

**UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS**

**FACULTAD DE CIENCIAS MATEMÁTICAS**

**E.A.P. DE ESTADÍSTICA**

**Representaciones gráficas de variables principales  
comunes: una aplicación a la botánica**

**TESIS**

Para optar el Título Profesional de Licenciado en Estadística

**AUTOR**

Carlos Díaz Gómez

**ASESOR**

Doris Gómez Ticerán

Lima - Perú

2011

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

---

**Dedico este trabajo  
A mis padres: Pablo y  
Fortunata, y mi hermano:  
Leonardo, grandes  
inspiradores de mi vida.**

## **AGRADECIMIENTO**

Agradezco a Dios, por darme la fuerza necesaria para cumplir mis objetivos, ya que el presente trabajo de investigación es el resultado de un gran esfuerzo que tiene por objetivo apoyar e incentivar a la investigación a todos aquellos estudiantes que quieran seguir investigando.

Por tal motivo, es preciso mencionar que la elaboración de este trabajo demandó tiempo, trabajo y dedicación; así como también necesitó de la cooperación de parte de algunas personas, y aquí expreso mi agradecimiento:

A mis padres, por todo su apoyo y dedicación en todo momento y cada instante de mi vida.

A la Dra. Doris Gómez Ticerán, mi asesora, por su orientación permanente y darme las herramientas necesarias para llevar a cabo la elaboración de mi trabajo de investigación, por brindarme su confianza, su consejo, su paciencia, apoyo incondicional y amistad.

Al Mg. Yakov Quinteros Gómez, estudiante de Doctorado de la Facultad de Ciencias Biológicas de la Universidad Nacional Mayor de San Marcos por permitirme el uso de los datos preliminares de su trabajo doctoral.

A la Facultad de Matemática, Escuela de Estadística y a sus profesores por formarme en esta carrera profesional.

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

---

**REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES  
COMUNES: UNA APLICACIÓN A LA BOTÁNICA**

Tesis presentada a consideración del Cuerpo Docente de la Facultad de Ciencias Matemáticas,  
de la Universidad Nacional Mayor de San Marcos, como parte de los requisitos para optar el  
Título Profesional de Licenciado en Estadística

Aprobada por:

Mg. Wilfredo Domínguez Cirilo  
Presidente

---

Dr. Erwin Kraneao  
Miembro

---

Dra. Doris Gómez Ticerán  
Miembro - Asesor

---

Lima – Perú  
Diciembre 2011

**RESUMEN**  
**REPRESENTACIONES GRÁFICAS DE VARIABLES**  
**PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA**

PRESENTADO POR: Bachiller. DÍAZ GÓMEZ, Juan Carlos

DIRIGIDO POR: Dra. GÓMEZ TICERÁN, Doris

DICIEMBRE 2011

El objetivo del presente trabajo de investigación, es estudiar en detalle la teoría y la obtención de las variables componentes principales comunes y mostrar sus correspondientes representaciones gráficas bidimensionales en  $R^2$ . Previamente, se presentan aspectos básicos del modelo de componentes principales para un grupo, la cual desde la perspectiva teórica, según Flury (1984), se generaliza al análisis de componentes principales comunes, este método se basa en la búsqueda de una estructura común, una rotación (común), que diagonalice las matrices de covarianza de los datos originales simultáneamente en todas las poblaciones. Luego se concluye el trabajo con una aplicación a la Botánica realizados con las muestras de un estudio desarrollado en 4 lugares del Departamento de Loreto.

**Palabras Claves**

Componentes Principales, Componentes Principales Comunes.

---

**ÍNDICE**

<b>ÍNDICE</b> .....	6
<b>INTRODUCCIÓN</b> .....	8
 <b>CAPÍTULO I. ANTECEDENTES: MODELO DE COMPONENTES PRINCIPALES PARA UN GRUPO</b>	
1.1. Introducción.....	13
1.2. Objetivo de la Investigación.....	14
1.3. Método de obtención de las combinaciones lineales.....	14
Componentes principales	
1.4. Estimación de las Componentes Principales.....	21
 <b>CAPÍTULO II. MODELOS DE COMPONENTES PRINCIPALES COMUNES</b>	
2.1. Introducción.....	25
2.2. Estimadores Máximo verosimilitud bajo el modelo.....	29
Componentes Principales Comunes	
2.3. Combinaciones lineales denominados.....	35
Componentes Principales Comunes	
2.4. Representaciones gráficas.....	42
 <b>CAPÍTULO III. CARACTERIZACION DE LOS AGUAJES DEL DEPARTAMENTO DE SAN DE SAN MARTÍN MEDIANTE COMPONENTES PRINCIPALES COMUNES</b>	
3.1. Introducción.....	44
3.2. Descripción de los datos.....	46
3.3. Resultados y discusión.....	47
3.4. Representaciones gráficas de las.....	56
Componentes Principales Comunes	

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

---

<b>CONCLUSIONES</b> .....	64
<b>REFERENCIAS BIBLIOGRÁFICAS</b> .....	66
<b>APÉNDICE</b> .....	68
A. Anexo 1.....	68
Datos Transformados, aplicando el método de Componentes Principales Comunes.	
B. Anexo 2.....	73
Se presenta las muestras de las dos primeras Componentes Principales Comunes para los Cuatro grupos.	
C. Anexo 3.....	78
Programa en R la cual se hizo todos los cálculos para llegar a los resultados óptimos	

## **INTRODUCCION**

Cuando el estudio se lleva a cabo en una única población con muchas variables, contenidas en el vector multivariante  $\vec{X}$ , y lo que se quiere es tener una representación parsimoniosa de los datos que facilite la interpretación y la obtención de conclusiones; la metodología de Componente Principales (CP), que mediante la descomposición espectral de la matriz de covarianzas reduce la dimensión original de los datos,  $p$ , en una dimensión menor,  $q$ , ( $q < p$ ), preservando la variabilidad del problema original, es usada con mucha frecuencia y éxito para el propósito descrito (Mardia, 1979; Manly, 2005; Flury, 1984, entre otros).

Si la investigación involucra dos poblaciones,  $k = 2$ , donde previamente se comprueba que las matrices de covarianzas de las poblaciones concurrentes, son diferentes, y el objetivo es clasificar nuevos individuos en una de las poblaciones, se puede abordar el problema mediante la metodología de Análisis de Componentes Principales Generalizados (Flury, 1983; Gómez, 2002). Se llegan a construir las variables Componentes Principales Generalizadas como una emulación de las Componentes Principales en un solo grupo, de las que se escogen las de varianzas extremas, pasando de una dimensión multivariante, ( $p > 2$ ), a un espacio de dimensión  $p = 2$ , donde las interpretaciones y las representaciones gráficas de los datos originales se simplifican.

Cuando la investigación involucra varias poblaciones ( $k \geq 3$ ) y se desea clasificar nuevos organismos o individuos a una de esas poblaciones pre establecidas, el análisis discriminante o el análisis de varianza multivariante son utilizadas para resolver el problema, para lo que se supone igualdad de las matrices de covarianzas de las diversas poblaciones (Mardia, 1976, entre otros) y si este supuesto no se verifica, es usual estimar las matrices de covarianza de los grupos por separado, lo que conlleva a un número excesivo de parámetros, atentando contra el principio de parsimonia.

Se observa frecuentemente que en muchas aplicaciones, en el campo biológico en particular, el supuesto de homoscedasticidad no es satisfactorio ya que las matrices de escala de las  $k$  poblaciones pueden presentar algún tipo de estructura común, como lo señaló Flury



## “REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”

Carlos Díaz Gómez

---

(1987). Por otro lado, el principio de parsimonia sugiere que se introduzcan parámetros solamente cuando sea necesario.

Varios autores han estudiado distintas formas de modelar matrices de covarianzas simultáneamente bajo ciertas restricciones. Flury (1988) proporciona una revisión de distintos modelos para matrices de covarianza de varias poblaciones independientes.

Una generalización de la igualdad de matrices de escala consiste en suponer que las matrices de covarianzas son proporcionales:

$\Sigma_k = \rho_k \Sigma_1 \quad 1 \leq k \leq g \quad \rho_1 = 1$ , donde  $\Sigma_k$  es la matriz de covarianzas de la  $k$ -ésima población.

El modelo de Componentes Principales Comunes (CPC) generaliza este supuesto considerando que las matrices de covarianzas son conmutables, es decir, que tienen distintos autovalores pero iguales autovectores, o sea:

$$\Sigma_k = B \Lambda_k B^T \quad 1 \leq k \leq g,$$

donde  $\Lambda_k$  son matrices diagonales y  $B$  es una matriz ortogonal de autovectores comunes. Este modelo puede verse como una generalización del modelo de Componentes Principales (CP) a  $g$  grupos, ya que la transformación principal (autovectores) es la misma en todas las poblaciones consideradas, mientras las varianzas asociadas con las direcciones principales varían entre los grupos.

Como en el análisis de CP en un grupo, el modelo de CPC puede usarse para reducir la dimensión de los datos, conservando en lo posible la variabilidad presente en cada una de las poblaciones. Aunque los ejes principales son los mismos para todas las poblaciones, traducido en los autovectores, el porcentaje de variabilidad explicada por los autovalores varía entre las  $g$  poblaciones.

El modelo de CPC fue introducido por Flury (1984), quien describió cómo obtener los estimadores máximo verosímiles de  $B$  y  $\Lambda_k$  y cómo docimar la hipótesis del modelo de

## “REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”

Carlos Díaz Gómez

---

CPC suponiendo que las variables son conjuntamente normales multivariantes. La distribución asintótica de los estimadores máximo verosímiles fue estudiada por Flury (1986). En Flury (1988) se analizan y discuten varios ejemplos, incluyendo conjuntos de datos biológicos.

En muchas aplicaciones biométricas, las CP (en un grupo o una población) se interpretan como factores independientes que determinan el crecimiento, tamaño o forma de partes de las plantas motivo de estudio, como pueden ser los frutos de las plantas. Es natural entonces considerar un modelo en el cual los mismos factores aparezcan en distintos géneros o, en los mismos géneros que se desarrollan en diferentes lugares. El modelo de CPC claramente sirve para tal propósito. Aplicaciones y/o interpretaciones de este modelo pueden verse en Flury (1988), Phillips & Arnold (1999), Houle, D.et.al. (2002) y Figueroa (2009) entre otros.

Desde la perspectiva teórica, según Flury (1984), El Análisis de Componentes Principales Comunes (*ACPC*) es un método de estadística multivariada relativamente reciente, que generaliza los conceptos del análisis de componentes principales en una población (*ACP*), a varios grupos o poblaciones. El método se basa en la búsqueda de una estructura común, una rotación (común), que diagonaliza simultáneamente las matrices de covarianza de los datos originales en todas las poblaciones, a partir de la comparación de las matrices de covarianzas. La hipótesis para la estructura básica común de las matrices de covarianza  $\Sigma_1, \Sigma_2, \dots, \Sigma_g$  (definidas positivas) para  $g$  poblaciones ( $k=1,2,\dots,g$ ) es:  $H_{CPC} : \mathbf{B}'\Sigma_k\mathbf{B} = \Lambda_k$ , donde:  $\mathbf{B} = (\vec{\beta}_1, \vec{\beta}_2, \dots, \vec{\beta}_p)$  es una matriz ortogonal ( $p \times p$ ) de autovectores de  $\Sigma_k$  cuyas columnas se llaman ejes principales comunes y los  $\Lambda_k$ , son las matrices diagonales de autovalores de  $\Sigma_k$ . El símbolo,  $H_{CPC}$ , significa la hipótesis que las matrices de covarianzas de las poblaciones concurrentes siguen el modelo de Componentes Principales Comunes(CPC).

Es decir, se postula la hipótesis que la descomposición de las matrices de covarianzas de los datos originales en sus autovectores y autovalores, en las diversas poblaciones, considera los mismos autovectores (rotación común) o ejes principales comunes

## **“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

---

para todas las poblaciones y que solo son diferentes los autovalores en las diversas poblaciones.

La Estadística se ha convertido en un instrumento al servicio de diversas áreas del conocimiento humano, es una ciencia interdisciplinaria, con lenguaje predominantemente matemático y que muchos de sus métodos o modelos tienen razón de ser si se sitúan adecuadamente en el contexto de la Biología, Psicología, Economía, etc. Son innumerables los trabajos de investigación que muestran el uso de métodos estadísticos en el análisis de datos en las diversas áreas del conocimiento humano.

La experiencia a través de los últimos años permite observar un crecimiento exponencial en lo que se refiere al desarrollo de una investigación, en particular en las ciencias de la vida, promovida por el interés de utilizar todo el material que aporta la información recogida en la búsqueda de nuevos avances. Estimar parámetros biológicos, cuya característica intrínseca es la variabilidad, nos hacen recurrir a métodos estadísticos multivariantes, que tienen en cuenta esa diferenciabilidad, lo que nos permitirán establecer leyes que expresen el comportamiento de los datos, llevando a conclusiones objetivas y ayudando a la toma de decisiones, Rius F. y Barón, J. (2005).

Por otro lado, la propia naturaleza es de carácter multidimensional. Intervienen e interactúan numerosos aspectos bióticos y abióticos, contribuyendo en la formación de patrones o estructuras espaciales o temporales en las comunidades biológicas. Detectar y describir esas estructuras de los datos, en particular cuando el número de poblaciones concurrentes es mayor a dos (Gómez, 1997) es de interés. Habiendo identificado estructuras que pueden ser similares o no, hacer las representaciones gráficas de los datos multidimensionales en un espacio de dimensión menor, constituyen los principales retos de la utilidad de los métodos multivariantes, en esta área tan importante del conocimiento humano.

En el contexto descrito, el objetivo de la presente tesis de licenciatura es estudiar en detalle la teoría y la obtención de las variables Componentes Principales Comunes y mostrar sus correspondientes representaciones gráficas en el espacio bidimensional. La correspondiente aplicación será hecha con los datos de muestras de mediciones de los

## **“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

---

frutos de aguaje recolectados en marzo del 2009, en los mercados de Mainas, Alto Amazonas Loreto y Requena del Departamento de Loreto, correspondientes a un estudio piloto.

Para tal fin, el trabajo tiene la siguiente estructura. En el primer capítulo se hace una revisión de la metodología Análisis de Componentes Principales en un grupo. En el capítulo 2 se presenta la metodología del Análisis de Componentes Principales Comunes y en el capítulo 3 se ilustra la metodología de Componentes Principales Comunes con las muestras de un estudio exploratorio desarrollado en cuatro lugares del Departamento de Loreto, correspondientes a mediciones de frutos de aguaje en los mercados de Mainas, Alto Amazonas, Loreto y Requena, donde las características evaluadas fueron diámetros del fruto y de la semilla del aguaje.

## **CAPITULO I**

### **ANTECEDENTES: MODELO DE COMPONENTES PRINCIPALES EN UN GRUPO**

#### **1.1 INTRODUCCION**

Los métodos estadísticos multivariantes factoriales, de los cuales forma parte el Análisis de Componentes Principales (ACP), son métodos multivariantes que permiten combinar variables originales u observables ( $X_i$ ) para obtener variables llamadas factores componentes principales, no observables directamente ( $f_i$ ) pero que son de muchos interés para los investigadores de diferentes áreas del conocimiento.

Con estos métodos, al reducir la dimensión de la población original, se gana en la interpretación de los datos con la menor pérdida de información. En este contexto, el ACP es un método multivariante de análisis de datos que permite la reducción de datos de alta dimensión ( $p > 2$ ) en un conjunto de menor dimensión, donde las representaciones gráficas y las interpretaciones de las relaciones entre las variables es mucho más sencilla. Este método de reducción de datos preserva en lo posible la información contenida en los datos originales, utilizando criterios de optimalidad geométrica y/o algebraica.

En general, a partir de los datos originales se hacen transformaciones, de manera que las nuevas variables denominadas Componentes Principales (CP), en el nuevo sistema, tengan propiedades deseables en términos de variancias y relaciones entre ellas.

Esta metodología inicialmente desarrollada por Pearson a finales del siglo XIX, posteriormente fue estudiada por Hotelling en los años 30 del siglo XX y su primera implementación computacional se dio en los años 60.

Cabe resaltar que es un método esencialmente descriptivo y no hay necesidad de postular ninguna estructura probabilística para los datos.

## **1.2 OBJETIVOS**

El objetivo fundamental del Análisis de Componentes Principales es:

*Reducir* la dimensión del problema que se está estudiando a través de ejes explicativos y representaciones gráficas como paso previo para otros análisis, pero también permite:

Construir nuevas variables no correlacionadas que expresen la misma información contenida en el conjunto de datos originales correlacionados. Se divide la variación total del problema original en porciones de varianzas decrecientes.

Comprender mejor las correlaciones entre las variables originales, seleccionando y/o eliminar variables cuando el aporte de información es mínimo.

Construir las variables con máxima y mínima varianza, para luego hacer otros análisis de datos.

Las nuevas variables generadas utilizando una adecuada transformación matemática sobre los datos originales, se denominan componentes principales (CP) y poseen características deseables tales como independencia, cuando se asume multinormalidad, y en todos los casos están no correlacionados. Esto significa que si las variables originales no están correlacionadas, el ACP no ofrece ventaja alguna.

## **1.3 METODO DE OBTENCION DE LAS COMBINACIONES LINEALES COMPONENTES PRINCIPALES**

Se justifica la presentación de este tema de manera detallada porque en el siguiente capítulo se emulará la presente metodología al caso de  $k \geq 3$  grupos.

Supongamos un vector  $p$ -variable,  $\vec{X} = (X_1, \dots, X_p)^T$  con vector de medias  $\mu$  y matriz de covarianzas  $\Sigma$  denotado:  $\vec{X} \sim (\mu, \Sigma)$ , sin pérdida de generalidad suponemos que el vector de medias es el vector cero,  $\mu = \vec{0}$ , puesto que todo análisis se centra en la matriz de varianzas y covarianzas o en la matriz de correlaciones.

## “REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”

Carlos Díaz Gómez

El método de obtención consiste en transformar las  $p$  variables  $X_1, \dots, X_p$  contenidas en el vector  $\vec{X} = (X_1, \dots, X_p)^T$ , mediante una matriz ortogonal  $B$ , en  $p$  nuevas variables  $\vec{Y} = (Y_1, \dots, Y_p)^T$  no correlacionadas. Luego, se escogen los  $Y_j$  variables,  $j = 1, \dots, q < p$  que contengan la máxima variabilidad de las  $X_1, \dots, X_p$ ,  $j = 1, \dots, p$ ; para lo cual se procede de la siguiente manera.

Se busca la combinación lineal  $\vec{Y} = \vec{b}' \vec{X}$ , donde  $\vec{b}' \in \mathfrak{R}^p$  y  $\vec{b}' \neq \vec{0}$ , cuya varianza  $Var(Y) = \vec{b}' \Sigma \vec{b}$  sea máxima. Maximizar  $Var(Y) = \vec{b}' \Sigma \vec{b}$ , es equivalente a maximizarla sujeta a la restricción  $\vec{b}' \vec{b} = 1$ , es decir, se busca maximizar la función  $\phi_1$ , donde:

$$\phi_1 = \vec{b}' \Sigma \vec{b} - \lambda (\vec{b}' \vec{b} - 1) \quad \text{donde } \lambda \text{ es el multiplicador de Lagrange.}$$

derivando con respecto a  $\vec{b}$ , se tiene:

$\frac{\partial \phi_1}{\partial \vec{b}} = 2 \Sigma \vec{b} - 2 \lambda \vec{b}$ , como  $\vec{b}' \Sigma \vec{b}$  y  $\vec{b}' \vec{b}$  tienen derivadas en todas las regiones que contiene  $\vec{b}' \vec{b} = 1$ , el vector que maximiza

$$\vec{b}' \Sigma \vec{b} \text{ debe satisfacer: } \begin{cases} 2 \Sigma \vec{b} - 2 \lambda \vec{b} = \vec{0} \\ \Rightarrow (\Sigma - \lambda I) \vec{b} = \vec{0} \end{cases}$$

Para que esta expresión tenga solución, sujeta a la restricción  $\vec{b}' \vec{b} = 1$ ,  $(\Sigma - \lambda I)$ , debe ser singular, es decir  $\lambda$  debe satisfacer

$$|\Sigma - \lambda I| = 0 \tag{1}$$

$|\Sigma - \lambda I| = 0$  es un polinomio en  $\lambda$  de grado  $p$  entonces la solución de la ecuación (1) tiene  $p$  raíces  $\lambda_1, \lambda_2, \dots, \lambda_p$  tales que  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$  que no son otra cosa que las raíces características o autovalores de la matriz de covarianzas  $\Sigma$ , de las variables  $X_1, X_2, \dots, X_p$  que es una matriz definida positiva.

