10
2ej

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

INTRODUCCIÓN A LAS TÉCNICAS DE COMPONENTES PRINCIPALES Y ANÁLISIS DISCRIMINANTE

T E S I S

QUE PARA OBTENER EL TITULO DE:

A C T U A R I A

P R E S E N T A :

YAZMIN ILIANA BARCENAS OROZCO

DIRECTOR DE TESIS:

DR. JOSÉ RODOLFO MENDOZA BLANCO

FACULTAD DE CIENCIAS
SECCION ESCOLAR

MÉXICO, D.F. 1997

VNIVERSIDAD NACIONAL
AVTÓNOMA DE
MEXICO

M. en C. Virginia Abrín Batule
Jefe de la División de Estudios Profesionales de la
Facultad de Ciencias
P r e s e n t e

Comunicamos a usted que hemos revisado el trabajo de Tesis:
INTRODUCCION A LAS TECNICAS DE COMPONENTES PRINCIPALES
Y ANALISIS DISCRIMINANTE
realizado por Yazmín Iliana Bárcenas Orozco

con número de cuenta 8935254-6 , pasante de la carrera de ACTUARIA

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	Dr. José Rodolfo Mendoza Blanco	José Rodolfo Mendoza Blanco
Propietario	Mat. Margarita Elvira Chávez Cano	M. E. Chávez C.
Propietario	M. en C. José Antonio Flores Díaz	
Suplente	M. en A.P. Pilar Alonso Reyes	
Suplente	F.M. Leticia Cañedo Suarez	Leticia Cañedo Suárez

Consejo Departamental de Matemáticas

M. en A.P. Pilar Alonso Reyes

A mis padres
Como respuesta a sus anhelos y esfuerzos.
Gracias por ayudarme a realizar este sueño.

Por que siempre me has dado lo mejor de tí, aún en los momentos difíciles. Gracias por estar a mi lado y creer en mí.
Para tí mami con todo mi cariño.

Para Cocito y la Baby:
Por que el amor que me brindan cada día, me impulsa a seguir adelante.

Para la Gordita:
Por estar aquí.

A mi abue y mi mamá Mago:
Por llenar mi vida con detalles de amor.

A mi Toty, a Mary y la Mu
Por su apoyo y su confianza incondicional, en especial por su cariño, para el que no existen palabras que expresen lo que ha significado para mi.

Para mis amigos: Laurita, Mayreshon, Karla, Luis, Marianelo, Alito y Jaime.

A Oscar:
Por que de muchas maneras me expresas tú cariño.

RECONOCIMIENTOS

De manera especial, deseo expresar mi agradecimiento al Dr. Mendoza, por sus consejos y por el apoyo brindados para la realización de este trabajo.
Gracias Profe, por darme parte de su valioso tiempo.
Con cariño, admiración y respeto:

Yazmín.

Doy las gracias a los Profesores:

Margarita Chávez Cano
Pilar Alonso Reyes
Leticia Cañedo Suárez y
José Antonio Flores

Quienes muy amablemente me proporcionaron su tiempo.

Introducción

En los últimos años las técnicas que comprende el Análisis Multivariado han cobrado un auge muy importante, ya que gracias al desarrollo de las computadoras ha sido posible promover su uso en casi todas las ramas de la ciencia. El Análisis Multivariado se puede definir como un conjunto de Técnicas que describen o modelan el comportamiento de dos o más características hechas sobre cada individuo proveniente de una o varias poblaciones. Dado que la teoría de cada una de las técnicas multivariadas es amplia, el presente trabajo hace una recopilación de los resultados más interesantes de componentes principales y análisis discriminante.

El Análisis de Componentes Principales está considerado como una técnica descriptiva, para reducir la dimensión del espacio en el que originalmente se encuentra la muestra y en la que ningún supuesto distribucional se hace inicialmente. El objetivo primario en ésta es construir combinaciones lineales con base en las variables originales, tal que acumulen la mayor variabilidad de la muestra inicial.

El Análisis Discriminante el cual clasifica individuos con base en su vector de mediciones o atributos, donde dicha asignación se hace mediante las funciones de clasificación las cuales producen una partición del espacio dimensional que en particular contiene a las observaciones de la muestra.

Debido a que en la Estadística Multivariada, las generalizaciones de las densidades de probabilidad juegan un papel muy importante, en el Capítulo 1 se presenta una síntesis de algunas densidades multivariadas que son utilizadas para hacer inferencias y contrastar juegos de hipótesis para el caso multivariado.

El Capítulo 2, presenta los resultados típicos del Análisis de Componentes Principales,

donde bajo el supuesto de normalidad de la muestra inicial se construyen intervalos de confianza y pruebas de hipótesis relacionadas con la proporción de variación explicada por un cierto número de componentes principales. Se presenta en esta parte el análisis de la muestra de los Irises de Fisher (1936) y el análisis para una muestra simulada en S-Plus, utilizando el modulo FACTOR ANALYSIS del paquete STATISTICA obteniendo en cada caso los resultados y las gráficas relevantes.

En el Capítulo 3, se presenta una introducción a la teoría del Análisis Discriminante. Este capítulo está dividido básicamente en dos partes: La discriminación paramétrica, cuando se supone que la muestra tiene asociada una distribución que depende de un conjunto de parámetros posiblemente desconocidos y la discriminación no paramétrica cuando se supone que la distribución de la muestra no sigue ninguna forma paramétrica conocida pero puede ser estimada mediante el método de los estimadores de núcleo (Kernel).

El Capítulo 4 hace una comparación de los métodos descritos en el Capítulo 3, donde se simularon 3 muestras con ayuda del S-Plus las cuales fueron analizadas en el paquete SAS. Algunos de las gráficas se elaboraron con ayuda del paquete STATISTICA. Adicionalmente se incluye el análisis de los Irises de Fisher (1936).

El Apéndice A, recopila los resultados de Algebra, Cálculo, Estadística y Probabilidad que fueron utilizados a lo largo de esta tesis. En el Apéndice B, se presentan las muestras correspondientes a los ejemplos dados en el Capítulo 2, y finalmente el Apéndice C, recopila las muestras y los programas de que fueron utilizados para la elaboración de los ejemplos presentados en el Capítulo 4.

INDICE

Capítulo 1

ALGUNAS DENSIDADES MULTIVARIADAS

En este primer Capítulo se hace una recopilación de las funciones de densidad multivariadas que tienen un gran uso dentro de la Teoría de la Estadística Multivariada; para estas densidades se hace una pequeña discusión de sus principales características y propiedades.

1.1 Resultados Preliminares

Por conveniencia se utilizarán las letras mayúsculas X, Y, Z, W para denotar a las matrices y a los vectores aleatorios, donde estos últimos son vectores columna. Las primeras letras mayúsculas del abecedario denotarán generalmente a las matrices y vectores de constantes. Las letras minúsculas denotan solamente a los elementos de los vectores y las matrices de constantes.

Siempre que se tenga un conjunto de variables aleatorias ($v.as$) será utilizada la notación vectorial o matricial.

Definición 1.1 *$X_{p\times 1}$ es un vector aleatorio si al menos una de sus componentes es una $v.a.$*

Definición 1.2 *$X_{n\times p}$ es una matriz aleatoria si al menos una de sus de sus componentes X_{ij} es una v.a.*

1.1.1 Operadores de Esperanza y Covarianza para Distribuciones Multivariadas

Las siguientes definiciones extienden al caso multivariado las medidas descriptivas correspondientes tanto a un vector aleatorio como a una matriz aleatoria.

Definición 1.3 *Sea X un vector aleatorio de dimensión p. La esperanza del vector X está definida como:*

$$E(X) = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{pmatrix}.$$

Definición 1.4 *Si $X_{n\times p} = \{X_{ij}\}$ es una matriz aleatoria, la esperanza de esta matriz está dada por:*

$$E(X) = \begin{pmatrix} E(X_{11}) & \cdots & E(X_{1p}) \\ \vdots & \ddots & \vdots \\ E(X_{n1}) & \cdots & E(X_{np}) \end{pmatrix}.$$

Al igual que en el caso univariado deben mencionarse las propiedades que cumple el operador esperanza para el caso múltiple, y que simplemente es generalizar el caso para vectores y matrices aleatorias.

Teorema 1.1 *Sean X, Y y Z vectores aleatorios de dimensión $(p \times 1)$ y $A_{r\times p}$, $B_{p\times r}$, $C_{p\times 1}$ matrices de constantes. Bajo estas condiciones las siguientes propiedades se cumplen:*

1. $E(A) = A$.
2. $E(AX) = AE(X)$.

3. $E(X+Y) = E(X) + E(Y)$.

4. $E(AX + BY + C) = AE(X) + BE(Y) + C$.

Demostración.

1. *Tómese* a_{ij} *el* $ij-$ *ésimo elemento de la matriz de constantes* A, *con* $i = 1, .., r$ *y* $j = 1, ..., p$. *Dado que*

$$E(a_{ij}) = a_{ij},$$

de la Definición 1.4 se sigue el resultado.□

2. *El* $i-$ *ésimo elemento de* $E(AX)$ *está dado por :*

$$E\left[\sum_{j=1}^{p} a_{ij}X_j\right].$$

Utilizando la linealidad del operador esperanza en $\Re$, *se sigue*

$$E\left[\sum_{j=1}^{p} a_{ij}X_j\right] = \sum_{j=1}^{p} a_{ij}E(X_j),$$

en donde el lado derecho de esta igualdad corresponde al i-ésimo elemento de $AE(X)$.□

3. *El* $i-$ *ésimo elemento del vector* $(X+Y)$ *está dado por* $(X_i + Y_i) \in \Re$, *por la linealidad del operador Esperanza en* $\Re$ *se sigue que:*

$$E(X_i + Y_i) = E(X_i) + E(Y_i), \quad i = 1, ..., p.$$

Obsérvese que $E(X_i) + E(Y_i)$ *corresponde al* $i-$ *ésimo elemento de* $E(X) + E(Y)$ *y por lo tanto se sigue el resultado.*□

4. *Se sigue de aplicar los incisos 1 , 2 y 3 de este Teorema.*

Teorema 1.2 *Sean* $X_{n\times p}, Y_{n\times p}$ *matrices aleatorias y* $A_{r\times n}$, $B_{p\times s}$, $C_{r\times n}$, $D_{p\times s}$, $F_{n\times p}$, $G_{r\times s}$ *matrices de constantes, entonces se cumplen las siguientes propiedades:*

1. $(E(X^t))^t = E(X)$.

2. $E(X+Y+F) = E(X) + E(Y) + F$.

3. $E(AX) = AE(X)$.

4. $E(AXB) = AE(X)B$.

5. $E(AXB + CYD + G) = AE(X)B + CE(Y)D + G$.

Demostración.

1. *El elemento ji de la matriz aleatoria X^t es X_{ij}, aplíquese el operador esperanza para obtener el elemento ji de $E(X^t)$, dado por $E(X_{ij})$. Al trasponer $E(X^t)$, se obtiene que el elemento ij corresponde al elemento ji de $E(X^t)$, dado por $E(x_{ij})$, lo que demuestra el resultado.* □

2. *Sea $(X_{ij} + Y_{ij} + f_{ij})$ el elemento ij de la matriz $(X+Y+F)$, utilizando la linealidad del operador esperanza en $\Re$ se sigue*

$$E(X_{ij} + Y_{ij} + f_{ij}) = E(X_{ij}) + E(Y_{ij}) + f_{ij}.$$

Nótese de esta ecuación que $E(X_{ij}) + E(Y_{ij}) + f_{ij}$ corresponde al elemento ij de la matriz $E(X) + E(Y) + F$, y por lo tanto

$$E(X+Y+F) = E(X) + E(Y) + F.\square$$

3. *El elemento ij de la matriz AX es $\sum_{k=1}^{n} a_{ik}X_{kj}$ y aplicando el operador esperanza a este elemento se tiene*

$$E\left(\sum_{k=1}^{n} a_{ik}X_{kj}\right) = \sum_{k=1}^{n} a_{ik}E(X_{kj}),$$

el elemento del lado derecho de la ecuación anterior corresponde al $ij-$ésimo elemento de $AE(X)$, siguiéndose el resultado. □

4. Por aplicación de los incisos 1 y 4 se sigue

$$\begin{aligned} E(AXB) &= AE(XB) \\ &= A\left[E(XB)^t\right]^t \\ &= A\left[E(B^tX^t)\right]^t \\ &= A\left\{B^tE(X^t)\right\}^t \\ &= AE(X)B.\square \end{aligned}$$

5. Se sigue de los incisos 2 y 4 de este Teorema.□

La definición del operador covarianza entre dos variables reales puede extenderse al caso multivariado cuando se desea obtener la matriz de covarianzas entre dos vectores aleatorios que no necesariamente son de la misma dimensión. Esta misma generalización puede hacerse con las propiedades vistas para la covarianzas en los cursos básicos de Estadística, estos resultados se enuncian a continuación:

Definición 1.5 *La matriz de covarianzas entre dos vectores aleatorios $X_{p\times 1}$ y $Y_{q\times 1}$ está definida como:*

$$\begin{aligned} C(X,Y) &= E\left\{[X-E(X)]\,[Y-E(Y)]^t\right\} \\ &= \begin{pmatrix} C(X_1,Y_1) & \cdots & C(X_1,Y_q) \\ \vdots & \ddots & \vdots \\ C(X_p,Y_1) & \cdots & C(X_p,Y_q) \end{pmatrix}. \end{aligned}$$

Teorema 1.3 *La matriz de covarianzas de un vector aleatorio X denotada por $V(X)$, está definida por:*

$$\begin{aligned} V(X) &= C(X,X) \\ &= E\left[(X-E(X))\,(X-E(X))^t\right] \\ &= \begin{pmatrix} V(X_1) & C(X_1,X_2) & \cdots & C(X_1,X_p) \\ C(X_2,X_1) & V(X_2) & \cdots & C(X_2,X_p) \\ \vdots & \vdots & \ddots & \vdots \\ C(X_p,X_1) & C(X_p,X_2) & \cdots & V(X_p) \end{pmatrix}. \end{aligned}$$

Note que $V(X)$ es una matriz simétrica.

Teorema 1.4 *Sean $X_{p\times 1}$, $Y_{q\times 1}$, $Z_{q\times 1}$ vectores aleatorios y matrices de constantes $A_{r\times p}$, $B_{s\times q}$, $D_{q\times 1}$ y $F_{p\times 1}$ entonces se satisfacen las siguientes propiedades:*

1. $C(X,Y) = E(XY^t) - E(X)E(Y^t)$.
2. $C(X,Y) = C(Y,X)^t$.
3. $C(X,F) = 0$.
4. $C(AX,BY) = AC(X,Y)B^t$.
5. $C(X,Y+D) = C(X,Y)$.
6. $C(X,Y+Z) = C(X,Y) + C(X,Z)$.

Demostración.

1. *Utilizando el Teorema 1.2 y la Definición 1.5 se cumple*

$$\begin{aligned} C(X,Y) &= E\left[(X-E(X))(Y-E(Y))^t\right] \\ &= E\left[XY^t - XE(Y)^t - E(X)Y^t + E(X)E(Y)^t\right] \\ &= E(XY^t) - E(X)E(Y^t) - E(X)E(Y^t) + E(X)E(Y^t) \\ &= E(XY^t) - E(X)E(Y^t). \square \end{aligned}$$

2. *El elemento ij de la matriz de covarianzas $C(X,Y)$ es $C(X_i,Y_j)$ y el elemento ji de $C(Y,X)$ es $C(Y_j,X_i) = C(X_i,Y_j)$, pero $C(X_i,Y_j) = C(Y_j,X_i)$ satisfaciéndose el resultado.* $\square$

3. *Por la Definición 1.5 se cumple*

$$\begin{aligned} C(X,F) &= E(XF^t) - E(X)E(F^t) \\ &= E(X)F^t - E(X)F^t \\ &= 0. \square \end{aligned}$$

4. Siguiendo el inciso 4 del Teorema 1.2 se tiene

$$\begin{aligned} C(AX, BY) &= E\left[(AX - E(AX))(BY - E(BY))^t\right] \\ &= E\left[A(X - E(X))(Y - E(Y))^t B^t\right] \\ &= AE\left[(X - E(X))(Y - E(Y))^t\right] B^t \\ &= AC(X,Y) B^t . \square \end{aligned}$$

5. Utilizando la Definición 1.5, se sigue

$$\begin{aligned} C(X, Y + D) &= E\left(X(Y + D)^t\right) - E(X) E(Y + D)^t \\ &= E\left(XY^t\right) + E(X) D^t - E(X) E\left(Y^t\right) - E(X) D^t \\ &= E\left(XY^t\right) - E(X) E\left(Y^t\right) \\ &= C(X,Y) . \square \end{aligned}$$

6. Por el resultado 1 de este Teorema se satisface

$$\begin{aligned} C(X, Y + Z) &= E\left[X(Y + Z)^t\right] - E(X) E\left[(Y + Z)^t\right] \\ &= E\left[(X)(Y^t + Z^t)\right] - E(X) E\left(Y^t\right) - E(X) E\left(Z^t\right) \\ &= E\left(XY^t\right) + E\left(XZ^t\right) - E(X) E\left(Y^t\right) - E(X) E\left(Z^t\right) \\ &= E\left(XY^t\right) - E(X) E\left(Y^t\right) + E\left(XZ^t\right) - E(X) E\left(Z^t\right) \\ &= C(X,Y) + C(X,Z) \\ &= C(X,Y) . \square \end{aligned}$$

Teorema 1.5 *Si los vectores X y Y son independientes entonces:*

$$C(X,Y) = E\left[(X - E(X))(Y - E(Y))^t\right] = 0.$$

Demostración

Por el Teorema 1.4 inciso 1 se sigue

$$C(X,Y) = E\left(XY^t\right) - E(X) E\left(Y^t\right),$$

por la hipótesis de independencia entre los vectores X y Y se satisface

$$C(X,Y) = 0.\square$$

Teorema 1.6 *Sean X y Y dos vectores aleatorios de la misma dimensión. Entonces:*

$$V(X+Y) = V(X) + V(Y) + 2C(X,Y).$$

Demostración

por la Definición 1.3 y los incisos 2 y 6 del Teorema 1.4 se tiene

$$\begin{aligned} V(X+Y) &= C(X+Y, X+Y) \\ &= C(X,X) + C(Y,Y) + C(X,Y) + C(Y,X)^t \\ &= V(X) + V(Y) + 2C(X,Y).\square \end{aligned}$$

Teorema 1.7 *Si X y Y son dos vectores aleatorios independientes entonces*

$$V(X+Y) = V(X) + V(Y).$$

Demostración.

Se sigue de aplicar los Teoremas 1.5 y 1.6.$\square$

Teorema 1.8 *Sea X un vector aleatorio de dimensión p y $A_{r\times p}$, $B_{p\times 1}$ y $C_{p\times 1}$ matrices de constantes. Entonces se cumplen las siguientes propiedades:*

1. $V(C) = 0.$
2. $V(AX) = AV(X)A^t.$
3. $V(X+B) = V(X).$

Demostración.

1. *Sea $c_i \in \Re$ el $i-$ésimo elemento del vector C, entonces se sabe*

$$V(c_i) = 0 \quad y \quad C(c_i, c_j) = 0,$$

Por lo tanto se sigue el resultado. □

2. *Siguiendo la Definición 1.3 se tiene*

$$\begin{aligned} V(AX) &= C(AX, AX) \\ &= AC(X, X)A^t \\ &= AV(X)A^t. \square \end{aligned}$$

3. *Por el Teorema 1.6 se sigue*

$$V(X+B) = V(X) + V(B) + 2C(X, B).$$

Utilizando el inciso 1 de este Teorema, la Definición 1.3 y el inciso3 del Teorema 1.4, se obtiene finalmente el resultado. □

1.2 Densidad Normal Multivariada

La distribución Normal en la Estadística univariada juega un papel muy importante ya que muchos de los métodos estadísticos están basados en el supuesto de normalidad en la población, y gracias a éste ha sido posible desarrollar resultados con base en algunas Teorías como las de Estimación y Prueba de Hipótesis que son de gran importancia en el análisis de una muestra. Conociendo la importancia de esta distribución no es de asombrarse que la Normal multivariada posea un papel fundamental dentro de la Estadística Multivariada para el desarrollo de las aplicaciones como la generalización de los procedimientos de estimación y pruebas de hipótesis; y es por esta razón que en este Capítulo se enuncian las propiedades básicas de la distribución Normal Multivariada así como de sus pruebas.

1.2.1 Función de Densidad

A continuación se enuncian las Definiciones y Teoremas que caracterizan a la función de densidad de una normal tanto para el caso univariado como para el multivariado:

Caso Univariado

Definición 1.6 *Una v.a. X se distribuye como una normal estándar univariada con media 0 y varianza 1, si su función de densidad está dada por:*

$$f_X(x;0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2},$$

con $x \in \Re$. Esto se denota como $X \sim N(0,1)$.

A partir de esta definición construyamos una *v.a.* Y con cualquier media μ y cualquier varianza σ^2 definida como:

$$Y = \sigma X + \mu,$$

en la que la *v.a.* $X \sim N(0,1)$ y la esperanza y varianza de Y se obtienen utilizando las propiedades de los Teoremas 1.1 y 1.8 respectivamente y están dados por:

$$\begin{aligned} E(Y) &= E(\sigma X + \mu) \\ &= E(\sigma X) + \mu \\ &= \sigma E(X) + \mu \\ &= \mu \end{aligned}$$

y

$$\begin{aligned} V(Y) &= V(\sigma X + \mu) \\ &= V(\sigma X) \\ &= \sigma^2 V(X) \\ &= \sigma^2. \end{aligned}$$

Donde μ y σ^2 representan los parámetros de la distribución de Y, y que satisfacen $-\infty < \mu < \infty$ y $\sigma^2 > 0$. Con esto se puede enunciar la siguiente definición:

Teorema 1.9 *Sea $Y = \sigma X + \mu$ una v.a. con media μ y varianza σ^2, donde $X \sim N(0,1)$. Se dice que Y se distribuye normal con media μ y varianza σ^2 si su función de densidad está dada por:*

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) \quad -\infty < \mu < \infty \quad \sigma > 0,$$

denotándose como $Y \sim N(\mu, \sigma^2)$.

Demostración.

Utilizando el Teorema del Cambio de Variable y se sigue

$$\begin{aligned} f_Y(y) &= f_X\left(\frac{y-\mu}{\sigma}\right)\left|\frac{1}{\sigma}\right| \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}. \square \end{aligned}$$

Caso Multivariado

Para generalizar el caso univariado, tómese ahora un vector aleatorio $X \in \Re^p$, donde cada una de las entradas $X_1, ..., X_p$ de X son *v.as.* independientes e idénticamente distribuidas (*v.a.i.i.d.*) como $N(0,1)$; con base en ello se establece la siguiente definición:

Definición 1.7 *Se dice* X *tiene densidad normal multivariada de media* 0 *y matriz de covarianzas* I_p *si* $X = (X_1, ..., X_p)$ *donde* $X_1, ..., X_p$ *son v.a.i.i.d. como* $N(0,1)$. *Esto se denota por* $X \sim N_p(0, I_p)$.

