion de
   los valores propios", ylab = "Valores Propios", xlab = "Indice
   ", ylim=c(0,4),col="darkblue")
legend("topright", c("Grafica de codo", "Modelo de palo roto"),
       pch=15, col=c("black","darkblue"), bty="6")
```

Figura 2.6

```
library(bpca)
p.comp <- princomp(datosS,cor=TRUE)
barplot(t(cbind(p.comp$sdev^2, estimaciones)), beside=TRUE,
        main="Valores propios vs Valores propios estimados", col=
           c("grey","darkblue"), las=2)
legend("topright", c("Valores Propios", "Modelo de palo roto"),
       pch=15, col=c("grey","darkblue"), bty="6")
```

Figura 2.7 y **Figura 2.8**

```
library(FactoMineR)
library(Factoshiny)
datosS<-scale(datosS,T,F)
datosS<-data.frame(datosS)
PCAshiny(datosS)
```

Figura 2.9

```
p.comp <- princomp(datosS,cor=TRUE)
biplot(p.comp,var.axes=TRUE)
```

Figura 2.10

```
par(mfrow= c(1,2))
boxplot(datosS[,1],col = "deepskyblue4")
title("THC")
boxplot(datosS2[,2], col= "firebrick1")
title("AC")
```

Figura 2.11

```
par(mfrow= c(1,2))
boxplot(datosS[,3], col= "deepskyblue4",ylim=c(-.5,0.9))
title("Range")
boxplot(datosS[,4], col= "firebrick1",ylim=c(-.5,0.9))
title("Ratio")
```

Figura 2.12

```
par(mfrow= c(1,2))
boxplot(datosS[,5], col= "deepskyblue4",ylim=c(-6.5,5.5))
title("TF")
boxplot(datosS[,6], col= "firebrick1",ylim=c(-6.5,5.5))
title("CI")
```

Figura 2.13

```
datosS<-scale(datosS,T,T)
scores<-datosS%*%p.comp$loadings
par(mfrow= c(1,1))
plot(scores[,c(1,2)],col = "deepskyblue4",pch = 16)
```

El documento “climate.csv” corresponde a la información de la tabla 3.2

Figura 3.1

```
datosM2<-read.csv("climate.csv",header=TRUE)
rownames(datosM2)<-datosM2[,1]
datosM2<-datosM2[,-1]
datosM2<-scale(datosM2,T,F)
cor(datosM2)
corrplot(cor(datosM2), type = "upper", order = "hclust", tl.col =
    "black", tl.srt = 90,addgrid.col = "black")
```

Figura 3.2

```
acp <- prcomp(datosM2, scale = TRUE)
valores.propios <- acp$sdev^2
cumsum(valores.propios / sum(valores.propios) * 100)
vp_porcentaje<-round((valores.propios / sum(valores.propios) *
   100),8)
names(valores.propios)<-seq(1:17)
barplot(valores.propios, main="Criterio de Kaiser-Guttman" ,xlab=
   "Componentes",ylab = "Valores propios")
abline(h=mean(valores.propios),col="red")
legend("topright", c("lambda = 1"),
       pch=15, col=c("red"),lty=1,box.lty = 0)
```

Figura 3.3

```
plot(acp$sdev^2, type = "o", pch = 19, main = "Grafica de codo",
   ylab = "Valores Propios", xlab = "Componente",xlim = c(1,17))
```

Figura 3.4

```
lambdaEst<-function(n)
{
  lambdaest<-0
  for(i in 1:n){
    lambdaest<-lambdaest+(1/(n-i+1))
  }
  return(lambdaest)
}
estimaciones<-c(rep(0,ncol(datosM2)))
for (i in 1:(ncol(datosM2))-1){
  estimaciones[1]<-lambdaEst(ncol(datosM2))
  estimaciones[i+1]<-lambdaEst(ncol(datosM2))-lambdaEst(i)
}
plot(acp$sdev^2, type = "o", pch = 19, main = "Comparacion de los
    valores propios", ylab = "Varianzas", xlab = "Componentes",
   ylim=c(0,6),xlim=c(1,14))
par(new=TRUE)
plot(estimaciones, type = "o", pch = 19, main = "Comparacion de
   los valores propios", ylab = "Varianzas", xlab = "Componentes"
   , ylim=c(0,6),xlim=c(1,14),col="darkblue")
legend("topright", c("Grafica de codo", "Modelo de palo roto"),
       pch=15, col=c("black","darkblue"), bty="6")
```

Figura 3.5