## “REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”

Carlos Díaz Gómez

Multiplicando la ecuación  $(\Sigma - \lambda I)\vec{b} = \vec{0}$  por  $\vec{b}'$  y usando la restricción  $\vec{b}'\vec{b} = 1$ , se tiene:

$$(\Sigma - \lambda I)\vec{b} = \vec{0}$$

$$\vec{b}'(\Sigma - \lambda I)\vec{b} = \vec{b}'\vec{0} = 0$$

$$\Rightarrow \vec{b}'\Sigma\vec{b} = \lambda\vec{b}'\vec{b} = \lambda$$

$$\Rightarrow \text{Var}(Y) = \text{Var}(\vec{b}'\vec{X}) = \vec{b}'\Sigma\vec{b} = \lambda$$

Para maximizar esta varianza se usa el máximo autovalor de  $\Sigma$ , es decir  $\lambda_1 \Rightarrow (\Sigma - \lambda_1 I)\vec{b} = \vec{0}$  tal que  $\vec{b}'\vec{b} = 1$  lo cual dará el vector propio de norma uno  $\vec{\beta}_1$ , con el que se define la combinación lineal de máxima varianza, es decir la primera componente principal.

$$Y_1 = \vec{\beta}_1'\vec{X}$$

que cumple

$$\vec{\beta}_1'\vec{\beta}_1 = 1 \quad \text{Var}(Y_1) = \vec{\beta}_1'\Sigma\vec{\beta}_1 = \lambda_1 \quad \Sigma\vec{\beta}_1 = \lambda_1\vec{\beta}_1.$$

El siguiente paso es obtener la combinación lineal  $\vec{Y} = \vec{b}'\vec{X}$  normalizada según  $\vec{b}'\vec{b} = 1$  que tiene mayor varianza entre todas las combinaciones lineales no correlacionadas con  $Y_1$ , o sea buscar la combinación lineal  $Y$  que cumple:

$$\Sigma\vec{\beta}_1 = \lambda_1\vec{\beta}_1$$

$$\text{Cov}(\vec{Y}, Y_1) = \text{Cov}(Y_1, \vec{Y}) = \text{Cov}(Y_1, \vec{b}'\vec{X})$$

$$= \text{Cov}(\vec{\beta}_1'\vec{X}, \vec{b}'\vec{X}) = \vec{\beta}_1'\Sigma\vec{b}$$

$$= \vec{b}'\Sigma\vec{\beta}_1 = \vec{b}'\lambda_1\vec{\beta}_1 = \lambda_1\vec{b}'\vec{\beta}_1 = 0$$



**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

El vector  $\vec{b}'\vec{X}$  es ortogonal a  $Y_1$  en el sentido estadístico y geométrico. La ortogonalidad en el sentido estadístico, está expresada en la falta de correlación y en el sentido geométrico en el producto interno  $\vec{b}'\vec{\beta}_1 = 0$

Esto es  $\lambda_1 \vec{b}'\vec{\beta}_1 = 0 \Leftrightarrow \vec{b}'\vec{\beta}_1 = 0$  cuando  $\lambda_1 = 0$  y  $\lambda_1 \neq 0$  si  $\Sigma > 0$ . Es decir, se busca la combinación lineal  $Y$ , con máxima varianza  $\vec{b}'\Sigma\vec{b}$  sujeto a  $\vec{b}'\vec{b} = 1$  y  $\vec{b}'\Sigma\vec{\beta}_1 = 0$ , lo que es equivalente a maximizar  $\phi_2$ ,  $\phi_2 = \vec{b}'\Sigma\vec{b} - \lambda(\vec{b}'\vec{b} - 1) - 2\nu_1(\vec{b}'\Sigma\vec{\beta}_1 - 0)$  donde:  $\lambda$  y  $\nu_1$  son multiplicadores de lagrange, así:

$$\frac{\partial \phi_2}{\partial \vec{b}} = 2\Sigma\vec{b}' - 2\lambda\vec{b}' - 2\nu_1\Sigma\vec{\beta}_1 = 0$$

multiplicando por  $\vec{\beta}_1'$ , se tiene: (2)

$$2\vec{\beta}_1'\Sigma\vec{b}' - 2\lambda\vec{\beta}_1'\vec{b}' - 2\nu_1\vec{\beta}_1'\Sigma\vec{\beta}_1 = 0$$

Usando los siguientes resultados:  $\vec{\beta}_1'\Sigma\vec{\beta}_1 = 1$ ,  $\vec{\beta}_1'\vec{b}' = 0$  y  $\vec{b}'\Sigma\vec{\beta}_1 = 0$ , la ecuación anterior se reduce a  $\nu_1 = 0$ , luego reemplazando este valor en (2) se tiene que  $\vec{b}'\Sigma\vec{b}' = \lambda$ .

Entonces debemos considerar  $\lambda_2$  y  $\vec{\beta}_2$  con lo cual, se define la segunda componente principal.

$$Y_2 = \vec{\beta}_2'\vec{X} \quad \text{tal que} \quad \vec{\beta}_2'\vec{\beta}_2 = 1$$

$$\text{Var}(Y_2) = \vec{\beta}_2'\Sigma\vec{\beta}_2 = \lambda_2 \quad \vec{\beta}_2'\Sigma\vec{\beta}_1 = 0$$

$$\Sigma\vec{\beta}_2 = \lambda_2\vec{\beta}_2$$

Continuando con el mismo procedimiento, supongamos definidas las  $q$  primeras componentes, de la siguiente manera:

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

$$Y_i = \vec{\beta}_i' \vec{X} \quad 1 \leq i \leq q < p, \text{ que satisfacen:}$$

$$\vec{\beta}_i' \vec{\beta}_i = 1$$

$$\text{Var}(Y_i) = \vec{\beta}_i' \Sigma \vec{\beta}_i = \lambda_i$$

$$\text{Cov}(Y_i, Y_j) = \vec{\beta}_i' \Sigma \vec{\beta}_j = 0 \quad i \neq j$$

$$\vec{\beta}_i' \vec{\beta}_j = 0$$

Se busca la  $(q+1)$ -ésima componente principal  $\vec{Y} = \vec{b}' \vec{X}$  con máxima varianza entre todas las combinaciones lineales no correlacionadas con las  $q$  primeras componentes. Es decir, debe maximizarse  $\vec{b}' \Sigma \vec{b}$  sujeto a:

$$\vec{b}' \vec{b} = 1 \quad \vec{b}' \Sigma \vec{\beta}_i = 0 \quad \vec{b}' \vec{\beta}_i = 0 \quad \Sigma \vec{\beta}_i = \lambda \vec{\beta}_i \quad (3)$$

O sea, buscar a  $\vec{Y} = \vec{b}' \vec{X}$ , un máximo  $\phi_{q+1}$

$$\phi_{q+1} = \vec{b}' \Sigma \vec{b} - \lambda (\vec{b}' \vec{b} - 1) - 2 \left[ \sum_{i=1}^q v_i \vec{b}' \Sigma \vec{\beta}_i - 0 \right]$$

donde:

$\lambda$  y  $v_i$ ,  $i = 1, \dots, q$  son multiplicadores de lagrange.

Derivando la ecuación anterior respecto a  $\vec{b}$  y multiplicando por  $\vec{\beta}_j'$ , ( $1 \leq j \leq q$ ) e igualando a cero, tenemos:

$$2 \vec{\beta}_j' \Sigma \vec{b} - 2 \lambda \vec{\beta}_j' \vec{b} - 2 \sum_{i=1}^q v_i \vec{\beta}_j' \Sigma \vec{\beta}_i = 0$$

Utilizando las condiciones dadas en (3),  $v_j = 0$ , entonces  $(\Sigma - \lambda I) \vec{b} = 0 \Rightarrow \vec{b}' \Sigma \vec{b} = 0$ , donde  $\vec{b}$  y  $\lambda$  deben de satisfacer la misma ecuación que las soluciones anteriores  $\vec{\beta}_i$  y  $\lambda_i$  ( $1 \leq i \leq q$ ).

## “REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”

Carlos Díaz Gómez

Observamos que los  $q$  primeros valores propios y sus vectores propios no cumplen la segunda y tercera condición dada en (3).

Por lo tanto los siguientes candidatos son  $\vec{\beta}_{q+1}$  y  $\lambda_{q+1}$  quienes satisfacen:

$$(\Sigma - \lambda_{q+1} I) \vec{\beta}_{q+1} = 0, \text{ multiplicando por } \vec{\beta}_i'$$

$$\vec{\beta}_i' (\Sigma - \lambda_{q+1} I) \vec{\beta}_{q+1} = 0$$

$$\vec{\beta}_i' \Sigma \vec{\beta}_{q+1} = \lambda_{q+1} \vec{\beta}_i' \vec{\beta}_{q+1}$$

$$\vec{\beta}_{q+1}' \Sigma \vec{\beta}_i = \lambda_i \vec{\beta}_{q+1}' \vec{\beta}_i$$

$$\Rightarrow \vec{\beta}_i' \Sigma \vec{\beta}_{q+1} = \lambda_i \vec{\beta}_i' \vec{\beta}_{q+1}$$

Si suponemos  $\vec{\beta}_i' \vec{\beta}_{q+1} \neq 0 \Rightarrow \lambda_i = \lambda_{q+1}$ , resultado que contradice la simplicidad de los autovalores; por tanto, se cumplen las condiciones segunda y tercera de (3). Con todo lo obtenido se asegura que la combinación lineal:

$$Y_{q+1} = \vec{\beta}_{q+1}' \vec{X} \quad 1 \leq i \leq q$$

con

$$\text{Var}(Y_{q+1}) = \vec{\beta}_{q+1}' \Sigma \vec{\beta}_{q+1} = \lambda_{q+1}$$

$$\text{Cov}(Y_{q+1}, Y_i) = \vec{\beta}_{q+1}' \Sigma \vec{\beta}_i = 0$$

$$\vec{\beta}_{q+1}' \vec{\beta}_i = 0$$

es la  $(q+1)$ -ésima componente principal.

Puesto que la matriz  $\Sigma$  es definida positiva, podemos continuar con el mismo procedimiento y en el último paso el valor propio más pequeño  $\lambda_p$  y el vector propio asociado  $\vec{\beta}_p$  son usados para definir la  $p$ -ésima componente principal.

$$Y_p = \vec{\beta}_p' \vec{X}$$

con

$$\text{Var}(Y_p) = \vec{\beta}_p' \Sigma \vec{\beta}_p = \lambda_p$$

$$\text{Cov}(Y_p, Y_i) = \vec{\beta}_p' \Sigma \vec{\beta}_i = 0 \quad 1 \leq i \leq p-1$$

$$\vec{\beta}_p' \vec{\beta}_i = 0$$

Si  $\vec{X}$  es un vector aleatorio con vector de medias  $\vec{\mu}$  y una matriz de covarianzas  $\Sigma$ , entonces las componentes principales son:

$$Y_i = \vec{\beta}_i' \vec{X} \quad i = 1, \dots, p$$

donde

$$\text{Var}(Y_i) = \vec{\beta}_i' \Sigma \vec{\beta}_i = \lambda_i$$

$$\text{Cov}(Y_i, Y_j) = \vec{\beta}_i' \Sigma \vec{\beta}_j = 0$$

$\lambda_i$  son los autovalores asociados a los autovectores  $\vec{\beta}_i$  de la matriz  $\Sigma$ .

Se han obtenido  $p$  nuevas variables  $\vec{Y} = (Y_1, \dots, Y_p)^T$  con matriz de varianzas y covarianzas  $\Lambda$ , donde  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$  y la matriz de covarianza del valor original  $\Sigma$  se ha descompuesto de la siguiente manera:  $\Sigma = B' \Lambda B$

donde:

la matriz de autovectores está representado por

$$B = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdot & \cdot & \beta_{1p} \\ \beta_{21} & \cdot & \cdot & \cdot & \beta_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \beta_{p1} & \beta_{p2} & \cdot & \cdot & \beta_{pp} \end{bmatrix} = (\vec{\beta}_1, \vec{\beta}_2, \dots, \vec{\beta}_p)$$

y la matriz de autovalores es:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & \cdot & 0 & \lambda_p \end{bmatrix}$$

#### 1.4 ESTIMACION DE LAS COMPONENTES PRINCIPALES(CP)

Generalmente los parámetros de la distribución del vector  $\vec{X} = (X_1, \dots, X_p)^T$  son desconocidos, en particular  $\Sigma$ , presentándose el problema de estimar los autovalores  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p (>0)$  y sus autovectores asociados  $\vec{\beta}_1, \vec{\beta}_2, \dots, \vec{\beta}_p$ .

En la práctica, la solución de este problema se enfoca de la siguiente manera:

- a) Obtener una muestra aleatoria multivariada donde cada elemento ha sido extraído al azar de una población de individuos. Sea  $\vec{x}_1, \dots, \vec{x}_n$  una muestra aleatoria desde la población con parámetros media  $\vec{\mu}$  y matriz de covarianzas  $\Sigma$  y cada  $\vec{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T \quad i = 1, 2, \dots, n$  es un vector fila que contiene los valores de las  $p$  variables del  $i$ -ésimo individuo, entonces, se tiene la matriz de datos multivariantes como el arreglo:

$$X = \begin{pmatrix} x_{11} & \cdot & x_{1i} & \cdot & x_{1p} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{i1} & \cdot & \cdot & \cdot & x_{ip} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{n1} & \cdot & x_{ni} & \cdot & x_{np} \end{pmatrix}$$

con

$x_{ij}$  medida de la  $j$ -ésima variable en el  $i$ -ésimo individuo.

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

b) Se estima el vector de medias  $\hat{\vec{\mu}}$ , la matriz de covarianzas  $\hat{\Sigma}$ , y la matriz de correlaciones  $\hat{\rho}$ . Así los estimadores son:

$$\hat{\vec{\mu}} = \vec{\bar{X}} = X^T \frac{1}{n} = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p]^T$$

$$\hat{\Sigma} = S = \left(\frac{1}{n}\right) X' X - \vec{\bar{X}} \vec{\bar{X}}$$

$$\hat{\rho} = \begin{bmatrix} 1 & \hat{\rho}_{12} & \cdot & \cdot & \hat{\rho}_{1p} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \hat{\rho}_{p1} & \hat{\rho}_{p2} & \cdot & \cdot & 1 \end{bmatrix}$$

Si las variables originales están dadas en diferentes unidades de medida o los valores de las variables tienen mucha variación, entonces se recomienda estandarizar los datos previamente y trabajar con la matriz de correlaciones.

c) Habiéndose decidido trabajar con la matriz de correlaciones o de covarianzas, se estiman los autovalores y autovectores que se denotan por  $\hat{\lambda}_i$  y  $\hat{\beta}_i$  respectivamente.

d) Se definen las Componentes Principales  $\hat{Y}_i = \hat{\beta}_i'(\vec{X} - \vec{\bar{X}})$   $i = 1, \dots, n$

Al hacer esta transformación se define una nueva matriz de datos de orden  $n \times p$  que denotamos con  $Y$ .

e) Es importante encontrar las correlaciones entre la  $i$ -ésima variable en la  $j$ -ésima componente principal donde:

$$Var(\hat{Y}_j) = \lambda_j$$

$$Var(X_i) = s_{ii}$$

## “REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”

Carlos Díaz Gómez

---

f) Si  $(Y_1, Y_2, \dots, Y_p)$  son las componentes principales del vector aleatorio  $X$ , tenemos que:

$$\sum_{i=1}^p \text{var}(\hat{Y}_i) = \sum_{i=1}^p \hat{\lambda}_i = \text{traza}(\Lambda) = \text{traza}(\vec{\beta} \Lambda \vec{\beta}') = \text{traza}(\Sigma) = \sum_{i=1}^p s_{ii}$$

El porcentaje de la variabilidad explicado por las  $m$  primeras componentes principales está

expresado por: 
$$\frac{\sum_{i=1}^m \hat{\lambda}_i}{\sum_{i=1}^p \hat{\lambda}_i} \quad m < p$$

Cuando pocas componentes explican gran porcentaje de la varianza total, entonces la estructura de asociación es muy fuerte.

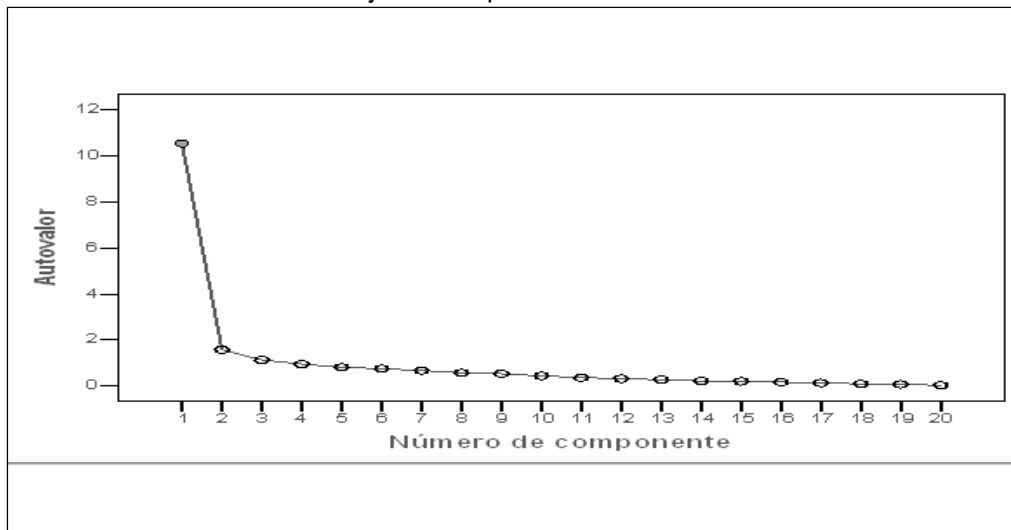
Un problema latente al aplicar la metodología de CP es, decidir cuántas CP retienen en una situación particular, la máxima variabilidad, de manera que se garantice la mínima pérdida de información o dicho de otra manera, se preserve la variabilidad contenida en las variables originales. Uno de los criterios recomienda hacer un gráfico donde en el eje de las abscisas van las componentes principales de la primera a la última y en la ordenada se representa la correspondiente variación explicada por la CP o las varianzas de las CP en forma decreciente (Ver Gráfico 1).

Otro aspecto importante a ser resaltado de las CP en un grupo, es hacer la representación gráfica en  $R^2$  de los datos originales que están en la dimensión  $p > 2$ , de manera que se facilite la lectura de los datos, puesto que su representación gráfica en la dimensión original que corresponde,  $p > 2$ , sería prácticamente imposible de realizar.

# “REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”

Carlos Díaz Gómez

Gráfico 1: Gráfico de las CP y sus respectivas varianzas



Basados en esta ventaja de las CP en un grupo, es que, se ha propuesto emular los gráficos en más de una población, para lo que se ha considerado 4 grupos, o sea,  $k = 4$ , para las muestras de agujajes de los mercados de Mainas, Alto Amazonas, Loreto y Requena del Departamento de Loreto, mediante los datos generados a través de un estudio piloto. Esta propuesta motivo el estudio de la correspondiente teoría, la misma que se presenta a continuación, es decir el Análisis de Componentes Principales Comunes.



## **CAPITULO II**

### **MODELO DE COMPONENTES PRINCIPALES COMUNES (CPC)**

#### **2.1. INTRODUCCIÓN**

Cuando el estudio se lleva a cabo en una única población con muchas variables, contenidas en el vector multivariante  $\vec{X}$ , y lo que se quiere es tener una representación parsimoniosa de los datos que facilite la interpretación y la obtención de conclusiones, la metodología de Componente Principales (CP), que mediante la descomposición espectral de la matriz de covarianzas reduce la dimensión original de los datos,  $p$ , en una dimensión menor,  $q$ , ( $q < p$ ), preservando la variabilidad del problema original, es usada con mucha frecuencia y éxito para el propósito descrito (Mardia, 1979; Manly, 2005; Flury, 1984, entre otros).

En CP para un grupo, los estimadores de las direcciones principales del vector multivariante  $\vec{X}$ , son los autovectores de la matriz de covarianzas muestral.

Cuando la investigación involucra varias poblaciones ( $k \geq 3$ ) y se desea clasificar nuevos organismos o individuos a una de esas poblaciones pre establecidas, el análisis discriminante o el análisis de varianza multivariante son utilizadas para resolver el problema, para lo que se supone igualdad de las matrices de covarianzas de las diversas poblaciones (Mardia, 1976, entre otros) y si este supuesto no se verifica, es usual estimar las matrices de covarianza de los grupos por separado, lo que conlleva a un número excesivo de parámetros, atentando el principio de parsimonia. Se observa frecuentemente que en muchas aplicaciones, en el campo biológico en particular, el supuesto de homoscedasticidad no se satisface ya que las matrices de escala de las  $g$  poblaciones pueden presentar algún tipo de estructura común, como lo señaló Flury (1987). Por otro lado, el principio de parsimonia sugiere que se introduzcan parámetros solamente cuando sea necesario. Varios autores han estudiado distintas formas de modelar matrices de covarianzas simultáneamente bajo ciertas restricciones. Flury (1988) proporciona una revisión de distintos modelos para matrices de covarianza de varias poblaciones independientes.

## “REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”

Carlos Díaz Gómez

---

Una generalización de la igualdad de matrices de escala consiste en suponer que las matrices de covarianzas son proporcionales:

$$\Sigma_k = \rho_k \Sigma_1 \quad 1 \leq k \leq g \quad \rho_1 = 1,$$

donde:

$\Sigma_k$  es la matriz de covarianzas de la  $k$ -ésima población.