Teorema 1.10 *Sea* $X = (X_1, \ldots, X_p)$ *un vector aleatorio de dimensión* p. X *se distribuye como una normal multivariada estándar con media* $\underline{0}$ *y varianza* I_p *si su función de densidad está dada por:*

$$\begin{aligned} f_{X_1,...,X_p}(x_1,...,x_p) &= f_X(x_1,...,x_p) \\ &= |2\pi I|^{-\frac{1}{2}}\exp\left\{-\tfrac{1}{2}x^t x\right\}. \end{aligned}$$

Demostración.

$$\begin{aligned} f_{X_1,...,X_p}(x_1,...,x_p) &= \prod_{i=1}^{p} f_{X_i}(x_i) \\ &= \textstyle\prod_{i=1}^{p}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}x^2\right\}, \end{aligned}$$

utilizando la independencia sobre $X_1, ..., X_p$ *se sigue*

$$\begin{aligned} f_{X_1,...,X_p}(x_1,...,x_p) &= (2\pi)^{-\frac{p}{2}}\exp\left\{-\tfrac{1}{2}\textstyle\sum_{i=1}^{p}x_i^2\right\} \\ &= |2\pi I_p|^{-\frac{1}{2}}\exp\left\{-\tfrac{1}{2}X^t X\right\}. \square \end{aligned}$$

Ahora si se toma cualquier vector de medias $\mu \in \Re^p$ y cualquier matriz $\Sigma_{p\times p} = \{\sigma_{ii}\} > 0$ con $i = 1, ..., p$, para definir $Y = \Sigma^{\frac{1}{2}} X + \mu$ un vector aleatorio de dimensión p, con base en un vector $X \sim N_p(0, I_p)$. Como se verá más adelante (ver Teorema 1.18 inciso 1) cada $Y_i \in Y$ $i = 1, ..., p$ es una *v.a.* con distribución $Y_i \sim N(\mu_i, \sigma_{ii})$. Luego para el vector Y se obtiene el vector de medias y la matriz de covarianzas, utilizando los Teoremas 1.1 y 1.8, así pues

$$\begin{aligned} E(Y) &= E\left(\Sigma^{\frac{1}{2}} X + \mu\right) \\ &= E\left(\Sigma^{\frac{1}{2}} X\right) + \mu \\ &= \Sigma^{\frac{1}{2}} E(X) + \mu \\ &= \mu. \\ V(Y) &= V\left(\Sigma^{\frac{1}{2}} X + \mu\right) \\ &= V\left(\Sigma^{\frac{1}{2}} X\right) \\ &= \Sigma^{\frac{1}{2}} V(X) \Sigma^{\frac{1}{2}} \\ &= \Sigma. \end{aligned}$$

Definición 1.8 *Se dice que* $Y \sim N_p(\mu, \Sigma)$ *si* $Y = \Sigma^{\frac{1}{2}} X + \mu$ *con* $X_p \sim N_p(0, I_p)$.

Teorema 1.11 *Se dice que* $Y = \Sigma^{\frac{1}{2}} X + \mu$ *tiene una distribución normal multivariada con media el vector* μ *y matriz de escala* $\Sigma > 0$, *si la función de densidad de* Y *está dada por:*

$$\begin{aligned} f_{Y_1,...,Y_p}(y_1, \ldots, y_p) &= f_Y(y_1, \ldots, y_p) \\ &= |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(y-\mu)^t \Sigma^{-1}(y-\mu)\right\}. \end{aligned}$$

donde Y, $\mu \in \Re^p$ *y* $\Sigma_{p\times p}$ *es una matriz definida positiva. En notación se escribe* $Y \sim N_p(\mu, \Sigma)$.

Demostración.

De acuerdo al Teorema del Cambio de Variable se sigue

$$\begin{aligned} f_Y(y) &= f_X\left(\Sigma^{-\frac{1}{2}}(Y-\mu)\right)\left|\Sigma^{-\frac{1}{2}}\right| \\ &= |2\pi I_p|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}\left[\Sigma^{-\frac{1}{2}}(Y-\mu)\right]^t \left[\Sigma^{-\frac{1}{2}}(Y-\mu)\right]\right\}\left|\Sigma^{-\frac{1}{2}}\right| \\ &= |2\pi I_p|^{-\frac{1}{2}}\left|\Sigma^{-\frac{1}{2}}\right| \exp\left\{-\tfrac{1}{2}(Y-\mu)^t \Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}(Y-\mu)\right\} \\ &= |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(Y-\mu)^t \Sigma^{-1}(Y-\mu)\right\}. \square \end{aligned}$$

Teorema 1.12 *Los vectores aleatorios $X_i \in \Re^p$, $i = 1, \ldots, p$ son independientes si y sólo si su densidad conjunta puede escribirse como:*

$$f_{X_1,\ldots,X_p}(x_1, \ldots, x_p) = \prod_{i=1}^{p} H_i(x_i) \quad \forall\ (x_1, \ldots, x_k) \in \Re^p,$$

donde cada $H_i(x_i)$ son funciones arbitrarias de X_i para $i = 1, ..., p$.

Demostración.

$\Rightarrow$] *Por hipótesis $X_1, ..., X_p$ son v.a.i.i.d. entonces*

$$f_{X_1,\ldots,X_p}(x_1, ..., x_p) = f_{X_1}(x_1) \cdots f_{X_p}(x_p),$$

luego si se toma $H_i(x_i) = f_{X_i}(x_i)$ se obtiene

$$\begin{aligned} f_{X_1,\ldots,X_p}(x_1, ..., x_p) &= \textstyle\prod_{i=1}^{p} f_{X_i}(x_i) \\ &= \textstyle\prod_{i=1}^{p} H_i(x_i). \square \end{aligned}$$

$\Leftarrow$]

$$\begin{aligned} f_{X_i}(x_i) &= \textstyle\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\ldots,X_p}(x_1, ..., x_p)\, dx_1 ....dx_p \\ &= \textstyle\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^{p} H_i(x_i)\, dx_1 ...dx_p \\ &= \left(\textstyle\prod_{j\neq i}^{p} \int_{-\infty}^{\infty} H_j(x_j)\, dx_j\right) H_i(x_i) \\ &= \left(\textstyle\prod_{j\neq i}^{p} h_j\right) H_i(x_i), \end{aligned}$$

donde $h_j = \int_{-\infty}^{\infty} H_j(x_j)\, dx_j$, entonces

$$\begin{aligned} \textstyle\prod_{i=1}^{p} f_{X_i}(x_i) &= \textstyle\prod_{i=1}^{p} \left(H_i(x_i) \prod_{j\neq i}^{p} h_j\right) \\ &= \textstyle\prod_{i=1}^{p} H_i(x_i)\, (h_1 ... h_p)^{p-1}, \end{aligned}$$

como

$$\begin{aligned} 1 &= \textstyle\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\ldots,X_p}(x_1, ..., x_p)\, dx_1 ....dx_p \\ &= \textstyle\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^{p} H_i(x_i)\, dx_1 ...dx_p \\ &= \textstyle\prod_{i=1}^{p} \int_{-\infty}^{\infty} H_i(x_i)\, dx_i \\ &= \textstyle\prod_{i=1}^{p} h_i(x_i), \end{aligned}$$

por lo tanto

$$\begin{aligned} \textstyle\prod_{i=1}^{p} f_{X_i}(x_i) &= \textstyle\prod_{i=1}^{p} H_i(x_i) \\ &= f_{X_1,\ldots,X_p}(x_1, ..., x_p). \square \end{aligned}$$

1.2.2 Función Generadora de Momentos

La función generadora de momentos juega un papel muy importante en la Teoría de Probabilidad y Estadística, ya que, ésta caracteriza de manera única a cada función de densidad de probabilidad. En otras palabras, si se conoce la expresión de la función generadora de momentos de alguna $m.a.$, se conoce también la distribución a la que pertenece. Los siguientes resultados establecen su definición:

Definición 1.9 *Sea X una v.a. con densidad $f_X(x)$. La Función Generadora de Momentos (F.G.M.) es el valor esperado de e^{tX}, si este valor existe para t en un intervalo $-h < t < h$, $h > 0$. La F.G.M que se denota por $M_X(t)$ o $M(t)$, se define por:*

$$\begin{aligned} M_X(t) &= E\left(e^{tX}\right) \\ &= \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx, \end{aligned}$$

si la v.a. X es continua, y como

$$\begin{aligned} M_X(t) &= E\left(e^{tX}\right) \\ &= \sum_x e^{tx} f_X(x), \end{aligned}$$

si la v.a. X es de tipo discreto.

Teorema 1.13 *Sea $X \in \Re$ una v.a., si $X \sim N(\mu, \sigma^2)$ la F.G.M. de X está dada por:*

$$M_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right), \text{ con } t \in \Re.$$

Demostración.

Por definición se sabe que

$$\begin{aligned} M_X(t) &= E\left(e^{tX}\right) \\ &= E\left[\exp\left(tx + t\mu - t\mu\right)\right] \\ &= \exp(t\mu)\, E\left[\exp\left(t(x-\mu)\right)\right] \\ &= \exp(t\mu) \int_{-\infty}^{\infty} \exp\left\{t(x-\mu)\right\} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx \\ &= \exp(t\mu) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left((x-\mu)^2 - 2\sigma^2 t(x-\mu)\right)\right\} dx, \end{aligned}$$

Completando el cuadrado en el exponente de X, se tiene

$$\begin{aligned}(x-\mu)^2 - 2\sigma^2 t\,(x-\mu) &= (x-\mu)^2 - 2\sigma^2 t\,(x-\mu) + \sigma^4 t^2 - \sigma^4 t^2 \\ &= ((x-\mu) - \sigma^2 t)^2 - \sigma^4 t^2,\end{aligned}$$

y de aquí puede obtenerse

$$M_X(t) = \exp(t\mu)\exp\left(\frac{\sigma^2 t^2}{2}\right)\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}\left(x-(\mu+\sigma^2 t)\right)^2\right\}dx,$$

Donde el integrando es una f.d. de $X \sim N(\mu+\sigma^2 t, \sigma^2)$, la cual integrada sobre todo su recorrido es igual a 1. y por lo tanto

$$M_X(t) = \exp\left(t\mu + \frac{\sigma^2 t^2}{2}\right).\square$$

Definición 1.10 *Sea $X \in \Re^p$ un vector aleatorio, si X se distribuye como $f_X(x_1,...,x_p)$, entonces la F.G.M. que se denota como $M_X(t) = E(\exp(t^t X))$, donde $t \in \Re^p$, se define por:*

$$M_X(t) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left(t'x\right) f_X(x_1,...,x_p)\,dx_1,...,dx_p.$$

Si X tiene una función de densidad continua y por:

$$M_X(t) = \sum_x \exp(t^t x)\, f_X(x_1,...,x_p).$$

Si X es discreta, donde la suma corre sobre todos los valores de $(x_1,...,x_p)$, y en ambos casos la función existe en un intervalo abierto que contiene al cero.

Teorema 1.14 *Sea X un vector aleatorio de dimensión p, donde cada uno de sus componentes X_i son v.a.i.i.d. como una normal estándar univariada y sea $t \in \Re^p$. La F.G.M. para X está dada por:*

$$M_X(t) = \exp\left(\frac{1}{2}t't\right) \text{ con } t \in \Re^p.$$

Demostración.

$$\begin{aligned}M_X(t) &= E\left(\exp\left(t'X\right)\right) \\ &= \int_{-\infty}^{\infty}\exp(t^t x)\,|2\pi I_p|^{-\frac{1}{2}}\exp\left\{-\tfrac{1}{2}x^t x\right\}dx.\end{aligned}$$

Completando la forma cuadrática en el exponente del vector X se cumple

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} |2\pi I_p|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(x-t)^t(x-t) + \tfrac{1}{2}t^t t\right\} dx \\ &= \exp\left(\tfrac{1}{2}t^t t\right) \int_{-\infty}^{\infty} |2\pi I_p|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(x-t)^t(x-t)\right\} dx, \end{aligned}$$

donde el integrando es una función de X que se distribuye como una normal con media el vector t y matriz de escala I_p, la cual integrada sobre todo su recorrido es igual a 1; por lo tanto

$$M_X(t) = \exp\left(\frac{1}{2}t^t t\right) \quad \forall\ t \in \Re^p. \square$$

Teorema 1.15 *Sea $Y_{p\times 1}$ un vector aleatorio con densidad $N(\mu, \Sigma)$, esto es, donde $Y = \Sigma^{\frac{1}{2}}X + \mu$ y $X \sim N_p(0, I_p)$. La F.G.M. de Y está dada como:*

$$M_Y(\delta) = \exp\left(\delta^t\mu + \frac{1}{2}\delta^t\Sigma\delta\right) \ \forall\ \delta \in \Re^p.$$

Demostración.

$$\begin{aligned} M_Y(\delta) &= E\left(\exp\left(\delta^t Y\right)\right) \\ &= E\left(\exp\left(\delta^t\left(\Sigma^{\frac{1}{2}}X + \mu\right)\right)\right) \\ &= \exp\left(\delta^t\mu\right) E\left(\exp\left(\delta^t\Sigma^{\frac{1}{2}}X\right)\right) \\ &= \exp\left(\delta^t\mu\right) E\left(\exp\left(t^t X\right)\right), \end{aligned}$$

donde $t = \Sigma^{\frac{1}{2}}\delta$. Ahora utilizando el resultado del Teorema 1.14 se sigue que:

$$\begin{aligned} M_Y(\delta) &= \exp\left(\delta^t\mu\right)\exp\left(\tfrac{1}{2}t^t t\right) \\ &= \exp\left(\delta^t\mu + \tfrac{1}{2}\delta^t\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}\delta\right) \\ &= \exp\left(\delta^t\mu + \tfrac{1}{2}\delta^t\Sigma\delta\right). \end{aligned}$$

El resultado se cumple para toda $\delta \in \Re^p$. $\square$

1.2.3 Propiedades de la Normal Multivariada

Los siguientes Teoremas establecen las propiedades que cumple una *m.a.* $X_1, ..., X_n$ que se distribuye normal multivariada o cualquier subconjunto de esta muestra.

Teorema 1.16 *Un vector aleatorio X de dimensión p tiene distribución normal multivariada, si y sólo si cualquier combinación de los componentes de X en donde no todos sus elementos son cero, se distribuye como una normal univariada.*

Demostración.

$\Rightarrow$] *P. D. que si $A \neq 0$ un vector de constantes en $\Re^p$, y X un vector con distribución normal, entonces $Y = A^t X = \sum_{i=1}^{p} a_i X_i \in \Re$ se distribuye como una normal univariada.*

Utilizando la expresión de la F.G.M. se tiene

$$\begin{aligned} M_Y(t) &= E\left(\exp\left\{t^t\left(A^t X\right)\right\}\right) \\ &= E\left(\exp\left\{t^t A^t X\right\}\right), \end{aligned}$$

donde

$$M_Y(t) = M_X(\delta), \tag{1.1}$$

con $\delta = (At)_{p\times 1}$ y $t \in \Re$. Siguiendo el resultado del Teorema 1.15 se tiene

$$\begin{aligned} M_Y(t) &= \exp\left\{\delta^t \mu + \tfrac{1}{2}\delta^t \Sigma \delta\right\} \\ &= \exp\left\{t A^t \mu + \tfrac{1}{2} t^2 A^t \Sigma A\right\}, \end{aligned}$$

y se concluye del Teorema 1.13 que la v.a. Y se distribuye como una normal univariada con media $(A\mu)_{1\times 1}$ y varianza $(A^t \Sigma A)_{1\times 1}$.

$\Leftarrow$] *Sea X un vector aleatorio y $Y = A^t X$ que se distribuye normal, entonces se mostrará que X tiene distribución normal multivariada.*

Como $Y = A^t X$ tiene densidad normal, denótese por θ y ρ^2 la media y la varianza de Y. Entonces la F.G.M de Y está dada por:

$$M_Y(t) = \exp\left\{t\theta + \frac{1}{2}t^2\rho^2\right\}.$$

Denotando por μ y Σ la media y la matriz de covarianzas de X se tiene

$$M_Y(t) = \exp\left\{tA^t\mu + \frac{1}{2}t^2 A^t \Sigma A\right\},$$

y por la ecuación (1.1) se obtiene

$$\begin{aligned} M_X(\delta) &= M_Y(t) \ \text{con}\ \delta = At \\ &= \exp\left\{t^t A^t \mu + \tfrac{1}{2} t^t A^t \Sigma A t\right\} \\ &= \exp\left\{\delta^t \mu + \tfrac{1}{2}\delta^t \Sigma \delta\right\}, \end{aligned}$$

lo cual implica que $X \sim N_p(\mu, \Sigma)$.□

Teorema 1.17 *Si* $X_1, ..., X_n$ *es una muestra aleatoria (m.a.) de la población* $N_p(\mu, \Sigma)$ *entonces la distribución de* $\bar{X}$ *es* $N_p\left(\mu, \frac{1}{n}\Sigma\right)$ *,donde* $\bar{X}^t = \left(\bar{X}_1, ..., \bar{X}_p\right)$ *con* $\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$.

Demostración.

Utilícese la F.G.M. para obtener la distribución de $\bar{X}$ *definiendo un vector* $t \in \Re^p$

$$\begin{aligned} M_{\bar{X}}(t) &= E\left(\exp\left\{t' \bar{X}\right\}\right) \\ &= E\left(\exp\left\{\tfrac{1}{n}\textstyle\sum_{i=1}^{n} t' X_i\right\}\right), \end{aligned}$$

con la hipótesis de independencia sobre las $X_i's$ *se obtiene*

$$\begin{aligned} M_{\bar{X}}(t) &= \textstyle\prod_{i=1}^{n} E\left(\exp\left\{\frac{1}{n} t' X\right\}\right) \\ &= \textstyle\prod_{i=1}^{n} M_{X_i}\left(\frac{1}{n} t\right). \end{aligned}$$

Como cada componente $X_i \sim N_p(\mu, \Sigma)$, *y el Teorema 1.15 se tiene*

$$\begin{aligned} M_{\bar{X}}(t) &= \textstyle\prod_{i=1}^{n} \exp\left\{\frac{1}{n} t' \mu + \frac{1}{2n^2} t' \Sigma t\right\} \\ &= \exp\left\{t' \mu + \tfrac{1}{2} t' \left(\tfrac{1}{n}\Sigma\right) t\right\}. \end{aligned}$$

Por lo tanto, se concluye que $\bar{X} \sim N_p\left(\mu, \frac{1}{n}\Sigma\right)$.□

Teorema 1.18 *Para* Y *un vector aleatorio normalmente distribuido con media* μ *y matriz de covarianzas* $\Sigma > 0$, *se cumplen las siguientes propiedades:*

1. *Las distribuciones marginales de normales son normales con parámetros respectivos.*
2. *Sea* $A_{q \times p}$ *una matriz de constantes y* $b \in \Re^q$ *si* $Z = AY + b \quad y \quad r(A) = q < p$, *entonces* $Z \sim N_q(A\mu + b, A\Sigma A^t)$.

Demostración.

1. *Sin perder generalidad (S.P.G.) puede tomarse un vector aleatorio Y definido como*

$$Y = \begin{pmatrix} Y_{1_{p1\times 1}} \\ Y_{1_{p2\times 1}} \end{pmatrix},$$

en donde $p_1 + p_2 = p$ y particiónese μ y Σ como:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad y \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

observe que Y_1, μ_1 tienen dimensión $p_1 \times 1$ y Y_2, μ_2 son de dimensión $p_2 \times 1$ y la matriz Σ es definida positiva.

Defínase un vector $M_{p\times 1}$ como:

$$M = \begin{pmatrix} t_{p_1\times 1} \\ m_{p_2\times 1} \end{pmatrix},$$

y obténgase la F.G.M. para Y_1 de la siguiente manera

$$\begin{aligned} M_{Y_1}(t) &= M_Y(M)\mid_{\eta=0} \\ &= E\left[\exp\left(M^tY\right)\right]\mid_{m=0} \\ &= \exp\left\{(t^t, m^t)\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix} + \frac{1}{2}(t^t, m^t)\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}\begin{pmatrix} t\\ m\end{pmatrix}\right\}\Bigg|_{m=0}, \end{aligned}$$

y de lo anterior se obtiene

$$M_{Y_1}(t) = \exp\left(t^t\mu_1 + \frac{1}{2}t^t\Sigma_{11}t\right).$$

Esta última expresión es la caracterización de la F.G.M. para una normal, en donde la media es el vector $\mu_1 \in \Re^{p1}$ y Σ_{11} es la matriz de covarianzas, lo único que resta por demostrar es que Σ_{11} sea definida positiva, esto es

$$t^t\Sigma_{11}t > 0 \quad \forall\ t \neq 0.$$

Por hipótesis se tiene que $r^t\Sigma r > 0 \ \forall \ r \neq 0$, *considérese un vector* r *dado como:*

$$r = \begin{pmatrix} t \\ 0 \end{pmatrix},$$

donde $t \neq 0$ *entonces*

$$\begin{aligned} r^t\Sigma r &= (t^t, 0) \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} t \\ 0 \end{pmatrix} \\ &= t^t\Sigma_{11}t > 0. \end{aligned}$$

por lo tanto, $\Sigma_{11} > 0$ *y* $Y_1 \sim N_{p1}(\mu_1, \Sigma_{11})$. □ *La demostración es análoga para* Y_2.

2. *La F.G.M. para* Z *está definida como:*

$$\begin{aligned} M_Z(t) &= E(\exp(t^tZ)) \\ &= E[\exp\{t^t(AY + b)\}] \\ &= \exp(t^tb)\, E[\exp(t^tAY)] \\ &= \exp(t^tb)\, M_Y(A^tt), \end{aligned}$$

como $M_Y(t)$ *es la F.G.M. para* Y, *utilizándose el resultado del Teorema 1.15 se obtiene*

$$\begin{aligned} M_Z(t) &= \exp(t^tb) \exp\left(t^tA\mu + \tfrac{1}{2}t^t(A\Sigma A^t)t\right) \\ &= \exp\left\{\left[t^t(A\mu + b) + \tfrac{1}{2}t^t(A\Sigma A^t)t\right]\right\}. \end{aligned}$$

Esta última expresión tiene la forma de la F.G.M. para una normal, en donde la media es el vector $(A\mu + b) \in \Re^q$ *y la matriz de escala está dada por la expresión* $A\Sigma A^t_{q\times q}$. *Por último tiene que demostrarse que* $A\Sigma A^t > 0$.

Sea $t \in \Re^q$ *con* $t \neq 0$ *debe probarse que* $t^tA\Sigma A^tt > 0$. *Para esto se tiene*

$$\begin{aligned} t^tA\Sigma A^tt &= (t^tA)\,\Sigma\,(A^tt) \\ &= s^t\Sigma s, \end{aligned} \tag{1.2}$$

donde $s = A^tt$ *es un vector no nulo se puede premultiplicar éste por la matriz* A *y se obtiene*

$$As = AA^tt,$$

ya que AA^t es una matriz cuadrada de dimensión $q \times q$ y recordando que el $r(AA^t) = r(A) = q$ (Ver Definición A.21), entonces se garantiza la existencia de la matriz $(AA^t)^{-1}$ lo cual permite encontrar una expresión para t que está dada por

$$t = (AA^t)^{-1} As.$$

Si $s = 0$ entonces $t = 0$ lo cual es una contradicción ya que por hipótesis $t \neq 0$ entonces $s \neq 0$ y de la ecuación (1.2) se concluye que $A\Sigma A^t > 0$, por lo tanto

$$Z \sim N_q\left(A\mu + b, A\Sigma A^t\right). \square$$

Teorema 1.19 *Si $X \sim N_p(\mu, \Sigma)$ entonces $Y = \Sigma^{-\frac{1}{2}}(X - \mu) \sim N_p(0, I_p)$.*

Demostración.

Por el Teorema 1.18 parte 2 se sabe que Y se distribuye como una normal de dimensión p, entonces

$$\begin{aligned} E(Y) &= E\left(\Sigma^{-\frac{1}{2}}(X - \mu)\right) \\ &= \Sigma^{-\frac{1}{2}} E(X) - \Sigma^{-\frac{1}{2}} E(\mu) \\ &= 0, \end{aligned}$$

y matriz de covarianzas

$$\begin{aligned} V(Y) &= \Sigma^{-\frac{1}{2}} V(X) \Sigma^{-\frac{1}{2}} \\ &= \Sigma^{-\frac{1}{2}} \Sigma \Sigma^{-\frac{1}{2}} \\ &= I_p. \square \end{aligned}$$

Ejemplo 1.1 *Sea X un vector aleatorio de dimensión 3, que se distribuye normalmente con media $\mu = \begin{pmatrix} -1 \\ 4 \\ -3 \end{pmatrix}$ y matriz de escala $\Sigma = \begin{pmatrix} 5 & -1 & 2 \\ -1 & 4 & 0 \\ 2 & 0 & 3 \end{pmatrix}$.*

Determinar la densidad de:

1. X_2.

2. $Z = \begin{pmatrix} X_3 \\ X_1 \end{pmatrix}$.

3. $Y = \begin{pmatrix} 2X_1 - 3X_2 \\ X_2 + X_3 \end{pmatrix}$.

4. $E(X_1^2 X_3)$.

Solución.