```
barplot(t(cbind(acp$sdev^2, estimaciones)), beside=TRUE,
```

```
        main="Valores propios vs Valores propios estimados", col=
           c("grey","darkblue"), las=2, xlab = "Componentes",ylab
            = "Varianzas")
legend("topright", c("Valores Propios", "Modelo de palo roto"),
       pch=15, col=c("grey","darkblue"), bty="6")
```

Figura 3.6

```
datosM2<-data.frame(datosM2)
pca.Bac=PCA(datosM2,scale.unit=TRUE,graph=TRUE)
plot( pca.Bac$ind$coord[ ,1],  pca.Bac$ind$coord[ ,2], col = "
   darkred",
      xlab = "Componente 1", ylab = "Componente 2")
abline(h = 0)
abline(v = 0)
rcolor<-c(rep("blue",14),rep("red",9))
text( pca.Bac$ind$coord[ ,1],  pca.Bac$ind$coord[ ,2], labels = (
   rownames(pca.Bac$ind$coord)), col = rcolor)
```

Figura 3.7

```
res.PCA<-PCA(datosM2,quali.sup=NULL,quanti.sup=NULL,ind.sup=NULL,
   scale.unit=TRUE,graph=FALSE,ncp=5)
plot.PCA(res.PCA,axes=c(1,2),choix='var',select=NULL,cex=1,cex.
   main=1,cex.axis=1,title='Biplot',unselect=0,col.quanti.sup='
   blue',col.var='#000000')
```

Figura 3.8

```
plot( pca.Bac$ind$coord[ ,2],  pca.Bac$ind$coord[ ,3], col = "
   darkred",
      xlab = "Componente 2", ylab = "Componente 3")
abline(h = 0)
abline(v = 0)
rcolor<-c(rep("blue",14),rep("red",9))
text( pca.Bac$ind$coord[ ,2],  pca.Bac$ind$coord[ ,3], labels = (
   rownames(pca.Bac$ind$coord)), col = rcolor)
```

Figura 3.9

```
plot.PCA(res.PCA,axes=c(2,3),choix='var',select=NULL,cex=1,cex.
   main=1,cex.axis=1,title='Biplot',unselect=0,col.quanti.sup='
   blue',col.var='#000000')
```

Figura 3.10

```
plot( pca.Bac$ind$coord[ ,2],  pca.Bac$ind$coord[ ,4], col = "
   darkred",
```

```
      xlab = "Componente 2", ylab = "Componente 4")
abline(h = 0)
abline(v = 0)
rcolor<-c(rep("blue",14),rep("red",9))
text( pca.Bac$ind$coord[ ,2],  pca.Bac$ind$coord[ ,4], labels = (
   rownames(pca.Bac$ind$coord)), col = rcolor)
```

Figura 3.11

```
plot.PCA(res.PCA,axes=c(2,4),choix='var',select=NULL,cex=1,cex.
   main=1,cex.axis=1,title='Biplot',unselect=0,col.quanti.sup='
   blue',col.var='#000000')
```

A.2. Resultados no gráficos del problema cromosómico de leguminosas

```
# Leemos la información
datosS<-read.csv("matriz1.csv",header=TRUE)
rownames(datosS)<-datosS[,1]
# Centramos la matriz de datos
datosS<-scale(datosS,T,F)
# Matriz de varianzas y covarianzas
var(datosS)
# Matriz de correlación muestral
cor(datosS)
# Medida global de dependencia
mg <- 1-det(cor(datosS))
# Análisis de componentes principales
p.comp <- princomp(datosS,cor=TRUE)
# Valores propios
valores.propios <- p.comp$sdev^2
# Proporción de varianza
vp_porcentaje<-round((valores.propios / sum(valores.propios) *
   100),8)
# Varianza acumulada
var_acum <- cumsum(valores.propios / sum(valores.propios) * 100)
# Criterio del palo roto
lambdaEst<-function(n)
{
    lambdaest<-0
    for(i in 1:n){
    lambdaest<-lambdaest+(1/(n-i+1))
    }
```