El modelo de Componentes Principales Comunes (CPC) generaliza este supuesto considerando que las matrices de covarianzas tienen distintos autovalores pero iguales autovectores, o sea:

$$\Sigma_k = \mathbf{B}' \Lambda_k \mathbf{B} \quad 1 \leq k \leq g,$$

donde

$\Lambda_k$  son matrices diagonales y  $\mathbf{B}$  es una matriz ortogonal de autovectores comunes. Este modelo puede verse como una generalización del modelo de Componentes Principales (CP) en un grupo, a  $g$  grupos, ya que la transformación principal es la misma en todas las poblaciones consideradas, mientras las varianzas asociadas con las direcciones principales varían entre los grupos.

Como en el análisis de CP en un grupo, el modelo de CPC puede usarse para reducir la dimensión de los datos, conservando en lo posible la variabilidad presente en cada una de las poblaciones. Aunque los ejes principales son los mismos para todas las poblaciones, el porcentaje de variabilidad explicada por los autovalores varía entre las  $g$  poblaciones.

El modelo de CPC fue introducido por Flury (1984), quien describió cómo obtener los estimadores máximo verosímiles de las matrices  $\mathbf{B}$  y  $\Lambda_k$  y cómo docimar la hipótesis del modelo de CPC suponiendo que las variables son conjuntamente normales multivariantes. La distribución asintótica de los estimadores máximos verosímiles fue estudiada por Flury (1986). En Flury (1988) se analizan y discuten varios ejemplos, incluyendo conjuntos de datos biológicos.

## “REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”

Carlos Díaz Gómez

---

En muchas aplicaciones biométricas, las CP (en un grupo o una población) se interpretan como factores independientes que determinan el crecimiento, tamaño o forma del organismo motivo de estudio. Es natural entonces considerar un modelo en el cual los mismos factores aparezcan en especies distintas pero relacionadas, o, en las mismas especies que se desarrollan en diferentes lugares. El modelo de CPC claramente sirve para tal propósito. Aplicaciones de este modelo puede verse en Flury (1988) y Phillips & Arnold (1999), Figueroa (2009) entre otros.

El modelo de CPC es uno de los niveles de jerarquía considerados por Flury (1984) quien definió los siguientes niveles de jerarquía entre las matrices de covarianzas.

**Nivel 1:** Las matrices de covarianzas son arbitrarias, es decir no guardan ninguna estructura entre ellas.

**Nivel 2:** Las matrices cumplen el modelo de CPC, o sea,  
 $\Sigma_k = B' \Lambda_k B \quad 1 \leq k \leq g,$

donde

$\Lambda_k$  son matrices diagonales que contienen los autovalores de la matriz de covarianza de cada población y  $B = (\vec{\beta}_1, \vec{\beta}_2, \dots, \vec{\beta}_p)$  es una matriz ortogonal de autovectores comunes a todas las matrices de covarianza, donde:

$$\Sigma_k = \begin{bmatrix} \sigma_{11k} & \sigma_{12k} & \cdot & \cdot & \sigma_{1pk} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \sigma_{p1k} & \sigma_{p2k} & \cdot & \cdot & \sigma_{ppk} \end{bmatrix}$$

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

$\sigma_{ijk}$  : varianzas entre las variables  $x_i$  y  $x_j$  en la población  $k$  .  
 $i, j = 1, \dots, p$   
 $k = 1, \dots, g$

$$B = \begin{bmatrix} \beta_{11} & \beta_{12} & \dots & \beta_{1p} \\ \beta_{21} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \beta_{p1} & \beta_{p2} & \cdot & \beta_{pp} \end{bmatrix}$$

es la matriz de autovectores de  $\Sigma_k$

$$\Lambda_k = \begin{bmatrix} \lambda_{1k} & 0 & \cdot & 0 & 0 \\ 0 & \lambda_{2k} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & \cdot & 0 & \lambda_{pk} \end{bmatrix}$$

matriz de autovalores de la matriz de  $\Sigma_k$  , donde  $\lambda_{ik}$  es el  $i$  –ésimo autovalor en la  $k$  –ésima población.

Es decir,

$$\begin{bmatrix} \sigma_{11k} & \sigma_{12k} & \cdot & \cdot & \sigma_{1pk} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \sigma_{p1k} & \sigma_{p2k} & \cdot & \cdot & \sigma_{ppk} \end{bmatrix} = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdot & \cdot & \beta_{1p} \\ \beta_{21} & \beta_{22} & \cdot & \cdot & \beta_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \beta_{p1} & \beta_{p2} & \cdot & \cdot & \beta_{pp} \end{bmatrix} \begin{bmatrix} \lambda_{1k} & 0 & \cdot & 0 & 0 \\ \cdot & \lambda_{2k} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & \cdot & 0 & \lambda_{pk} \end{bmatrix} \begin{bmatrix} \beta_{11} & \beta_{21} & \cdot & \cdot & \beta_{p1} \\ \beta_{12} & \beta_{22} & \cdot & \cdot & \beta_{p2} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \beta_{1p} & \beta_{2p} & \cdot & \cdot & \beta_{pp} \end{bmatrix}$$

**Nivel 3:** Las matrices de covarianzas son proporcionales entre si, es decir,

$\Sigma_k = \rho_k \Sigma_1 \quad 1 \leq k \leq g \quad \rho_1 = 1$  , donde  $\rho_k$  es una constante de proporcionalidad.

**Nivel 4:**  $\Sigma_k = \Sigma_1 \quad 2 \leq k \leq g$

Por el objetivo de la tesis, a continuación se detalla la metodología para obtener los estimadores máximo verosímiles en el supuesto que se cumple, el modelo de CPC y luego se describe la metodología para docimar la hipótesis que los datos siguen el modelo de CPC.

## **2.2. ESTIMADORES DE MAXIMA VEROSIMILITUD BAJO EL MODELO DE COMPONENTES PRINCIPALES COMUNES**

Como se verá en el siguiente ítem, la hipótesis de *CPC* equivale a probar que las matrices de covarianzas de las poblaciones concurrentes  $\Sigma_1, \Sigma_2, \dots, \Sigma_g$  tienen los ejes principales comunes. Para probar la hipótesis de que los datos siguen el modelo de componentes principales comunes,  $H_{CPC}$ , se usará la estadística de razón de log-verosimilitud (Flury, 1988; Rodríguez M., Boente G. y Pires, A. M. (2005))

$$RV = f(H_0)/f(H_1) \rightarrow -2 \ln \lambda = -2 \ln \frac{L(\Sigma_1, \Sigma_2, \dots, \Sigma_g)}{L(S_1, S_2, \dots, S_g)} \quad (4)$$

donde:

$L(\Sigma_1, \Sigma_2, \dots, \Sigma_g)$  es la función de verosimilitud de  $\Sigma_1, \Sigma_2, \dots, \Sigma_g$  bajo  $H_{CPC}$  y  $L(S_1, S_2, \dots, S_g)$  es la función de verosimilitud el máximo sin restricción. Entonces, se busca encontrar los estimadores máximo verosímiles de los ejes de las componentes principales comunes (*CPC*).

Se tiene las muestras aleatorias independientes de las poblaciones normales multivariantes los -vectores  $\vec{X}_{ik} = (x_{i1k}, x_{i2k}, \dots, x_{ipk})^T$ ,  $i = 1, \dots, n_g$   $k = 1, 2, \dots, g$  tienen distribución normal  $p$ -variante  $\vec{X}_{ik} \sim N_p(\vec{\mu}_k, \Sigma_k)$ , donde  $\vec{\mu}_k \in \mathfrak{R}^p$  y  $\Sigma_k$  son matrices de covarianzas simétricas y definidas positivas.

Para las muestras de tamaño  $N = \sum_{k=1}^g n_k$ ,  $k = 1, 2, \dots, g$ , denotaremos por  $S_k$

( $k = 1, 2, \dots, g$ ) a las matrices de covarianzas muestrales insesgadas, entonces:

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

$M_k = n_k S_k$ , tiene distribución wishart con parámetros  $n_k$  y  $\Sigma_k$  y se denota  $W_p(n_k, \Sigma_k)$

(Mardia, 1979). Por lo que la función de densidad de  $M_k$  es:

$$f(M_k) = \frac{|n_k S_k|^{(n_k-p-1)/2} \exp\left\{-\frac{n_k}{2} \text{tr}(\Sigma_k^{-1} S_k)\right\}}{2^{(n_k p/2)} \Pi^{p(p-1)/4} |\Sigma_k|^{n_k/2} \prod_{j=1}^p \Gamma((n+1-j)/2)} \quad (5)$$

donde la función de verosimilitud de  $\Sigma_1, \Sigma_2, \dots, \Sigma_g$  dado  $S_1, S_2, \dots, S_g$  es:

$$L(\Sigma_1, \Sigma_2, \dots, \Sigma_g) = \prod_{k=1}^g f(M_k) = \prod_{k=1}^g \left\{ \frac{|n_k S_k|^{(n_k-p-1)/2} \exp\left\{-\frac{n_k}{2} \text{tr}(\Sigma_k^{-1} S_k)\right\}}{2^{(n_k p/2)} \Pi^{p(p-1)/4} |\Sigma_k|^{n_k/2} \prod_{j=1}^p \Gamma((n+1-j)/2)} \right\} \quad (6)$$

entonces, la función de verosimilitud común de  $\Sigma_1, \Sigma_2, \dots, \Sigma_g$  dado  $S_1, S_2, \dots, S_g$  es:

$$L(\Sigma_1, \Sigma_2, \dots, \Sigma_g) = C \prod_{k=1}^g \exp\left\{-\frac{1}{2} n_k \text{tr}(\Sigma_k^{-1} S_k)\right\} |\Sigma_k|^{-n_k/2} \quad (7)$$

donde:

$C$  es una constante que no depende del  $\Sigma_k$ , ni de la función exponencial de la traza.

$$C = \frac{\prod_{k=1}^g |n_k S_k|^{(n_k-p-1)/2}}{2^{\sum_{k=1}^g n_k p/2} \Pi^{g p(p-1)/4} \prod_{k=1}^g \prod_{j=1}^p \Gamma((n+1-j)/2)}$$

dividiendo la ecuación (7) entre  $C$ , tomando  $\ln$  y multiplicando por  $(-2)$ , obtendremos lo siguiente.

$$\begin{aligned}
 g(\Sigma_1, \dots, \Sigma_g) &= -2 \left( \ln \frac{L(\Sigma_1, \dots, \Sigma_g)}{C} \right) \\
 g(\Sigma_1, \dots, \Sigma_g) &= -2 \ln L(\Sigma_1, \dots, \Sigma_g) + 2 \ln C \\
 g(\Sigma_1, \dots, \Sigma_g) &= (-2) \left( \frac{-1}{2} \sum_{k=1}^g n_k (tr \Sigma_k^{-1} S_k) - \frac{n_k}{2} \ln |\Sigma_k| \right) \quad (8) \\
 g(\Sigma_1, \dots, \Sigma_g) &= (-2) \left( \frac{-1}{2} \right) \sum_{k=1}^g n_k (\ln |\Sigma_k| + tr \Sigma_k^{-1} S_k) \\
 g(\Sigma_1, \dots, \Sigma_g) &= \sum_{k=1}^g n_k (\ln |\Sigma_k| + tr (\Sigma_k^{-1} S_k))
 \end{aligned}$$

Ahora asumamos que se cumple el modelo de CPC, para una cierta matriz ortogonal  $B$ , esto es

$\Sigma_k = B' \Lambda_k B$   $k = 1, 2, \dots, g$ , donde:

$$\Lambda_k = \text{diag}(\lambda_{1k}, \lambda_{2k}, \dots, \lambda_{pk}), \quad k = 1, 2, \dots, g;$$

entonces,

$$|\Sigma_k| = |B' \Lambda_k B| = \prod_{j=1}^p \lambda_{jk} \quad y,$$

$$\text{tomando } \ln : \ln |\Sigma_k| = \sum_{j=1}^p \ln \lambda_{jk} \quad k = 1, 2, \dots, g \quad (9)$$

Además bajo la misma suposición:  $\Sigma_k^{-1} = B' \Lambda_k^{-1} B$ ; ahora, a esta ecuación multiplicando por

$S_k$  y tomando su traza, se tiene:

$$tr(\Sigma_k^{-1} S_k) = tr(B' \Lambda_k^{-1} B S_k) = tr(\Lambda_k^{-1} B' S_k B) = \sum_{j=1}^p \left( \vec{\beta}_j' S_k \vec{\beta}_j \right) / \lambda_{jk} \quad (10)$$

donde:

$k = 1, 2, \dots, g$  y  $\vec{\beta}_j$  es la  $j$ -ésima columna de  $B$ ; entonces de (9) y (10) en (8), se tiene:

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

$$g(\Sigma_1, \dots, \Sigma_g) = g\left(\vec{\beta}_1, \dots, \vec{\beta}_p, \lambda_{11}, \dots, \lambda_{p1}, \dots, \lambda_{1g}, \dots, \lambda_{pg}\right)$$

$$g(\Sigma_1, \dots, \Sigma_g) = \sum_{k=1}^g n_k \left[ \sum_{j=1}^p \left( \ln \lambda_{jk} + \left( \vec{\beta}_j' S_k \vec{\beta}_j \right) / \lambda_{jk} \right) \right] \quad (11)$$

Suponiendo de que existe una matriz B ortogonal ( $B'B = I_p$ ), la función g debe ser reducido al mínimo bajo las restricciones:

$$\vec{\beta}_j' \vec{\beta}_h = \begin{cases} 0, & \text{Si } j \neq h \\ 1, & \text{Si } j = h \end{cases} \quad j = 1, \dots, p \quad (12)$$

así, si se desea minimizar la función, (Flury y Gautschi W. (1984))

$$G(\Sigma_1, \dots, \Sigma_g) = g(\Sigma_1, \dots, \Sigma_g) - \sum_{j=1}^p \eta_j \left( \vec{\beta}_j' \vec{\beta}_j - 1 \right) - 2 \sum_{j=1}^p \sum_{\substack{h=1 \\ j < h}}^p \eta_{jh} \left( \vec{\beta}_j' \vec{\beta}_h \right)$$

$$G(\Sigma_1, \dots, \Sigma_g) = \sum_{k=1}^g n_k \left[ \sum_{j=1}^p \left( \ln \lambda_{jk} + \left( \vec{\beta}_j' S_k \vec{\beta}_j \right) / \lambda_{jk} \right) \right] \quad (13)$$

$$- \sum_{j=1}^p \eta_j \left( \vec{\beta}_j' \vec{\beta}_j - 1 \right) - 2 \sum_{j=1}^p \sum_{\substack{h=1 \\ j < h}}^p \eta_{jh} \left( \vec{\beta}_j' \vec{\beta}_h \right)$$

donde:  $\eta_j (1 \leq j \leq p)$  y  $\eta_{jh} (1 \leq j < h \leq p)$  son los multiplicadores de Lagrange, es decir se

tiene p multiplicadores de Lagrange de la forma  $\eta_j$  y  $\frac{p(p-1)}{2}$  multiplicadores de Lagrange

de la forma  $\eta_{jh}$ , en total tenemos  $\frac{p(p+1)}{2}$  multiplicadores de Lagrange, para minimizar la

función (13) se debe hacer:

$$\frac{\partial G(\Sigma_1, \dots, \Sigma_g)}{\partial \lambda_{jk}} = 0 \quad j = 1, 2, \dots, p$$

$$k = 1, 2, \dots, g$$



**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

$$\Rightarrow \frac{\partial G}{\partial \lambda_{jk}} = n_k \left[ (1/\lambda_{jk}) - \left( \vec{\beta}_j' S_k \vec{\beta}_j \right) / \lambda_{jk}^2 \right] = 0$$

$$\Rightarrow \frac{1}{\lambda_{jk}} - \frac{\vec{\beta}_j' S_k \vec{\beta}_j}{\lambda_{jk}^2} = 0$$

$$\text{así: } \lambda_{jk} = \vec{\beta}_j' S_k \vec{\beta}_j; \quad \begin{array}{l} k = 1, 2, \dots, g \\ j = 1, 2, \dots, p \end{array} \quad (14)$$

$$\frac{\partial G(\Sigma_1, \dots, \Sigma_g)}{\partial \vec{\beta}_j} = 0 \quad j = 1, 2, \dots, p$$

$$\frac{\partial G}{\partial \vec{\beta}_j} = \sum_{k=1}^g n_k \left( S_k \vec{\beta}_j / \lambda_{jk} \right) - \eta_j \vec{\beta}_j - \sum_{\substack{h=1 \\ j < h}}^p \eta_{jh} \vec{\beta}_h = 0 \quad (15)$$

$\Rightarrow$  multiplicando esta ecuación, por  $\vec{\beta}_j'$  tenemos :

$$\sum_{k=1}^g n_k \left( \vec{\beta}_j' S_k \vec{\beta}_j \right) / \lambda_{jk} - \eta_j \vec{\beta}_j' \vec{\beta}_j - \sum_{\substack{h=1 \\ j < h}}^p \eta_{jh} \vec{\beta}_j' \vec{\beta}_h = 0 \quad (16)$$

$$\text{por (12) y } \lambda_{jk} = \vec{\beta}_j' S_k \vec{\beta}_j, \text{ se reduce a: } \sum_{k=1}^g n_k = \eta_j \quad (17)$$

Luego, usando la ecuación (14) en (10), obtenemos la siguiente identidad

$$\text{tr}(\Sigma_k^{-1} S_k) = \sum_{j=1}^p \left( \frac{\lambda_{jk}}{\lambda_{jk}} \right) = p; \quad k = 1, 2, \dots, g \quad (18)$$

ahora resolviendo la ecuación (15), obtenemos

$$\frac{\partial G}{\partial \vec{\beta}_j} = \sum_{k=1}^g n_k \left( S_k \vec{\beta}_j / \lambda_{jk} \right) - \eta_j \vec{\beta}_j - \sum_{\substack{h=1 \\ j < h}}^p \eta_{jh} \vec{\beta}_h = 0$$

y multiplicando por  $\vec{\beta}_l$  ( $l \neq j = 1, 2, \dots, p$ ), tenemos:

$$\sum_{k=1}^g n_k \left( \vec{\beta}_l S_k \vec{\beta}_j \right) / \lambda_{jk} - \eta_j \vec{\beta}_l \vec{\beta}_j - \sum_{\substack{h=1 \\ j < h}}^p \eta_{jh} \vec{\beta}_l \vec{\beta}_h = 0 \quad (19)$$

por la restricción (12), el segundo termino se hace cero, además como  $l \neq j$ , y en el tercer término tenemos  $l \neq h$ , entonces tenemos dos posibilidades: que  $l = h$  y en este caso  $\vec{\beta}_l \vec{\beta}_h = 1$  o bien  $l \neq h$ , entonces  $\vec{\beta}_l \vec{\beta}_h = 0$ , por lo tanto la ecuación (19) se reduce a:

$$\sum_{k=1}^g n_k \left( \vec{\beta}_l S_k \vec{\beta}_j \right) / \lambda_{jk} = \eta_{jl} \quad (j=1, \dots, p) \quad j \neq l \quad (20)$$

intercambiando los índices  $j$  y  $l$  en la ecuación (20) y notando que:

$$\vec{\beta}_j S_k \vec{\beta}_l = \vec{\beta}_l S_k \vec{\beta}_j \quad \text{y} \quad \eta_{jl} = \eta_{lj} \quad \text{se tiene que:}$$

$$\sum_{k=1}^g n_k \left( \vec{\beta}_l S_k \vec{\beta}_j \right) / \lambda_{lk} = \eta_{jl} \quad (j=1, \dots, p) \quad j \neq l \quad (21)$$

igualando (20) y (21), obtenemos:

$$\vec{\beta}_l \left[ \sum_{k=1}^g n_k \frac{S_k}{\lambda_{jk}} \right] \vec{\beta}_j - \vec{\beta}_l \left[ \sum_{k=1}^g n_k \frac{S_k}{\lambda_{lk}} \right] \vec{\beta}_j = 0$$

acomodando las sumatorias en la última ecuación tendríamos finalmente que:

$$\vec{\beta}_l \left[ \sum_{k=1}^g n_k \left( \frac{\lambda_{lk} - \lambda_{jk}}{\lambda_{jk} \lambda_{lk}} \right) S_k \right] \vec{\beta}_j = 0 \quad (j=1, \dots, p) \quad j \neq l \quad (22)$$

## “REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”

Carlos Díaz Gómez

esta última ecuación tiene  $p(p-1)$  ecuaciones, pero hay que notar que la  $(l, j)$ -ésima ecuación es la misma que la  $(j, l)$ -ésima ecuación, por lo tanto esto se reduce a  $p(p-1)/2$  ecuaciones con  $1 \leq l < j \leq p$ , en lugar de  $p(p-1)$ .