1. *Por el resultado 1 del Teorema 1.18, se sigue que* X_2 *se distribuye normalmente con parámetros*

$$\mu = 4 \quad y \quad \sigma^2 = 4.$$

En notación : $X_2 \sim N_1(4, 4)$.

2. *Por el inciso 2 del Teorema 1.18,* Z *se distribuye como una normal bivariada con parámetros:*

$$\mu_Z = \begin{pmatrix} -3 \\ -1 \end{pmatrix} \quad y \quad \Sigma_Z = \begin{pmatrix} 3 & 2 \\ 2 & 5 \end{pmatrix}.$$

Por lo tanto

$$Z \sim N_2 \left(\begin{pmatrix} -3 \\ -1 \end{pmatrix}, \begin{pmatrix} 3 & 2 \\ 2 & 5 \end{pmatrix} \right).$$

3. *Se sabe que* Y *es una transformación lineal de las entradas del vector* X*; y utilizándose el inciso 2 del Teorema 1.18, se cumple que*

$$Y \sim N_2(A\mu, A\Sigma A').$$

En particular se puede descomponer la transformación lineal

$$Y = \begin{pmatrix} 2X_1 - 3X_2 \\ X_2 + X_3 \end{pmatrix},$$

como

$$\begin{aligned} Y &= \begin{pmatrix} 2 & -3 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \\ &= AX. \end{aligned}$$

En donde la media puede obtenerse como:

$$\begin{aligned} E(Y) &= E(AX) \\ &= A\mu \\ &= \begin{pmatrix} 2 & -3 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ 4 \\ -3 \end{pmatrix} \\ &= \begin{pmatrix} -14 \\ 1 \end{pmatrix}, \end{aligned}$$

y la matriz de dispersión :

$$\begin{aligned} V(Y) &= V(AX) \\ &= A\Sigma A^t \\ &= \begin{pmatrix} 2 & -3 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 5 & -1 & 2 \\ -1 & 4 & 0 \\ 2 & 0 & 3 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ -3 & 1 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 68 & -10 \\ -10 & 7 \end{pmatrix}. \end{aligned}$$

Finalmente

$$Y \sim N_2 \left(\begin{pmatrix} -14 \\ 1 \end{pmatrix}, \begin{pmatrix} 68 & -10 \\ -10 & 7 \end{pmatrix} \right).$$

4. *Sea* $W = \begin{pmatrix} X_1 \\ X_3 \end{pmatrix}$ *un vector aleatorio de dimensión 2, el cual por el resultado 1 del Teorema 1.18 se distribuye como una normal con parámetros*

$$\mu_W = \begin{pmatrix} -1 \\ -3 \end{pmatrix} \quad y \quad \Sigma_W = \begin{pmatrix} 5 & 2 \\ 2 & 3 \end{pmatrix}.$$

Utilizando la expresión de la F.G.M. de una normal (ver Teorema 1.15), se obtiene:

$$\begin{aligned} M_W(t) &= \exp\left(t'\mu + \frac{1}{2}t'\Sigma_W t\right) \quad \forall\ t \neq 0. \\ &= \exp\left((t_1,t_3)\begin{pmatrix} -1 \\ -3 \end{pmatrix} + \frac{1}{2}(t_1,t_3)\begin{pmatrix} 5 & 2 \\ 2 & 3 \end{pmatrix}\begin{pmatrix} t_1 \\ t_3 \end{pmatrix}\right) \\ &= \exp\left(-t_1 - 3t_3 + \frac{1}{2}(5t_1 + 2t_3, 2t_1 + 3t_3)\begin{pmatrix} t_1 \\ t_3 \end{pmatrix}\right) \\ &= \exp\left(-t_1 - 3t_3 + \frac{1}{2}\left(5t_1^2 + 4t_1t_3 + 3t_3^2\right)\right) \\ &= \exp\left(-t_1 - 3t_3 + \frac{5}{2}t_1^2 + 2t_1t_3 + \frac{3}{2}t_3^2\right). \end{aligned}$$

Ahora calculando la primera y segunda derivada parcial con respecto de t_1 *y la primera derivada con respecto de* t_3*, se tiene*

$$\frac{\partial M_W(t)}{\partial t_1} = M_W(t)\left(-1 + 5t_1 + 2t_3\right).$$

derivando nuevamente con respecto de t_1 :

$$\frac{\partial^2 M_W(t)}{\partial t_1^2} = M_W(t)\left(5 + (-1 + 5t_1 + 2t_3)^2\right),$$

y por último derivando con respecto de t_3

$$\frac{\partial^3 M_W(t)}{\partial t_1^2 \partial t_3} = M_W(t)\left[4(-1 + 5t_1 + 2t_3) + \left(5 + (-1 + 5t_1 + 2t_3)^2\right)(-3 + 2t_1 + 3t_3)\right],$$

evaluando esta última expresión en $t_1 = t_3 = 0$*, se obtiene:*

$$\begin{aligned} \frac{\partial^3 M_W(t)}{\partial t_1^2 \partial t_3} \Big|_{t_1=t_3=0} &= 1 \cdot \left[4(-1) + \left(5 + (-1)^2\right)(3)\right] \\ &= 14. \end{aligned}$$

Por lo tanto

$$E\left(X_1^2 X_3\right) = 14.$$

1.2.4 Independencia

El concepto de Independencia se analizará bajo el supuesto de normalidad.

Teorema 1.20 *Sea* $X \sim N_p(\mu, \Sigma)$ *y considérese las siguientes particiones*

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

donde $X_i, \mu_i \in \Re^{p_i}$, $i = 1, 2$ *y* Σ_{ij} *tiene dimensión* $p_i \times p_j$. *Las variables* X_1 *y* X_2 *son independientes si y sólo si* $C(X_1, X_2) = \Sigma_{12} = 0$

Demostración.

$\Rightarrow$] *Esto es fácil demostrarlo ya que la covarianza puede ser escrita como*

$$\begin{aligned} C(X_1, X_2) &= E(X_1 X_2^t) - E(X_1) E(X_2^t) \\ &= E(X_1) E(X_2) - E(X_1) E(X_2) \\ &= 0. \end{aligned}$$

$\Leftarrow$] *Si* $\Sigma_{12} = 0$ *entonces la forma cuadrática del exponente en la f.d. de* X *puede ser vista como*

$$\begin{aligned} Q &= (X-\mu)^t \Sigma^{-1} (X-\mu) \\ &= \left[(X_1-\mu_1)^t \ , \ (X_2-\mu_2)^t\right] \begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix} \begin{bmatrix} X_1-\mu_1 \\ X_2-\mu_2 \end{bmatrix} \\ &= (X_1-\mu_1)^t \ \Sigma_{11}^{-1} (X_1-\mu_1) + (X_2-\mu_2)^t \Sigma_{22}^{-1} (X_2-\mu_2). \end{aligned}$$

Definiendo $Q_1 = (X_1-\mu_1)^t \ \Sigma_{11}^{-1} (X_1-\mu_1)$ *y* $Q_2 = (X_2-\mu_2)^t \Sigma_{22}^{-1} (X_2-\mu_2)$, *entonces* Q *puede expresarse como:*

$$Q = Q_1 + Q_2.$$

Dado que la f.d.p. de X está dada por

$$f_X(x;\mu,\Sigma) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}Q\right\},$$

si se particiona la matriz Σ y la forma cuadrática Q, se puede entonces reescribir la f.d.p. como

$$f_X(x;\mu,\Sigma) = |2\pi\Sigma_1|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}Q_1\right\} \cdot |2\pi\Sigma_2|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}Q_2\right\},$$

donde por el resultado del Teorema 1.12 si

$$H_1(x_1) = |2\pi\Sigma_1|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}Q_1\right\},$$

y

$$H_2(x_2) = |2\pi\Sigma_2|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}Q_2\right\},$$

la función de densidad conjunta de X puede escribirse como

$$f_X(x;\mu,\Sigma) = H_1(x_1)\, H_2(x_2),$$

por lo tanto los vectores X_1 y X_2 son independientes. □

Cabe mencionar que no basta que la covarianza entre X_1 y X_2 sea cero para que éstos se distribuyan independientemente, es necesario además que la distribución conjunta de X_1 y X_2 sea una normal multivariada.

Las condiciones bajo las cuales se obtiene la independencia entre dos transformaciones lineales Y y Z que provienen de una muestra $X_1, ..., X_n$ con distribución normal quedan establecidos en el siguiente Teorema.

Teorema 1.21 *Sea $X^t_{p\times n} = (X_1, ..., X_n)$ donde $X_1, ..., X_n$ es una m.a. de $N_p(\mu,\Sigma)$ y además*

$$Y_{q\times r} = A_{q\times n} X_{n\times p} B_{p\times r} \quad y \quad Z_{s\times t} = C_{s\times n} X_{n\times p} D_{p\times t},$$

donde $qr + st \leq np$. Las matrices Y y Z son independientes si y sólo si $B^t\Sigma D = 0$ o $AC^t = 0$.

Demostración. Ver Mendoza (1987).

1.2.5 Distribuciones Condicionales

La distribución condicional entre un conjunto de variables que se distribuye normal dado otro conjunto de variables con la misma distribución se obtiene en el siguiente Teorema:

Teorema 1.22 *Sea X un v.a. que se distribuye como $N_p(\mu, \Sigma)$, defínase $X^t = (X_1^t, X_2^t)$, $\mu^t = (\mu_1^t, \mu_2^t)$ y $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ en donde X_i , $\mu_i \in \Re^{p_i}$ y Σ_{ij} tiene dimensión $p_i \times p_j$ con $i = 1, 2$ y $p_1 + p_2 = p$. La densidad condicional de $X_2 \mid X_1$ es normal con media y matriz de covarianzas dadas por:*

$$\begin{aligned} E(X_2 \mid X_1 = x_1) &= \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1) \\ &= \mu_{2.1}. \end{aligned}$$

$$\begin{aligned} V(X_2 \mid X_1 = x_1) &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \\ &= \Sigma_{22.1} \\ &= \Sigma_{2.1} \end{aligned}$$

Demostración.

Por el Teorema 1.18 se cumple que $X_1 \sim N_p(\mu_1, \Sigma_{11})$ y dando la matriz M como

$$M = \begin{pmatrix} I_{p1} & 0 \\ -\Sigma_{21}\Sigma_{11}^{-1} & I_{p2} \end{pmatrix},$$

puede verificarse fácilmente que

$$M^{-1} = \begin{pmatrix} I_{p1} & 0 \\ \Sigma_{21}\Sigma_{11}^{-1} & I_{p2} \end{pmatrix}.$$

La función de densidad conjunta de X_1 y X_2 tiene la expresión

$$f_X(x) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(X - \mu)^t \Sigma^{-1}(X - \mu)\right\}. \tag{1.3}$$

y de ésta obsérvese que la forma cuadrática del exponente puede expresarse como:

$$\begin{aligned}(X-\mu)^t \Sigma^{-1} (X-\mu) &= (X-\mu)^t M^t \left(M^t\right)^{-1} \Sigma^{-1} M^{-1} M (X-\mu) \\ &= (X-\mu)^t M^t \left(M \Sigma M^t\right)^{-1} M (X-\mu),\end{aligned}$$

Desarrollando $(M\Sigma M^t)^{-1}$ *se obtiene*

$$(X-\mu)^t \Sigma^{-1} (X-\mu) = \begin{pmatrix} X_1-\mu_1 \\ X_2-\mu_2 \end{pmatrix}^t M^t \begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22.1}^{-1} \end{pmatrix} M \begin{pmatrix} X_1-\mu_1 \\ X_2-\mu_2 \end{pmatrix}, \quad (1.4)$$

Resolviendo los productos de matrices se llega a dos formas cuadráticas, sean éstas Q_1 *y* $Q_{2.1}$ *definidas como:*

$$Q_1 = (X_1-\mu_1)^t \Sigma_{11}^{-1} (X_1-\mu_1).$$

y

$$Q_{2.1} = \left[(X_2-\mu_2) - \Sigma_{21}\Sigma_{11}^{-1}(X_1-\mu_1)\right]^t \Sigma_{22.1}^{-1} \left[(X_2-\mu_2) - \Sigma_{21}\Sigma_{11}^{-1}(X_1-\mu_1)\right],$$

entonces la (1.4) puede escribirse como

$$(X-\mu)^t \Sigma^{-1} (X-\mu) = Q_1 + Q_{2.1}, \quad (1.5)$$

y así la función de densidad conjunta (1.3) de X puede ser reescrita como:

$$f_X(x) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(Q_1 + Q_{2.1})\right\}.$$

Por definición, la función de densidad condicional $X_2 \mid X_1$ *está dada por*

$$f_{X_2|X_1}(x_2 \mid x_1) = \frac{|2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(X-\mu)^t \Sigma^{-1}(X-\mu)\right\}}{|2\pi\Sigma_{11}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(X-\mu)^t \Sigma_{11}^{-1}(X-\mu)\right\}}.$$

Utilizando (1.5) se puede expresar este cociente como

$$f_{X_2|X_1}(x_2 \mid x_1) = \frac{|2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(Q_1+Q_{2.1})\right\}}{|2\pi\Sigma_{11}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(Q_1)\right\}},$$

Y utilizando las propiedades de los determinantes se sabe que

$$\begin{aligned}|\Sigma| &= |\Sigma_{11}| \left|\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right| \\ &= |\Sigma_{11}| \, |\Sigma_{22.1}|,\end{aligned}$$

entonces

$$\begin{aligned} f_{X2|X1}(x_1,x_2) &= \frac{(2\pi)^{-\frac{p1}{2}}|\Sigma_{11}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}Q_1\right\}\cdot(2\pi)^{-\frac{p2}{2}}|\Sigma_{22.1}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}Q_{2.1}\right\}}{(2\pi)^{-\frac{p1}{2}}|\Sigma_{11}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}Q_1\right\}} \\ &= (2\pi)^{-\frac{p2}{2}}|\Sigma_{22.1}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}Q_{2.1}\right\} \\ &= (2\pi)^{-\frac{p2}{2}}|\Sigma_{22.1}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(x_2-\mu_{2.1})^t\,\Sigma_{22.1}^{-1}(x_2-\mu_{2.1})\right\}. \end{aligned}$$

de donde se sigue que $X_2 \mid X_1$ *tiene densidad normal de parámetros*

$$E(X_2 \mid X_1 = x_1) = \mu_{2.1},$$

y

$$V(X_2 \mid X_1 = x_1) = \Sigma_{2.1}.\square$$

Ejemplo 1.2 *Sea* X *un vector aleatorio de dimensión 3, que se distribuye como*

$$N_3\left(\begin{pmatrix} 5.5 \\ 3 \\ 3.5 \end{pmatrix}, \begin{pmatrix} 5 & -3 & 1 \\ -3 & 4 & 0 \\ 1 & 0 & 3 \end{pmatrix}\right)$$

Encuentre la distribución condicional de $X_2 \mid X_1 = x_1, X_3 = x_3$, *cuando*

$$\begin{pmatrix} x_1 \\ x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 4 \end{pmatrix}.$$

Solución.

Definase el vector $Y^t = (Y_1^t, Y_2^t)$ *donde la entrada* $Y_1^t = (X_1, X_3)$ *y* $Y_2^t = X_2$; *la esperanza del vector* Y *está dada por:*

$$\begin{aligned} \mu_Y &= \begin{pmatrix} E(Y_1) \\ E(Y_2) \end{pmatrix} \\ &= \begin{pmatrix} 5.5 \\ 3.5 \\ 3 \end{pmatrix} \end{aligned}$$

La matriz de escala para el vector Y según el Teorema 1.22 tiene la siguiente expresión:

$$\Sigma_Y = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

En donde las entradas de las matrices Σ_{11}, Σ_{12}, Σ_{21} y Σ_{22} son tomadas de la matriz de escala de X como:

$$\begin{aligned} \Sigma_{11} &= \begin{pmatrix} 5 & 1 \\ 1 & 3 \end{pmatrix} \\ \Sigma_{12} &= \begin{pmatrix} -3 \\ 0 \end{pmatrix} \\ \Sigma_{21} &= \begin{pmatrix} -3 & 0 \end{pmatrix} \\ \Sigma_{22} &= (4) \end{aligned}$$

$$\Sigma_{11}^{-1} = \begin{pmatrix} \frac{3}{14} & -\frac{1}{14} \\ -\frac{1}{14} & \frac{5}{14} \end{pmatrix}$$

De tal suerte que la matriz Σ_Y está definida como:

$$\Sigma_Y = \begin{pmatrix} 5 & 1 & -3 \\ 1 & 3 & 0 \\ -3 & 0 & 4 \end{pmatrix}$$

De acuerdo al Teorema 1.22 la función de densidad condicional de $Y_2 \mid Y_1$ es normal con parámetros

$$\begin{aligned} E(Y_2 \mid Y_1) &= E\left(Y_2 \mid \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 6 \\ 4 \end{pmatrix}\right) \\ &= (3) + (-3, 0) \begin{pmatrix} \frac{3}{14} & -\frac{1}{14} \\ -\frac{1}{14} & \frac{5}{14} \end{pmatrix} \begin{pmatrix} 6 - 5.5 \\ 4 - 3.5 \end{pmatrix} \\ &= (3) + \left(-\frac{9}{14}, \frac{3}{14}\right) \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} \end{aligned}$$

$$
\begin{aligned}
&= (3) + \left(-\frac{9}{28} + \frac{3}{28}\right) \\
&= (3) + \left(-\frac{6}{28}\right) \\
&= \frac{39}{14},
\end{aligned}
$$

y varianza

$$
\begin{aligned}
V(Y_2 \mid Y_1) &= V\left(Y_2 \mid \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 6 \\ 4 \end{pmatrix}\right) \\
&= (4) - \begin{pmatrix} -3 & 0 \end{pmatrix} \begin{pmatrix} \frac{3}{14} & -\frac{1}{14} \\ -\frac{1}{14} & \frac{5}{14} \end{pmatrix} \begin{pmatrix} -3 \\ 0 \end{pmatrix} \\
&= (4) - \begin{pmatrix} -\frac{9}{14} & \frac{3}{14} \end{pmatrix} \begin{pmatrix} -3 \\ 0 \end{pmatrix} \\
&= (4) - \left(\frac{27}{14}\right) \\
&= \frac{29}{14}.
\end{aligned}
$$

Por lo tanto la distribución condicional de $Y_2 \mid Y_1 = \begin{pmatrix} 6 \\ 4 \end{pmatrix}$ *es* $N_1\left(\frac{39}{14}, \frac{29}{14}\right)$. □

1.2.6 Estimación de Parámetros

Con frecuencia los parámetros de una muestra que se distribuye normal son desconocidos, pero pueden ser estimados mediante el método de máxima verosimilitud utilizando la información de la muestra. El siguiente Teorema proporciona las expresiones matemáticas que caracterizan al estimador para la media μ y la matriz de covarianzas a la cual se ha identificado como Σ.

Teorema 1.23 *Sea* $X \sim N_p(\mu, \Sigma)$ *y sea* $X_1, \ldots, X_n$ *una m.a. de esta densidad. Los estimadores máximo verosímiles de* μ *y* Σ *son*

$$\hat{\mu} = \bar{X}.$$

$$\hat{\Sigma}=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(X_i-\bar{X}\right)^t.$$

Demostración.

Sea $X_1,...,X_n$ una *m.a.* de la densidad $N(\mu,\Sigma)$, cuya función de verosimilitud está dada por

$$\begin{aligned}
L(\mu,\Sigma) &= \prod_{i=1}^{n} f_{X_i}(x_i) \\
&= \prod_{i=1}^{n}\left[|2\pi\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(X_i-\mu)^t\Sigma^{-1}(X_i-\mu)\right\}\right] \\
&= |2\pi\Sigma|^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)^t\Sigma^{-1}(X_i-\mu)\right\}.
\end{aligned}$$

Esta expresión puede simplificarse según el Teorema A.17 como

$$\begin{aligned}
L(\mu,\Sigma) &= |2\pi\Sigma|^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}tr\left[\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^t\Sigma^{-1}\left(X_i-\bar{X}\right)+n\left(\bar{X}-\mu\right)^t\Sigma^{-1}\left(\bar{X}-\mu\right)\right]\right\} \\
&= |2\pi\Sigma|^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}tr\left(X_i-\bar{X}\right)^t\left(\Sigma^{-1}\left(X_i-\bar{X}\right)\right)+n\left(\bar{X}-\mu\right)^t\Sigma^{-1}\left(\bar{X}-\mu\right)\right]\right\}.
\end{aligned}$$

Aplicando las propiedades de la traza en esta última ecuación se tiene

$$\begin{aligned}
L(\mu,\Sigma) &= |2\pi\Sigma|^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}tr\left\{\Sigma^{-1}\left(X_i-\bar{X}\right)\left(X_i-\bar{X}\right)^t\right\}+n\left(\bar{X}-\mu\right)^t\Sigma^{-1}\left(\bar{X}-\mu\right)\right]\right\} \\
&= |2\pi\Sigma|^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\left[tr\Sigma^{-1}\left\{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(X_i-\bar{X}\right)^t\right\}+n\left(\bar{X}-\mu\right)^t\Sigma^{-1}\left(\bar{X}-\mu\right)\right]\right\}.
\end{aligned}$$

Dado que la matriz de covarianzas muestral está dada por $nS=\Sigma_{i=1}^{n}\left(X_i-\bar{X}\right)\left(X_i-\bar{X}\right)^t$, la función de verosimilitud puede escribirse como

$$L(\mu,\Sigma)=|2\pi\Sigma|^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\left[tr\Sigma^{-1}nS+n\left(\bar{X}-\mu\right)^t\Sigma^{-1}\left(\bar{X}-\mu\right)\right]\right\},$$

Aplicando la función logaritmo a la ecuación anterior, se tiene

$$\ln L(\mu,\Sigma)=-\frac{np}{2}\ln(2\pi)-\frac{n}{2}\ln|\Sigma|-\frac{n}{2}\left[tr\Sigma^{-1}nS+\left(\bar{X}-\mu\right)^t\Sigma^{-1}\left(\bar{X}-\mu\right)\right]. \qquad (1.6)$$

El problema de maximización es equivalente a encontrar un mínimo de $-\ln L(\mu, \Sigma)$con respecto a μ y Σ. Entonces basta con minimizar la forma cuadrática $\left(\bar{X} - \mu\right)^t \Sigma^{-1} \left(\bar{X} - \mu\right)$ con respecto de μ, alcanzando un mínimo cuando $\hat{\mu}=\bar{X}$, por lo que la función

$$\left(\bar{X} - \hat{\mu}\right)^t \Sigma^{-1} \left(\bar{X} - \hat{\mu}\right) = 0,$$

lo cual ocurre si y sólo si $\hat{\mu}=\bar{X}$. De esta manera se concluye que el estimador máximo verosímil para la media es $\hat{\mu}=\bar{X}$.

Para obtener el estimador máximo verosímil de Σ, basta maximizar $L\left(\hat{\mu}, \Sigma\right)$ que de acuerdo al Teorema A.18 el máximo es $L\left(\hat{\mu}, \hat{\Sigma}\right)$, donde $\hat{\Sigma}= S$ y según este resultado se debe probar que $\Sigma^{-\frac{1}{2}} S \Sigma^{-\frac{1}{2}} \geq 0$, pero $\Sigma^{-1} > 0$ de lo que se deduce que $\Sigma^{-\frac{1}{2}} > 0$ y sólo debe mostrarse que la matriz S es definida positiva. La matriz $S_X \geq 0$ si $(n-1) \geq p$ (ver Teoremas 1.25 y 1.27) lo cual ocurre si y sólo si $n \geq p+1$, y se cumple que

$$L\left(\hat{\mu}, \hat{\Sigma}\right) \geq L\left(\hat{\mu}, \Sigma\right).$$

Se puede entonces concluir que los estimadores máximo verosímiles para μ y Σ están dados por

$$\begin{aligned} \hat{\mu} &= \bar{X} \\ \hat{\Sigma} &= S_X \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right) \left(X_i - \bar{X}\right)^t . \square \end{aligned}$$

1.3 Densidad Wishart

En esta Sección se enuncia la generalización de la distribución χ^2 y que está dada por la forma cuadrática del tipo $X'CX$, donde C es una matriz simétrica y los renglones de la matriz X corresponden a una $m.a.$ de la distribución normal. En especial se analiza el caso de la estadística nS, la cual tiene asociada esta distribución.