```
    return(lambdaest)
}

estimaciones<-c(lambdaEst(6),lambdaEst(6)-lambdaEst(1),lambdaEst
   (6)-lambdaEst(2),lambdaEst(6)-lambdaEst(3),lambdaEst(6)-
   lambdaEst(4),lambdaEst(6)-lambdaEst(5))
p.comp$sdev^2
# Procedimiento Bootstrap
 p<-ncol(datosS)
 n<-nrow(datosS)[1]

matriz_r<-matrix(nrow=10000*n,ncol=6)
vp.acum_bootstrap<-matrix(nrow = 10000,ncol=6)
set.seed(1)
for(i in 1:(n*10000))
{
 matriz_r[i,]<-datosS[sample(1:(n),1,FALSE),]
}
for(i in 1:10000)
{
  p.comp.bootstrap<-princomp(matriz_r[(1*i):(n*i),])
  x2<-cor(matriz_r[(1*i):(n*i),])
  vp_boots<-eigen(x2)
  vp.acum_bootstrap[i,]<-round(vp_boots$values / sum(
     vp_boots$values) * 100,8)
}
save(vp.acum_bootstrap,file="vp.acum_bootstrap.RData")
# Estimación Bootstrap
est.bootstrap<-round(apply(vp.acum_bootstrap, 2, mean),8)
# Desviación estándar
err_est.boot<-vector(mode = 'numeric',length = p)
for(i in 1:p){
  err_est.boot[i]<-round(sqrt(sum((vp.acum_bootstrap[,i]-est.
     bootstrap[i])^2)/((10000-1))),8)
}
# Matriz de correlación
x3<-cor(datosS)
# Obtención de vectores y valores propios
vp<-eigen(x3)
# Porcentaje de variación de cada componente
vp_porcentaje<-round(vp$values / sum(vp$values) * 100,8)
# Estadísticas T
h0<-(est.bootstrap-vp_porcentaje)/err_est.boot

# Procedimiento Jackknife
componentesN<-matrix(nrow = n, ncol = p)
```

```
pseudovalores<-matrix(nrow = n, ncol = p)

set.seed(2)
for(i in 1:n){
  datosn1<-datosS
  datosn1<-datosn1[-i,]
  p.comp <- princomp(datosn1,cor=TRUE);
  valores.propios <- p.comp$sdev^2;
  componentesN[i,]<-round(valores.propios / sum(valores.propios)
      * 100,8)
  pseudovalores[i,]<-(n*vp_porcentaje)-(n-1)*componentesN[i,]
}
# Estimación Jackknife
estimadorijackknife<-round(apply(pseudovalores, 2, mean),8)
# Desviación estandar
err_est<-vector(mode = 'numeric',length = p)
for(i in 1:p){
  err_est[i]<-round(sqrt(sum((pseudovalores[,i]-
      estimadorijackknife[i])^2)/(n*(n-1))),8)
}
# Estadísticas T
h0jackk<-(estimadorijackknife-vp_porcentaje)/err_est
```

A.3. Resultados no gráficos del problema del clima

```
# Leemos la información
datosM2<-read.csv("climate.csv",header=TRUE)
rownames(datosM2)<-datosM2[,1]
datosM2<-datosM2[,-1]
# Centramos la matriz de datos
datosM2<-scale(datosM2,T,F)
# Matriz devarianzas y covarianzas
var(datosM2)
# Matriz de correlación muestral
cor(datosM2)
# Medida global de dependencia
det(cor(datosM2))
# Análisis de componentes principales
acp <- prcomp(datosM2, scale = TRUE)
# Valores propios
valores.propios <- acp$sdev^2
# Proporción de varianza
vp_porcentaje<-round((valores.propios / sum(valores.propios) *
    100),8)
```

```
# Varianza acumulada
cumsum(valores.propios / sum(valores.propios) * 100)
# Criterio de palo roto
lambdaEst<-function(n)
{
  lambdaest<-0
  for(i in 1:n){
    lambdaest<-lambdaest+(1/(n-i+1))
  }
  return(lambdaest)
}

estimaciones<-c(rep(0,ncol(datosM2)))
for (i in 1:(ncol(datosM2))-1){
  estimaciones[1]<-lambdaEst(ncol(datosM2))
  estimaciones[i+1]<-lambdaEst(ncol(datosM2))-lambdaEst(i)
}
estimaciones
acp$sdev^2