La ecuación (22) debe ser resuelta bajo las siguientes condiciones:

i)  $B'B = I_p$

ii)  $\lambda_{jk} = \vec{\beta}_j' S_k \vec{\beta}_j \quad (j=1, \dots, p) \quad \text{y} \quad (k=1, \dots, g)$

La primera es la condición de ortonormalidad y la segunda nos dice que  $\lambda_{jk}$  ( $j=1, \dots, p; k=1, \dots, g$ ) es igual a las varianzas de las combinaciones lineales  $\vec{\beta}_j' X^{(k)}$ . La solución conduce a la obtención de los estimadores máximo verosímiles de los ejes principales comunes  $\vec{\beta}_1, \dots, \vec{\beta}_p$  de los  $\lambda_{jk}$ . Como podemos observar, la ecuación (22) y de las condiciones i) y ii), para hallar los autovectores  $\vec{\beta}_j$ , necesitamos de los autovalores  $\lambda_{jk}$ , y para obtener estos autovalores  $\lambda_{jk}$ , necesitamos de los autovectores  $\vec{\beta}_j$ . Por lo tanto, en las aplicaciones, solo a través de un proceso iterativo, se encontraran las estimaciones.

### 2.3. COMBINACIONES LINEALES DENOMINADAS COMPONENTES PRINCIPALES COMUNES (CPC)

En el supuesto que las matrices de covarianzas cumplen el modelo de componentes principales comunes, emulando el caso de las combinaciones lineales de las componentes principales en un grupo, se definen las combinaciones lineales  $y_{jk}$ ,  $j=1, \dots, p$ ,  $k=1, \dots, g$ , denominadas Componentes Principales Comunes.

Se tienen  $k$  poblaciones normales multivariantes con parámetros media poblacional de cada grupo  $\vec{\mu}_k$  y matriz de covarianzas de cada población  $\Sigma_k$  en la que las matrices de covarianzas de las poblaciones tienen la siguiente estructura:  $\Sigma_k = B' \Lambda_k B$ , donde  $\Lambda_k$  son

## “REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”

Carlos Díaz Gómez

matrices diagonales que contienen los autovalores de las matrices de covarianzas de cada población y  $B = (\vec{\beta}_1, \vec{\beta}_2, \dots, \vec{\beta}_p)$  es una matriz ortogonal de autovectores comunes a todas las poblaciones.

Las  $p$  combinaciones lineales, denominados componentes principales comunes, en la población 1 son

$$Y_{11} = \vec{\beta}_1' X^{(1)} \quad \text{Var}(Y_{11}) = \vec{\beta}_1' \Sigma_1 \vec{\beta}_1 = \lambda_{11}$$

$$Y_{21} = \vec{\beta}_2' X^{(1)} \quad \text{Var}(Y_{21}) = \vec{\beta}_2' \Sigma_1 \vec{\beta}_2 = \lambda_{21}$$

.

$$Y_{p1} = \vec{\beta}_p' X^{(1)} \quad \text{Var}(Y_{p1}) = \vec{\beta}_p' \Sigma_1 \vec{\beta}_p = \lambda_{p1}$$

En la población 2:

$$Y_{12} = \vec{\beta}_1' X^{(2)} \quad \text{Var}(Y_{12}) = \vec{\beta}_1' \Sigma_2 \vec{\beta}_1 = \lambda_{12}$$

$$Y_{22} = \vec{\beta}_2' X^{(2)} \quad \text{Var}(Y_{22}) = \vec{\beta}_2' \Sigma_2 \vec{\beta}_2 = \lambda_{22}$$

.

$$Y_{p2} = \vec{\beta}_p' X^{(2)} \quad \text{Var}(Y_{p2}) = \vec{\beta}_p' \Sigma_2 \vec{\beta}_p = \lambda_{p2}$$

Generalizando, en la población  $g$ :

$$Y_{1g} = \vec{\beta}_1' X^{(g)} \quad \text{Var}(Y_{1g}) = \vec{\beta}_1' \Sigma_g \vec{\beta}_1 = \lambda_{1g}$$

$$Y_{2g} = \vec{\beta}_2' X^{(g)} \quad \text{Var}(Y_{2g}) = \vec{\beta}_2' \Sigma_g \vec{\beta}_2 = \lambda_{2g}$$

.

$$Y_{pg} = \vec{\beta}_p' X^{(g)} \quad \text{Var}(Y_{pg}) = \vec{\beta}_p' \Sigma_g \vec{\beta}_p = \lambda_{pg}$$

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

Donde la representación matricial de las matrices de covarianza, seguirá el modelo de componentes principales comunes:

$$\begin{bmatrix} \sigma_{11k} & \sigma_{12k} & \cdot & \cdot & \sigma_{1pk} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \sigma_{p1k} & \sigma_{p2k} & \cdot & \cdot & \sigma_{ppk} \end{bmatrix} = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdot & \cdot & \beta_{1p} \\ \beta_{21} & \beta_{22} & \cdot & \cdot & \beta_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \beta_{p1} & \beta_{p2} & \cdot & \cdot & \beta_{pp} \end{bmatrix} \begin{bmatrix} \lambda_{1k} & 0 & \cdot & 0 & 0 \\ \cdot & \lambda_{2k} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & \cdot & 0 & \lambda_{pk} \end{bmatrix} \begin{bmatrix} \beta_{11} & \beta_{21} & \cdot & \cdot & \beta_{p1} \\ \beta_{12} & \beta_{22} & \cdot & \cdot & \beta_{p2} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \beta_{1p} & \beta_{2p} & \cdot & \cdot & \beta_{pp} \end{bmatrix}$$

Si siguiendo la metodología presentada en el (2.2), se encuentran los estimadores máximo verosímiles, los que denotamos como sigue:

$$\hat{Y}_{jk} = \hat{\beta}_j' X^{(k)} ; \quad j = 1, \dots, p, \quad k = 1, \dots, g$$

$$Var(\hat{Y}_{jk}) = \hat{\beta}_j' \hat{\Sigma}_k \hat{\beta}_j = \hat{\lambda}_{jk}$$

$$Cov(\hat{Y}_{ik}, \hat{Y}_{lk}) = 0 \quad \text{con } i \neq l$$

que se usa en forma matricial

$$\hat{\Lambda}_k = \hat{B}' \hat{\Sigma}_k \hat{B}$$

Un aspecto esencial corresponde a presentar de manera sencilla, la metodología para el ajuste del modelo de componentes principales comunes

Sean  $\Sigma_1, \Sigma_2, \dots, \Sigma_g$  las matrices de covarianza definidas positivas de las  $g$  poblaciones concurrentes.

**Paso 1:** Contrastar las hipótesis más restrictiva, en la que se postula que las matrices de covarianzas de las poblaciones concurrentes, son iguales, frente a la hipótesis de que las matrices de covarianzas son diferentes, es decir:

$$H_0: \Sigma_1 = \Sigma_2 = \Sigma_3 = \dots = \Sigma_g$$

$$\mathbf{H11:} \Sigma_1 \neq \Sigma_2 \neq \Sigma_3 \neq \dots \neq \Sigma_g$$

Según el principio de razón de verosimilitud (Krzanowski, 1988; Mardia, 1976), la estadística usada para este test es la estadística,  $-2 \ln \lambda$ :

$$-2 \ln \lambda = n \ln \left| \hat{\Sigma} \right| - \sum_{k=1}^g n_k \ln \left| \hat{\Sigma}_k \right| \quad (23)$$

Cuya distribución asintótica es ji cuadrado  $\frac{p(p-1)(g-1)}{2}$  grados de libertad,

que se denota

$$-2 \ln \lambda \approx \chi^2 \left( \frac{1}{2} p(p-1)(g-1) \right),$$

donde  $\hat{\Sigma}_k = S_k$  son los estimadores máximo verosímiles de las matrices de covarianzas de cada población,  $\hat{\Sigma} = S$  es el estimador máximo verosímil de la matriz de covarianzas combinada de todas las poblaciones,  $n_k$  es el tamaño de cada una de las muestras y  $\ln$  es el logaritmo neperiano.

Si la decisión es rechazar la hipótesis nula al nivel de significación propuesto por el investigador, se pasa a postular que los datos siguen el modelo de CPC, que las matrices de covarianzas  $\Sigma_k$ ,  $k = 1, \dots, g$  asociados a las combinaciones lineales  $y_{jk}$  cumple el modelo de componentes principales comunes  $\Sigma_k = B' \Lambda_k B$

**Paso 2:** Contrastar la hipótesis que se cumple el modelo de CPC, es decir

$$\mathbf{Ho2} = H_{CPC}: \Sigma_k = B' \Lambda_k B \quad 1 \leq k \leq g, \quad (24)$$

**H12:** C.C

Para resolver el problema, se presenta un resumen de la metodología desarrollada por Flury (1984). Suponiendo que las muestras provienen de distribuciones normales multivariadas y

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

que son de tamaño  $n_k$ , se denota la solución que maximiza la verosimilitud de  $L(\Sigma_1, \Sigma_2, \dots, \Sigma_g)$  por:

$\hat{B} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)$ ,  $\hat{\Lambda}_k = \text{diag}(\hat{\lambda}_{1k}, \hat{\lambda}_{2k}, \dots, \hat{\lambda}_{pk})$  son los estimadores máximo verosímiles de la matriz de autovectores comunes, y de la matriz de covarianzas de los componentes principales comunes de cada población respectivamente; resultando el estimador máximo verosímil de  $\hat{\Sigma}_k = \hat{B}' \hat{\Lambda}_k \hat{B}$

donde  $\hat{\lambda}_{jk} = \hat{\beta}_j' S_k \hat{\beta}_j$  es el estimador máximo verosímil de la varianza de la combinación lineal de  $Y_{jk} = \hat{\beta}_j' \bar{X}^{(k)}$ ,  $S_k$  es el estimador máximo verosímil de la matriz de covarianza de cada población.

Así, el máximo de  $L$  bajo la hipótesis CPC se obtiene de reemplazar la ecuación (18) en la ecuación (7) y es:

$$L(\hat{\Sigma}_1, \dots, \hat{\Sigma}_g) = C \prod_{k=1}^g \exp \left\{ - \left( \frac{n_k p}{2} \right) |\hat{\Sigma}_k|^{-n_k/2} \right\} \quad (25)$$

mientras que el máximo sin restricción, sin suponer el modelo de CPC es:

$$L(S_1, \dots, S_g) = C \prod_{k=1}^g \exp \left\{ - \left( \frac{n_k p}{2} \right) |S_k|^{-n_k/2} \right\} \quad (26)$$

reemplazando (26) y (25) en (4) que es:

$$-2 \ln \lambda = -2 \ln \frac{L(\Sigma_1, \Sigma_2, \dots, \Sigma_g)}{L(S_1, S_2, \dots, S_g)}$$

se obtiene la estadística de prueba

$\sum_{k=1}^g n_k \ln \frac{|\hat{\Sigma}_k|}{|S_k|}$ , para probar la hipótesis de componentes principales comunes.

La distribución asintótica de la estadística de prueba

$$\chi^2 = \sum_{k=1}^g n_k \ln \frac{|\hat{\Sigma}_k|}{|S_k|}, \quad (27)$$

es distribución ji-cuadrado con  $\frac{p(p-1)(g-1)}{2}$  grados de libertad Flury (1984).

Si la decisión es no rechazar la hipótesis del modelo de componentes principales comunes al nivel de significación propuesto por el investigador, se concluye que los datos siguen el modelo de componentes principales comunes, es decir, las matrices de covarianzas tiene la siguiente estructura:

$$\Sigma_k = B \Lambda_k B' \quad 1 \leq k \leq g,$$

donde B es la matriz de autovectores comunes y  $\Lambda_k$  la matriz de covarianzas de las CPC en la población k

Por otro lado: bajo la hipótesis de componentes principales comunes, se cumple:

$$|\hat{\Sigma}_k| = |\hat{B}' \hat{\Lambda}_k \hat{B}| = |\hat{\Lambda}_k| = \prod_{j=1}^p \hat{\lambda}_{jk}$$

$$|S_k| = \prod_{j=1}^p \ell_{jk}; \text{ donde:}$$

$\ell_{jk}$  son los autovalores de la matriz  $S_k$

$\hat{\lambda}_{jk}$  son las varianzas de las combinaciones lineales  $\hat{Y}_k = \hat{B}' X^{(k)}$

por tanto la estadística de prueba (27), es equivalente a la siguiente:



$$\chi^2 = \sum_{k=1}^g n_k \left[ \ln \frac{\prod_{j=1}^p \hat{\lambda}_{jk}}{\prod_{j=1}^p \ell_{jk}} \right] \quad (28)$$

Entonces, si las matrices de covarianza, no son exactamente idénticas en todos los grupos ( $\Sigma_1 \neq \Sigma_2 \neq, \dots, \neq \Sigma_k$ ); pero se justifica que estas matrices de covarianzas, siguen una estructura básica común, podemos hallar las componentes principales comunes para una muestra dada, así como también la estimación de máximo verosimilitud de las matrices de covarianzas poblacionales de la siguiente manera

$$\hat{Y}_{jk} = \hat{\beta}_j^{\rightarrow} X^{(k)} \quad (k = 1, \dots, g) \text{ Componentes principales comunes muestrales.}$$

$$F_k = \hat{B}' S_k \hat{B}, \quad (29)$$

matrices de covarianzas muestrales de las componentes principales comunes transformados  $\hat{Y}_{jk}$ , donde  $S_k$ , son las matrices de covarianzas de cada muestra.

$\hat{\Sigma} = \hat{B}' \hat{\Lambda}_k \hat{B}$ , estimaciones de máxima verosimilitud de las matrices de covarianzas poblacionales.

dado que  $|F_k| = \left| \hat{B}' S_k \hat{B} \right|$

La estadística de prueba presentada en la ecuación (27) también se puede expresar de la siguiente manera:

$$\chi^2 = \sum_{k=1}^g n_i \text{Ln} \frac{|diag F_k|}{|F_k|} \quad (30)$$

## **2.4. REPRESENTACIONES GRÁFICAS DE LAS COMPONENTES PRINCIPALES COMUNES**

Conforme ya se señaló antes, en muchas aplicaciones biométricas es importante representar los datos originales en las dos primeras Componentes principales, tratándose de un estudio en un solo un grupo; por lo que resulta natural extrapolar dicho gráfico al caso de tres o más grupos.

Luego de aplicar la metodología del análisis se obtienen las Componentes Principales en cada uno de los grupos, a las que se denominan Componentes Principales Comunes, CPC, por las consideraciones ya descritas. Sin embargo, a fin de poder visualizar los datos de cada una de las muestras concurrentes, en el espacio bidimensional, debemos escoger de cada uno de los grupos, las variables con las mayores varianzas a fin de retener, en lo posible, las variabilidades contenidas en las variables originales de cada grupo.

Así, en la población 1 se generan las dos primeras CPC:

$$Y_{11} = \vec{\beta}'_1 X^{(1)} \quad \text{Var}(Y_{11}) = \vec{\beta}'_1 \Sigma_1 \vec{\beta}_1 = \lambda_{11}$$

$$Y_{21} = \vec{\beta}'_2 X^{(1)} \quad \text{Var}(Y_{21}) = \vec{\beta}'_2 \Sigma_1 \vec{\beta}_2 = \lambda_{21}$$

con las mayores varianzas.

En la población 2, se genera también las dos primeras CPC:

$$Y_{12} = \vec{\beta}'_1 X^{(2)} \quad \text{Var}(Y_{12}) = \vec{\beta}'_1 \Sigma_2 \vec{\beta}_1 = \lambda_{12}$$

$$Y_{22} = \vec{\beta}'_2 X^{(2)} \quad \text{Var}(Y_{22}) = \vec{\beta}'_2 \Sigma_2 \vec{\beta}_2 = \lambda_{22},$$

con las mayores varianzas.

y generalizando, en la población  $g$  también se generan las dos primeras CPC:

$$Y_{1g} = \vec{\beta}'_1 X^{(g)} \quad \text{Var}(Y_{1g}) = \vec{\beta}'_1 \Sigma_g \vec{\beta}_1 = \lambda_{1g}$$

## “REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”

Carlos Díaz Gómez

---

$$Y_{2g} = \vec{\beta}_2' X^{(g)}, \quad \text{Var}(Y_{2g}) = \vec{\beta}_2' \Sigma_g \vec{\beta}_2 = \lambda_{2g}$$

con las mayores varianzas.

Evaluando en dichas funciones, los datos originales de dimensión  $p \geq 3$ , se pasa a tener en cada grupo las matrices de datos con  $n_i, (i = 1, 2, \dots, k)$ , filas y 2 columnas respectivamente, cuyas representaciones gráficas serán en el espacio de dimensión 2, es decir en  $R^2$ , donde se facilita la lectura de los datos originales y como será comprobado en el siguiente capítulo a través de la aplicación.

## **CAPITULO III**

### **CARACTERIZACION DE MUESTRAS DE AGUJES DEL DEPARTAMENTO DE LORETO MEDIANTE COMPONENTES PINCIPALES COMUNES**

#### **3.1 INTRODUCCION**

El aguaje es una de las plantas más representativas de la amazonia peruana, crece en los plátanos de las zonas inundables y en los vallecitos de quebradas de las zonas de altura, los ecosistemas donde predomina la planta, son conocidas locamente como aguajales.

El aguaje es una palmera polígamo dioica (palmas con flores masculinas, femeninas o bisexuales), tiene una copa esférica, y en condiciones naturales puede alcanzar una altura de 35m. El tallo o estípite es recto, liso, cilíndrico. Las raíces primarias profundizan hasta 60 cm. y luego desarrollan horizontalmente hasta 40m, tienen raíces secundarias aeríferas o neumatóforos que le permiten respirar a las raíces en condiciones hidromorfas.

Lo más valioso del aguaje es su fruto, porque es rico en fitoestrógenos, en particular vitaminas y minerales que mejoran la salud interna y externa de nuestro organismo, la piel y cabello (<http://www.aguajes.com>).

Los principales componentes de los frutos de aguaje son:

- Vitamina caroteno (provitamina A)
- Tocoferoles (vitamina E)
- Acido ascórbico (vitamina C)

En estudios, hechos en Bélgica por la Facultad de Ciencias Agronómicas de L'Etat, Gembloux, conjuntamente con la Universidad Nacional de la Amazonía Peruana (1987) y otro hecho por la Universidad Federal del Estado de Pará, en Brasil (1998) quedó demostrado que el Aguaje contiene beta caroteno (Provitamina A) de 173 a 300 mg/100g y tocoferoles (Vitamina E) de 80 a 100 mg/100g.

## **“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

---

También la pulpa del Aguaje contiene ácido ascórbico (Vitamina C) de 50 a 60 mg/100g contiene fitoestrógenos y ácidos oleicos que son muy importantes en la alimentación humana sobre todo de la mujer durante toda su vida y en especial en el climaterio.

Entre sus propiedades figuran el hecho de ser modelador de la figura femenina, por la presencia de sitosterol puede ser precursor de hormonas femeninas del tipo progestágenos. Debido a esto, muchas empresas comercializan cápsulas de aguaje con la esperanzadora frases de “para aumentar el tamaño de los senos, perder la velloidad de la piel y aumentar la belleza femenina”; sin embargo, esto no está respaldado por estudios científicos. Posee hormonas femeninas como el fitoestrógenos, razón por la que muchos atribuyen al aguaje la belleza y exuberancia de las mujeres de la selva peruana, quienes la consumen con regularidad.

- Los frutos del aguaje presentan elevadas cantidades de vitamina A, la misma que es indispensable para la alimentación humana. Por la presencia de vitamina C, es recomendable para quienes padecen estas enfermedades y el escorbuto y enfermedades que desestabilizan el metabolismo orgánico como tuberculosis y disentería. Por la presencia de cantidades importantes de alfatocoferol (vitamina E) es considerado como antioxidante.

La fructificación se inicia entre 12 y 24 meses después de la plantación y ocurre al año; la producción comercial al 5to. año y se incrementa hasta el año 12, momento de el cual alcanza su máximo desarrollo

Los rendimientos estimados de frutos frescos en monoculturas de 3 x3 m son: 2,5 t/ha el segundo año; 9,1 t/ha el tercer. año; 9,8 t/ha el cuarto. año; 21,5 t/ha el quinto. año y 40,6 t/ha el sexto. año.

La cosecha es anual, directa de la planta, y se realiza cuando el fruto está verde al estado "pintón". Los frutos cosechados continuarán su madurez fisiológica, y no les afectarán los manipuleos de cosecha y de transporte, que sí dañan a los frutos maduros que son muy delicados.

La teoría desarrollada y descrita en el capítulo anterior, es aplicada a 4 muestras aleatorias que corresponden frutos recolectados en los mercados de Mainas, Alto Amazonas Loreto y Requena en el Departamento de Loreto (Quinteros, 2009).

En este caso, se hacen las pruebas de hipótesis para la comparación de las matrices de covarianza de cada muestra y las pruebas de hipótesis correspondientes al modelo de componentes principales comunes que viene a ser el tema central de dicha tesis.

Cabe indicar que para el modelo de las componentes principales comunes, se han elaborado los programas estadísticos a través del software libre “R”, también se hará uso de los softwares SPSS y MATLAB, para el desarrollo de las estadísticas descriptivas y así poder aplicar el modelo de componentes principales comunes la cual es el objetivo de dicha tesis.