Definición 1.11 *Se dice que una matriz $M_{p\times p}$ tiene densidad Wishart de dimensión p con matriz de escala Σ y de n grados si M puede ser escrita como*

$$M = \sum_{i=1}^{n} X_i X_i^t.$$

donde $X_1, ..., X_n$ es una m.a. de la densidad $N_p(0, \Sigma)$ y se denota por $M \sim W_p(\Sigma, n)$.

Nota. Si se define $X^t = (X_1, ..., X_n)$, M puede escribirse como $M = X^t X$.

Teorema 1.24 *Si $n \geq p$ la variable M tiene densidad Wishart con matriz de escala Σ y n grados de libertad si su densidad está dada por*

$$f_M(m) = \frac{|m|^{-\frac{(n-p-1)}{2}} \exp\left\{-\frac{1}{2} tr\left(\Sigma^{-1} m\right)\right\}}{2^{\frac{np}{2}} \pi^{\frac{p(p-1)}{4}} |\Sigma|^{\frac{n}{2}} \prod_{i=1}^{p} \Gamma\left[\frac{1}{2}(n+1-i)\right]}.$$

si $m > 0$, $\Sigma > 0$;

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx \quad t > 0.$$

Demostración. Ver Anderson (1994).

Teorema 1.25 *Si $M \sim W_p(\Sigma, n)$, entonces $r(M) = p \Leftrightarrow n \geq p$.*

Demostración

Por definición $M = X^t X$, pero M es una matriz invertible si y sólo si $r(X^t X) = p$, pero dado que $r(X^t X) = r(X)$, ésto ocurre si y sólo si $r(X) = p$. Dado que las p columnas de la matriz X tienen densidad normal en n dimensiones, la condición $n \geq p$ garantiza $r(X^t X) = p$. Por otro lado, si $n < p$, $r(X) \leq \min\{n, p\} = n < p$, lo cual demuestra el resultado.

Teorema 1.26 *Si $X_{p\times n}^t = (X_1, ..., X_n)$ donde $X_1, ..., X_n$ es una m.a. de la densidad $N_p(\mu, \Sigma)$ y $A_{n\times n}$ es simétrica, entonces $X^t A X \sim W_p(\Sigma, r)$ donde $r = r(A) = tr(A)$ si y sólo si*

1. *A es una matriz idempotente.*

2. $\mu = 0$ *o* $A1_{n\times 1} = 0$,*donde* $1^t_{1\times n} = (1, 1, ..., 1)$.

Demostración. Ver Mendoza (1987).

Teorema 1.27 *Sea* $X_1, ..., X_n$ *una m.a. de* $N_p(\mu, \Sigma)$, *si* $S = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^t$ *entonces* $nS \sim W_p(\Sigma, n-1)$.

Demostración

Por el Teorema A.14, la matriz nS se puede escribir como

$$nS = X^t H X,$$

con $H = I_n - \frac{1}{n}1_n 1_n^t$ *una matriz simétrica e idempotente (ver Teorema A.14) y* $X^t = (X_1, ..., X_n)$. *Los grados de libertad están dados por*

$$r(H) = n - 1,$$

donde la última igualdad se sigue por el Teorema A.14. □

Teorema 1.28 *Si* $M \sim W_p(\Sigma, n)$ *y* $B_{q\times p}$ *es constante y* $r(B) = q \leq p$ *entonces* $BMB^t \sim W_p(B\Sigma B^t, n)$.

Demostración

Por la Definición 1.11, $X_1, ..., X_n$ *es una m.a. de la densidad* $N_p(0, \Sigma)$ *y* $M = \sum_{i=1}^{n} X_i X_i^t$, *entonces*

$$\begin{aligned} BMB^t &= B\left(\sum_{i=1}^{n} X_i X_i^t\right) B^t \\ &= \sum_{i=1}^{n} \left(BX_i X_i^t B^t\right) \\ &= \sum_{i=1}^{n} Z_i Z_i^t. \end{aligned}$$

en donde cada $Z_i = BX_i$ *con* $i = 1, ..., n$.

Como $X_1, ..., X_n$ *es una m.a., eso implica que los* $Z_1, ..., Z_n$ *son independientes y por el Teorema 1.18 se sigue* $Z_i \sim N_q(0, B\Sigma B^t)$. *Lo cual implica de acuerdo a la Definición 1.11 que*

$$BMB^t \sim W_q(B\Sigma B^t, n). \square$$

Corolario 1.1 *Submatrices diagonales de* M *también tienen una distribución Wishart.*

Corolario 1.2 $\Sigma^{-\frac{1}{2}} M \Sigma^{-\frac{1}{2}} \sim W_p(I_n, n)$

Demostración

Por definición $M = X^t X$ *entonces*

$$\Sigma^{-\frac{1}{2}} M \Sigma^{-\frac{1}{2}} = \Sigma^{-\frac{1}{2}} X^t X \Sigma^{-\frac{1}{2}}.$$

Sea $Y = X\Sigma^{-\frac{1}{2}}$ *de lo cual se sigue que*

$$\Sigma^{-\frac{1}{2}} M \Sigma^{-\frac{1}{2}} = Y^t Y,$$

además se sabe que $X \sim N_p(0, \Sigma)$ *y siguiendo el Teorema 1.18,* $Y \sim N_p(0, I_p)$ *y se concluye que*

$$\Sigma^{-\frac{1}{2}} M \Sigma^{-\frac{1}{2}} \sim W_p(I_p, n). \square$$

Corolario 1.3 *Si* $M \sim W_p(I_p, n)$ *y* $B_{p\times q}$ *satisface que* $B^t B = I_q$, *entonces* $B^t M B \sim W_q(I_q, n)$.

Demostración.

Por el Teorema 1.28 se sigue que

$$B^t M B \sim W(B^t I_p B, n),$$

observando que $B^t I_p B = B^t B = I_q$, *se sigue el resultado.* $\square$

Corolario 1.4 *Si* $M \sim W_p(\Sigma, n)$, $a \in \Re^p$ *y* $\Sigma > 0$ *entonces:*

$$\left(\frac{a^t M a}{a^t \Sigma a}\right) \sim \chi^2_n.$$

Demostración.

M puede escribirse como $M = \sum_{i=1}^{n} X_i X_i^t$ *donde* $X_1, ..., X_n$ *es una m.a. de la densidad* $N_p(0, \Sigma)$, *entonces*

$$\begin{aligned} \frac{a^t M a}{a^t \Sigma a} &= \frac{1}{a^t \Sigma a}\left[\sum_{i=1}^{n} a^t X_i X_i^t a\right] \\ &= \frac{1}{a^t \Sigma a}\left[\sum_{i=1}^{n} Z_i^2\right], \end{aligned}$$

con $Z_i = X_i^t a = a^t X_i \sim N_1(0, a^t \Sigma a)$ *para* $i = 1, ..., n$.

Estandarizando cada Z_i, *se obtiene que*

$$\begin{aligned} \frac{Z_i}{\sqrt{a^t \Sigma a}} &\sim N(0,1) \quad \forall\, i \\ \frac{Z_i^2}{a^t \Sigma a} &\sim \chi^2_1, \quad \forall\, i, \end{aligned}$$

sumando estas nuevas variables y considerando la independencia de $X_1, ..., X_n$

$$\sum_{i=1}^{n} \frac{Z_i^2}{a^t \Sigma a} \sim \chi^2_n,$$

y por lo tanto se concluye que

$$\frac{a^t M a}{a^t \Sigma a} \sim \chi^2_n. \square$$

Teorema 1.29 *Si* $M_j \sim W_p(\Sigma, n_j)$ $j = 1, ..., r$ *son independientes, entonces* $\sum_{j=1}^{r} M_j \sim W_p(\Sigma, \sum_{i=1}^{r} n_j)$.

Demostración.

Por hipótesis $M_1, ..., M_r$ *son independientes y cada una de las* $M_j \sim W_p(\Sigma, n_j)$, *entonces se puede escribir*

$$M_j = \sum_{k=1}^{nj} X_{jk} X_{jk}^t \quad j = 1, ..., r$$

donde $X_{11}, ..., X_{1n}, ..., X_{r1}, ..., X_{rn}$ *es una m.a. de* $N_p(0, \Sigma)$ *y la suma de las* M_j *queda expresada como:*

$$\sum_{j=1}^{r} M_j = \sum_{j=1}^{r} \sum_{k=1}^{nj} X_{jk} X_{jk}^t,$$

entonces $\sum_{j=1}^{r} M_j \sim W_p\left(\Sigma, \sum_{i=1}^{r} n_j\right)$.□

Teorema 1.30 *Sea* $X_1, ..., X_n$ *una m.a. de una normal* $N_p(\mu, \Sigma)$,

$$\begin{aligned} \bar{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ &= \frac{1}{n} X_i^t 1_n \ y \\ S_X &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^t \\ &= \frac{1}{n} X^t H X, \end{aligned}$$

donde $X = \begin{pmatrix} X_1^t \\ \vdots \\ X_n^t \end{pmatrix}$ *y* $H = I_n - \frac{1}{n} 1_n 1_n^t$. *Las estadísticas* $\bar{X}$ *y* S_X *son independientes.*

Demostración

$$S_X = \frac{1}{n} X^t H X,$$

y por el Teorema A.14 se sabe que la matriz H *es idempotente y simétrica entonces se cumple*

$$S_X = \frac{1}{n}\left(X^t H\right)\left(HX\right),$$

donde HX *es una transformación lineal, entonces siguiendo el Teorema 1.21 basta demostrar que* $\bar{X}$ *y* HX *son independientes, así*

$$\begin{aligned} \left(\frac{1}{n} 1_n^t\right) H &= \frac{1}{n} 1_n^t \left(I - \frac{1}{n} 1_n 1_n^t\right) \\ &= \frac{1}{n}\left(1_n^t - \frac{1}{n} 1_n^t 1_n 1_n^t\right) \\ &= 0. \end{aligned}$$

y por lo tanto $\bar{X}$ *y* S_X *son independientes.*□

1.4 Densidad T^2 de Hotelling

En esta parte se examinan las funciones que son de la forma $X^t W^{-1} X$, donde X tiene distribución normal , W tiene distribución Wishart y X y W son independientes.

Definición 1.12 *Si la variable Z puede escribirse como $Z = mX^t M^{-1} X$ donde X y M son independientes tales que $X \sim N_p(0, I)$ y $M \sim W_p(I, m)$ $m \geq p$, se dice que Z tiene densidad T^2 de Hotelling de parámetros p y m y se denota por $Z \sim T^2(p, m)$.*

Teorema 1.31 *Si $X \sim N_p(\mu, \Sigma)$ y $M \sim W_p(\Sigma, m)$ con $m \geq p$ son independientes, entonces*

$$m(X-\mu)^t M^{-1}(X-\mu) \sim T^2(p, m),$$

Demostración

Sea $Y = \Sigma^{-\frac{1}{2}}(X-\mu) \sim N_p(0, I_p)$ y defínase la variable $Z = \Sigma^{-\frac{1}{2}} M \Sigma^{-\frac{1}{2}}$ entonces por el Teorema 1.28, $Z \sim W_p(I, m)$ y de la Definición 1.12 se cumple

$$w = mY^t Z^{-1} Y \sim T^2(p, m).$$

Observando que

$$\begin{aligned} w &= m(X-\mu)^t \Sigma^{-\frac{1}{2}} \left(\Sigma^{-\frac{1}{2}} M^{-1} \Sigma^{-\frac{1}{2}}\right) \Sigma^{-\frac{1}{2}} (X-\mu) \\ &= m(X-\mu)^t \Sigma^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} M^{-1} \Sigma^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} (X-\mu) \\ &= m(X-\mu)^t I_p M^{-1} I_p (X-\mu), \end{aligned}$$

se sigue que

$$w = m(X-\mu)^t M^{-1}(X-\mu) \sim T^2(p, m). \square$$

Teorema 1.32 *Sean $X_1, ..., X_n$ una m.a. de $N_p(\mu, \Sigma)$*

$$(n-1)\left(\bar{X}-\mu\right)^t S^{-1}\left(\bar{X}-\mu\right) \sim T^2(p, n-1).$$

Demostración.

Por los resultados de los Teoremas 1.17 y 1.28 se cumple que $\bar{X} \sim N_p\left(\mu, \frac{1}{n}\Sigma\right)$ *y* $nS \sim W_p\left(\Sigma, n-1\right)$ *respectivamente, entonces*

$$\left(\frac{1}{\sqrt{n}}I\right)^t nS\left(\frac{1}{\sqrt{n}}I\right) \sim W_p\left(\frac{1}{n}\Sigma, n-1\right),$$

entonces

$$S \sim W_p\left(\frac{1}{n}\Sigma, n-1\right),$$

por el Teorema 1.30 *se sabe que* $\bar{X}$ *y* S_X *son independientes y por el Teorema 1.31 se obtiene*

$$(n-1)\left(\bar{X}-\mu\right)^t S^{-1}\left(\bar{X}-\mu\right) \sim T^2\left(p, n-1\right).\square$$

En este punto es conveniente aclarar que existe una relación entre la distribución T^2 de Hotelling y la distribución F, esta relación queda establecida en el siguiente Teorema:

Teorema 1.33 *Si* $Z \sim T^2\left(p, m\right)$ *entonces la distribución de* Z *es igual a la distribución de* $\frac{mp}{m-p+1}Y$ *donde* $Y \sim F\left(p, m-p+1\right)$. *Esto se denota como:*

$$T^2\left(p, m\right) = \frac{mp}{m-p+1}F\left(p, m-p+1\right).$$

Demostración. Ver Mendoza (1987).

Corolario 1.5 *Si* $X \sim N_p\left(\mu, \Sigma\right)$ *y* $M \sim W_p\left(\Sigma, m\right)$ *tienen distribución independiente entonces:*

$$\frac{(m-p+1)}{p}\left(X-\mu\right)^t M^{-1}\left(X-\mu\right) \sim F_{(p,m-p+1)}.$$

Demostración.

Utilizando el resultado de la Definición 1.12 se sigue

$$m\left(X-\mu\right)^t M^{-1}\left(X-\mu\right) \sim T^2_{(p,m)},$$

y por el Teorema 1.33 se cumple

$$\frac{(m-p+1)}{p}\left(X-\mu\right)^t M^{-1}\left(X-\mu\right) \sim F_{(p,m-p+1)}.\square$$

Teorema 1.34 *Si $X_1, ..., X_n$ es una m.a. de la densidad $N_p(\mu, \Sigma)$, entonces*

$$\frac{n-p}{p}\left(\bar{X}-\mu\right)^t S^{-1}\left(\bar{X}-\mu\right) \sim F(p, n-p).$$

Demostración

Por el Teorema 1.32 se sabe que

$$(n-1)\left(\bar{X}-\mu\right)^t S^{-1}\left(\bar{X}-\mu\right) \sim T^2(p, n-1),$$

luego utilizando el Teorema 1.33 se sigue

$$\frac{(n-p)}{p}\left(\bar{X}-\mu\right)^t S^{-1}\left(\bar{X}-\mu\right) \sim F(p, n-p). \square$$

1.5 Densidad Λ de Wilks

La distribución Λ de Wilks tiene un uso importante en las técnicas de Análisis Multivariado, ya que por ejemplo, el cociente de verosimilitudes en la prueba de igualdad de medias sigue esta distribución.

Definición 1.13 *Sean W_1 y W_2 matrices aleatorias independientes donde $W_1 \sim W_p(I_p, m)$ y $W_2 \sim W_p(I_p, n)$ con $m \geq p$. Si X puede escribirse como:*

$$X = \frac{|W_1|}{|W_1 + W_2|} = |I_p + W_1^{-1} W_2|^{-1}.$$

se dice que X tiene densidad Λ de Wilks de parámetros p, m y n. Se denota por $X \sim \Lambda(p, m, n)$.

La relación que existe entre la distribución Λ de *Wilks* y la distribución F se establece en el siguiente Teorema.

Teorema 1.35 *1.* $\frac{1-\Lambda(p,m,1)}{\Lambda(p,m,1)} \sim \frac{p}{m-p+1} F(p, m-p+1)$.

2. $\frac{1-\Lambda(1,m,n)}{\Lambda(1,m,n)} \sim \frac{n}{m} F(n,m)$.

3. $\frac{1-\sqrt{\Lambda(p,m,2)}}{\sqrt{\Lambda(p,m,2)}} \sim \frac{p}{m-p+1} F(2p, 2(m-p+1))$.

4. $\frac{1-\sqrt{\Lambda(2,m,n)}}{\sqrt{\Lambda(2,m,n)}} \sim \frac{n}{m-1} F(2n, 2(m-1))$.

Demostración. Ver Anderson.

1.6 Pruebas de Hipótesis Multivariadas

En esta Sección se presentan los contrastes de hipótesis relacionados con los parámetros de una población Normal Multivariada. Dichas pruebas consideran en algunos casos g poblaciones o grupos que se denotarán por $\prod_i$, $i = 1, ..., g$, distibuídos como $N_p(\mu_i, \Sigma_i)$, en donde asociada a cada población $\prod_i$ se tiene una matriz $\mathbf{X}_i$ de dimensión $n_i \times p$ y cuyos elementos $X_{rs} \in \Re^p$ con $r = 1, ..., g$ y $s = 1, ..., n_i$. Estas g poblaciones pueden agruparse en una muestra total $\mathbf{X}$ como:

$$\mathbf{X}_{n\times p} = \begin{pmatrix} \mathbf{X}_{1(n_1\times p)} \\ \mathbf{X}_{2(n_2\times p)} \\ \vdots \\ \mathbf{X}_{g(n_g\times p)} \end{pmatrix},$$

donde $\sum_{i=1}^{g} n_i = n$.

Los contrastes que se consideran en esta Sección son los siguientes:

1. Igualdad de medias y matrices de covarianzas. La hipótesis nula supone que los g grupos $\prod_1, ..., \prod_g$ provienen de una misma población Normal contra la hipótesis alternativa la cual postula que existe por lo menos un grupo Π_i que no proviene de dicha población Normal.

2. Igualdad de Matrices de Covarianzas. La hipótesis nula supone que los g grupos comparten la misma matriz de dispersión Σ mientras que la hipótesis alternativa postula que por lo menos un grupo posee una matriz de escala diferente.

3. Igualdad de medias condicionada a igualdad de matrices de covarianzas. La hipótesis nula postula que los g grupos están centrados alrededor de la misma media dado que tienen la misma matriz de dispersión y la hipótesis alternativa supone que existe por lo menos un grupo cuya media es diferente.

1.6.1 Igualdad de Medias y Matrices de Covarianzas

El primer juego de hipótesis a probar es el siguiente:

$$H_1 \ : \ \mu_i = \mu_k, \ \Sigma_i = \Sigma_k$$

$$vs$$

$$H_{1.1} \ : \ \mu_i \neq \mu_k, \ \Sigma_i \neq \Sigma_k \text{ para al menos una } i.$$

Bajo la hipótesis H_1 las n observaciones tienen la misma distribución $N_p(\mu, \Sigma)$, y de acuerdo al Teorema 1.23 pueden obtenerse los estimadores máximo verosímiles como:

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{n_i} X_{ij}$$

$$\hat{\Sigma} = T = \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X} \right) \left(X_{ij} - \bar{X} \right)^t .$$

Bajo la hipótesis $H_1 \cup H_{1.1}$ el máximo de la función de verosimilitud se obtiene maximizando las funciones de verosimilitud de los parámetros correspondientes a cada población, y que de acuerdo al Teorema 1.23 estos estimadores máximo verosímiles están dados como:

$$\hat{\mu}_i = \bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \quad i = 1, ..., g.$$

$$\hat{\Sigma}_i = S_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_i \right) \left(X_{ij} - \bar{X}_i \right)^t .$$

Así el máximo de la función de verosimilitud bajo H_1 está dada por la expresión:

$$L\left(\hat{\mu}, \hat{\Sigma}; \mathbf{X}_1, ..., \mathbf{X}_g \right) = (2\pi)^{-\frac{np}{2}} |T|^{-\frac{n}{2}} \exp\left\{ -\frac{np}{2} \right\}. \tag{1.7}$$

Mientras que la función verosimilitud bajo $H_1 \cup H_{1.1}$ resulta ser:

$$L\left(\hat{\mu}_1, ..., \hat{\mu}_g, \hat{\Sigma}_1, ..., \hat{\Sigma}; \mathbf{X}_1, ..., \mathbf{X}_g \right) = (2\pi)^{-\frac{np}{2}} \left(\prod_{i=1}^{g} |S_i|^{-\frac{n_i}{2}} \right) \exp\left\{ -\frac{np}{2} \right\}. \tag{1.8}$$

De la expresiones (1.7) y (1.8), se obtiene el cociente de verosimilitudes generalizado y que toma la siguiente expresión:

$$\begin{aligned}\Lambda_1 &= \frac{\sup_{H_1} L(\mu, \Sigma)}{\sup_{H_1 \cup H_{1,1}} L(\mu_1, ..., \mu_g, \Sigma_1, ..., \Sigma_g)}, \\ &= \frac{(2\pi)^{-\frac{np}{2}} |T|^{-\frac{n}{2}} \exp\left\{-\frac{np}{2}\right\}}{(2\pi)^{-\frac{np}{2}} \prod_{i=1}^{g} |S_i|^{-\frac{n_i}{2}} \exp\left\{-\frac{np}{2}\right\}}, \\ &= \frac{|T|^{-\frac{n}{2}}}{\prod_{i=1}^{g} |S_i|^{-\frac{n_i}{2}}}.\end{aligned}$$

La estadística:

$$-2 \ln \Lambda_1, \tag{1.9}$$

sigue asintóticamente una distribución χ^2_r (ver Teorema A.19) donde los grados de libertad son obtenidos del número de restricciones impuestas en los parámetros bajo H_1, y que se derivan del vector de medias y de la matriz de covarianzas. Ya que cada vector μ_i está en $\Re^p$ se obtienen p parámetros para cada una de las g poblaciones, mientras que la igualdad $\mu_i = \mu_k$ impone $(g-1)$ restricciones obteniéndose finalmente $p(g-1)$ restricciones en los vectores de medias. Ahora como la matriz de covarianzas es simétrica basta contar el número de restricciones que existen debajo o por encima de la diagonal de dicha matriz, con lo cual se obtiene $\frac{1}{2}p(p+1)$ y la igualdad $\Sigma_i = \Sigma_k$ impone $(g-1)$ restricciones, haciendo un total de $\frac{1}{2}p(p+1)(g-1)$ restricciones sobre la matriz de covarianzas. Finalmente el número de grados de libertad en la ecuación (1.9) se reduce a

$$\begin{aligned}r &= p(g-1) + \frac{1}{2}(g-1)p(p+1) \\ &= \frac{1}{2}(g-1)p(p+3).\end{aligned}$$

Entonces asintóticamente se tiene que

$$-2 \ln \Lambda_1 = n \ln |T| - \sum_{i=1}^{g} n_i \ln |S_i| \sim \chi^2_r,$$

definiendo

$$\lambda_1 = n \ln |T| - \sum_{i=1}^{g} n_i \ln |S_i|,$$

la región de rechazo es el conjunto de muestras dado por:

$$C_1 = \{\lambda_1 \mid \lambda_1 \geq \gamma\},$$

con γ el cuantil $(1-\alpha)$ de una χ^2 con $r = \frac{1}{2}(g-1)p(p+3)$ grados de libertad.