# Procedimiento Bootstrap
p<-ncol(datosM2)
n<-dim(datosM2)[1]
matriz_r<-matrix(nrow=10000*n,ncol=p)
vp.acum_bootstrap<-matrix(nrow = 10000,ncol=p)
set.seed(1)
for(i in 1:(n*10000))
{
  matriz_r[i,]<-datosJackk[sample(1:(n),1,FALSE),]
}
 for(i in 1:10000)
 {
   p.comp.bootstrap<-princomp(matriz_r[(1*i):(n*i),])
   x2<-cor(matriz_r[(1*i):(n*i),])
   vp_boots<-eigen(x2)
   vp.acum_bootstrap[i,]<-round(vp_boots$values / sum(
      vp_boots$values) * 100,8)
 }
save(vp.acum_bootstrap,file="vp.acum_bootstrap.RData")
load("vp.acum_bootstrap.RData")
# Estimación Bootstrap
est.bootstrap<-round(apply(vp.acum_bootstrap, 2, mean),8)
# Desviación estándar
err_est.boot<-vector(mode = 'numeric',length = p)
for(i in 1:p){
```

```
  err_est.boot[i]<-round(sqrt(sum((vp.acum_bootstrap[,i]-est.
     bootstrap[i])^2)/((10000-1))),8)
}
# Matriz de correlación
x3<-cor(datosM2)
# Obtención de vectores y valores propios
vp<-eigen(x3)
# Porcentaje de variación de cada componente
vp_porcentaje<-round(vp$values / sum(vp$values) * 100,8)
# Estadísticas T
h0<-(est.bootstrap-vp_porcentaje)/err_est.boot

# Procedimiento Jackknife

componentesN<-matrix(nrow = n, ncol = p)
pseudovalores<-matrix(nrow = n, ncol = p)

set.seed(2)
componentesN[1,]<-round(vp_porcentaje,8)
pseudovalores[1,]<-(n*vp_porcentaje)-(n-1)*componentesN[1,]

for(i in 1:n){
  datosn1<-datosM2
  datosn1<-datosn1[-i,]
  p.comp <- princomp(datosn1,cor=TRUE);
  valores.propios <- p.comp$sdev^2;
  componentesN[i,]<-round(valores.propios / sum(valores.propios)
     * 100,8)
  pseudovalores[i,]<-(n*vp_porcentaje)-(n-1)*componentesN[i,]
}
# Estimación Jackknife
estimadorijackknife<-round(apply(pseudovalores, 2, mean),8)
# Desviación estándar
err_est<-vector(mode = 'numeric',length = p)
for(i in 1:p){
  err_est[i]<-round(sqrt(sum((pseudovalores[,i]-
     estimadorijackknife[i])^2)/(n*(n-1))),8)
}
# Estadística t
h0jackk<-(estimadorijackknife-vp_porcentaje)/err_est
```

Bibliografía

[1] Carroll J. D., Green P. E. & Chaturvedi A. (1997). *Mathematical Tools for Applied Multivariate Analysis*. Academic Press.

[2] De la Fuente, F. Santiago (2011b) *Análisis Componentes Principales-ACP. Material de apoyo docente, Santiago de la Fuente Fernández, Fac. Económica y de Ciencias Empresariales, Univ. Autónoma de Madrid (UAM).*

[3] Digby P. G. N. & Kempton R. A.(1987). *Multivariate Analysis of Ecological Communities*. Champman and Hall: London, New York.

[4] Everitt B. & Hothorn Torsten (2011). *An Introduction to Applied Multivariate Analysis with R*. Springer, New York.

[5] Gentleman, R., Hornik, K., & Parmigiani, G. (2011). Use R!.

[6] Greenacre, M., & Primicerio, R. (2014). *Multivariate analysis of ecological data*. Fundacion BBVA.

[7] Tapia-Pastrana, F., Delgado-Salinas, A., & Caballero, J. (2020). Patterns of chromosomal variation in Mexican species of Aeschynomene (Fabaceae, Papilionoideae) and their evolutionary and taxonomic implications. *Comparative Cytogenetics, 14*(1), 157–182

[8] Jeffers, J. N. R. (1967). Two case studies in the application of principal component analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 16*(3), 225-236.

[9] Lang, S. (1976). *Introducción al álgebra lineal*. Addison-Wesley Iberoamericana.

[10] McGarigal K., Cushman S. & Stafford S. (2000). *Multivariate Statistics for Wildlife and Ecology Research*. Springer, New York.

[11] Oksanen, J. (2004). Multivariate Analysis in Ecology-Lecture Notes. *Department of Biology, University of Oulu.*

[12] Peña, D. (2002). *Análisis de datos multivariantes.* McGraw-Hill.

[13] Rencher A. C. (2002). *Methods of Multivariate Analysis.* Wiley-Interscience. John Wiley & Sons, Inc. Publication.