### **3.2 DESCRIPCION DE LOS DATOS:**

Los datos que se usan en el presente trabajo, son el resultado de una investigación realizada en el Departamento de Loreto entre enero y marzo del año 2009 (Quinteros 2009) y corresponden a 4 muestras aleatorias de frutos de aguaje recolectados en mercados de Mainas, Alto Amazonas Loreto y Requena en el Departamento de Loreto (Quinteros, 2009). Las muestras fueron de tamaños 27, 30, 21 Y 57 y las mediciones de cada uno de los frutos de aguaje, corresponden a las siguientes variables:

X1: Diámetro ecuatorial del fruto (cm.)

X2: Diámetro longitudinal del fruto (cm.)

X3: Diámetro ecuatorial de la semilla (cm.)

X4: Diámetro longitudinal de la semilla (cm.)

X5: Diámetro ecuatorial de la escama (cm.)

X6: Diámetro longitudinal de la escama (cm.)

Los datos se presentan en el Anexo\_1.

Grafico 2:- Frutos de aguajes



### **3.3 RESULTADOS Y DISCUSIÓN**

Al realizar el análisis descriptivo se encontró mucha variabilidad entre las variables (diámetro), por lo que los datos originales fueron transformados mediante la transformación logarítmica, con lo que se corrigió el problema descrito. Con los datos transformados se realizó todo el análisis de componentes principales generalizados.

A continuación se presenta en detalle los resultados encontrados:

1. En la Tabla 1 se presenta el resumen de algunas medidas descriptivas para los logaritmos de los diámetros del fruto, sin considerar el lugar de procedencia de los frutos. Los datos transformados según la transformación logarítmica se presentan en el Anexo\_2.

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

Tabla 1 Estadísticas descriptivas del logaritmo del diámetro de los frutos en cm.

Estadísticos descriptivos					
	Tamaño de la muestra total	Mínimo	Máximo	Media aritmética	Desviación estándar.
Diámetro ecuatorial del fruto	135	1	2	1,37	,076
Diámetro longitudinal del fruto	135	1,49	1,97	1,7204	,12413
Diámetro ecuatorial de la semilla	135	,88	1,37	1,1473	,09888
Diámetro longitudinal de la semilla	135	1,28	1,80	1,5476	,14767
Diámetro ecuatorial de la escama	135	-1,61	-,56	-,9021	,15266
Diámetro longitudinal de la escama	135	-1,39	-,49	-,8164	,14346
N válido (según lista)	135				

2. A partir de los resultados de cada una de las muestras, para los logaritmos de los datos, se estimaron los parámetros globales del vector de medias y de la matriz de covarianzas.

$$\hat{\mu} = \bar{X} = \begin{bmatrix} 1.3680 \\ 1.7204 \\ 1.1473 \\ 1.5476 \\ 0.9021 \\ 0.8164 \end{bmatrix}, \text{ la matriz de Covarianzas } S,$$

$$\hat{\Sigma} = S = \begin{bmatrix} 0.0058 & 0.0065 & 0.0067 & 0.0076 & 0.0056 & 0.0061 \\ 0.0065 & 0.0154 & 0.0076 & 0.0178 & 0.0078 & 0.0082 \\ 0.0067 & 0.0076 & 0.0098 & 0.0086 & 0.0061 & 0.0071 \\ 0.0076 & 0.0178 & 0.0086 & 0.0218 & 0.0082 & 0.0085 \\ 0.0056 & 0.0078 & 0.0061 & 0.0082 & 0.0233 & 0.0190 \\ 0.0061 & 0.0082 & 0.0071 & 0.0085 & 0.0190 & 0.0206 \end{bmatrix}$$

3. Los vectores de medias y las matrices de covarianzas de cada una de las muestras, fueron:



# “REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”

Carlos Díaz Gómez

Muestra 1: Muestra de Mainas

$$\hat{\mu}_1 = \bar{\bar{X}}_1 = \begin{bmatrix} 1.3566 \\ 1.6700 \\ 1.1333 \\ 1.4800 \\ 1.0714 \\ 0.9485 \end{bmatrix} \quad \hat{\Sigma}_1 = S_1 = \begin{bmatrix} 0.0035 & -0.0032 & 0.0030 & -0.0033 & 0.0041 & 0.0039 \\ -0.0032 & 0.0054 & -0.0025 & 0.0064 & -0.0043 & -0.0037 \\ 0.0030 & -0.0025 & 0.0032 & -0.0023 & 0.0037 & 0.0029 \\ -0.0033 & 0.0064 & -0.0023 & 0.0086 & -0.0073 & -0.0064 \\ 0.0041 & -0.0043 & 0.0037 & -0.0073 & 0.0311 & 0.0210 \\ 0.0039 & -0.0037 & 0.0029 & -0.0064 & 0.0210 & 0.0194 \end{bmatrix}$$

Muestra 2: Muestra de Alto Amazonas

$$\hat{\mu}_2 = \bar{\bar{X}}_2 = \begin{bmatrix} 1.3090 \\ 1.5740 \\ 1.0870 \\ 1.3780 \\ 0.9016 \\ 0.8123 \end{bmatrix} \quad \hat{\Sigma}_2 = S_2 = \begin{bmatrix} 0.0027 & 0.0008 & 0.0030 & 0.0014 & 0.0010 & 0.0011 \\ 0.0008 & 0.0021 & 0.0021 & 0.0024 & 0.0003 & 0.0010 \\ 0.0030 & 0.0021 & 0.0059 & 0.0022 & 0.0011 & 0.0023 \\ 0.0014 & 0.0024 & 0.0022 & 0.0044 & -0.0012 & 0.0001 \\ 0.0010 & 0.0003 & 0.0011 & -0.0012 & 0.0078 & 0.0051 \\ 0.0011 & 0.0010 & 0.0023 & 0.0001 & 0.0051 & 0.0075 \end{bmatrix}$$

Muestra 3: Muestra de Loreto

$$\hat{\mu}_3 = \bar{\bar{X}}_3 = \begin{bmatrix} 1.4457 \\ 1.8643 \\ 1.2414 \\ 1.7200 \\ 0.8029 \\ 0.7248 \end{bmatrix} \quad \hat{\Sigma}_3 = S_3 = \begin{bmatrix} 0.0010 & 0.0011 & 0.0021 & 0.0013 & 0.0011 & 0.0006 \\ 0.0011 & 0.0016 & 0.0021 & 0.0017 & 0.0018 & 0.0013 \\ 0.0021 & 0.0021 & 0.0051 & 0.0029 & 0.0016 & 0.0008 \\ 0.0013 & 0.0017 & 0.0029 & 0.0022 & 0.0015 & 0.0010 \\ 0.0011 & 0.0018 & 0.0016 & 0.0015 & 0.0095 & 0.0083 \\ 0.0006 & 0.0013 & 0.0008 & 0.0010 & 0.0083 & 0.0092 \end{bmatrix}$$

Muestra 4: Muestra de Requena

$$\hat{\mu}_4 = \bar{\bar{X}}_4 = \begin{bmatrix} 1.3757 \\ 1.7684 \\ 1.1510 \\ 1.6052 \\ 0.8587 \\ 0.7896 \end{bmatrix} \quad \hat{\Sigma}_4 = S_4 = \begin{bmatrix} 0.0061 & 0.0068 & 0.0074 & 0.0073 & 0.0064 & 0.0084 \\ 0.0068 & 0.0098 & 0.0072 & 0.0105 & 0.0084 & 0.0110 \\ 0.0074 & 0.0072 & 0.0116 & 0.0072 & 0.0068 & 0.0100 \\ 0.0073 & 0.0105 & 0.0072 & 0.0129 & 0.0087 & 0.0115 \\ 0.0064 & 0.0084 & 0.0068 & 0.0087 & 0.0144 & 0.0147 \\ 0.0084 & 0.0110 & 0.0100 & 0.0115 & 0.0147 & 0.0207 \end{bmatrix}$$

4. La prueba de hipótesis para docimar la multinormalidad de los datos arrojó los siguientes resultados:

Cabe decir que se aplicó el test de Mardia, la cual se aplica para datos multivariados, es decir, Mardia (1970) define que sea  $n$  observaciones del vector aleatorio  $X$ , entonces los correspondientes coeficientes de asimetría y curtosis muestrales son:

$$b_{1,p} = \frac{1}{n^2} \sum_i^n \sum_j^n \left\{ (X_i - \bar{X})^T S^{-1} (X_j - \bar{X}) \right\}^3 \quad \begin{array}{l} i = 1, 2, \dots, n \\ j = 1, 2, \dots, n \end{array}$$
$$b_{2,p} = \frac{1}{n} \sum_i^n \left\{ (X_i - \bar{X})^T S^{-1} (X_j - \bar{X}) \right\}^2$$

Donde  $\bar{X}$  es el vector de medias y  $S$  es la matriz de covarianzas del grupo  $k$ . Mardia (1974) derivó las distribuciones asintóticas bajo la hipótesis nula de normal multivariada:

$$\frac{n}{6} b_{1,p} \sim \chi_{\frac{p(p+1)(p+2)}{6}}^2 \quad \text{y} \quad b_{2,p} \sim N\left(p(p+2), \frac{8p(p+2)}{n}\right)$$

se resolvió la prueba de hipótesis, al nivel de significancia de 0.05. Para contrastar,

$H_0$  : Los datos siguen una distribución normal multivariada

$H_1$  : Caso contrario.

La cual se menciona lo datos para cada uno de los 4 grupos  $k = 1, 2, 3, 4$

Grupo1 - Mainas:

El p-valor para el grupo1 es 0.949 la cual nos dice que no rechazamos la hipótesis de que los datos siguen una distribución normal multivariada.

Grupo2 – Alto Amazonas:

El p-valor para el grupo2 es 0.917 la cual nos dice que no rechazamos la hipótesis de que los datos siguen una distribución normal multivariada.

Grupo3 - Loreto:

El p-valor para el grupo3 es 0.910 la cual nos dice que no rechazamos la hipótesis de que los datos siguen una distribución normal multivariada.

Grupo4 - Loreto:

El p-valor para el grupo4 es 0.620 la cual nos dice que no rechazamos la hipótesis de que los datos siguen una distribución normal multivariada.

5. A continuación se realizó la prueba de hipótesis para comparar las matrices de covarianzas. Es decir, se resolvió la hipótesis de que las matrices de covarianzas son diferentes, al nivel de significación 0.05.

Para contrastar,  $H_0 : \Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma_4$ , con un nivel de significación  $\alpha=0.05$ , se usó el Test

de Razón de Verosimilitud, que según la fórmula (23), la estadística de prueba

$-2 \ln \lambda = N \ln \left| \hat{\Sigma} \right| - \sum_{k=1}^g n_k \ln \left| \hat{\Sigma}_k \right|$ , asintóticamente tiene distribución Ji-cuadrado con grados de libertad  $\left( \frac{1}{2} p(p-1)(g-1) \right)$ .

Con las  $p = 6$  variables, todas las muestras,  $N = 135$

$n = n_1 + n_2 + n_3 + n_4 = (27 + 30 + 21 + 57)$  y  $g = 4$  grupos:

El número de los grados de libertad resultante fue 45,

$$\sum_{k=1}^g n_k \ln \left| \hat{\Sigma}_k \right| = -4884.319,$$

$$N \ln \left| \hat{\Sigma} \right| \approx -4.3602e + 003 = -4360.2,$$

por lo que, el valor de la estadística de prueba según la ecuación (23) fue,

$$-2 \ln \lambda = 524.1186.$$

## “REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”

Carlos Díaz Gómez

Al comparar dicho valor con el valor teórico para el nivel de significación 0.05 y 45 grados de libertad, 61.05, se rechaza la hipótesis nula, es decir, se rechaza la hipótesis que las matrices de covarianzas son iguales al nivel de significación 0.05.

6. A continuación se postula que los datos se ajustan al Modelo de Componentes Principales Comunes (Flury, 1984), es decir, se resuelve todos los pasos necesarios para contrastar la hipótesis de que los datos siguen el modelo de CPC, que se detalla para esta situación en particular.

Para contrastar la hipótesis

$$H_{CPC}: B' \Lambda_k B = \Sigma_k \text{ versus } k = 1, 2, 3, 4$$

$$H_1: \Sigma_k,$$

que las matrices de covarianzas son arbitrarias  $\forall k \in \{1, \dots, g\}$  grupos, según la ecuación (24) del Cap. II y teniendo en cuenta el valor de la estadística de prueba según las fórmulas (27) y (28), se realizaron los siguientes cálculos previos:

Matrices de autovectores,  $V$ , y de autovalores,  $D$ , de la matriz de covarianza  $\hat{\Sigma} = S$

$$V = \begin{bmatrix} -0.2415 & -0.0994 & -0.4433 & -0.1273 & 0.7911 & 0.3050 \\ -0.4166 & -0.4068 & 0.2170 & 0.0391 & -0.3248 & 0.7118 \\ -0.2851 & -0.1315 & -0.7752 & -0.1542 & -0.4795 & -0.2160 \\ -0.4769 & -0.5364 & 0.3086 & 0.0212 & 0.1921 & -0.0220 \\ -0.4843 & 0.5540 & 0.2392 & -0.6319 & -0.0374 & -0.0220 \\ -0.4741 & 0.4609 & -0.0534 & 0.7474 & 0.0157 & -0.0316 \end{bmatrix}$$

$$D = \begin{bmatrix} 0.0631 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.0229 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0066 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.0026 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0007 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.0004 \end{bmatrix}$$

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

Según el modelo de Componentes principales Comunes, la matriz ortogonal  $\hat{B}$ , es el estimador de la matriz de autovectores comunes será:

$$\hat{B} = \text{autovec}(\hat{\Sigma}) = \begin{bmatrix} -0.2415 & -0.0994 & -0.4433 & -0.1273 & 0.7911 & 0.3050 \\ -0.4166 & -0.4068 & 0.2170 & 0.0391 & -0.3248 & 0.7118 \\ -0.2851 & -0.1315 & -0.7752 & -0.1542 & -0.4795 & -0.2160 \\ -0.4769 & -0.5364 & 0.3086 & 0.0212 & 0.1921 & -0.5933 \\ -0.4843 & 0.5540 & 0.2392 & -0.6319 & -0.0374 & -0.0220 \\ -0.4741 & 0.4609 & -0.0534 & 0.7474 & 0.0157 & -0.0316 \end{bmatrix}$$

Esta matriz diagonaliza en forma simultánea a las matrices  $S_k$ ,  $k = 1, 2, 3, 4$ .

La transpuesta de la matriz ortogonal  $\hat{B}$  es:

$$\hat{B}' = \begin{bmatrix} -0.2415 & -0.4166 & -0.2851 & -0.4769 & -0.4843 & -0.4741 \\ -0.0994 & -0.4068 & -0.1315 & -0.5364 & 0.5540 & 0.4609 \\ -0.4433 & 0.2170 & -0.7752 & 0.3086 & 0.2392 & -0.0534 \\ -0.1273 & 0.0391 & -0.1542 & 0.0212 & -0.6319 & 0.7474 \\ 0.7911 & -0.3248 & -0.4795 & 0.1921 & -0.0374 & 0.0157 \\ 0.3050 & 0.7118 & -0.2160 & -0.5933 & -0.0220 & -0.0316 \end{bmatrix}$$

Las estimaciones de las matrices  $F_k = \hat{B}' S_k \hat{B}$ ,  $k = 1, 2, 3, 4$ , según la ecuación (29), fueron las siguientes:

$$F_1 = \begin{bmatrix} 0.0192 & -0.0198 & -0.0007 & 0.0021 & 0.0000 & 0.0000 \\ -0.0198 & 0.0388 & -0.0066 & -0.0037 & 0.0022 & 0.0009 \\ -0.0006 & -0.0066 & 0.0084 & 0.0005 & -0.0024 & 0.0001 \\ 0.0021 & -0.0037 & 0.0005 & 0.0031 & 0.0000 & 0.0003 \\ 0.0000 & 0.0022 & -0.0024 & 0.0000 & 0.0010 & 0.0000 \\ 0.0000 & 0.0009 & 0.0001 & 0.0003 & 0.0000 & 0.0003 \end{bmatrix}$$

$$F_2 = \begin{bmatrix} 0.0124 & -0.0024 & 0.0033 & -0.0004 & 0.0006 & 0.0001 \\ -0.0024 & 0.0097 & 0.0010 & -0.0002 & -0.0002 & 0.0010 \\ -0.0006 & -0.0066 & 0.0084 & 0.0005 & -0.0024 & 0.0001 \\ -0.0004 & -0.0002 & 0.0001 & 0.0024 & -0.0002 & -0.0002 \\ 0.0006 & -0.0002 & 0.0001 & -0.0002 & 0.0001 & -0.0000 \\ 0.0001 & 0.0010 & 0.0000 & -0.0002 & -0.0000 & 0.0003 \end{bmatrix}$$

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

$$F_3 = \begin{bmatrix} 0.0156 & -0.0057 & 0.0012 & 0.0003 & 0.0011 & 0.0005 \\ -0.0057 & 0.0085 & 0.0027 & 0.0014 & 0.0002 & 0.0002 \\ 0.0012 & 0.0027 & 0.0025 & 0.0008 & 0.0005 & 0.0004 \\ 0.0003 & 0.0014 & 0.0008 & 0.0014 & 0.0001 & 0.0000 \\ 0.0011 & 0.0002 & 0.0005 & 0.0001 & 0.0002 & 0.0001 \\ 0.0005 & 0.0002 & 0.0004 & 0.0000 & 0.0001 & 0.0002 \end{bmatrix}$$

$$F_4 = \begin{bmatrix} 0.0578 & -0.0024 & 0.0062 & -0.0041 & -0.0010 & -0.0008 \\ -0.0024 & 0.0073 & 0.0007 & 0.0005 & -0.0005 & 0.0001 \\ 0.0062 & 0.0007 & 0.0065 & -0.0002 & 0.0010 & -0.0000 \\ -0.0041 & 0.0005 & -0.0002 & 0.0026 & 0.0000 & 0.0000 \\ 0.0010 & -0.0005 & 0.0010 & 0.0000 & 0.0007 & -0.0000 \\ -0.0008 & 0.0001 & -0.0000 & -0.0000 & -0.0000 & 0.0006 \end{bmatrix}$$

Entonces las diagonales de las matrices  $F_k$  son las varianzas muestrales  $\hat{\lambda}_{jk}$  ( $k = 1, \dots, 4; j = 1, \dots, 6$ ) y se representan de la siguiente manera:

$$\hat{\lambda}_{j1} = [0.0192 \quad 0.0388 \quad 0.0084 \quad 0.0037 \quad 0.0010 \quad 0.0003]$$

$$\hat{\lambda}_{j2} = [0.0124 \quad 0.0097 \quad 0.0044 \quad 0.0024 \quad 0.0010 \quad 0.0003]$$

$$\hat{\lambda}_{j3} = [0.0156 \quad 0.0085 \quad 0.0025 \quad 0.0014 \quad 0.0002 \quad 0.0002]$$

$$\hat{\lambda}_{j4} = [0.0578 \quad 0.0073 \quad 0.0065 \quad 0.0026 \quad 0.0007 \quad 0.0006]$$

mientras que los autovalores de la matriz de covarianzas  $S_k$ , para los grupos  $k = 1, 2, 3, 4$  son denotados por  $\ell_{jk}$ ,  $j = 1, 2, 3, 4, 5, 6$  y fueron las siguientes:

$$\ell_{j1} = [0.0523 \quad 0.0119 \quad 0.0034 \quad 0.0033 \quad 0.0002 \quad 0.0001]$$

$$\ell_{j2} = [0.0145 \quad 0.0095 \quad 0.0031 \quad 0.0022 \quad 0.0009 \quad 0.0002]$$

$$\ell_{j3} = [0.01889 \quad 0.0079 \quad 0.0010 \quad 0.0006 \quad 0.0001 \quad 0.0001]$$

$$\ell_{j4} = [0.0590 \quad 0.0077 \quad 0.0055 \quad 0.0023 \quad 0.0006 \quad 0.0003]$$

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

7. Luego, el valor de la estadística de prueba, para contrastar la hipótesis que los datos siguen en modelo de CPC según la ecuación (28) fue:

$$\chi_{CPC}^2 = 27 * \ln\left(\frac{0.0714}{0.0712}\right) + 30 * \ln\left(\frac{0.0302}{0.0304}\right) + 21 * \ln\left(\frac{0.0284}{0.0286}\right) + 57 * \ln\left(\frac{0.0755}{0.0754}\right)$$

$$\chi_{CPC}^2 = 0.4967;$$

Al nivel de significación 0.05, el valor,  $\chi_{CPC}^2 = 0.4967$ , comparado con el valor teórico, 61.05, conduce a no rechazar la hipótesis de que las matrices de covarianzas de las poblaciones concurrentes siguen el modelo de Componentes Principales Comunes. Es decir, al nivel de significación 0.05 se cumple el supuesto de que los datos siguen el modelo de CPC para el nivel de significación 0.05.