1.6.2 Igualdad de Matrices de Covarianzas

El segundo juego corresponde a la hipótesis sobre la igualdad de matrices de covarianzas. Este juego puede escribirse como:

$$H_2 \quad : \quad \Sigma_1 = \Sigma_2 = \cdots = \Sigma_g.$$

$$vs$$

$$H_{2.1} \quad : \quad \Sigma_i \neq \Sigma_k \quad i \neq k \text{ para al menos una } i.$$

Bajo la hipótesis H_2 cada grupo $\mathbf{X}_i$ proviene de una población $N_p(\mu_i, \Sigma)$, entonces los estimadores $\hat{\mu}_1, ..., \hat{\mu}_g$ que maximizan la función de verosimilitud en términos de $\mu_1, ..., \mu_g$ de acuerdo al Teorema 1.23 están dados por:

$$\hat{\mu}_i = X_i,$$

y la función de verosimilitud maximizando respecto a los vectores de medias es

$$\begin{aligned} L\left(\hat{\mu}_1, ..., \hat{\mu}_g, \Sigma; \mathbf{X}_1, ..., \mathbf{X}_g\right) &= \prod_{i=1}^{g} |2\pi\Sigma|^{-\frac{n_i}{2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^t \Sigma^{-1}\left(X_{ij} - \bar{X}_i\right)\right\} \\ &= |2\pi\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{n}{2} tr\left(\Sigma^{-1} W\right)\right\}, \end{aligned}$$

y siguiendo el Teorema 1.23 se tiene

$$\begin{aligned} \hat{\Sigma} &= W \\ &= \frac{1}{n_i}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)\left(X_{ij} - \bar{X}_i\right)^t \\ &= \frac{1}{n}\sum_{i=1}^{g} n_i S_i. \end{aligned}$$

Mientras tanto bajo la hipótesis $H_2 \cup H_{2.1}$ los estimadores de la función de verosimilitud según el Teorema 1.23 tienen las siguientes expresiones:

$$\begin{aligned} \hat{\mu}_i &= \bar{X}_i. \\ \hat{\Sigma}_i &= S_i. \end{aligned}$$

El cociente de verosimilitudes generalizado para este contraste está dado por:

$$\Lambda_2 = \frac{\sup_{H_2} L\left(\mu_1, ..., \mu_g, \Sigma\right)}{\sup_{H_2 \cup H_{2.1}} L\left(\mu_1, ..., \mu_g, \Sigma_1, ..., \Sigma_g\right)}$$

$$
\begin{aligned}
&= \frac{\prod_{i=1}^{g}\left\{|2\pi W|^{-\frac{n_i}{2}}\exp\left\{-\frac{1}{2}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)^t W^{-1}\left(X_{ij}-\bar{X}_i\right)\right\}\right\}}{\prod_{i=1}^{g}\left\{|2\pi S_i|^{-\frac{n_i}{2}}\exp\left\{-\frac{1}{2}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)^t S_i^{-1}\left(X_{ij}-\bar{X}_i\right)\right\}\right\}} \\
&= \frac{|W|^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)^t W^{-1}\left(X_{ij}-\bar{X}_i\right)\right\}}{\prod_{i=1}^{g}|S_i|^{-\frac{n_i}{2}}\exp\left\{-\frac{1}{2}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)^t S_i^{-1}\left(X_{ij}-\bar{X}_i\right)\right\}} \\
&= \frac{|W|^{-\frac{n}{2}}\exp\left\{-\frac{np}{2}\right\}}{\prod_{i=1}^{g}|S_i|^{-\frac{n_i}{2}}\exp\left\{-\frac{np}{2}\right\}} \\
&= \frac{|W|^{-\frac{n}{2}}}{\prod_{i=1}^{g}|S_i|^{-\frac{n_i}{2}}}.
\end{aligned}
$$

Utilizando la distribución asintótica $-2\ln\Lambda \sim \chi_s^2$ se sabe

$$-2\ln\Lambda_2 = n\ln|W| - \sum_{i=1}^{g} n_i \ln|S_i| \sim \chi_s^2.$$

El número de grados de libertad se obtiene del número de restricciones impuestas a los parámetros bajo la hipótesis nula. El cual corresponde a $(g-1)$ restricciones impuestas en la matriz de covarianzas por el número de parámetros libres en cada una de las matrices, es decir, $\frac{1}{2}p(p+1)$siendo éstos:

$$s = \frac{1}{2}(g-1)\,p(p+1),$$

grados de libertad, por lo tanto

$$-2\ln\Lambda_2 = n\ln|W| - \sum_{i=1}^{g} n_i \ln|S_i|,$$

tiene una distribución asintótica χ_s^2. Si

$$\lambda_2 = n\ln|W| - \sum_{i=1}^{g} n_i \ln|S_i|,$$

la región de rechazo para esta hipótesis toma la expresión

$$C_2 = \{\lambda_2 : \lambda_2 \geq \gamma\},$$

donde γ es el cuantil $(1-\alpha)$ de la distribución χ^2 con $s = \frac{1}{2}(g-1)\,p(p+1)$ grados de libertad.

1.6.3 Igualdad de Medias Condicionada a Matrices de Covarianzas.

El siguiente juego de hipótesis a contrastar es:

$$H_3 \ : \ \mu_i = \mu_k, \text{ dado que } \Sigma_i = \Sigma_k \ \forall i, k = 1, ..., g.$$

$$vs$$

$$H_{3.1} \ : \ \mu_i \neq \mu_k, \text{ dado que } \Sigma_i = \Sigma_k \text{ para al menos una } i.$$

En donde la función de verosimilitud de los parámetros está dada como:

$$L\left(\mu_i, ..., \mu_g, \Sigma; \mathbf{X}_1, ..., \mathbf{X}_g\right) = |2\pi\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(X_{ij}-\mu_i\right)^t \Sigma^{-1}\left(X_{ij}-\mu_i\right)\right\}.$$

La muestra bajo la hipótesis H_3 proviene de la población normal $N_p(\mu, \Sigma)$, y según el Teorema 1.23 se deduce que los estimadores correspondientes a la media y a la matriz de covarianzas están dados por:

$$\begin{aligned}\hat{\mu} &= \bar{X} = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i} X_{ij} \\ \hat{\Sigma} &= T = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}\right)\left(X_{ij}-\bar{X}\right)^t.\end{aligned}$$

Mientras tanto la muestra bajo $H_3 \cup H_{3.1}$, proviene de una población $N(\mu_i, \Sigma)$ y de acuerdo al Teorema 1.23 el estimador máximo verosímil para la media está dado por:

$$\hat{\mu}_i = \bar{X}_i.$$

Por lo que la función de verosimilitud está dada por:

$$\begin{aligned}L\left(\hat{\mu}_1, ..., \hat{\mu}_g, \Sigma; \mathbf{X}_1, ..., \mathbf{X}_g\right) &= \prod_{i=1}^{p}|2\pi\Sigma|^{-\frac{n_i}{2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}\right)\Sigma^{-1}\left(X_{ij}-\bar{X}\right)^t\right\} \\ &= |2\pi\Sigma|^{-\frac{n_i}{2}} \exp\left\{-\frac{n}{2}tr\Sigma^{-1}W\right\},\end{aligned}$$

y por el Teorema 1.23 se sigue que

$$\begin{aligned}\hat{\Sigma} &= \frac{1}{n_i}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)\left(X_{ij}-\bar{X}_i\right)^t \\ &= W \\ &= \frac{1}{n}\sum_{i=1}^{g} n_i S_i.\end{aligned}$$

Siguiendo las expresiones de los estimadores máximo verosímiles bajo cada una de las hipótesis, el cociente de verosimilitud es

$$\begin{aligned}\Lambda_3 &= \frac{\prod_{i=1}^{g}\left[|2\pi T|^{-\frac{n_i}{2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{g}\left(X_{ij}-\bar{X}\right)^t T^{-1}\left(X_{ij}-\bar{X}\right)\right\}\right]}{\prod_{i=1}^{g}\left[|2\pi W|^{-\frac{n_i}{2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{g}\left(X_{ij}-\bar{X}_i\right)^t W^{-1}\left(X_{ij}-\bar{X}_i\right)\right\}\right]} \\ &= \frac{|2\pi T|^{-\frac{np}{2}}\exp\left\{-\frac{np}{2}\right\}}{|2\pi W|^{-\frac{np}{2}}\exp\left\{-\frac{np}{2}\right\}} \\ &= \frac{|T|^{-\frac{n}{2}}}{|W|^{-\frac{n}{2}}} \\ &= \left|\frac{W}{T}\right|^{\frac{n}{2}}.\end{aligned}$$

En donde la región de rechazo es el conjunto definido como:

$$\begin{aligned}C &= \{\Lambda_3 \mid \Lambda_3 \leq \delta\} \\ &= \left\{\Lambda_3 = \left|\frac{W}{T}\right| \leq \delta\right\}.\end{aligned}$$

La distribución de Λ_3 puede encontrarse haciendo el desarrollo siguiente sobre la matriz T :

$$\begin{aligned}T &= \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}\right)\left(X_{ij}-\bar{X}\right)^t \\ &= \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i+\bar{X}_i-\bar{X}\right)\left(X_{ij}-\bar{X}_i+\bar{X}_i-\bar{X}\right)^t \\ &= \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)\left(X_{ij}-\bar{X}_i\right)^t+\frac{1}{n}\sum_{i=1}^{g}n_i\left(\bar{X}_i-\bar{X}\right)\left(\bar{X}_i-\bar{X}\right)^t.\end{aligned}$$

Recordando que la matriz W tiene la expresión

$$W=\frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)\left(X_{ij}-\bar{X}_i\right)^t,$$

y si se define la matriz B como:

$$B=\frac{1}{n}\sum_{i=1}^{g}n_i\left(\bar{X}_i-\bar{X}\right)\left(\bar{X}_i-\bar{X}\right)^t,$$

se obtiene entonces la igualdad

$$T = W + B.$$

De esta manera el cociente de verosimilitudes tiene la siguiente expresión:

$$\Lambda_3 = \frac{|W|}{|W+B|}. \tag{1.10}$$

Se probará que las matrices W y B tienen asociada una distribución Wishart con sus parámetros respectivos, y para ello defínase una matriz $\mathbf{X}^t = (\mathbf{X}_1^t, ..., \mathbf{X}_g^t)$ que contenga la información sobre cada muestra y en la que cada componente $\mathbf{X}_i$ de $\mathbf{X}$ es una matriz de dimensión $(n_i \times p)$ que está asociada a la $i - ésima$ población. Construyamos además un vector 1_i de dimensión $(n \times 1)$ formado por *unos* en la $i - ésima$ población y ceros en otro lado y una matriz $I_{i(n\times n)} = diag\,(1_i)$ obteniéndose con estas definiciones los siguientes resultados:

$$1_n = \sum_{i=1}^{g} 1_i.$$

$$I_n = \sum_{i=1}^{g} I_i.$$

Ya que la matriz nW es una forma cuadrática, puede expresarse como $nW = X^t M_1 X$, en donde la matriz M_1 puede obtenerse con base en la relación $nW = \sum_{i=1}^{n} n_i S_i$, donde la componente S_i es la matriz de covarianzas del $i - ésimo$ grupo de poblaciones.

Entonces si se considera el resultado del Teorema A.14 puede definirse una matriz H_i con la siguiente expresión:

$$H_i = I_i - \frac{1}{n_i} 1_i 1_i^t,$$

cumpliéndose $\forall\ i = 1, ..., g$, lo cual conduce a expresar $n_i S_i$ como:

$$n_i S_i = X^t H_i X$$

y consecuentemente la matriz M_1 queda expresada como:

$$M_1 = \sum_{i=1}^{g} H_i.$$

De acuerdo al Teorema 1.26 se debe probar que M_1 es idempotente y que $M_1 1_{n\times 1} = 0$.

1. M_1 es idempotente.

   Demostración.

$$M_1 M_1 = \sum_{i=1}^{g} H_i H_i + \sum_{i \neq j}^{g} H_i H_j,$$

   para el primer sumando

$$\begin{aligned}
\sum_{i=1}^{g} H_i H_i &= \sum_{i=1}^{g} \left(I_i - \frac{1}{n_i} 1_i 1_i^t \right) \left(I_i - \frac{1}{n_i} 1_i 1_i^t \right) \\
&= \sum_{i=1}^{g} \left(I_i - \frac{1}{n_i} 1_i 1_i^t - \frac{1}{n_i} 1_i 1_i^t + \frac{1}{n_i^2} 1_i 1_i^t 1_i 1_i^t \right) \\
&= \sum_{i=1}^{g} \left(I_i - \frac{1}{n_i} 1_i 1_i^t - \frac{1}{n_i} 1_i 1_i^t + \frac{1}{n_i^2} 1_i \left(1_i^t 1_i \right) 1_i^t \right) \\
&= \sum_{i=1}^{g} \left(I_i - \frac{1}{n_i} 1_i 1_i^t - \frac{1}{n_i} 1_i 1_i^t + \frac{1}{n_i} 1_i 1_i^t \right) \\
&= \sum_{i=1}^{g} \left(I_i - \frac{1}{n_i} 1_i 1_i^t \right) \\
&= \sum_{i=1}^{g} H_i,
\end{aligned}$$

   para el segundo sumando

$$\begin{aligned}
\sum_{i \neq j}^{g} H_i H_j &= \sum_{i=1}^{g} \left(I_i - \frac{1}{n_i} 1_i 1_i^t \right) \left(I_j - \frac{1}{n_j} 1_j 1_j^t \right) \\
&= 0.
\end{aligned}$$

   Finalmente puede concluirse que la matriz M_1 es idempotente.□

2. $M_1 1_{n \times 1} = 0$.

   Demostración.

$$\begin{aligned}
M_1 1_n &= \left(\sum_{i=1}^{g} H_i \right) 1_{n \times 1} \\
&= \sum_{i=1}^{g} I_i 1_n - \sum_{i=1}^{g} \frac{1}{n_i} 1_i 1_i^t 1_{n \times 1} \\
&= 1_n - 1_n \\
&= 0.
\end{aligned}$$

Ya que M_1 es idempotente su rango coincide con la traza, entonces se obtienen los grados de libertad como:

$$\begin{aligned} r(M_1) &= tr(M_1) \\ &= \sum_{i=1}^{g} trH_i \\ &= \sum_{i=1}^{g} tr\left(I_i - \frac{1}{n_i}1_i 1_i^t\right) \\ &= \sum_{i=1}^{g}\left(n_i - \frac{n_i}{n_i}\right) \\ &= n-1. \end{aligned}$$

Por lo tanto $nW \sim W_p(\Sigma, n-1)$.□

Obsérvese que la matriz nB toma la siguiente expresión:

$$\begin{aligned} nB &= \sum_{i=1}^{g} n_i\left(\bar{X}_i - \bar{X}\right)\left(\bar{X}_i - \bar{X}\right)^t \\ &= \sum_{i=1}^{g} n_i\, \bar{X}_i \bar{X}_i^{\,t} - n\,\bar{X}\bar{X}^t. \end{aligned}$$

Tomando $C_i = \frac{1}{n_i}1_i 1_i^t - \frac{1}{ng}11^t$, asociada a la $i-ésima$ población para $i = 1, ..., g$, se puede definir

$$X^t C_i X = \frac{1}{n_i}X^t 1_i 1_i^t X - \frac{1}{ng}X^t 11^t X.$$

Análogamente la forma cuadrática de nB es igual a la suma de $\sum_{i=1}^{g} X^t C_i X$ y la matriz M_2 coincide con $\sum_{i=1}^{g} C_i$, basándose en estas expresiones se puede escribir a nB como: Considerando el resultado del Teorema 1.26, debe probarse que M_2 es idempotente y $M_2 1 = 0$; las cuales son semejantes al caso de la matriz nW por lo cual se omite su demostración, y como conclusión se obtiene que $nB \sim W_p(\Sigma, g-1)$ en cuyo caso sólo debe obtenerse el número de grados de libertad asociados a esta distribución, lo que no es más que demostrar $r(M_2) = g-1$, es decir

$$\begin{aligned} r(M_2) &= tr(M_2) \\ &= tr\sum_{i=1}^{g} C_i \end{aligned}$$

$$\begin{aligned} &= \sum_{i=1}^{g} tr\left(\frac{1}{n_i}1_i1_i^t - \frac{1}{ng}11^t\right) \\ &= g-1. \end{aligned}$$

Por lo tanto $nB \sim W_p(\Sigma, g-1)$.

Por último se debe demostrar que nW y nB tienen una distribución independiente, por lo cual de acuerdo al Teorema 1.21 debe cumplirse que $M_1M_2 = 0$.

$$\begin{aligned} M_1M_2 &= \left(\sum_{i=1}^{g} H_i\right)\left(\sum_{i=1}^{g} C_i\right) \\ &= \left(I - \sum_{i=1}^{g}\frac{1}{n_i}1_i1_i^t\right)\left(\sum_{j=1}^{g}\frac{1}{n_j}1_j1_j^t - \frac{1}{n}11^t\right) \\ &= \sum_{j=1}^{g}\frac{1}{n_j}1_j1_j^t - \frac{1}{n}11^t - \sum_{i=1}^{g}\sum_{j=1}^{g}\frac{1}{n_in_j}1_i1_i^t1_j1_j^t + \sum_{i=1}^{g}\frac{1}{n_in}1_i1_i^t11^t \\ &= \sum_{j=1}^{g}\frac{1}{n_j}1_j1_j^t - \frac{1}{n}11^t - \sum_{j=1}^{g}\frac{1}{n_j}1_j1_j^t + \frac{1}{n}11^t \\ &= 0.\square \end{aligned}$$

De la Definición 1.13 y suponiendo que $n \geq p+g$ bajo la hipótesis H_3 se puede expresar la ecuación (1.10) como:

$$\Lambda_3 = \left|I + W^{-1}B\right|^{-1} \sim \Lambda\left(p, n-g, g-1\right).$$

Por lo tanto, la región de rechazo toma la expresión:

$$\begin{aligned} C_3 &= \{\Lambda_3 : \Lambda_3 \leq \lambda\} \\ &= \left\{\Lambda_3 = \left|I + W^{-1}B\right|^{-1} \leq \lambda\right\}. \end{aligned}$$

Donde λ es el cuantil $(1-\alpha)$ asociado a la distribución Λ de Wilks con $(p, n-g, g-1)$ grados de libertad.

Capítulo 2

ANALISIS DE COMPONENTES PRINCIPALES

En ocasiones si se observa un conjunto de p variables que están correlacionadas es necesario transformarlas en un conjunto con menor número de variables no correlacionadas llamadas Componentes Principales, que guardan la información relevante de la muestra. El Análisis de Componentes Principales ($A.C.P.$), transforma las variables originales en un conjunto de combinaciones lineales tal que estas acumulen la mayor proporción de varianza del conjunto original. Esta transformación es de hecho una rotación ortogonal en un espacio de p dimensiones.

El objetivo principal de este Análisis es tomar sólo algunos componentes, de tal manera que estos acumulen una proporción significativa de la varianza del conjunto original. Si esto puede hacerse entonces se concluye que la dimensión puede reducirse.

El $A.C.P.$ es una técnica matemática que no requiere de ningún modelo estadístico. En particular, ningún supuesto se hace inicialmente acerca de la distribución de probabilidad de las variables originales.

Sin embargo si se supone que la población tiene una distribución normal, la muestra observada podrá ser utilizada para realizar inferencias estadísticas a partir de pruebas de hipótesis que contribuyan a conocer la estructura de la población original.

2.1 Definición y Propiedades de los Componentes Principales en la Población

En esta Sección se da la expresión matemática que define a los componentes principales así como las propiedades que dichos componentes satisfacen. Estos resultados consideran una muestra de n individuos a los cuales se les denota por X_i, donde cada uno de estos individuos tiene asociadas p características sobre las cuales está basado el *A.C.P.*, esta muestra puede agruparse en una matriz $\mathbf{X}$ de dimensión $n \times p$ con la forma

$$\mathbf{X} = \left.\begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2j} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nj} & \cdots & X_{np} \end{pmatrix}\right\} n \text{ individuos.} \tag{2.11}$$

Adicionalmente se puede suponer que cada individuo X_i tiene vector de medias y matriz de covarianzas totalmente definidos y que están dados por

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} \quad \text{y} \quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}. \tag{2.12}$$

2.1.1 Componentes Principales Basados en la Matriz de Covarianzas Poblacional

Si la matriz de covarianzas asociada a un vector aleatorio X es conocida, la transformación que lleva a esta variable a su correspondiente vector Y de componentes principales está dada en la siguiente definición:

Definición 2.1 *Si X es un vector aleatorio con media μ y matriz de covarianzas Σ, entonces la transformación de los componentes principales es la transformación*

$$X \rightarrow Y = \Gamma^t (X - \mu), \tag{2.13}$$

donde Γ es una matriz ortogonal, $\Gamma^t \Sigma \Gamma = \Lambda$ es una matriz diagonal de los valores propios de la matriz Σ, los cuales cumplen que $\lambda_1 \geq ... \geq \lambda_p \geq 0$; estos valores son estrictamente positivos si la matriz $\Sigma > 0$.

El $j-$ésimo componente principal de X está definido como el $j-$ésimo elemento del vector Y, cuya expresión es la siguiente:

$$Y_j = \Gamma_{(j)}^t (X - \mu).$$

donde $\Gamma_{(j)}$ es la $j-$ésima columna de Γ, y es llamado el $j-$ésimo vector correspondiente al $j-$ésimo componente principal.

Teorema 2.1 *Si X es un vector de parámetros μ y Σ, donde la matriz Σ puede descomponerse como $\Sigma = \Gamma \Lambda \Gamma^t$ con $\Gamma^t \Gamma = I_p$ y $Y = \Gamma^t (X - \mu)$ entonces se satisfacen las siguientes propiedades:*

1. $E(Y_j) = 0 \quad \forall\, j$
2. $V(Y) = \Lambda$.
3. $V(Y_1) \geq ... \geq V(Y_p) \geq 0$.
4. $\sum_{j=1}^{p} V(Y_j) = tr\Sigma$.
5. $\prod_{j=1}^{p} V(Y_j) = |\Sigma|$.

Demostración

1.

$$\begin{aligned} E(Y_j) &= E\left[\Gamma_{(j)}^t (X-\mu)\right] \\ &= E\left(\Gamma_{(j)}^t X\right) - E\left(\Gamma_{(j)}^t \mu\right) \\ &= \Gamma_{(j)}^t E(X) - \Gamma_{(j)}^t \mu \\ &= 0. \ \square \end{aligned}$$

Teorema 2.2

$$\begin{aligned} V(Y) &= V\left(\Gamma^t (X-\mu)\right) \\ &= V\left(\Gamma^t X\right) \\ &= \Gamma^t V(X)\, \Gamma \\ &= \Gamma^t \Gamma \Lambda \Gamma^t \Gamma \\ &= \Lambda. \ \square \end{aligned}$$

1. Se sigue de la Definición 2.1 y del inciso anterior. $\square$

2.

$$\begin{aligned} tr(\Sigma) &= tr\left(\Gamma \Lambda \Gamma^t\right) \\ &= tr\left(\Gamma^t \Gamma \Lambda\right) \\ &= tr(\Lambda) \\ &= \sum_{j=1}^{p} \lambda_j \\ &= \sum_{j=1}^{p} V(Y_j). \ \square \end{aligned}$$

3.

$$\begin{aligned} |\Sigma| &= \left|\Gamma \Lambda \Gamma^t\right| \\ &= |\Gamma|\,|\Lambda|\,\left|\Gamma^t\right| \\ &= \left|\Gamma^t \Gamma\right|\,|\Lambda| \\ &= |\Lambda| \end{aligned}$$

$$= \prod_{j=1}^{p} \lambda_j$$
$$= \prod_{j=1}^{p} V(Y_j). \square$$

Definición 2.2 *Una combinación lineal estándarizada (CLE) de* $X \in \Re^p$ *se define como:*

$$Y = a^t X.$$

donde $a^t a = 1$.

Teorema 2.3 *Sea* $Y = a^t X$ *una CLE de* X, *donde* $E(X) = \mu$ *y* $V(X) = \Sigma \geq 0$, *entonces* $V(Y) \leq \lambda_1$ *donde la descomposición espectral de la matriz* Σ *se define por* $\Sigma = \Gamma \Lambda \Gamma^t$, $\Lambda = diag(\lambda_1, ..., \lambda_p)$, $\Gamma^t \Gamma = I_p$ *y se satisface que* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. *Adicionalmente* $V\left(\Gamma_{(1)}^t X\right) = \lambda_1$.

Demostración.

Como las columnas de Γ *son linealmente independientes, se puede escribir*

$$a = \Gamma b,$$

con $b^t = (b_1, ..., b_p)$

$$a = \left(\Gamma_{(1)}, ..., \Gamma_{(p)}\right) b$$
$$= \sum_{j=1}^{p} b_j \Gamma_{(j)}.$$

Por la Definición 2.2 se sabe que

$$1 = a^t a$$
$$= (\Gamma b)^t (\Gamma b)$$
$$= b^t b.$$

Por lo que la $V(Y)$ toma la expresión:

$$\begin{aligned} V(Y) &= V(a^t X) \\ &= a^t V(X)\, a \\ &= b^t \Lambda b \\ &= \sum_{j=1}^{p} b_j^2 \lambda_j. \end{aligned}$$

Entonces la $V(Y)$ se maximiza si

$$\begin{aligned} b &= e_{(1)} \\ &= \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \end{aligned}$$

y con esto

$$\begin{aligned} a &= \Gamma b \\ &= \Gamma_{(1)}. \end{aligned}$$

Finalmente la $V(Y)$ se calcula como:

$$\begin{aligned} V(Y) &= \Gamma_{(1)}^t \Lambda \Gamma_{(1)} \\ &= \lambda_1. \square \end{aligned}$$

Teorema 2.4 *Si $Y = a^t X$ es una CLE, la cual está no correlacionada con los primeros k componentes principales de X, definidos por:*

$$Y_j = \Gamma_{(j)}^t (X - \mu) \quad j = 1, ..., k.$$

Entonces $V(Y)$ se maximiza si $a = \Gamma_{(k+1)}$.

Demostración.

Dado que las columnas de Γ son linealmente independientes, se puede escribir

$$a = \Gamma b,$$

donde $b^t b = 1$, *utilizando este resultado se tiene*

$$\begin{aligned} 0 &= C\left(Y, \Gamma_{(j)}^t (X-\mu)\right) \\ &= C\left(a^t X, \Gamma_{(j)}^t X\right) \\ &= a^t C(X,X)\,\Gamma_{(j)} \\ &= a^t \Gamma \Lambda \Gamma^t \Gamma_{(j)} \\ &= b^t \Lambda \Gamma^t \Gamma_{(j)}. \end{aligned} \tag{2.14}$$

Dado que

$$\begin{pmatrix} \Gamma_{(1)} \\ \Gamma_{(2)} \\ \vdots \\ \Gamma_{(p)} \end{pmatrix} \Gamma_{(j)} = e_{(j)},$$

se puede escribir la ecuación (2.14) como:

$$\begin{aligned} b^t \Lambda e_{(j)} &= 0 \\ b_j \lambda_j &= 0. \end{aligned}$$

Aquí se puede suponer que $\lambda_j > 0 \ \forall\ j = 1, ..., k$, *ya que en el momento en que alguna* $\lambda_j = 0$, *esto implicaría que las* $\lambda_r = 0 \ \forall\ r \neq j,\ r = j, ..., p$.

Esto implica que

$$b_j = 0 \quad j = 1, ..., k.$$

Por el desarrollo de la demostración del Teorema 2.3 se cumple que

$$\begin{aligned} V(Y) &= \sum_{j=1}^{p} b_j^2 \lambda_j \\ &= \sum_{j=k+1}^{p} b_j^2 \lambda_j. \end{aligned}$$

Por lo tanto, $V(Y)$ *se maximiza si* $b_j = e_{(k+1)}$, *entonces*

$$\begin{aligned} a &= \Gamma b \\ &= \Gamma e_{(k+1)} \\ &= \Gamma_{(k+1)}. \end{aligned}$$

Entonces

$$\begin{aligned} V(Y) &= V\left(\Gamma_{(k+1)}^t X\right) \\ &= \Gamma_{(k+1)}^t \Sigma \Gamma_{(k+1)} \\ &= e_{(k+1)}^t \Lambda e_{(k+1)} \\ &= \lambda_{k+1}. \square \end{aligned}$$

2.1.2 Análisis de Componentes Principales con Base en la Matriz de Correlación Poblacional

Si las variables bajo estudio están medidas en unidades diferentes es recomendable aplicarles una estandarización para que los efectos de la escala no influyan en la determinación de los componentes principales. Dicha estandarización supone que la matriz $\mathbf{X}_{n\times p}$ dada por (2.11), tiene asociado un vector de medias y una matriz de covarianzas como en (2.12) de modo que la estandarización para la $j-ésima$ variable del elemento X_i toma la expresión

$$\widetilde{Z}_{=} \frac{X_{ij}-\mu_j}{\sqrt{\sigma_{jj}}}\ , \quad \forall\, j$$

Sea $\Delta = diag\Sigma$, entonces la ecuación anterior puede escribirse en forma matricial como

$$\widetilde{Z}_{=} \left(\Delta^{\frac{1}{2}}\right)^{-1}(X-\mu)\,,$$

y se sigue claramente que $E\left(\widetilde{Z}\right)=0$, y

$$\begin{aligned} V\left(\widetilde{Z}\right) &= V\left[\left(\Delta^{\frac{1}{2}}\right)^{-1}(X-\mu)\right] \\ &= \left(\Delta^{\frac{1}{2}}\right)^{-1}\Sigma\left(\Delta^{\frac{1}{2}}\right)^{-1} \\ &= \rho. \end{aligned}$$

En donde ρ denota la matriz de correlación poblacional que puede descomponerse como $\rho = \widetilde{\Gamma}\widetilde{\Lambda}\widetilde{\Gamma}^t$ y las columnas de la matriz $\widetilde{\Gamma}$ son los vectores propios de ρ, los cuales son ortogonales. Por lo tanto, la transformación que define a los componentes principales está dada por:

$$Y = \widetilde{\Gamma}^t \widetilde{Z},$$

y

$$\begin{aligned} V(Y) &= V\left(\tilde{\Gamma}^t \tilde{Z}\right) \\ &= \tilde{\Gamma}\, \rho\, \tilde{\Gamma}^t \\ &= \tilde{\Lambda}\,. \end{aligned}$$

2.2 Componentes Principales Generados a partir de una Muestra

Generalmente la matriz de covarianzas Σ es desconocida pero puede ser estimada mediante la muestra observada. En esta Sección se analizarán los componentes principales que son generados de la matriz de covarianzas y de la matriz de correlación. En ambos casos se supone que los valores de las p variables $X_1, ..., X_p$ que son obtenidos de una muestra observada de n individuos, se concentran en una matriz $\mathbf{X}$ de dimensión $n \times p$ con la misma forma que la dada en la ecuación (2.11).

2.2.1 Análisis de Componentes Principales Basado en la Matriz de Covarianzas

Dado que en este punto la matriz de covarianzas es desconocida, el *A.C.P.* se basa en una matriz de covarianzas muestral S_X donde:

$$\bar{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij} \quad \text{con} \quad j = 1, ..., p,$$

es la media de los valores observados para la $j - \acute{e}sima$ variable sobre los n individuos, y

$$S_{jk} = \frac{1}{n} \sum_{i=1}^{n} \left(X_{ij} - \bar{X}_j\right) \left(X_{ik} - \bar{X}_k\right)^t.$$

es la covarianza muestral entre las variables X_j y X_k. De modo que $S_X = \{S_{jk}\}$ es la matriz de covarianzas de las p variables.

Definición 2.3 *Sea U una matriz ortogonal cuyos elementos en la diagonal son positivos y tales que*

$$U^t S_X U = \widetilde{\Lambda} .$$

y $U^t U = I_p$ donde $\widetilde{\lambda}_1, ..., \widetilde{\lambda}_p$ son los valores propios ordenados y que están asociados a la matriz S_X. La transformación de los componentes principales de un vector $X \in \Re^p$ está definida como:

$$Y_{p\times 1} = U^t \left(X - \bar{X} \right),$$

en donde la $i-$ésima nueva observación está dada por:

$$Y_i = U^t \left(X_i - \bar{X} \right) \quad i = 1, ..., n.$$

La media muestral de estas nuevas observaciones está dada como:

$$\begin{aligned} \bar{Y} &= \frac{1}{n} \sum_{i=1}^{n} Y_i \\ &= \frac{1}{n} \sum_{i=1}^{n} U^t \left(X_i - \bar{X} \right) \\ &= \frac{1}{n} U^t \sum_{i=1}^{n} \left(X_i - \bar{X} \right) \\ &= 0, \end{aligned}$$

y la matriz de covarianzas muestral S_Y puede definirse por:

$$\begin{aligned} S_Y &= \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \bar{Y} \right) \left(Y_i - \bar{Y} \right)^t, \\ &= \frac{1}{n} \sum_{i=1}^{n} U^t \left(X_i - \bar{X} \right) \left(X_i - \bar{X} \right)^t U, \\ &= U^t S_X U, \\ &= L. \end{aligned}$$

En forma matricial las nuevas observaciones pueden escribirse de la forma siguiente:

$$Y_{n\times p} = \begin{pmatrix} Y_1^t \\ Y_2^t \\ \vdots \\ Y_n^t \end{pmatrix} = \begin{pmatrix} \left(X_1 - \bar{X}\right)^t \\ \left(X_2 - \bar{X}\right)^t \\ \vdots \\ \left(X_n - \bar{X}\right)^t \end{pmatrix} U.$$

En donde el $j-ésimo$ elemento de Y_i, dado por Y_{ij} representa el puntaje del $j-ésimo$ componente sobre el $i-ésimo$ individuo. De tal forma que en términos de este individuo se puede escribir la transformación del componente principal como

$$Y_{ij} = U_{(j)}^t \left(X_i - \bar{X}\right) \quad \text{con} \quad i = 1, ..., n \quad \text{y} \quad j = 1, ..., p.$$

2.2.2 Componentes Principales Generados Mediante Variables Estandarizadas

En esta Sección se considera un análisis similar al de la Sección (2.1.2), donde se supone que las variables bajo estudio están medidas en unidades sumamente diferentes y que los efectos de la escala puedan influir en la composición o derivación de los Componentes Principales. La estandarización se aplica a una matriz $\mathbf{X}_{n\times p}$ dada en (2.11), y que tiene asociado un vector de medias $\bar{X}_{p\times 1}$ y una matriz de covarianzas dados por

$$\bar{X} = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_p \end{pmatrix} \quad \text{y} \quad S_X = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}.$$

La estandarización de la matriz $\mathbf{X}$ queda establecida en la siguiente relación

$$Z = \left(D^{\frac{1}{2}}\right)^{-1} \left(X - \bar{X}\right), \text{ donde } X \in \Re^p$$

donde la matriz $D = diagS_X$ y Z tiene media cero y varianza

$$\begin{aligned} V(Z) &= V\left[\left(D^{\frac{1}{2}}\right)^{-1}\left(X-\bar{X}\right)\right] \\ &= \left(D^{\frac{1}{2}}\right)^{-1} V(X)\left(D^{\frac{1}{2}}\right)^{-1} \\ &= R. \end{aligned}$$

En donde R denota a la matriz de correlación muestral, cuya descomposición espectral se define como $R = \tilde{U}\tilde{L}\tilde{U}^t$. Finalmente la transformación por la que se obtienen los componentes principales está dada por

$$Y = \tilde{U}^t Z,$$

y varianza

$$\begin{aligned} V(Y) &= V\left(\tilde{U}^t Z\right) \\ &= \tilde{U}^t R \tilde{U} \\ &= \tilde{L}. \end{aligned}$$

2.2.3 Estructura de Correlación.

Examínese ahora la correlación entre el punto X y el vector de componentes principales Y, definido como en la ecuación (2.13). La covarianza entre el punto X y la variable Y se calcula como:

$$\begin{aligned} C(X,Y) &= C(X, \Gamma^t(X-\mu)) \\ &= C(X, \Gamma^t X) \\ &= C(X, X\Gamma) \\ &= V(X)\Gamma \\ &= \Gamma\Lambda\Gamma^t\Gamma \\ &= \Gamma\Lambda. \end{aligned}$$

Entonces la covarianza entre X_k y Y_j está dada como:

$$C(X_k, Y_j) = \Gamma_{kj}\lambda_j.$$

Ahora si se define a la $V(\mathbf{X}) = \Sigma = \{\sigma_{kj}\}$, $k, j = 1, ..., p$ y Λ la matriz diagonal de las varianzas de Y, la correlación entre las variables X_k y Y_j se obtiene como:

$$r_{kj} = \frac{\Gamma_{kj}\lambda_j}{\sqrt{\sigma_{kk}\lambda_j}}$$

$$= \frac{\Gamma_{kj}\sqrt{\lambda_j}}{\sqrt{\sigma_{kk}}}.$$

Entonces se dice que la proporción de variabilidad explicada de X_k por la componente Y_j es r^2_{kj}, donde

$$r^2_{kj} = \frac{\Gamma^2_{kj}\lambda_j}{\sigma_{kk}}.$$

Ya que los elementos de Y son no correlacionados, cualquier subconjunto I de componentes principales explica una proporción

$$\begin{aligned} r^2_{kI} &= \sum_{j\in I} r^2_{kj} \\ &= \frac{1}{\sigma_{kk}} \sum_{j\in I} \lambda_j \Gamma^2_{kj}. \end{aligned}$$

de la variación de X_k. El denominador de esta última expresión representa la variación de X_k que va a ser explicada, y el numerador proporciona la variación acumulada por el conjunto I. Cuando I incluye todos los componentes principales, la proporción acumulada es lógicamente uno.

La proporción de X_i explicada por Y_j con $j \in I$ puede concentrarse en la siguiente Tabla 1

r^2_{ij}	Y_1	$\cdots$	Y_p
X_1	$\frac{\lambda_1 U^2_{11}}{\sigma_{11}}$	$\cdots$	$\frac{\lambda_p U^2_{1p}}{\sigma_{11}}$
X_2	$\frac{\lambda_2 U^2_{21}}{\sigma_{22}}$	$\cdots$	$\frac{\lambda_p U^2_{2p}}{\sigma_{22}}$
$\vdots$			
X_p	$\frac{\lambda_p U^2_{p1}}{\sigma_{pp}}$	$\cdots$	$\frac{\lambda_p U^2_{pp}}{\sigma_{pp}}$

Tabla 1 . Estructura de correlación

2.2.4 Algunas Propiedades sobre los Componentes Principales

Las propiedades más importantes de los Componentes Principales están dadas en el Teorema 2.1 y en la Definición 2.1. En esta Sección se analizarán algunas propiedades que son de gran ayuda para interpretar los resultados obtenidos.

1. **Componentes principales bajo cambios de escala de las variables.** Los componentes principales de un vector aleatorio no son invariantes respecto a la escala.

2. **La media de las nuevas observaciones.** Si se define a $\bar{X}$ como la media muestral de las observaciones originales y se utiliza la matriz de covarianzas muestral S_X, entonces la transformación general es:

$$Y = U^t \left(X - \bar{X} \right),$$

donde esta transformación consta de una translación seguida de una rotación; para la cual la media de los componentes principales es igual a cero.

Si $\tilde{U}$ denota la matriz de los vectores propios asociada a la matriz de correlación R, entonces dicha transformación sólo puede ser usada después de estandarizar las variables $\left(X - \bar{X} \right)$ de tal suerte que cada variable tenga varianza unitaria.

3. **Proporción de variabilidad explicada.** El cociente

$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j},$$

representa la proporción de variabilidad explicada por los primeros k componentes principales.

4. **Valores propios iguales a cero.** Esto ocurre cuando algunas de las variables originales son linealmente dependientes, entonces algunos de los valores propios de Σ son iguales a cero. Por el corolario A.3 se sabe que la dimensión del espacio que contiene a las observaciones es igual al rango de la matriz Σ, y éste está dado por $(p - q)$ donde q representa el número de valores propios iguales a cero.

5. **El rango de la matriz X.** Si la matriz de covarianzas de la muestra $\mathbf{X}_{n \times p}$ es de rango $k < p$, entonces la variabilidad de $\mathbf{X}$ puede ser explicada totalmente por los primeros k componentes principales.

6. **Valores propios repetidos.** En algunas ocasiones los valores propios de Σ son iguales y si ocurre que

$$\lambda_{q+1} = \cdots = \lambda_{q+k} = \lambda,$$

entonces se dice que la raíz λ es de multiplicidad k. Los vectores propios correspondientes a las raíces múltiples no son únicos y sus correspondientes componentes tendrán la misma varianza.

2.3 Interpretación Geométrica de los Componentes Principales bajo Normalidad de las Observaciones

En la Sección 1.2.1 del Capítulo 1, se mencionó que la función de densidad de un vector $X \in \Re^p$ puede escribirse como:

$$f_X(X) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(X-\mu)^t \Sigma^{-1}(X-\mu)\right\},$$

siempre que $X \sim N_p(\mu, \Sigma)$, $\Sigma > 0$.

Nótese que $|2\pi\Sigma|^{-\frac{1}{2}}$ es una constante que no depende del vector X y además que la forma cuadrática

$$(X-\mu)^t \Sigma^{-1}(X-\mu) = c, \tag{2.15}$$

define un elipsoide en un espacio de dimensión p. Se genera una familia de estos elipsoides haciendo variar la constante c.

Obsérvese que la matriz Σ es definida positiva entonces por el Corolario A.6 se sigue que Σ^{-1} también lo es y se puede utilizar el Teorema A.10 para descomponer esta matriz como:

$$\Sigma^{-1} = U\Lambda^{-1}U^t.$$

La ecuación (2.15) puede reescribirse como:

$$(X-\mu)^t U\Lambda^{-1}U^t(X-\mu) = c, \tag{2.16}$$

y los ejes principales de este elipsoide son simplemente los vectores propios de la matriz Σ. Tomando $Y = U^t(X-\mu)$, la ecuación (2.16) puede escribirse de la siguiente manera:

$$Y^t\Lambda^{-1}Y = c,$$

$\Leftrightarrow$

$$\sum_{j=1}^{p} \frac{Y_j^2}{\lambda_j} = c.$$

La magnitud del $j - ésimo$ eje principal está dado por:

$$y_j = \pm\sqrt{\lambda_j c}.$$

Ejemplo 2.1 *Encontrar la elipse del 95% de concentración para un vector aleatorio* $X^t = (X_1, X_2)$, *con distribución Normal bivariada, de parámetros*

$$\mu = \begin{pmatrix} -4 \\ 12 \end{pmatrix} \quad y \quad \Sigma = \begin{pmatrix} 4 & 5 \\ 5 & 9 \end{pmatrix}. \tag{2.17}$$

La forma distribucional del vector X *está dada como:*

$$f_{X_1,X_2}(x_1,x_2) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(X-\mu)^t \Sigma^{-1}(X-\mu)\right\}, \tag{2.18}$$

donde la descomposición espectral de la matriz de covarianzas es:

$$\Sigma = U\Lambda U^t,$$

con

$$U = \begin{pmatrix} 0.525731 & 0.850651 \\ 0.850651 & -0.525731 \end{pmatrix} \quad y \quad \Lambda = \begin{pmatrix} 12.0902 & 0 \\ 0 & 0.90983 \end{pmatrix}.$$

Entonces la forma cuadrática del exponente en (2.17) es

$$\begin{aligned} \delta &= \begin{pmatrix} X_1+4 \\ X_2-12 \end{pmatrix}^t U\Lambda^{-1}U^t \begin{pmatrix} X_1+4 \\ X_2-12 \end{pmatrix} \qquad (2.19) \\ &= \begin{pmatrix} X_1+4 \\ X_2-12 \end{pmatrix}^t \begin{pmatrix} 0.525731 & 0.850651 \\ 0.850651 & -0.525731 \end{pmatrix} \Lambda^{-1} \begin{pmatrix} 0.525731 & 0.850651 \\ 0.850651 & -0.525731 \end{pmatrix} \\ &= \begin{bmatrix} 0.525731(X_1+4)+0.850651(X_2-12) \\ 0.850651(X_1+4)-0.525731(X_2-12) \end{bmatrix}^t \begin{pmatrix} 12.0902 & 0 \\ 0 & 0.90983 \end{pmatrix}^{-1} \\ &\quad \times \begin{bmatrix} 0.525731(X_1+4)+0.850651(X_2-12) \\ 0.850651(X_1+4)-0.525731(X_2-12) \end{bmatrix}, \end{aligned}$$

y las curvas de nivel definen elipses de concentración en un espacio de dos dimensiones. Si se define un vector $Y^t = (Y_1, Y_2) = (X - \mu)^t U$ *la forma cuadrática (2.19) se simplifica como*

$$\delta = Y^t \Lambda^{-1} Y.$$

Entonces para encontrar la elipse de concentración del 95% se debe calcular

$$\begin{aligned} P\left[Y^t \Lambda^{-1} Y \leq \delta\right] &= 0.95 \\ P\left[\sum_{j=1}^{2} \frac{Y_j}{\lambda_j} \leq \delta\right] &= 0.95, \end{aligned}$$

donde $\delta = 5.99$ *es el cuantil 0.95 de una* χ^2 *con dos grados de libertad.*

Finalmente las intersecciones con los nuevos ejes coordenados están dados haciendo

$$Y_1 = 0 \text{ lo cual implica que } Y_2 = \pm\sqrt{\lambda_2 \delta} = \pm\sqrt{(0.90983)\,5.99} = 2.3345$$

y

$$Y_2 = 0 \text{ lo cual implica que } Y_1 = \pm\sqrt{\lambda_1 \delta} = \pm\sqrt{(12.09021)\,5.99} = 8.5100,$$

por lo tanto la magnitud del primer eje principal es $Y_1 = 8.5100$, *mientras que la magnitud correspondiente al segundo eje principal es* $Y_2 = 2.3345$. *Gráficamente la elipse de concentración en dos dimensiones se observa como:*

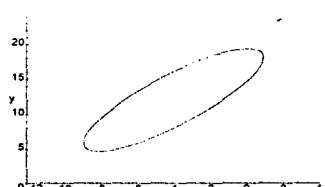


Figura 1. Elipse de concentración del 95% para la normal bivariada de parámetros definidos en (2.17).

2.4 Inferencia sobre los Componentes Principales

2.4.1 Estimación Máximo Verosímil para Datos Normales

La distribución de una muestra pequeña de valores y vectores propios de una matriz de covarianzas S, es extremadamente complicada aún cuando no exista correlación. Una de las razones, es por que los valores propios son funciones no racionales de los elementos de S.

Sin embargo para una muestra grande los resultados son conocidos, y algunas de las propiedades usuales de la muestra de componentes principales para datos normales están contenidos en los resultados de máxima verosimilitud que a continuación se citan.

Teorema 2.5 *Para datos normales cuando los valores propios de Σ son distintos, los componentes principales y valores propios muestrales, son los estimadores máximo verosímiles de los parámetros poblacionales correspondientes.*

Demostración. Se sigue de la propiedad de invarianza de los estimadores máximo verosímiles.

Teorema 2.6 *Para datos normales, cuando $k > 1$, los valores propios de Σ son iguales y toman un valor común $\bar{\lambda}$, se cumple que*

1. *El estimador máximo verosímil de $\bar{\lambda}$ es $\bar{l}$, la media aritmética muestral correspondiente a los vectores propios con $\bar{\lambda}$ común.*
2. *Los vectores propios muestrales correspondientes a $\bar{\lambda}$, son estimadores máximo verosímiles, sin embargo no son únicos.*

   Demostración. Ver Anderson (1984).

Teorema 2.7 *Sea Σ una matriz definida positiva con valores propios distintos y sean $M \sim W(\Sigma, m)$ y $W = m^{-1}M$. Considérese la descomposición espectral $\Sigma = \Gamma\Lambda\Gamma^t$ y $W = GLG^t$*

y sean $\Upsilon = diag(\Lambda)$ *y* $\Phi = diag(L)$. *Entonces las siguientes distribuciones asintótica se satisfacen siempre que* $m \to \infty$.

1. $\Phi \sim N_p\left(\Upsilon, \frac{2\Upsilon^2}{m}\right)$, *esto es, los valores propios de W son asintóticamente normales, insesgados e independientes.*

2. $g_{(i)} \sim N_p\left(\Gamma_{(i)}, \frac{V_i}{m}\right)$, *donde*

$$V_i = \lambda_i \sum_{j \neq i} \frac{\lambda_j}{(\lambda_j - \lambda_i)} \Gamma_{(i)} \Gamma'_{(j)},$$

Con $\Gamma_{(i)}$ *el vector correspondiente a la* $i-$*ésima columna de la matriz* Γ. *Los vectores propios de W son asintóticamente normales e insesgados, con matriz de covarianza asintótica* $\frac{V_i}{m}$.