8. Puesto que los datos se ajustan al modelo de CPC, el estimador de la matriz de autovectores comunes  $\hat{B}$ , fue:

$$\hat{B} = \text{autovec}(\hat{\Sigma}) = \begin{bmatrix} -0.2415 & -0.0994 & -0.4433 & -0.1273 & 0.7911 & 0.3050 \\ -0.4166 & -0.4068 & 0.2170 & 0.0391 & -0.3248 & 0.7118 \\ -0.2851 & -0.1315 & -0.7752 & -0.1542 & -0.4795 & -0.2160 \\ -0.4769 & -0.5364 & 0.3086 & 0.0212 & 0.1921 & -0.5933 \\ -0.4843 & 0.5540 & 0.2392 & -0.6319 & -0.0374 & -0.0220 \\ -0.4741 & 0.4609 & -0.0534 & 0.7474 & 0.0157 & -0.0316 \end{bmatrix}$$

9. Luego, se obtienen las Componentes Principales Comunes que en el presente caso son 6, por ser las variables evaluadas en cada uno de los 4 grupos, las que quedan representadas de la siguiente manera:

$$\begin{aligned} \hat{Y}_1^{(k)} &= -0.2415 X_1^{(k)} - 0.4166 X_2^{(k)} - 0.2851 X_3^{(k)} - 0.4769 X_4^{(k)} - 0.4843 X_5^{(k)} - 0.4741 X_6^{(k)} \\ \hat{Y}_2^{(k)} &= -0.0994 X_1^{(k)} - 0.4068 X_2^{(k)} - 0.1315 X_3^{(k)} - 0.5354 X_4^{(k)} + 0.5540 X_5^{(k)} + 0.4609 X_6^{(k)} \\ \hat{Y}_3^{(k)} &= -0.4433 X_1^{(k)} + 0.2170 X_2^{(k)} - 0.7752 X_3^{(k)} + 0.3086 X_4^{(k)} + 0.2392 X_5^{(k)} - 0.0534 X_6^{(k)}, \\ \hat{Y}_4^{(k)} &= -0.1273 X_1^{(k)} + 0.0391 X_2^{(k)} - 0.1542 X_3^{(k)} + 0.0212 X_4^{(k)} - 0.6319 X_5^{(k)} + 0.7474 X_6^{(k)} \\ \hat{Y}_5^{(k)} &= 0.7911 X_1^{(k)} - 0.3248 X_2^{(k)} - 0.4795 X_3^{(k)} + 0.1921 X_4^{(k)} - 0.0374 X_5^{(k)} + 0.0157 X_6^{(k)} \\ \hat{Y}_6^{(k)} &= 0.3050 X_1^{(k)} + 0.7118 X_2^{(k)} - 0.2160 X_3^{(k)} - 0.5933 X_4^{(k)} - 0.0220 X_5^{(k)} - 0.0316 X_6^{(k)} \end{aligned}$$

donde:  $k = 1, 2, 3, 4$  grupos;

## “REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”

Carlos Díaz Gómez

la primera ecuación representa la primera Componente Principal Común en cada uno de los 4 grupos, y así sucesivamente hasta la sexta Componente Principal Común.

En el Anexo\_3 se presentan los datos transformados según el modelo de CPC para cada una de las muestras.

### 3.4 REPRESENTACIONES GRÁFICAS DE LAS COMPONENTES PRINCIPALES COMUNES

Así como en el caso de las Componentes Principales en un Grupo, se emulan los resultados más importantes según el modelo de Componentes Principales Comunes en los 4 grupos.

En la Tabla 1: se presentan los autovectores comunes y los autovalores de cada grupo.

Tabla 1: Autovectores comunes y autovalores de cada grupo.

Autovectores comunes						Autovalores			
						Muestra1	Muestra2	Muestra3	Muestra 4
-0,2415	-0,0994	-0,4433	-0,1273	0,7911	0,3050	0.0523	0.0145	0,0188	0.059
-0,4166	-0,4068	0,2170	0,0391	-0,3248	0,7118	0.0119	0.0095	0,0078	0.0077
-0,2851	-0,1315	-0,7752	-0,1542	-0,4795	-0,2160	0.0034	0.0031	0,001	0.0055
-0,4769	-0,5364	0,3086	0,0212	0,1921	-0,5933	0.0033	0.0022	0,0006	0.0023
-0,4843	0,5540	0,2392	-0,6319	-0,0374	-0,0220	0.0002	0.0009	0,0001	0.0006
-0,4741	0,4609	-0,0534	0,7474	0,0157	-0,0316	0.0001	0.0002	0,0001	0.0003
Porcentaje variación explicada por las 2 primeras CPC						90.17%	78.95%	93.66%	88.46%

Para la primera muestra: agujajes de Mainas, el porcentaje de variación explicada por las 2 primeras componentes fue del 90.1%

Para la segunda muestra agujajes de Alto Amazonas, el porcentaje de variación explicada por las 2 primeras componentes fue del 78.9%.

Para la tercera muestra: agujajes de Loreto, el porcentaje de variación explicada por las 2 primeras componentes fue del 93.6%.

Para la cuarta muestra: agujajes de Requena, el porcentaje de variación explicada por las 2 primeras componentes fue del 88.4%.

Las dos primeras Componentes Principales en cada uno de los 4 grupos ( $k=1,2,3,4$ ) se representan a través de las siguientes funciones:



**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

$$\hat{Y}_1^{(k)} = -0.2415 X_1^{(k)} - 0.4166 X_2^{(k)} - 0.2851 X_3^{(k)} - 0.4769 X_4^{(k)} - 0.4843 X_5^{(k)} - 0.4741 X_6^{(k)}$$

$$\hat{Y}_2^{(k)} = -0.0994 X_1^{(k)} - 0.4068 X_2^{(k)} - 0.1315 X_3^{(k)} - 0.5354 X_4^{(k)} + 0.5540 X_5^{(k)} + 0.4609 X_6^{(k)}$$

Las estadísticas resumen para cada una de las muestras en las dos primeras variables denominadas Componentes Principales Comunes con las mayores varianzas, que se presentan en la Tabla 2.

Tabla 2 Estadísticas descriptivas de las dos primeras CPC en los 4 lugares

Muestras	VARIABLES CPC	Media	Varianza	Desviación estándar	Coficiente de variación
Mainas	$Y_1$	1.0838	0,0192	0.1387	0,1280
	$Y_2$	0,1471	0,0033	0,0576	0,3916
Alto Amazonas	$Y_1$	1.1173	0,0124	0,1114	0,0997
	$Y_2$	0,1429	0,0015	0,0388	0,2715
Loreto	$Y_1$	1,4243	0,0618	0,2485	0,1745
	$Y_2$	0,1969	0,0031	0,0554	0,2814
Requena	$Y_1$	1,3726	0,0579	0,2406	0,1753
	$Y_2$	0,1979	0,0015	0,0382	0,1930

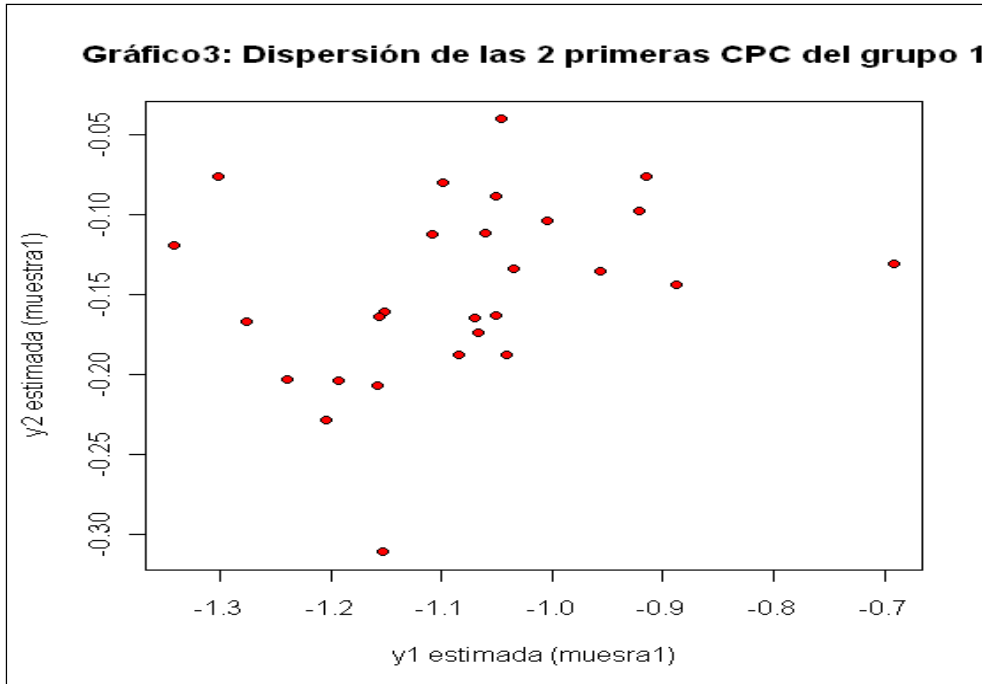
A continuación se representan cada una de las muestras originales a través de las 2 primeras combinaciones lineales, CPC, en cada uno de los lugares, Mainas, Alto Amazonas, Loreto y Requena, o sea para:  $k = 1,2,3,4$ .

En el Gráfico 3 y 4 se presentan los agujajes de Mainas, en el Gráfico 5 y 6 los agujajes de Alto Amazonas, en el Gráfico 7 y 8 los agujajes de Loreto y en el Gráfico 9 y 10 los agujajes de Requena.

Se logran hacer las representaciones gráficas en el espacio bidimensional, imposibles de realizar en el espacio dimensional original y que según el porcentaje de variabilidad contenido en estas CPC (90%, 79%, 93% y 88%), en cada uno de los casos, los datos estarían muy bien representados en dichos gráficos.

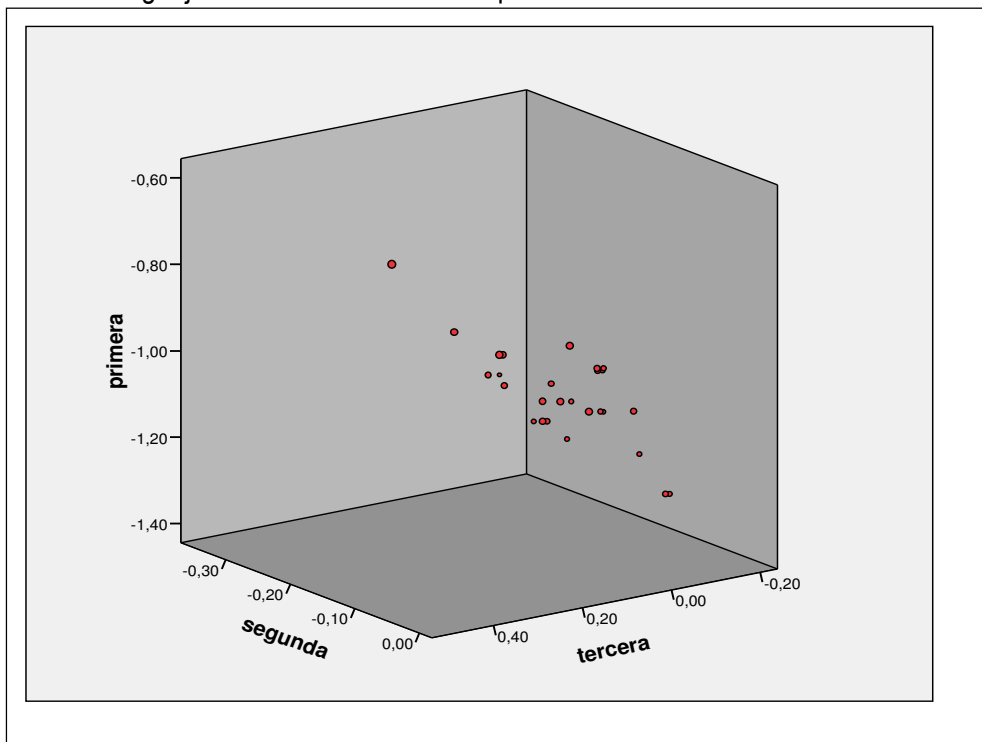
De la comparación de los Gráficos 3, 5, 7 y 9 se puede percibir que las estructuras internas de los datos, en las dos primeras CPC, se muestran relativamente diferentes, lo que se confirma mediante las diferencias en las varianzas de las mediciones de cada uno de los lugares.

Gráfico 3: Aguajes de Mainas en las dos primeras CPC



En este gráfico podemos observar para el primer Grupo que la dispersión de los datos están ligeramente agrupados.

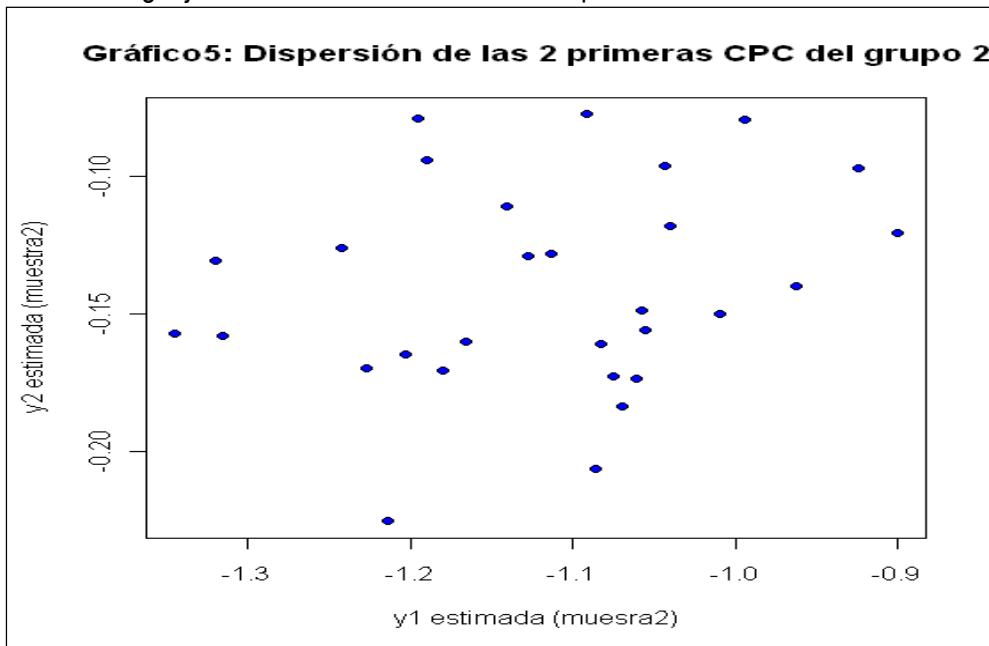
Gráfico 4: Aguajes de Mainas en las tres primeras CPC



# “REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”

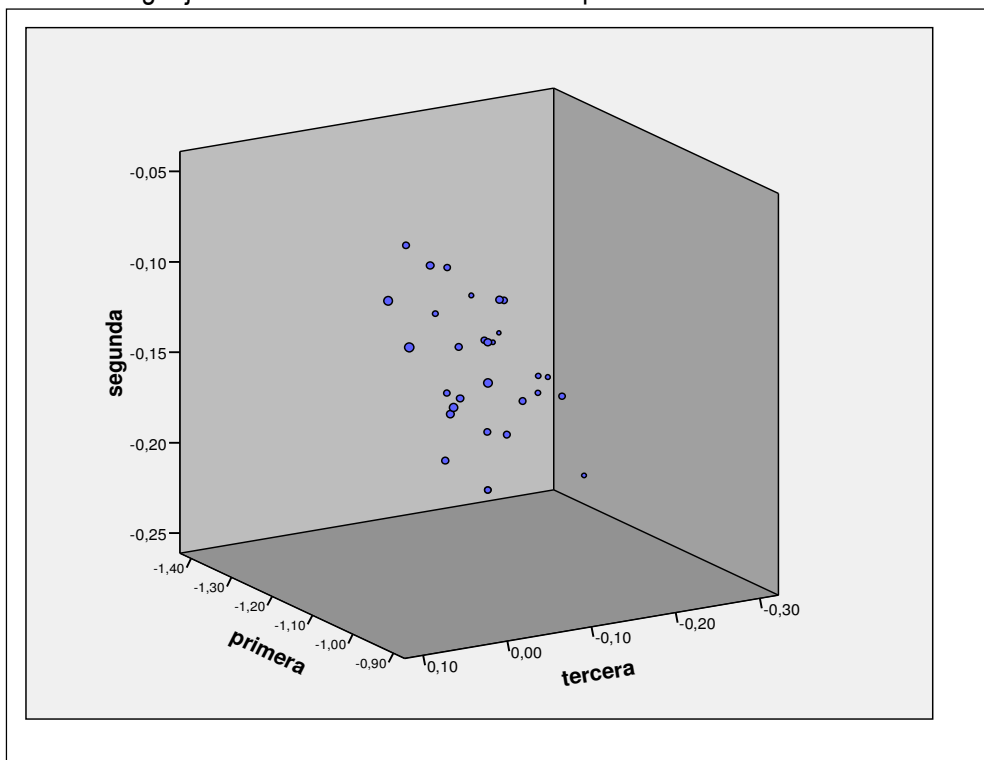
Carlos Díaz Gómez

Gráfico 5: Aguajes de Alto Amazonas en las dos primeras CPC



Para el segundo Grupo observamos que los datos están un poco dispersos ya que los puntos están más distanciados.

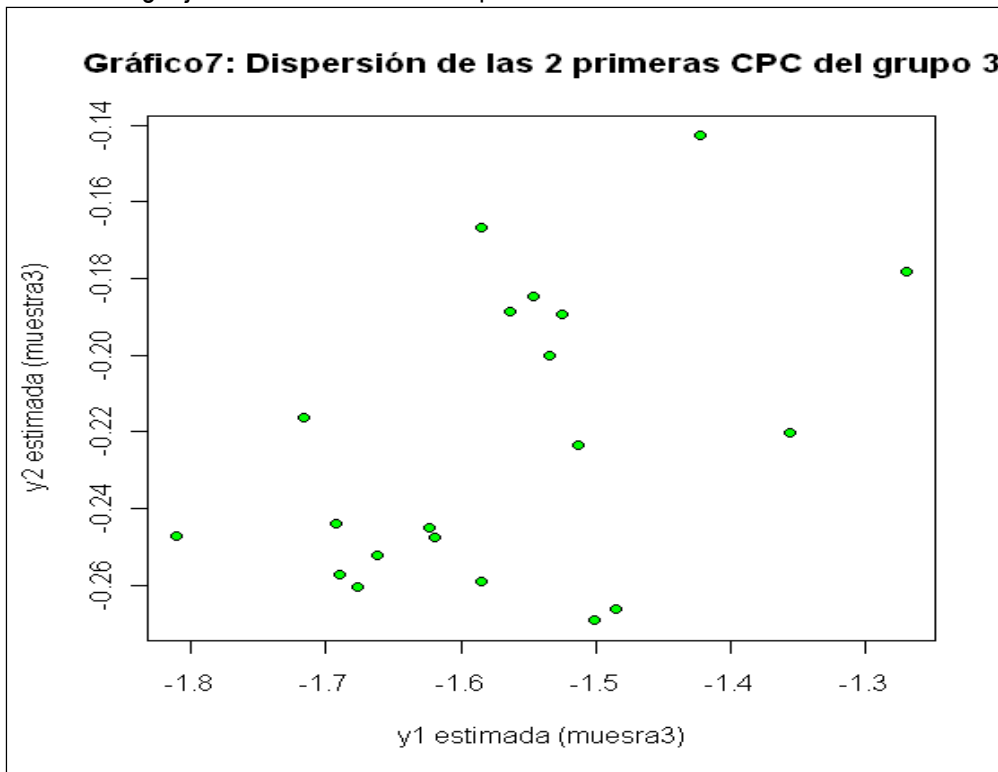
Gráfico 6: Aguajes de Alto Amazonas en las tres primeras CPC



**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

Gráfico 7: Aguajes de Loreto en las dos primeras CPC



Para el tercer grupo se observa que los datos están más dispersos, este grupo es el que tiene mayor dispersión de los grupos.