3. *La covarianza entre el r-ésimo elemento de* $g(i)$ *y el t-ésimo elemento de* $g(j)$ *es*

$$-\frac{\lambda_i \lambda_j \gamma_{ri} \gamma_{tj}}{m(\lambda_i - \lambda_j)^2}.$$

4. *Los elementos de L son asintóticamente independientes de los elementos de G.*

Demostración. Ver Mardia (1984)

Teorema 2.8 *Sea* $\hat{\Sigma}$ *el estimador máximo verosímil para* Σ, *basado en una muestra de tamaño n, de una población* $N(\mu, \Sigma)$. *Sean* $\hat{\lambda} = \left(\hat{\lambda_1}, ..., \hat{\lambda_p}\right)^t$ *y* $\lambda = (\lambda_1, ..., \lambda_p)^t$, *en donde* $\hat{\lambda_i}$ *y* λ_i *son los valores propios de* $\hat{\Sigma}$ *y* Σ *respectivamente. Sea* $\Lambda = diag(\lambda_1, ..., \lambda_p)$, *entonces si* $\Sigma > 0$ *y todos los valores propios son distintos, i.e.,* $\lambda_1 > \cdots > \lambda_p > 0$, $\sqrt{n-1}\left(\hat{\lambda} - \lambda\right)$ *se distribuye asintóticamente como una* $N(0, 2\Lambda^2)$.

Demostración. Ver Girshick (1939).

Teorema 2.9 *Sean las mismas definiciones y condiciones que en el Teorema 2.8, sean además* $\hat{\gamma}$ *y* γ *los vectores propios normalizados de* $\hat{\Sigma}$ *y* Σ *respectivamente.*

Entonces $\sqrt{n-1}\left(\hat{\gamma}_i - \gamma_i\right)$, $i = 1, ..., p$ *se distribuye asintóticamente como una* $N(0, L_i)$, *donde*

$$L_i = \lambda_i \sum_{j=1, j\neq i}^{p} \frac{\lambda_j}{(\lambda_j - \lambda_i)^2} \gamma_{(j)} \gamma_{(j)}^t.$$

Demostración. Girshick (1939).

2.4.2 Intervalo de Confianza para un Valor Propio

Como consecuencia de la Teoría anterior puede construirse un intervalo de confianza para algún λ_j, que hable de la dispersión de este valor; dicho intervalo puede obtenerse haciendo el siguiente desarrollo:

Por el Teorema 2.8 la expresión $\sqrt{n-1}\left(\hat{\lambda}_j - \lambda_j\right)$ sigue asintóticamente ($\approx$) una distribución $N(0, 2\lambda_j^2)$ $j = 1, ..., p$. Estandarizando se obtiene:

$$\frac{\sqrt{n-1}\left(\hat{\lambda}_j - \lambda_j\right)}{\sqrt{2}\lambda_j} \approx N(0,1),$$

esta expresión es una cantidad pivotal con la que puede obtenerse el intervalo deseado, mediante la siguiente probabilidad

$$P\left[z_1 < \frac{\sqrt{n-1}\left(\hat{\lambda}_j - \lambda_j\right)}{\sqrt{2}\lambda_j} < z_2\right] = (1-\alpha),$$

$\Leftrightarrow$

$$P\left[z_1 < \frac{\frac{\sqrt{n-1}\left(\hat{\lambda}_j - \lambda_j\right)}{\lambda_j}}{\frac{\sqrt{2}\lambda_j}{\lambda_j}} < z_2\right] = (1-\alpha),$$

$\Leftrightarrow$

$$P\left[z_1\sqrt{2} < \sqrt{n-1}\frac{\hat{\lambda}_j}{\lambda_j} - \sqrt{n-1} < z_2\sqrt{2}\right] = (1-\alpha),$$

$\Leftrightarrow$

$$P\left[z_1\sqrt{2} + \sqrt{n-1} < \sqrt{n-1}\frac{\hat{\lambda}_j}{\lambda_j} < z_2\sqrt{2} + \sqrt{n-1}\right] = (1-\alpha),$$

$$\Leftrightarrow$$

$$P\left[\frac{z_1\sqrt{2}}{\sqrt{n-1}}+1<\frac{\hat{\lambda}_j}{\lambda_j}<\frac{z_2\sqrt{2}}{\sqrt{n-1}}+1\right]=(1-\alpha),$$

$$\Leftrightarrow$$

$$P\left[\frac{1}{1+\frac{z_2\sqrt{2}}{\sqrt{n-1}}}<\frac{\lambda_j}{\hat{\lambda}_j}<\frac{1}{1+\frac{z_1\sqrt{2}}{\sqrt{n-1}}}\right]=(1-\alpha),$$

$$\Leftrightarrow$$

$$P\left[\frac{\hat{\lambda}_j}{1+\frac{z_2\sqrt{2}}{\sqrt{n-1}}}<\lambda_j<\frac{\hat{\lambda}_j}{1+\frac{z_1\sqrt{2}}{\sqrt{n-1}}}\right]=(1-\alpha),$$

Por lo tanto, el intervalo de confianza para λ_j al $(1-\alpha)\times 100\%$ está definido por:

$$\left(\frac{\hat{\lambda}_j}{1+\frac{z_2\sqrt{2}}{\sqrt{n-1}}},\frac{\hat{\lambda}_j}{1+\frac{z_1\sqrt{2}}{\sqrt{n-1}}}\right). \tag{2.20}$$

La longitud de este intervalo se minimiza, tomando $z_1=-z_2$, donde z_1 es el cuantil $(1-\frac{\alpha}{2})$ de una población normal estándar. Finalmente el intervalo resultante está dado por:

$$\left(\frac{\hat{\lambda}_j}{1+\frac{z_2\sqrt{2}}{\sqrt{n-1}}},\frac{\hat{\lambda}_j}{1-\frac{z_2\sqrt{2}}{\sqrt{n-1}}}\right).$$

Para muestras grandes Anderson sugiere la siguiente relación:

$$\frac{t_1}{\sum_{j=1}^{p}\hat{\lambda}_j}\leq\frac{\lambda_j}{\sum_{j=1}^{p}\lambda_j}\leq\frac{t_2}{\sum_{j=1}^{p}\hat{\lambda}_j}.$$

(Ver Anderson, 1982 p.314).

2.4.3 Pruebas de Hipótesis sobre los Componentes Principales

Con frecuencia se debe contar con un procedimiento para decidir cuando k componentes principales incluyen la variación que se considera importante de la matriz de observaciones $\mathbf{X}$. Claramente uno esperaría ignorar $(p-k)$ componentes si sus correspondientes valores propios son iguales a cero, pero esto ocurre sólo si la matriz Σ asociada a la muestra es de rango $(p-k)$; i.e., $(p-k)$ valores propios son iguales a cero; caso que generalmente en la práctica no ocurre.

Una segunda alternativa sería hacer una prueba de hipótesis, en la que la proporción de variabilidad explicada por las k componentes sea menor que un cierto valor π. Otra hipótesis que puede probarse es cuando los últimos $(p-k)$ valores propios son iguales. Esto implica que la variación es igual en todas las direcciones del espacio generado por los últimos $(p-k)$ vectores propios, esta situación es denominada variación isotrópica e implica que si alguna componente es eliminada, entonces deben ser eliminadas todas las restantes. Estas pruebas pueden realizarse suponiendo normalidad en la muestra original.

Proporción de Variación Explicada por los Primeros k Componentes Principales

Sean $\lambda_1, ..., \lambda_p$ los valores propios de Σ y $\hat{\lambda}_1, ..., \hat{\lambda}_p$ los valores propios muestrales de S. El juego de hipótesis a probar es el siguiente:

$$H_0 \ : \ \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j} = \pi \text{ con } k < p \ .$$

$$vs$$

$$H_a \ : \ no\, H_0.$$

Sea $\hat{\pi}$ el estimador muestral de π, y por el teorema 2.8 se sabe que los elementos $\hat{\lambda}_j$ tienen una distribución normal asintótica, y $\hat{\pi}$ tiene una distribución normal (ver Mardia p.234) con media π y varianza

$$V\left(\hat{\pi}\right) = \frac{2tr\left(\Sigma^2\right)}{(n-1)\left(tr\left(\Sigma\right)\right)^2}\left(\pi^2 - 2c\pi + c\right), \tag{2.21}$$

donde el número c en (2.21) está definido como:

$$c = \frac{\sum_{j=1}^{k} \lambda_j^2}{\sum_{j=1}^{p} \lambda_j^2}. \tag{2.22}$$

La estimación de la $V\left(\hat{\pi}\right)$ puede hacerse utilizando la matriz S_X y los p valores propios de ésta, i.e.

$$\begin{aligned} \hat{\Sigma} &= S \\ tr\left(\hat{\Sigma}\right) &= \sum_{j=1}^{p} \hat{\lambda}_j \ . \end{aligned}$$

Como consecuencia

$$\hat{V}(\pi) = \frac{2tr(S^2)}{(n-1)(tr(S))^2}\left(\pi^2 - 2\hat{c}\,\pi + \hat{c}\right).$$

Entonces

$$\hat{\pi} \approx N\left(\pi, V\left(\hat{\pi}\right)\right). \tag{2.23}$$

Alternativamente puede utilizarse un intervalo estandarizando la expresión (2.23), mismo que tiene un nivel de confianza del $(1-\alpha)\times 100\%$ y que se obtiene mediante el siguiente desarrollo:

$$\frac{\left(\hat{\pi} - \pi\right)}{\sqrt{V\left(\hat{\pi}\right)}} \sim N(0,1),$$

calculando

$$P\left(z_1 < \frac{\left(\hat{\pi} - \pi\right)}{\sqrt{V\left(\hat{\pi}\right)}} < z_2\right) = (1-\alpha),$$

$\Leftrightarrow$

$$P\left(z_1\sqrt{V\left(\hat{\pi}\right)} < \left(\hat{\pi} - \pi\right) < z_2\sqrt{V\left(\hat{\pi}\right)}\right) = (1-\alpha),$$

$\Leftrightarrow$

$$P\left(z_1\sqrt{V\left(\hat{\pi}\right)} - \hat{\pi} < -\pi < z_2\sqrt{V\left(\hat{\pi}\right)} - \hat{\pi}\right) = (1-\alpha),$$

$\Leftrightarrow$

$$P\left(\hat{\pi} - z_2\sqrt{V\left(\hat{\pi}\right)} < \pi < \hat{\pi} - z_1\sqrt{V\left(\hat{\pi}\right)}\right) = (1-\alpha),$$

Si se define $z_1 = -z_2$ para minimizar la longitud del intervalo, esta definición conduce a la expresión final la cual está dada por:

$$\left(\hat{\pi} - z_2\sqrt{V\left(\hat{\pi}\right)}, \hat{\pi} + z_2\sqrt{V\left(\hat{\pi}\right)}\right). \tag{2.24}$$

En donde z_2 es el cuantil $\left(1 - \frac{\alpha}{2}\right)$ de una normal estándar.

Prueba de Esfericidad

Esta prueba es usada para determinar el número de componentes principales que serán utilizados para describir el comportamiento de los datos. En ésta se desea probar que los

últimos $(p-k)$ valores propios son iguales, i.e., las últimas $(p-k)$ componentes principales tienen la misma varianza; esto significa que si se incluye una de ellas deben incluirse todas las demás.

El juego de hipótesis a probar está dado por:

$$\begin{aligned} H_0 &: \lambda_{k+1} = \cdots = \lambda_p. \\ &\quad vs \\ H_a &: no\, H_0. \end{aligned}$$

La estadística de prueba se obtiene por el Método de la razón de verosimilitud, el cual tiene asociada la siguiente expresión:

$$-2\ln\Lambda = np\left[a - \ln(g) - 1\right],$$

donde la constante a se define como en (2.22). De acuerdo al Teorema $A.5$, a y g corresponden a la media aritmética y geométrica de los valores propios de $\hat{\Sigma}^{-1}S$, donde $\hat{\Sigma}$ es el estimador máximo verosímil de Σ bajo la hipótesis nula. Sea

$$a_0 = \frac{\sum_{j=k+1}^{p} \hat{\lambda}_j}{(p-k)},$$

que denota la media aritmética de los valores propios $\hat{\lambda}_{k+1}, ..., \hat{\lambda}_p$ asociados a $\hat{\Sigma}$ y

$$g_0 = \left(\prod_{j=k+1}^{p} \hat{\lambda}_j\right)^{\frac{1}{(p-k)}},$$

la media geométrica. Por lo tanto, la estadística de prueba es

$$-2\ln\Lambda = np\left[a_0 - \ln(g_0) - 1\right]. \tag{2.25}$$

En donde n es el tamaño de muestra y p es la dimensión, así $-2\ln\Lambda$ en la ecuación (2.25) se distribuye como una χ^2_r, donde r es el número de grados de libertad . Siguiendo la aproximación de Bartlett, citada por Mardia (1982, p.236) la ecuación (2.25) puede escribirse como:

$$\left(n - \frac{2p+11}{6}\right)(p-k)\ln\left(\frac{a_0}{g_0}\right) \sim \chi^2_r, \tag{2.26}$$

donde $r = \frac{1}{2}(p-k+2)(p-k-1)$.

2.4.4 Reglas de Corte

Además de las pruebas de hipótesis mencionadas anteriormente, existen "reglas" que a pesar de ser subjetivas pueden ser de gran ayuda para determinar el número de componentes principales que deben ser retenidos. Dichas "reglas de corte" son las siguientes:

1. Una forma práctica de observar empíricamente la contribución de varios componentes principales (Cattell, 1966) es observar la gráfica conocida en la literatura como "Scree plot", la cual consiste en graficar el valor propio λ_j contra j. Dicho diagrama puede indicar claramente dónde terminan los valores propios grandes y en que punto empiezan los valores pequeños.

2. Incluir los componentes principales que en conjunto acumulen un 90% de la variación total.

3. (Kaisser) Excluir aquellos componentes cuyos valores propios sean menores que la media, i.e., menores que la unidad si es que se ha utilizado la matriz de correlación.

2.5 Ejemplos

Con el objetivo de ilustrar la Técnica de Componentes Principales, en esta Sección se analizarán 2 Ejemplos: El primero de ellos con datos reales basado en los Irises de Fisher, y el segundo con datos simulados mediante el paquete estadístico S-PLUS. En ambos casos se utiliza el paquete STATISTICA versión 4.5 para la obtención de los resultados.

Ejemplo 2.2 *La muestra observada consta de tres tipos de flor de Iris, las cuales son: Iris Setosa, Iris Virgínica e Iris Versicolor. Para cada uno de estos tipos de flor se hicieron 50 observaciones y para cada observación se registró 4 características que son:*

Largo del Sépalo (Sepallen,SL)
Ancho del Sépalo (Sepalwid,SW)
Largo del Pétalo (Petallen,PL)
Ancho del Pétalo (Petalwid,PW)

y a cada una de ellas se les asignó una variable:

$$\begin{aligned} SL &\rightarrow X_1 \\ SW &\rightarrow X_2 \\ PL &\rightarrow X_3 \\ PW &\rightarrow X_4 \end{aligned}$$

Finalmente, la muestra se concentra en una matriz X de dimensión 150×4, la cual se anexa en el Apéndice B.

La matriz de correlación a la que se le denotará como R está dada en la siguiente Tabla

Correlations (irisdat.sta)

Continue...

Casewise deletion of MD N=150

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.00	-.12	.87	.82
SEPALWID	-.12	1.00	-.43	-.37
PETALLEN	.87	-.43	1.00	.96
PETALWID	.82	-.37	.96	1.00

Tabla 2. Matriz de correlación de los Irises.

Obsérvese de esta matriz que las variables que presentan una alta correlación, son las parejas (X_1, X_3), (X_1, X_4) y (X_3, X_4). Esto significa que las flores con las proporciones más altas (más pequeñas) en la longitud de los pétalos también manifiestan las mayores (menores) dimensiones en las variables ancho de pétalo y largo de sépalo.

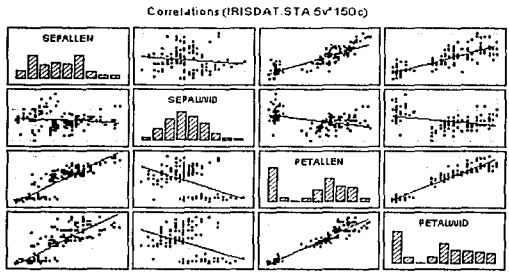


Figura 2. Gráfica de las 4 variables de los Irises de Fisher.

Del Apéndice B se tiene que los vectores y valores propios asociados a la matriz de correlación muestral están dados en la siguiente Tabla:

U_1	U_2	U_3	U_4	λ
0.5211	−0.3774	−.7196	0.2613	2.9184
−.2693	−0.9233	0.2444	−.1235	0.9140
0.5804	−0.0245	0.1421	−.8014	0.1467
0.5648	−0.0669	0.6343	0.5236	0.0207

Tabla 3. Vectores y valores propios para la muestra de Iris

donde los componentes principales con base en estos vectores están dados por:

$$\begin{aligned} Y_1 &= 0.5211X_1 - 0.2693X_2 + 0.5804X_3 + 0.5648X_4, \\ Y_2 &= -.3774X_1 - 0.9233X_2 - 0.0245X_3 - 0.0669X_4, \\ Y_3 &= -.7196X_1 + 0.2444X_2 + 0.1421X_3 + 0.6343X_4, \\ Y_4 &= 0.2613X_1 - 0.1235X_2 - 0.8014X_3 + 0.5236X_4. \end{aligned}$$

La asimilación de la varianza para cada componente principal , así como la varianza acumulada queda comprendida en la Tabla 4.

Eigenvalues (irisdat.sta)

Continue

Extraction: Principal components

Value	Eigenval	%total Variance	Cumul Eigenval	Cumul %
1	2.918498	72.96245	2.918498	72.9624
2	.914030	22.85076	3.832528	95.8132
3	.146757	3.66092	3.979285	99.4821
4	.020715	.51787	4.000000	100.0000

Tabla 4. Acumulación de la varianza de cada componente principal.

De la Tabla anterior se puede observar que es suficiente considerar sólo 2 componentes principales para explicar alrededor del 96% de la variación total de los datos.

La gráfica de los 2 primeros componentes principales se muestra en la siguiente Figura

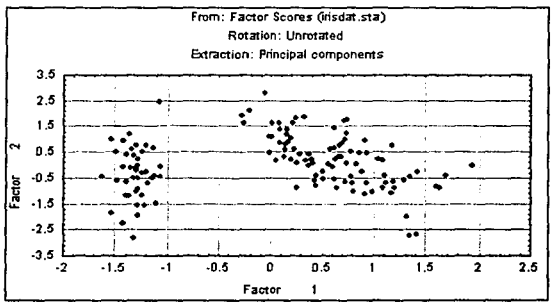


Figura 3. Gráfica de la componente 2 vs 1 para los Irises.

En ésta, se aprecia que alrededor del primer componente principal (Factor 1) las observaciones tienen una mayor variabilidad y disminuye para la componente principal 2 (Factor 2).

En la gráfica "Scree plot" , se aprecia que la pendiente que definen los segmentos de recta se empieza a estabilizar a partir del valor propio λ_2, lo cual nos hace pensar que es suficiente retener únicamente a los 2 primeros componentes principales.

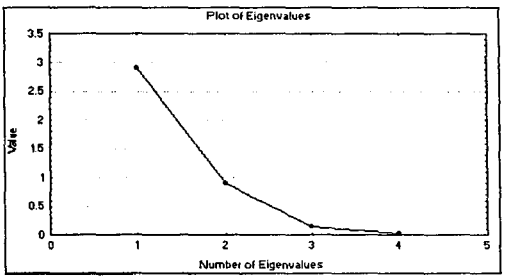


Figura 4. Gráfica de los valores propios para los Irises de Fisher.

Este hecho puede corroborarse si se analiza la estructura de correlación (valores de r^2) que existe entre cada uno de los componentes con cada una de las variables X_j. Para ello se observó la proporción de variabilidad que es acumulada para cada X_j con respecto a un subconjunto de componentes Y_j. Dichos valores están concentrados en la Tabla 5 que a continuación se presenta:

Communalities (irisdat sta)

Continue...

Extraction: Principal components
Rotation: Unrotated

Variable		From 2 Factors	From 3 Factors	From 4 Factors	Multiple R-Square
	.792400	.922599	.998586	1.000000	.858612
SEPALWID	.211731	.990919	.999684	1.000000	.524007
PETALLEN	.983182	.983730	.986694	1.000000	.968012
PETALWID	.931184	.935280	.994321	1.000000	.937850

Tabla 5. Estructura de correlación para los Irises.

Así se confirma que es suficiente tomar 2 componentes principales, ya que en conjunto acumulan una proporción de variación aproximadamente mayor al 90% sobre cada una de las variables X_i.

Hasta aquí se tiene un análisis descriptivo, adicionalmente se puede construir un intervalo de confianza para la proporción de varianza acumulada por los 2 primeros componentes principales y que indicarían entre qué valores fluctúa esta proporción. Este intervalo se construye utilizando el desarrollo presentado en la Sección 2.4.3 de este Capítulo.

El intervalo del 95% de confianza para la proporción π acumulada por las dos primeras componentes, está dado por la siguiente expresión:

$$\left(\hat{\pi} - t^{0.975}\sqrt{V\left(\hat{\pi}\right)}, \hat{\pi} + t^{0.975}\sqrt{V\left(\hat{\pi}\right)}\right),$$

en donde t es el cuantil de orden 0.975 de una normal estándar y

$$V\left(\hat{\pi}\right) = \frac{2tr\left(S^2\right)}{(n-1)\left(tr\left(S\right)\right)^2}\left(\pi^2 - 2c\pi + c\right),$$

donde la constante c se define como en (2.22).

En este caso

$$\begin{aligned}
\hat{\pi} &= 95.8132, \\
c &= 0.9976, \\
n &= 150, \\
t^{(0.975)} &= 1.9599, \\
tr\left(S^2\right) &= 9.3750, \\
\left(trS\right)^2 &= 16, \\
V\left(\hat{\pi}\right) &= 70.7060,
\end{aligned}$$

Finalmente el intervalo de confianza para $\pi = 95.8132$ queda definido como:

$$(79.3324, 100).$$

En otras palabras, aunque el punto estimado para los 2 primeros componentes principales muestra que la variación explicada es aproximadamente del 96%, el intervalo de confianza indica que el verdadero valor de esta estimación fluctúa entre el 79% y el 100%. Aunque el límite derecho resulta ser 112.2939 este valor se trunca en 100 ya que es la proporción máxima de varianza que puede ser explicada.

De los resultados anteriores se concluye que es suficiente considerar la aportación de los 2 primeros componentes principales.

Cabe mencionar que estos cálculos fueron hechos tomando la cola completa de decimales sin hacer ningún tipo de redondeo, pero para efectos prácticos sólo se han considerado los cuatro primeros decimales, estos cálculos se encuentran recopilados en el Apéndice B.

Ejemplo 2.3 *La simulación de los datos consiste de una muestra a la que se denotará por* $\tilde{X} = (X_{(1)}, X_{(2)})$ *de dimensión* (100×2), *que fue generada de tal modo que su distribución sea normal bivariada de parámetros*

$$\mu = \begin{pmatrix} 5 \\ 3 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 10 & 2 \\ 2 & 2 \end{pmatrix}.$$

A esta muestra se le agregó una tercera componente X_3 *de la forma*

$$X_3 = 3X_1 - 5X_2 + \varepsilon,$$

donde el error fue generado en forma independiente como una normal estándar de dimensión (100×1). *Denótese por* $X = \left(\tilde{X}, X_{(3)}\right)$ *la matriz de la muestra total, cuya dimensión asociada es* (100×3).

El objetivo de este análisis, es encontrar una transformación tal que se pueda reducir la dimensión en la que la nube de datos de la matriz X *se encuentra, con base en los criterios citados en la Sección 2.4.*

La matriz de correlación de X, *dada en la Tabla 6 que a continuación se presenta:*

Correlations (cpej2.sta)

Continue	Casewise deletion of MD N=100		
Variable	X1	X2	X3
X1	1.00	.45	.75
X2	.45	1.00	-.24
X3	.75	-.24	1.00

Tabla 6. Matriz de correlación para la muestra de la tabla B2.