Gráfico 8: Aguajes de Loreto en las tres primeras CPC

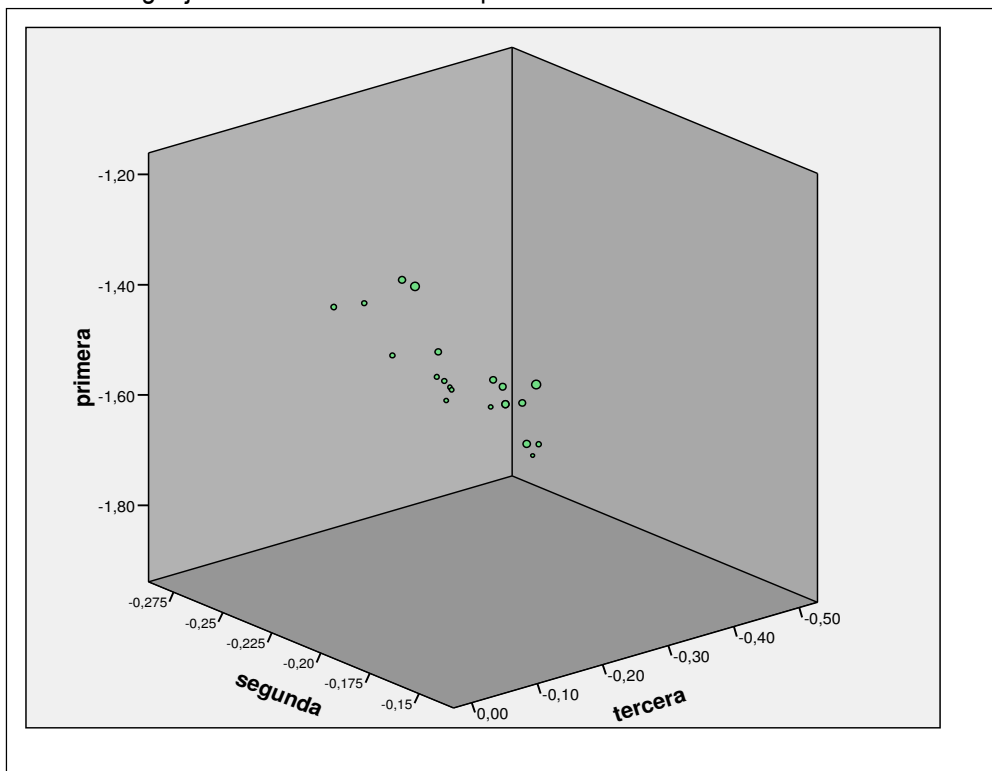
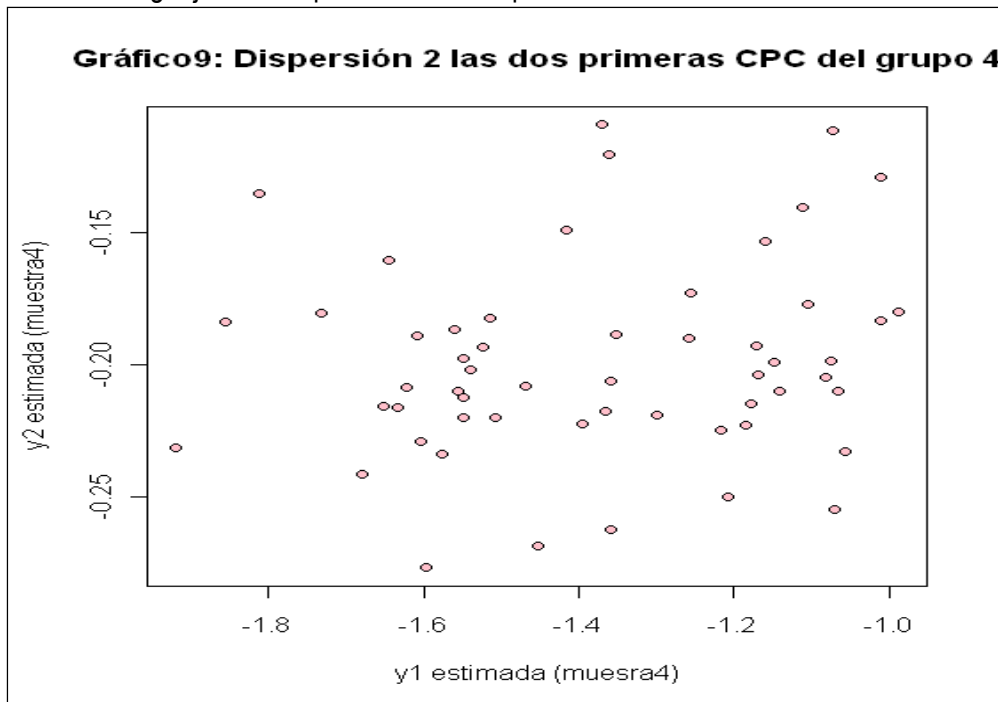
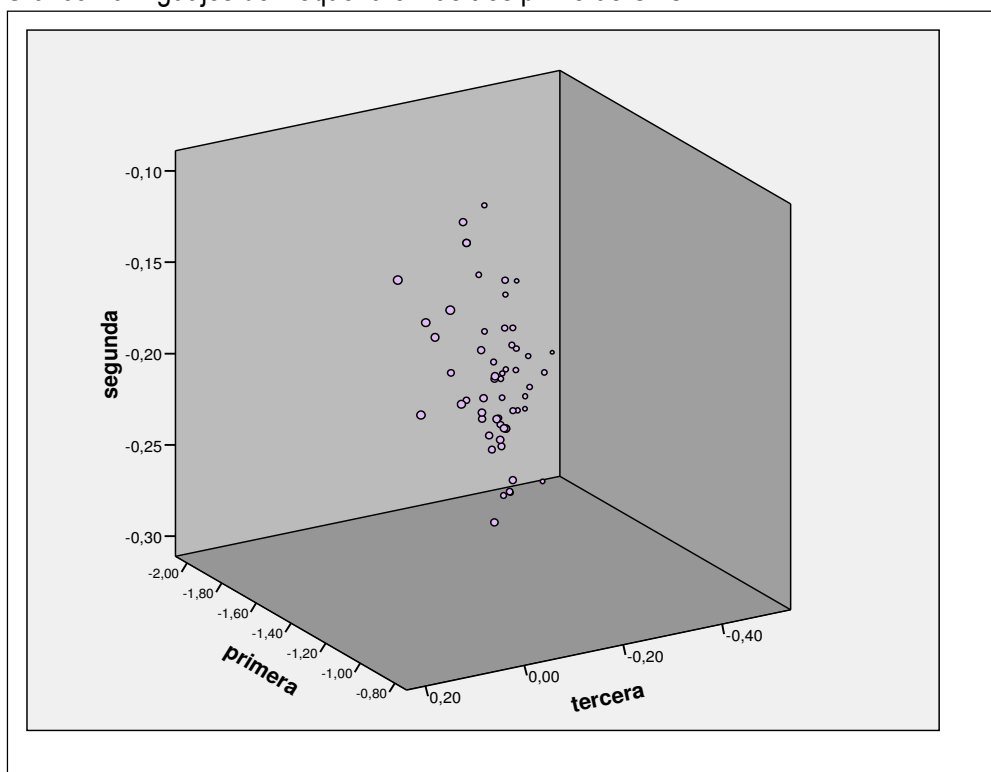


Gráfico 9: Aguajes de Requena en las dos primeras CPC



Para el cuarto grupo, se observa que están agrupados en su mayoría, que guarda una cierta concentración de los datos.

Gráfico 10: Aguajes de Requena en las tres primeras CPC



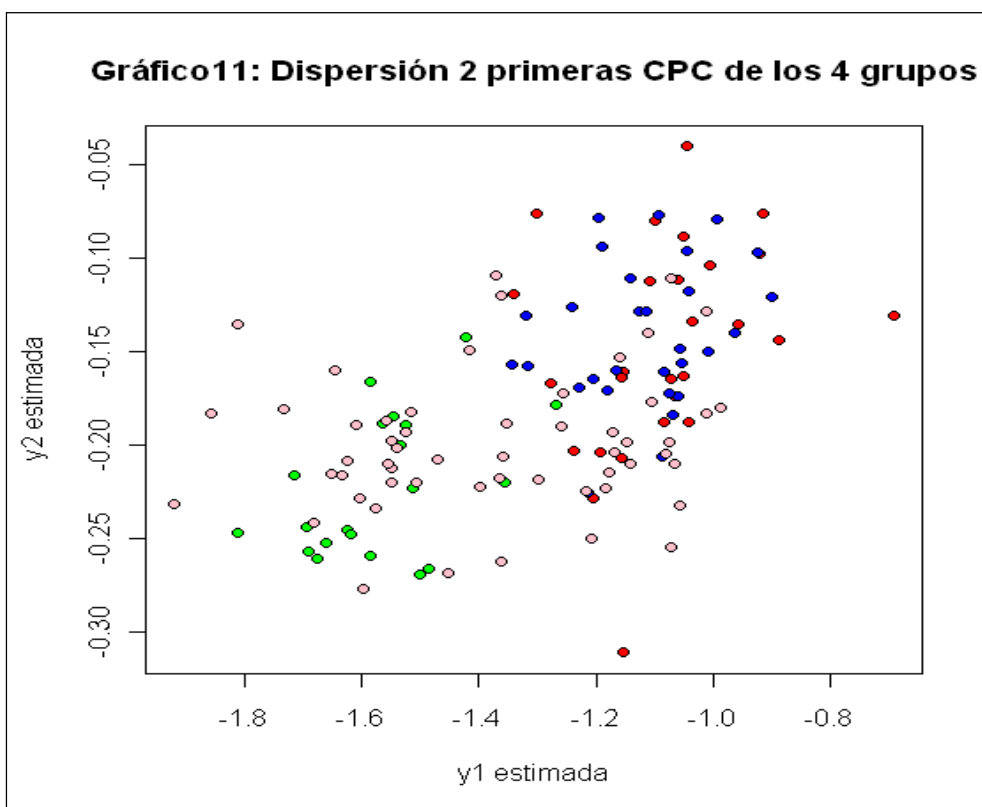
## “REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”

Carlos Díaz Gómez

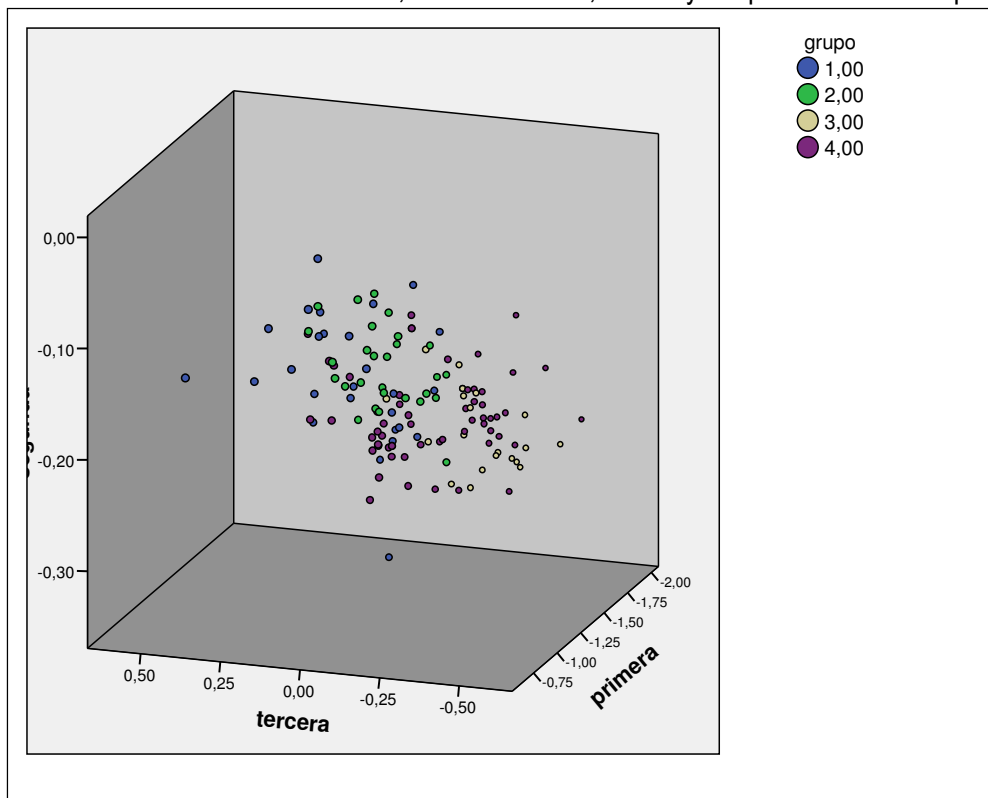
En el Gráfico 11 se presentan las muestras de los 4 lugares, Mainas, Alto Amazonas, Loreto y Requena, en las dos primeras Componentes Principales Comunes.

Se usó el color rojo para representar las muestras de Mainas, el color azul para representar las muestras de Alto Amazonas, el color verde para las muestras de Loreto y el color rosado para representar las muestras de Requena.

Gráfico 11: Muestras de Mainas, Alto Amazonas, Loreto y Requena en las dos primeras CPC



**Grafico 12:** Muestras de Mainas, Alto Amazonas, Loreto y Requena en las dos primeras CPC



Al observar el gráfico 11 se puede afirmar que las estructuras de covarianzas de las 4 muestras no son iguales, lo que ya ha sido comprobado a través de la correspondiente prueba de hipótesis, razón por la cual se postuló que las matrices de covarianzas de los datos seguían el modelo de Componentes Principales Comunes, es decir, que las matrices de covarianzas, sin llegar a ser completamente diferentes, guardan una estructura común, lo que queda expresado a continuación.

Se puede observar que los frutos de Mainas y Alto Amazonas (colores rojo y azul) tienen estructuras similares, es decir, son más parecidos entre ellos que con los frutos de los otros dos lugares y que están agrupados al lado derecho del gráfico. Los frutos de Loreto y Requena (colores verde y rosado) también tienen estructuras similares y se agrupan al lado izquierdo del gráfico.

## **CONCLUSIONES**

- El método de Componentes Principales Comunes que tiene como objetivo encontrar estructuras comunes de matrices de covarianzas de los datos cuando se tiene como mínimo 3 muestras, permite la obtención de nuevas variables o combinaciones lineales, en cada una de las poblaciones concurrentes, con propiedades similares a las variables Componentes Principales (CP) en un grupo. A estas variables se les denomina Componentes Principales Comunes (CPC).
- Luego de obtener las CPC se pueden elegir las dos primeras en cada una de las poblaciones, por que serán las combinaciones lineales que retengan la máxima variabilidad contenida en cada una de de las muestras multivariantes, en espacios vectoriales cuyas dimensiones son superiores a 3.
- Es posible realizar los gráficos de los datos originales, en las dos primeras CPC, para cada una de las poblaciones concurrentes.
- Es posible realizar en un solo gráfico de dimensión 2, los datos de todos los grupos o muestras, usando las dos primeras variables CPC.
- Para ilustrar la metodología de CPC se consideró las mediciones de 6 variables en agujajes de Mainas, Alto Amazonas, Loreto y Requena. Se comprobó a través de la prueba de hipótesis que los datos siguen el modelo de CPC, es decir las matrices de covarianzas guardan una estructura común, lo que permitió la construcción de las nuevas variables o combinaciones lineales denominadas CPC.
- El gráfico de las 4 muestras en las dos primeras CPC permite decir que los datos tienen una estructura común.
- A través del gráfico 11 podemos decir que:

En el Grafico 11 se puede observar que los frutos de Mainas y Alto Amazonas (colores rojo y azul) tienen estructuras similares, es decir, son mas parecidos entre ellos que con los frutos de los otros dos lugares y que están agrupados al lado



## **“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

---

derecho del gráfico. Los frutos de Loreto y Requena (colores verde y rosado) también tienen estructuras similares y se agrupan al lado izquierdo del gráfico.

Es decir, gráficamente se puede comprobar el resultado al que habíamos llegado a través de la prueba de hipótesis del modelo de CPC, es decir existe una estructura común y las CPC estarían representando muy bien a cada una de las muestras.

## **REFERENCIAS BIBLIOGRÁFICAS**

1. Anderson, T. W. (1963). “*Asimptotic Theory for Principal Component Analysis*”. *Annals Mathematical Statistics*, 34, 122-148.
2. Boente, G. (2008). “*Comportamiento asintótico de los estimadores projection - pursuit para el modelo de componentes principales*”. Tesis para optar el título profesional de Licenciada. Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires – Argentina.
3. David Houle, Jason Mezey, Paul Galpern (2002). “*Interpretation of the results of common principal components analyses*”. *Evolution* 56(3): 433 – 440. 2002
4. Figueroa, J. (2010). “*Medida de influencia en el análisis de componentes principales comunes. Una aplicación en botánica*”. Tesis para optar el título de profesional de Licenciada de Estadística. EAP Estadística, Facultad de Ciencias Matemáticas UNMSM. Lima – Perú.
5. Flury B. (1983) “*Some relations between the comparison of covariance matrices and principal component analysis*”. *Computational Statistics & Data Analysis*. 1: 97- 109.
6. Flury, B. (1984) “*Common Principal in k groups*”. *JASA*.79, 892-898.
7. Flury, Bernhard N. and Gautschi W. (1984). “*An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form*”. *Sci. Statist. Comput.* 7, pp. 169-184.
8. Flury, B.(1986). “*Asymptotic theory of common principal components analysis*”. *Annals of Statistics*. **14**, 418-430.
9. Flury, B. (1988). “*Common principal components and related multivariate models*”. New York: John Wiley & Sons.
10. Gómez, D. “*Componentes principales generalizadas. 1. ed. Lima Perú*”: Fondo Editorial de la Universidad Nacional Mayor de San Marcos Lima - Perú, 2002. 89 p.

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

---

11. Gómez D; Solano, O; Albán J; Vásquez C; Adriazola Y; Quinteros Y. (2008). *Determinación de patrones de variación morfológica del género Minthostachys en Unchus y Cajatambo, mediante métodos estadísticos multivariantes de reducción de datos*. Facultad de Matemática, UNMSM. Pesquimat. Vol.XI-N 1Pgs.45-56. Lima-Perú.
12. Klingenberg, C. & Neuenschwander, B. & Flury, B. “*Ontogeny and individual variation: Analysis of patterned covariance matrices with common principal components*” Vol 45(2) Pgs. 135-150 Año 1996.
13. Manly, B. (2005). “*Multivariate statistical methods. New York*”: Chapman & Hall/CRC. Third edition.
14. Mardia, K. V., Kent, J. T., Bibby, J. M. (1979). “*Multivariate Analysis. Academic Press*”.
15. Phillips, P. & Arnold, S. (1999). “*Hierarchical comparison of genetics variance-covariance matrices*”. I. Using the Flury hierarchy. *Evolution*. **53**, 506-1515.
16. Quinteros G. Y. (2010). “*Datos no publicados de un estudio piloto*”. Unidad de Post Grado. Facultad de Ciencias Biológicas. UNMSM.
17. Rodríguez M., Boente G. y Pires, A. M. (2005). “*Robust tests for the common principal components model*”. Dpto. de Matemática and Instituto de Cálculo, Fac. de Ciencias Exactas y Naturales, Buenos Aires, Argentina, pp. 234-258.

# APÉNDICE

## **ANEXO 1: DATOS TRASFORMADOS, APLICANDO EL MÉTODO DE COMPONENTES PRINCIPALES COMUNES**

Datos transformados el método de Componentes Principales Comunes para el Primer grupo “Mainas”.

Nº Muestra	1era Componente	2da Componente	3era Componente	4ta Componente	5ta Componente	6ta Componente
1	-1,2044	-0,2284	-0,0502	-0,3473	-0,7769	-0,4180
2	-1,1525	-0,3111	-0,0934	-0,3265	-0,7188	-0,4583
3	-1,0411	-0,1877	0,1115	-0,3355	-0,9008	-0,4510
4	-0,9219	-0,0978	0,2168	-0,3688	-0,9619	-0,3848
5	-1,0611	-0,1116	0,0996	-0,3826	-0,8775	-0,3502
6	-0,9558	-0,1353	0,1544	-0,3660	-0,9065	-0,3870
7	-1,0513	-0,1628	0,1111	-0,3040	-0,9373	-0,4060
8	-0,8882	-0,1440	0,2511	-0,2883	-1,0389	-0,4458
9	-0,6915	-0,1312	0,4101	-0,2677	-1,1512	-0,4968
10	-1,1085	-0,1125	0,0981	-0,3774	-0,9091	-0,3673
11	-1,1521	-0,1608	0,0179	-0,3739	-0,8379	-0,3701
12	-1,2388	-0,2029	-0,0882	-0,3765	-0,7513	-0,3590
13	-1,0836	-0,1879	-0,1463	-0,1455	-0,8169	-0,3309
14	-1,0352	-0,1342	-0,0572	-0,1494	-0,8961	-0,3278
15	-0,9148	-0,0764	0,0895	-0,1457	-1,0157	-0,3434
16	-1,0038	-0,1039	-0,0120	-0,1604	-0,9225	-0,3047
17	-1,0994	-0,0804	-0,0599	-0,1757	-0,8982	-0,2706
18	-1,0667	-0,1737	-0,1279	-0,1553	-0,8197	-0,3088
19	-1,0452	-0,0404	0,0984	-0,2011	-1,0494	-0,3080
20	-1,2764	-0,1671	-0,1990	-0,2055	-0,8036	-0,2828
21	-1,1566	-0,1638	-0,1066	-0,1921	-0,8674	-0,3153
22	-1,0506	-0,0887	0,0927	-0,2960	-0,9426	-0,3368
23	-1,0705	-0,1647	0,0027	-0,2848	-0,8552	-0,3549
24	-1,1572	-0,2067	-0,1034	-0,2874	-0,7685	-0,3438
25	-1,3019	-0,0761	-0,1256	-0,2146	-0,9000	-0,2686
26	-1,1929	-0,2035	-0,1706	-0,1791	-0,8253	-0,3388
27	-1,3407	-0,1190	-0,1969	-0,2114	-0,8367	-0,2712

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

Datos transformados el método de Componentes Principales Comunes para el segundo grupo “Alto Amazonas”.

Nº Muestra	1era Componente	2da Componente	3era Componente	4ta Componente	5ta Componente	6ta Componente
1	-1,0437	-0,0960	-0,0730	-0,1667	-0,8453	-0,2639
2	-0,9001	-0,1205	0,0097	-0,1455	-0,8929	-0,3116
3	-1,0826	-0,1608	-0,1660	-0,1597	-0,7597	-0,2731
4	-1,0604	-0,1737	-0,0898	-0,2151	-0,7822	-0,3394
5	-1,1270	-0,1289	-0,0951	-0,2309	-0,7942	-0,3068
6	-1,2142	-0,2254	-0,2555	-0,2239	-0,6518	-0,3125
7	-1,1795	-0,1705	-0,1840	-0,2411	-0,7161	-0,2842
8	-1,2274	-0,1696	-0,2188	-0,2469	-0,6927	-0,2705
9	-1,2034	-0,1646	-0,1960	-0,2450	-0,7100	-0,2757
10	-1,0404	-0,1180	-0,0576	-0,2360	-0,7987	-0,3029
11	-1,1410	-0,1107	-0,1253	-0,2490	-0,7553	-0,2725
12	-1,0551	-0,1559	-0,1061	-0,2309	-0,7526	-0,3106
13	-0,9240	-0,0969	0,0917	-0,2941	-0,8502	-0,2958
14	-0,9628	-0,1398	0,0204	-0,2910	-0,7869	-0,2984
15	-1,0916	-0,0774	-0,0137	-0,3172	-0,7854	-0,2429
16	-1,3204	-0,1305	-0,1725	-0,1886	-0,8640	-0,3002
17	-1,3158	-0,1578	-0,1961	-0,1832	-0,8384	-0,3100
18	-1,3446	-0,1573	-0,2170	-0,1867	-0,8244	-0,3018
19	-1,2418	-0,1260	-0,0919	-0,2287	-0,8324	-0,3452
20	-1,0698	-0,1837	-0,0210	-0,1983	-0,8605	-0,4112
21	-1,1654	-0,1601	-0,0689	-0,2137	-0,8362	-0,3772
22	-1,1950	-0,0788	-0,0346	-0,3211	-0,8216	-0,2826
23	-1,0860	-0,2062	-0,0795	-0,2856	-0,7469	-0,3528
24	-1,0860	-0,2062	-0,0795	-0,2856	-0,7469	-0,3528
25	-1,1134	-0,1283	-0,0581	-0,2588	-0,8132	-0,2727
26	-1,0755	-0,1726	-0,0738	-0,2464	-0,7872	-0,2971
27	-1,1898	-0,0942	-0,0812	-0,2738	-0,8094	-0,2407
28	-1,0094	-0,1498	0,0019	-0,2731	-0,8001	-0,3040
29	-0,9944	-0,0792	0,0829	-0,2840	-0,8797	-0,2862
30	-1,0573	-0,1489	-0,0329	-0,2789	-0,7768	-0,2903

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

Datos transformados el método de Componentes Principales Comunes para el tercer grupo “Loreto”.

Nº Muestra	1era Componente	2da Componente	3era Componente	4ta Componente	5ta Componente	6ta Componente
1	-1,5466	-0,1846	-0,2081	-0,4262	-0,7422	-0,3112
2	-1,5134	-0,2234	-0,2218	-0,4154	-0,7195	-0,3326
3	-1,5466	-0,1846	-0,2081	-0,4262	-0,7422	-0,3112
4	-1,6926	-0,2440	-0,3637	-0,3534	-0,7020	-0,3289
5	-1,5011	-0,2693	-0,2463	-0,3265	-0,7729	-0,3903
6	-1,7163	-0,2163	-0,3539	-0,3611	-0,7182	-0,3136
7	-1,5852	-0,2591	-0,2588	-0,2946	-0,7992	-0,4126
8	-1,4846	-0,2664	-0,1912	-0,2816	-0,8426	-0,4430
9	-1,6906	-0,2572	-0,3353	-0,3073	-0,7479	-0,3825
10	-1,8106	-0,2471	-0,4372	-0,3533	-0,6654	-0,3185
11	-1,6765	-0,2605	-0,3507	-0,3352	-0,7195	-0,3602
12	-1,6237	-0,2451	-0,2961	-0,3318	-0,7619	-0,3702
13	-1,2691	-0,1782	-0,0509	-0,3412	-0,8579	-0,3558
14	-1,4221	-0,1427	-0,1296	-0,3654	-0,8168	-0,3020
15	-1,3557	-0,2202	-0,1570	-0,3438	-0,7713	-0,3447
16	-1,5345	-0,2000	-0,2357	-0,3689	-0,7375	-0,3184
17	-1,5248	-0,1892	-0,2178	-0,3697	-0,7533	-0,3178
18	-1,5631	-0,1885	-0,2457	-0,3743	-0,7346	-0,3068
19	-1,5845	-0,1665	-0,1864	-0,3884	-0,8198	-0,3265
20	-1,6620	-0,2523	-0,3290	-0,3821	-0,6932	-0,3316
21	-1,6188	-0,2476	-0,2922	-0,3779	-0,7198	-0,3422

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

Datos transformados el método de Componentes Principales Comunes para el cuarto grupo “Requena”.

Nº Muestra	1era Componente	2da Componente	3era Componente	4ta Componente	5ta Componente	6ta Componente
1	-1,5499	-0,1977	-0,2698	-0,3141	-0,7659	-0,3158
2	-1,5594	-0,1866	-0,2659	-0,3172	-0,7724	-0,3097
3	-1,3968	-0,2223	-0,1802	-0,2918	-0,8182	-0,3663
4	-1,3583	-0,2060	-0,1017	-0,3905	-0,8183	-0,3525
5	-1,6029	-0,2288	-0,3063	-0,4150	-0,6715	-0,2912
6	-1,5070	-0,2197	-0,2259	-0,4054	-0,7293	-0,3152
7	-1,6523	-0,2155	-0,2846	-0,4136	-0,6889	-0,2745
8	-1,6332	-0,2159	-0,2707	-0,4113	-0,6982	-0,2799
9	-1,6089	-0,1890	-0,2262	-0,4133	-0,7377	-0,2784
10	-1,5968	-0,2769	-0,3395	-0,2423	-0,7510	-0,4066
11	-1,5772	-0,2337	-0,2821	-0,2477	-0,8050	-0,3986
12	-1,4529	-0,2687	-0,2242	-0,2270	-0,8321	-0,4442
13	-1,4696	-0,2080	-0,2983	-0,2759	-0,7214	-0,3103
14	-1,2585	-0,1901	-0,1235	-0,2545	-0,8463	-0,3637
15	-1,4163	-0,1490	-0,2003	-0,2802	-0,8084	-0,3068
16	-1,6460	-0,1602	-0,2274	-0,4350	-0,7772	-0,2989
17	-1,7325	-0,1805	-0,3117	-0,4415	-0,7129	-0,2810
18	-1,6804	-0,2414	-0,3332	-0,4245	-0,6772	-0,3146
19	-1,5481	-0,2201	-0,2961	-0,3756	-0,6824	-0,3165
20	-1,5142	-0,1825	-0,2338	-0,3783	-0,7378	-0,3143
21	-1,5239	-0,1933	-0,2516	-0,3776	-0,7220	-0,3150
22	-1,5486	-0,2123	-0,2380	-0,4061	-0,7068	-0,3202
23	-1,5389	-0,2016	-0,2202	-0,4069	-0,7226	-0,3196
24	-1,3519	-0,1887	-0,0682	-0,3874	-0,8303	-0,3679
25	-1,8560	-0,1835	-0,3775	-0,4102	-0,7402	-0,2878
26	-1,8124	-0,1352	-0,2973	-0,4137	-0,8114	-0,2850
27	-1,9188	-0,2314	-0,4716	-0,4090	-0,6597	-0,2852
28	-1,3602	-0,2625	-0,1773	-0,3253	-0,7713	-0,4056
29	-1,5562	-0,2099	-0,2711	-0,3576	-0,7259	-0,3344
30	-1,6233	-0,2086	-0,3198	-0,3656	-0,6933	-0,3153
31	-1,0717	-0,2547	-0,0579	-0,3049	-0,7494	-0,3803
32	-1,1718	-0,1929	-0,0712	-0,3276	-0,7618	-0,3331
33	-1,1479	-0,1988	-0,0592	-0,3238	-0,7679	-0,3416
34	-1,0121	-0,1289	0,0523	-0,3134	-0,8438	-0,3380
35	-1,0655	-0,2097	-0,0675	-0,3052	-0,7345	-0,3482
36	-1,0750	-0,1986	-0,0636	-0,3083	-0,7410	-0,3421
37	-1,2169	-0,2247	-0,1234	-0,3573	-0,7126	-0,3455
38	-1,1784	-0,2145	-0,0847	-0,3546	-0,7424	-0,3531
39	-1,1687	-0,2038	-0,0669	-0,3554	-0,7582	-0,3524
40	-0,9884	-0,1800	0,0380	-0,2282	-0,8870	-0,3785

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

Nº Muestra	1era Componente	2da Componente	3era Componente	4ta Componente	5ta Componente	6ta Componente
41	-1,1420	-0,2099	-0,1059	-0,2408	-0,7789	-0,3448
42	-1,1420	-0,2099	-0,1059	-0,2408	-0,7789	-0,3448
43	-1,2089	-0,2498	-0,1369	-0,3327	-0,7054	-0,3646
44	-1,2562	-0,1726	-0,0957	-0,3519	-0,7602	-0,3273
45	-1,1847	-0,2230	-0,0924	-0,3346	-0,7449	-0,3630
46	-1,1049	-0,1772	-0,0708	-0,2921	-0,7633	-0,3373
47	-1,0575	-0,2326	-0,0903	-0,2766	-0,7308	-0,3678
48	-1,0812	-0,2049	-0,0805	-0,2844	-0,7471	-0,3525
49	-1,1589	-0,1531	0,0315	-0,4946	-0,7211	-0,2780
50	-1,0723	-0,1110	0,1375	-0,4920	-0,8077	-0,2892
51	-1,0106	-0,1830	0,1121	-0,4719	-0,7655	-0,3288
52	-1,3612	-0,1201	-0,1030	-0,3171	-0,8411	-0,3237
53	-1,3612	-0,1201	-0,1030	-0,3171	-0,8411	-0,3237
54	-1,3707	-0,1090	-0,0991	-0,3202	-0,8476	-0,3176
55	-1,2988	-0,2188	-0,1498	-0,2889	-0,7488	-0,3739
56	-1,3659	-0,2176	-0,1985	-0,2969	-0,7162	-0,3548
57	-1,1111	-0,1404	0,0673	-0,2809	-0,9234	-0,4020



**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

**ANEXO 2: SE PRESENTA LAS MUESTRAS DE LAS DOS PRIMERAS COMPONENTES  
PRINCIPALES COMUNES PARA LOS CUATRO GRUPOS.**

<b>GRUPO Nº1: MAINAS</b>		
<b>Nº Muestra</b>	<b>1era Componente</b>	<b>2da Componente</b>
1	-1,2044	-0,2284
2	-1,1525	-0,3111
3	-1,0411	-0,1877
4	-0,9219	-0,0978
5	-1,0611	-0,1116
6	-0,9558	-0,1353
7	-1,0513	-0,1628
8	-0,8882	-0,1440
9	-0,6915	-0,1312
10	-1,1085	-0,1125
11	-1,1521	-0,1608
12	-1,2388	-0,2029
13	-1,0836	-0,1879
14	-1,0352	-0,1342
15	-0,9148	-0,0764
16	-1,0038	-0,1039
17	-1,0994	-0,0804
18	-1,0667	-0,1737
19	-1,0452	-0,0404
20	-1,2764	-0,1671
21	-1,1566	-0,1638
22	-1,0506	-0,0887
23	-1,0705	-0,1647
24	-1,1572	-0,2067
25	-1,3019	-0,0761
26	-1,1929	-0,2035
27	-1,3407	-0,1190

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

<b>GRUPO Nº2: ALTO AMAZONAS</b>		
<b>Nº Muestra</b>	<b>1era Componente</b>	<b>2da Componente</b>
1	-1,0437	-0,0960
2	-0,9001	-0,1205
3	-1,0826	-0,1608
4	-1,0604	-0,1737
5	-1,1270	-0,1289
6	-1,2142	-0,2254
7	-1,1795	-0,1705
8	-1,2274	-0,1696
9	-1,2034	-0,1646
10	-1,0404	-0,1180
11	-1,1410	-0,1107
12	-1,0551	-0,1559
13	-0,9240	-0,0969
14	-0,9628	-0,1398
15	-1,0916	-0,0774
16	-1,3204	-0,1305
17	-1,3158	-0,1578
18	-1,3446	-0,1573
19	-1,2418	-0,1260
20	-1,0698	-0,1837
21	-1,1654	-0,1601
22	-1,1950	-0,0788
23	-1,0860	-0,2062
24	-1,0860	-0,2062
25	-1,1134	-0,1283
26	-1,0755	-0,1726
27	-1,1898	-0,0942
28	-1,0094	-0,1498
29	-0,9944	-0,0792
30	-1,0573	-0,1489

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

---

<b>GRUPO N°3: LORETO</b>		
<b>N° Muestra</b>	<b>1era Componente</b>	<b>2da Componente</b>
1	-1,5466	-0,1846
2	-1,5134	-0,2234
3	-1,5466	-0,1846
4	-1,6926	-0,2440
5	-1,5011	-0,2693
6	-1,7163	-0,2163
7	-1,5852	-0,2591
8	-1,4846	-0,2664
9	-1,6906	-0,2572
10	-1,8106	-0,2471
11	-1,6765	-0,2605
12	-1,6237	-0,2451
13	-1,2691	-0,1782
14	-1,4221	-0,1427
15	-1,3557	-0,2202
16	-1,5345	-0,2000
17	-1,5248	-0,1892
18	-1,5631	-0,1885
19	-1,5845	-0,1665
20	-1,6620	-0,2523
21	-1,6188	-0,2476

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

<b>GRUPO Nº4: REQUENA</b>		
<b>Nº Muestra</b>	<b>1era Componente</b>	<b>2da Componente</b>
1	-1,5499	-0,1977
2	-1,5594	-0,1866
3	-1,3968	-0,2223
4	-1,3583	-0,2060
5	-1,6029	-0,2288
6	-1,5070	-0,2197
7	-1,6523	-0,2155
8	-1,6332	-0,2159
9	-1,6089	-0,1890
10	-1,5968	-0,2769
11	-1,5772	-0,2337
12	-1,4529	-0,2687
13	-1,4696	-0,2080
14	-1,2585	-0,1901
15	-1,4163	-0,1490
16	-1,6460	-0,1602
17	-1,7325	-0,1805
18	-1,6804	-0,2414
19	-1,5481	-0,2201
20	-1,5142	-0,1825
21	-1,5239	-0,1933
22	-1,5486	-0,2123
23	-1,5389	-0,2016
24	-1,3519	-0,1887
25	-1,8560	-0,1835
26	-1,8124	-0,1352
27	-1,9188	-0,2314
28	-1,3602	-0,2625
29	-1,5562	-0,2099
30	-1,6233	-0,2086
31	-1,0717	-0,2547
32	-1,1718	-0,1929
33	-1,1479	-0,1988
34	-1,0121	-0,1289
35	-1,0655	-0,2097
36	-1,0750	-0,1986
37	-1,2169	-0,2247
38	-1,1784	-0,2145
39	-1,1687	-0,2038
40	-0,9884	-0,1800

**“REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES:  
UNA APLICACIÓN A LA BOTÁNICA”**

**Carlos Díaz Gómez**

---

<b>GRUPO Nº4: REQUENA</b>		
<b>Nº Muestra</b>	<b>1era Componente</b>	<b>2da Componente</b>
41	-1,1420	-0,2099
42	-1,1420	-0,2099
43	-1,2089	-0,2498
44	-1,2562	-0,1726
45	-1,1847	-0,2230
46	-1,1049	-0,1772
47	-1,0575	-0,2326
48	-1,0812	-0,2049
49	-1,1589	-0,1531
50	-1,0723	-0,1110
51	-1,0106	-0,1830
52	-1,3612	-0,1201
53	-1,3612	-0,1201
54	-1,3707	-0,1090
55	-1,2988	-0,2188
56	-1,3659	-0,2176
57	-1,1111	-0,1404

**ANEXO 3: PROGRAMAS EN R, LA CUAL SE HIZO TODOS LOS CÁLCULOS PARA LLEGAR A  
LOS RESULTADOS OPTIMOS.**

Programa donde realizamos paso a paso los diferentes cálculos, como se detalla, observamos el cálculo para poder desarrollar el test de Mardia, para probar la Multinormalidad, cálculo de las matrices ortogonales y la estadística de prueba, cálculos del test de hipótesis de componentes principales comunes, realizar los gráficos para cada uno de los 4 grupos y todos los grupos juntos con las dos primeras componentes principales comunes en  $R^2$

```
#TEST DE MARDIA
#llamamos los datos transformados al software
CORK=read.table("m4.txt",header=T)
CORK
#nombramos a las variables para que el programa pueda cogerlos
attach(CORK)
p=length(CORK)
n=length(x1)
media=mean(CORK)
S=((n-1)/n)*cov(CORK)
S
grs=matrix(nr=n,nc=n,0)
for (i in 1:n)
for (j in 1:n)
{
grs[i,j]=t(solve(S)%*%t(CORK[j,]-media))%*%t(CORK[i,]-media)
}

b1p=(1/n^2)*sum(grs^3)
b1p
b2p=(1/n)*sum(diag(grs)^2)
b2p
f=(1/6)*p*(p+1)*(p+2)
pvalor1=1-pchisq(((1/6)*n*b1p,f)
pvalor1
z=(b2p-p*(p+2))/sqrt(8*p*(p+2)/n)
pvalor2=1-pnorm(z)
pvalor2

#llamamos la base de datos de los 4 grupos-
#juntos, al Software.
aguaj=read.table("datos.txt",header=T)
aguaj
```

## “REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”

Carlos Díaz Gómez

---

attach(aguaj)

#obtenemos los datos en forma matricial

#de los 4 grupos diferentes.

X=as.matrix(aguaj[,1:6])

X

X1=as.matrix(aguaj[1:27,1:6])

X1

X2=as.matrix(aguaj[28:57,1:6])

X2

X3=as.matrix(aguaj[58:78,1:6])

X3

X4=as.matrix(aguaj[79:135,1:6])

X4

#Obtenemos las media general y las 4 medias

#de los grupos diferentes.?pair

media=mean(aguaj[,1:6])

media

media1=mean(aguaj[1:27,1:6])

media1

media2=mean(aguaj[28:57,1:6])

media2

media3=mean(aguaj[58:78,1:6])

media3

media4=mean(aguaj[79:135,1:6])

media4

#Obtenemos las matrices de covarianza general

#y las matrices de covarianzas de los 4 grupos.

covar=cov(aguaj[,1:6])

covar

covar1=cov(aguaj[1:27,1:6])

covar1

covar2=cov(aguaj[28:57,1:6])

covar2

covar3=cov(aguaj[58:78,1:6])

covar3

covar4=cov(aguaj[79:135,1:6])

covar4

#Tamaño de muestra de cada uno de los 4 grupos.

n1=length(X1[,1])

n2=length(X2[,1])

n3=length(X3[,1])

n4=length(X4[,1])

## “REPRESENTACIONES GRÁFICAS DE VARIABLES PRINCIPALES COMUNES: UNA APLICACIÓN A LA BOTÁNICA”

Carlos Díaz Gómez

---

n=n1+n2+n3+n4  
k=4

```
#Aquí desarrollamos el Estadístico de Prueba
B=n1*log(det(covar1))+n2*log(det(covar2))+n3*log(det(covar3))+n4*log(det(covar4))
B
A=n*log(det(covar))
A
EP=A-B
EP
```

```
eigen(covar)$values
Beta=eigen(covar)$vectors
Beta
diag(eigen(covar)$values)
```

```
F1=t(Beta)%*%covar1)%*%Beta
F1
F2=t(Beta)%*%covar2)%*%Beta
F2
F3=t(Beta)%*%covar3)%*%Beta
F3
F4=t(Beta)%*%covar4)%*%Beta
F4
```

```
Lamda1=diag(F1)
Lamda1
Lamda2=diag(F2)
Lamda2
Lamda3=diag(F3)
Lamda3
Lamda4=diag(F4)
Lamda4
```

```
L1=eigen(covar1)$values
L1
L2=eigen(covar2)$values
L2
L3=eigen(covar3)$values
L3
L4=eigen(covar4)$values
L4
```

```
#GRUPO1
y11=c(-1.2044,-1.1525,-1.0411,-0.9219,-1.0611,-0.9558,-1.0513,-0.8882,-0.6915,
-1.1085,-1.1521,-1.2388,-1.0836,-1.0352,-0.9148,-1.0038,-1.0994,-1.0667,-1.0452,
-1.2764,-1.1566,-1.0506,-1.0705,-1.1572,-1.3019,-1.1929,-1.3407)
```