Obsérvese que la correlación más alta se encuentra en la pareja (X_1, X_3) *con* 0.75, *lo cual indica que existe una relación lineal entre estas dos variables: este hecho no debe parecer extraño ya que la variable* X_3 *es una combinación lineal de las componentes de* $\widetilde{X}$. *Gráficamente la relación que guardan estas tres variables por parejas puede apreciarse en la Figura 5*

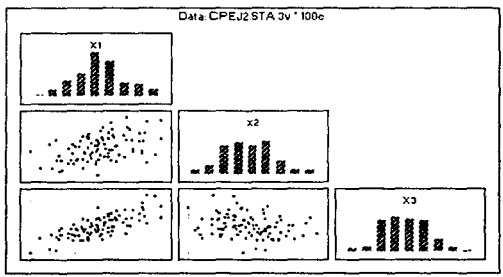


Figura 5. Matriz de Correlación para las variables X *de la muestra B2.*

La dispersión de los datos se puede apreciar conjuntamente en una gráfica en tres dimensiones que está dada en la Figura 6

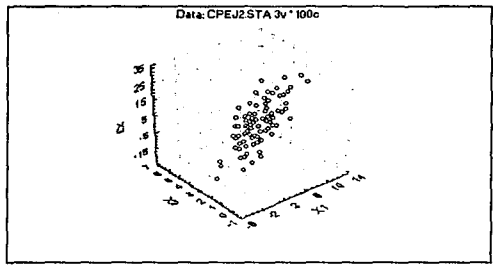


Figura 6. Gráfica en tres dimensiónes de las variables X de la muestra B2.

En esta gráfica se aprecia la elongación de la nube de datos y se puede observar que muy pocas observaciones están dispersas. Lo anterior puede ser visto en una gráfica de dos dimensiones

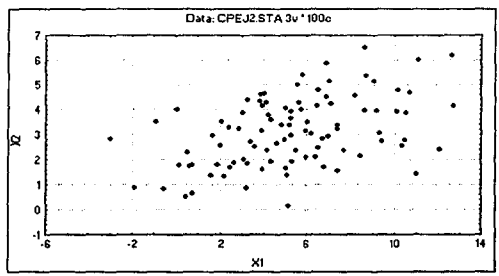


Figura 7. Gráfica de las variables X_2 vs X_1 para la muestra B2.

La siguiente Tabla contiene a los vectores propios asociados a la matriz de correlación muestral

U_1	U_2	U_3	λ_j
-0.7374	*0.1410*	*-.6605*	*1.7909*
-0.2313	*0.8660*	*0.4432*	*1.2042*
-0.6346	*-.4796*	*0.6060*	*0.0048*

Tabla 7. Vectores y valores propios para la muestra B2.

Con base en la Tabla anterior se obtienen las combinaciones lineales que definen a los tres componentes principales, los cuales están dados por

$$\begin{aligned} Y_1 &= -.7374X_1 - 0.2313X_2 - 0.6346X_3, \\ Y_2 &= 0.1410X_1 + 0.8660X_2 - 0.4796X_3, \\ Y_3 &= -.6605X_1 + 0.4432X_2 + 0.6060X_3, \end{aligned}$$

Los valores propios de la matriz de correlación, así como la varianza asociada a cada uno de los componentes principales Y_j con $j = 1, 2, 3$, pueden observarse en la Tabla 8.

Eigenvalues (cpe)2.sta

Continue...

Extraction: Principal components

Value	Eigenval	Total Variance	Cumul Eigenval	Cumul %
1	1.790947	59.69825	1.790947	59.6982
2	1.204206	40.14021	2.995154	99.8385
3	.004846	.16155	3.000000	100.0000

Tabla 8. Varianza por cada componente para la simulación B2.

En esta Tabla se aprecia que es suficiente retener los dos primeros componentes principales, ya que en conjunto acumulan el 99.8% de la varianza total, siendo posible entonces eliminar la componente Y_3, argumentando además que X_3 es una variable que no aporta información nueva y que el último valor propio es aproximadamente nulo. En otras palabras, la presencia de la combinación lineal X_3 provoca que la varianza asimilada por la

componente Y_3 sea pobre. Esto también puede observarse en la gráfica denominada "Scree plot" que a continuación se presenta:

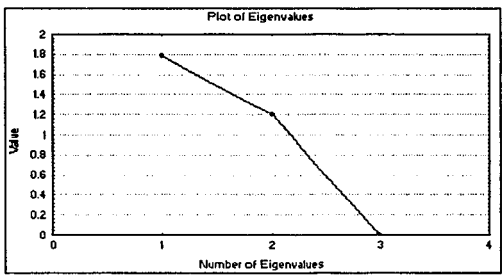


Figura 8: Gráfica de los valores propios de la matriz de correlación para la muestra de la Tabla B2.

Se observa de la gráfica que la caída más pronunciada se da en el segmento de recta formado por el segundo y tercer valor propio; lo cual nos induce a pensar que se puede despreciar la tercera componente principal.

La estructura de correlación existente entre alguna variable X_j (con $j = 1, 2, 3$) con algún subconjunto de componentes principales, está concentrada en la Tabla 9

Communalities (cpej?.sta)

Continue...

Extraction: Principal components
Rotation: Unrotated

Variable	From 1 Factor	From 2 Factors	From 3 Factors	Multiple R-Square
X1	.973920	.997886	1.000000	.988931
X2	.095825	.999048	1.000000	.975720
X3	.721203	.998220	1.000000	.986876

Tabla 9. Estrusctura de correlación para la muestra B2.

Nótese de la Tabla anterior que la proporción de variabilidad explicada para cada una de las X_j por las dos primeras componentes Y_1, Y_2 es aproximadamente mayor al 99%, eso significa que la información que guarda cada componente principal sobre cada variable X es considerable.

Es suficiente retener sólo dos componentes y con ellos conservar el 99.8% de la variabilidad total en X. Esta reducción nos lleva a modelar la matriz X en sólo dos dimensiones, lo cual puede visualizarse de la siguiente manera con base en los primeros dos componentes

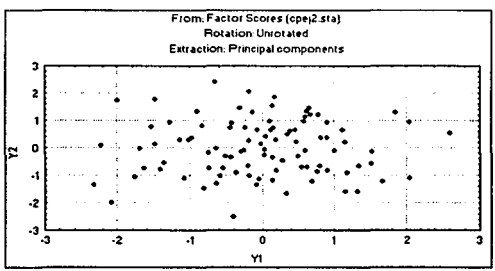


Figura 9. Gráfica de los componentes Y_2 vs Y_1 para la muestra B2.

Es bueno hacer notar que esta gráfica se parece mucho a la presentada en la Figura 7, dado que la variable X_3 ya no proporciona información relevante en la muestra y puede ser eliminada sin mayor problema.

Finalmente, se construye un intervalo del 95% confianza para λ_3, siguiendo el desarrollo citado con anterioridad en la sección 2.4.2. Este intervalo puede ser útil para conocer en que rango está el verdadero valor de λ_3 y así poder decidir si la varianza de Y_3 puede despreciarse en el análisis. Este intervalo se define por:

$$\left(\frac{\hat{\lambda_j}}{1 + \frac{q\sqrt{2}}{\sqrt{n-1}}}, \frac{\hat{\lambda_j}}{1 - \frac{q\sqrt{2}}{\sqrt{n-1}}} \right),$$

en donde q es el cuantil de orden 0.95 de una normal estándar con base en la muestra se obtiene

$$\hat{\lambda}_j = 0.0048,$$
$$q = 1.6448,$$
$$\sqrt{99} = 9.9498.$$

Por lo tanto, el intervalo resultante para λ_3 está dado como:

$$(0.0039, 0.0063)\,.$$

Concluyendo entonces que dicho valor propio aunque es diferente de cero, es muy pequeño para considerarlo en el análisis y por lo tanto la varianza de Y_3 puede despreciarse.

Capítulo 3

Análisis Discriminante

El Análisis Discriminante, es una técnica que tiene por objetivo clasificar individuos en uno y sólo uno de los g grupos o poblaciones que se tiene como alternativas. La asignación se basa en los supuestos distribucionales que se hacen sobre la muestra observada como: Discriminación Normal, la cual supone que las poblaciones tiene distribución normal multivariada; Discriminante Logístico, en donde se supone que la forma específica de las densidades es desconocida pero el logaritmo del cociente de las densidades es lineal en los parámetros asociados, Discriminante no Paramétrico, el cual supone que la distribución de la muestra no es conocida y que debe ser estimada.

Es conveniente hacer notar que en este trabajo se analiza por separado dos técnicas del análisis discriminante que son: Discriminación Normal y Discriminación no Paramétrica. Para ambos casos se debe considerar g poblaciones o grupos que se denotarán por Π_i con $i = 1, 2, ..., g$. Supóngase que asociada a cada población Π_i existe una función de densidad $f_i(x)$ en $\Re^p$ tal que el vector de observaciones X de un individuo proveniente de la población Π_i tiene $f.d.p.$ $f_i(x)$. Entonces la finalidad del análisis discriminante es asignar un individuo con vector de atributos X a uno y sólo uno de esos g grupos con base en sus p características, por lo que de ahora en adelante se hará referencia indistintamente al vector de características X que al individuo, ya que sus atributos están contenidos en este vector.

Claramente, es deseable considerar reglas de asignación que cometan los menos errores,

en el sentido de ser más precisos en la clasificación. Desde el punto de vista matemático una regla de asignación puede establecerse como sigue

Definición 3.1 *Una regla discriminante corresponde a una división del espacio* $\Re^p$ *en regiones disjuntas (o mutuamente excluyentes),* $\mathbf{R}_1, ..., \mathbf{R}_g$ $(\cup_{i=1}^{g} \mathbf{R}_i \equiv \Re^p)$, *donde la regla está dada como:*

$$asignar\ X\ a\ \Pi_i,\ \ si\ X \in \mathbf{R}_i\ \ con\ i = 1, ..., g.$$

La discriminación es más exacta en la medida en que Π_i tenga la mayor parte de su probabilidad concentrada en $\mathbf{R}_i$ para cada i.

Aunque en la práctica es difícil encontrar casos en los que la $f.d.p$ $f_i(x)$ es totalmente conocida, una variante de uso común de esta situación ocurre cuando la forma funcional de la $f.d.p$ para cada población es conocida, pero existen parámetros que deben ser estimados. La estimación está basada en una matriz muestral $\mathbf{X}_{n \times p}$ cuyos renglones están particionados en g grupos.

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_g \end{pmatrix}, \tag{3.27}$$

donde la matriz $\mathbf{X}_i$ de $n_i \times p$ representa una muestra de n_i individuos de la población Π_i, definida por:

$$\mathbf{X}_i = \begin{pmatrix} X_{i1}^t \\ \vdots \\ X_{in_i}^t \end{pmatrix},$$

con $i = 1, ..., g$. Nótese que los individuos están asignados en los renglones de la matriz $\mathbf{X}$, los cuales a su vez están agrupados en categorías.

3.1 Discriminación Normal

La discriminación normal, como se mencionó anteriormente, se basa en el supuesto de normalidad de la matriz $\mathbf{X}$, es decir, se supone que cada una de las g poblaciones tienen

asociada una densidad normal multivariada. Si además se sabe que los parámetros de la $f.d.p.$ $f_i(x)$ son conocidos para cada i, entonces la discriminación se centra en encontrar una regla que maximice la función de verosimilitud. Cuando la $f.d.p.$ es conocida pero sus parámetros deben ser estimados, estos se reemplazan por los estimadores máximo verosímil y la técnica de nuevo maximiza la función de verosimilitud.

Bajo el supuesto de normalidad de las observaciones existen dos técnicas generales conocidas como Discriminación Lineal y Discriminación Cuadrática. Dichas técnicas se discuten a continuación por separado para el caso de densidades totalmente conocidas y el caso en que existen parámetros que deben ser estimados.

3.1.1 Discriminación Cuando las Poblaciones son Conocidas

Regla de Asignación por Máxima Verosimilitud

Supóngase que cada una de las poblaciones Π_i se distribuyen como normales multivariadas, es decir, si X proviene de Π_i, entonces $X \sim N(\mu_i, \Sigma_i)$. Sea la $f.d.p.$ del $i-ésimo$ grupo $f_i(x)$ y $L_i(x) = f_i(x)$ la verosimilitud del vector de observaciones X. La siguiente definición establece la forma de construir las regiones de clasificación bajo el método de máxima verosimilitud.

Definición 3.2 *La regla discriminante de máxima verosimilitud (r.d.m.v) que asigna una observación X, en una de las poblaciones $\Pi_1, ..., \Pi_g$, consiste en asignar X a la población que tiene la verosimilitud más grande.*

Es decir, la $r.d.m.v$ dice que se debe asignar X a Π_i si:

$$L_i(x) = \max_k L_k(x).$$

En otras palabras se debe encontrar la población Π_i que maximice la verosimilitud del vector X, por lo que las regiones de clasificación pueden escribirse en este caso como

$$\mathbf{R}_i = \{X \in \Re^p \mid L_i(x) \geq L_k(x) \quad k = 1, ..., g.\},$$

para $i = 1, ..., g$.

Discriminación Lineal. En este apartado se considera que las matrices de covarianzas Σ es común para las g poblaciones, es decir, cada muestra $\mathbf{X}_i$ se distribuye normal de parámetros μ_i y Σ para todo índice $i = 1, ..., g$.

Teorema 3.1 *En el caso en el que Π_i tiene asociada una densidad $N(\mu_i, \Sigma)$, de parámetros conocidos la regla de asignación máximo verosímil asigna X a Π_i si*

$$PL_i(x) = a_i^t \left(X - \frac{1}{2}\mu_i \right) \quad con\ i = 1, ..., g, \tag{3.28}$$

donde $PL_i(X) = \max_j PL_j(X)$. PL denotará al puntaje lineal.

Demostración.

Maximizar la función de verosimilitud $L_i(x)$, es igual a maximizar la expresión

$$|2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(X - \mu_i)^t \Sigma^{-1}(X - \mu_i) \right\},$$

y es equivalente a maximizar la expresión

$$-\frac{1}{2}(X - \mu_i)^t \Sigma^{-1}(X - \mu_i), \tag{3.29}$$

Desarrollando la forma cuadrática (3.29), se obtiene la equivalencia a maximizar

$$\mu_i^t \Sigma^{-1} X - \frac{1}{2}\mu_i^t \Sigma^{-1} \mu_i,$$

donde $-\frac{1}{2}X^t\Sigma^{-1}X$ no depende de i. Si se define $a_i = \Sigma^{-1}\mu_i$, la ecuación anterior se puede escribir como:

$$a_i^t \left(X - \frac{1}{2}\mu_i \right).$$

Entonces la regla de asignación máximo verosímil asigna X a la población Π_i si

$$PL_i(x) = \max_j \left\{ a_j^t \left(X - \frac{1}{2}\mu_i \right) \right\}. \square$$

De esta manera las regiones $\mathbf{R}_i$ mutuamente excluyentes (o disjuntas con probabilidad 1), en las que se divide el espacio $\Re^p$ están definidas por:

$$\text{asignar } X \text{ a } \Pi_i \text{ si } X \in \mathbf{R}_i,$$

donde

$$\mathbf{R}_i = \{X \in \Re^p \mid PL_i(x) \geq PL_k(x) \quad \forall\, k\},$$

$i = 1, ..., g$ y $PL_i(x)$ se define como en (3.28).

Si varias verosimilitudes tomarán el mismo valor máximo, entonces cualquiera de esas verosimilitudes puede ser tomada. Este caso no es importante ya que desde el punto de vista práctico la probabilidad de que dos verosimilitudes tomen el mismo valor máximo es cero.

Teorema 3.2 **(1)** *Si X proviene de una población Π_i con densidad asociada $N_p(\mu_i, \Sigma)$, $i = 1, ..., g$, y $\Sigma > 0$, donde los parámetros asociados son conocidos entonces la r.d.m.v. que asigna X a Π_k, donde $k \in \{1, ..., g\}$, es aquel valor de i que minimiza la distancia de Mahalanobis*

$$(X - \mu_i)^t \Sigma^{-1} (X - \mu_i).$$

(2) *Para $g = 2$ grupos, la regla*

$$asigna\ X\ a = \begin{cases} \Pi_1 \text{ si } \alpha^t (X - \mu) > 0, \\ \Pi_2 \text{ en otro caso,} \end{cases}$$

donde $\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$ y $\mu = \frac{1}{2}(\mu_1 + \mu_2)$.

Demostración.

(1) *La $i-$ésima verosimilitud está dada por:*

$$L_i(x) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(X - \mu_i)^t \Sigma^{-1} (X - \mu_i)\right\}.$$

Maximizar esta función es equivalente a minimizar la forma cuadrática del exponente

$$(X - \mu_i)^t \Sigma^{-1} (X - \mu_i).$$

lo cual demuestra la primera parte de este Teorema. □

(2) *Para asignar X a Π_1 debe cumplirse que*

$$L_1(x) > L_2(x).$$

Esta desigualdad se satisface por la parte 3.2 si y sólo si

$$(X-\mu_1)^t \Sigma^{-1} (X-\mu_1) < (X-\mu_2)^t \Sigma^{-1} (X-\mu_2).$$

Desarrollando esta última desigualdad se llega fácilmente a la siguiente expresión:

$$(\mu_1-\mu_2)^t \Sigma^{-1} \left(X - \frac{1}{2}(\mu_1+\mu_2) \right) > 0, \tag{3.30}$$

haciendo $\alpha = \Sigma^{-1}(\mu_1-\mu_2)$ y $\mu = \frac{1}{2}(\mu_1+\mu_2)$, la ecuación (3.30) puede escribirse como:

$$\alpha^t (X-\mu) > 0,$$

lo que demuestra el resultado. □

Discriminación Cuadrática En esta Sección se considera el caso en el que g poblaciones Π_i tienen asociadas una densidad normal de parámetros μ_i y Σ_i $i = 1, ..., g$. La *r.d.m.v.* queda establecida en el siguiente Teorema.

Teorema 3.3 *La r.d.m.v. asigna X a Π_i si $L_i(x) \geq L_k(x)$ $\forall k \neq i$, lo que ocurre si y sólo si*

$$\begin{aligned} \mathbf{R}_i &= \{X \in \Re^p : (X-\mu_i)^t \Sigma_i^{-1} (X-\mu_i) \leq (X-\mu_k)^t \Sigma_k^{-1} (X-\mu_k) \\ k &= 1, ..., g\} \text{ con } i = 1, ..., g . \end{aligned} \tag{3.31}$$

Demostración.

La desigualdad de las verosimilitudes $L_i(x)$ y $L_k(x)$ se puede escribir como:

$$|2\pi\Sigma_i|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(X-\mu_i)^t \Sigma_i^{-1} (X-\mu_i) \right\} \geq |2\pi\Sigma_k|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(X-\mu_k)^t \Sigma_k^{-1} (X-\mu_k) \right\},$$

lo cual ocurre si y sólo si

$$\exp\left\{-\frac{1}{2}(X-\mu_i)^t\,\Sigma_i^{-1}(X-\mu_i)\right\} \geq \exp\left\{-\frac{1}{2}(X-\mu_k)^t\,\Sigma_k^{-1}(X-\mu_k)\right\},$$

si y sólo si

$$(X-\mu_i)^t\,\Sigma_i^{-1}(X-\mu_i) \leq (X-\mu_k)^t\,\Sigma_k^{-1}(X-\mu_k)\,. \tag{3.32}$$

Por lo que las regiones de clasificación se defínen como en (3.31) y se demuestra el resultado. □

3.1.2 Discriminación Bajo Estimación

Regla de Asignación Muestral por Máxima Verosimilitud

La regla discriminante muestral de máxima verosimilitud (*r.d.m.m.v.*), es utilizada cuando la forma de las distribuciones de los grupos $\Pi_1,...,\Pi_g$ son conocidas, pero sus parámetros deben ser estimados con base en la matriz de datos $\mathbf{X}_{n\times p}$. Supóngase que los renglones de $\mathbf{X}$ están particionados como en (3.27) y que la densidad asociada a Π_i es $f_i(x \mid \theta_i)$ donde θ_i es un vector de parámetros desconocidos. Entonces la *r.d.m.m.v* queda establecida como

Definición 3.3 *La regla discriminante muestra de máxima verosimilitud (r.d.m.m.v) que asigna una observación X, en una de las poblaciones $\Pi_1,...,\Pi_g$, consiste en asignar X a la población que tiene la verosimilitud muestral más grande.*

La *r.d.m.m.v* dice que se debe asignar X a Π_i si:

$$\hat{L}_i(x) = \max_k \hat{L}_k(x)\,.$$

Entonces las regiones de clasificación pueden escribirse en este caso como

$$\mathbf{R}_i = \{X \in \Re^p \mid \hat{L}_i(x) \geq \hat{L}_k(x) \quad k = 1,...,g.\},$$

para $i = 1,...,g$.

En esta sección se discutirán por separado dos tipos de discriminación: Discriminante Lineal, el cual considera que los g grupos tienen medias distintas pero que comparten la misma matriz de covarianzas; y Discriminante Cuadrático, en donde se considera que a cada grupo le corresponde una media μ_i y una matriz de covarianzas Σ_i.

Considérese en primer lugar que los g grupos son muestras tomadas de una población normal multivariada con vector de medias distinto y la misma matriz de covarianzas. Sean μ_i y Σ la media y la matriz de covarianzas asociadas al $i-ésimo$ grupo y sean los estimadores máximo verosímiles de estos parámetros dados como (en el Teorema 1.23)

$$\begin{aligned} \hat{\mu}_i &= \bar{X}_i, \ i = 1, \ldots, g. \\ &= \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \ i = 1, \ldots, g. \end{aligned} \tag{3.33}$$

y

$$W = \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_i \right) \left(X_{ij} - \bar{X}_i \right)^t .$$

En el caso en que los g grupos tienen asociadas poblaciones normales cada una con distinto vector de medias y distinta matriz de covarianzas, y sí para la $i-ésima$ población estos parámetros están dados por μ_i y Σ_i, entonces los estimadores máximo verosímiles están definidos como:

$$\hat{\mu}_i = \bar{X}_i \ \ i = 1, \ldots, g,$$

y

$$\begin{aligned} \hat{\Sigma}_i &= S_i, \\ &= \frac{1}{n_i} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_i \right) \left(X_{ij} - \bar{X}_i \right)^t . \end{aligned} \tag{3.34}$$

La relación que existe entre las matrices W y S_i está dada por la siguiente igualdad:

$$W = \frac{1}{n} \sum_{i=1}^{g} n_i S_i .$$

Discriminación Lineal

En este caso se considera que la matriz de covarianzas Σ es común para las g poblaciones aunque desconocida y que ha sido estimada mediante la matriz W. Los siguientes Teoremas establecen formalmente las expresiones que definen la $r.d.m.m.v.$ y que se obtienen

reemplazando los parámetros de la $f.d.p$ por los estimadores máximo verosímiles.

Teorema 3.4 *En el caso en el que Π_i tiene asociada una densidad $N_p(\mu_i, \Sigma)$, $i = 1, ..., g$ de parámetros desconocidos, la regla muestral por máxima verosimilitud asigna X a Π_i si $PL_i(x) = \max PL_j(x)$, donde $PL_i = \left\{ a^t \left(X - \frac{1}{2} \bar{X}_i \right) \right\}$ $i = 1, ..., g$, $a = W^{-1} \bar{X}_i$ y $\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$.*

Demostración

El resultado se sigue del Teorema 3.1. □

Esto es las regiones de clasificación con base en el Teorema anterior están dadas por

$$\mathbf{R}_i = \{ X \in \Re^p : PL_i(x) \geq PL_k(x) \quad \forall\, i \neq k \}.$$

Teorema 3.5 **(1)** *Si $\mathbf{X}_i$ proviene de una población $N_p(\mu_i, \Sigma)$, $i = 1, ..., g$, entonces la r.d.m.m.v. que asigna X a Π_k, donde $k \in \{1, ..., g\}$, es aquel valor de i que minimiza la distancia de Mahalanobis estimada*

$$\left(X - \bar{X}_i \right)^t W^{-1} \left(X - \bar{X}_i \right).$$

(2) *Para $g = 2$ grupos, la r.d.m.m.v.*

$$\text{asigna } X \text{ a} = \begin{cases} \Pi_1 \text{ si y sólo si } a^t \left(X - \bar{X} \right) > 0, \\ \Pi_2 \text{ en otro caso,} \end{cases} \tag{3.35}$$

donde $a = W^{-1} \left(\bar{X}_1 - \bar{X}_2 \right)$ y $\bar{X} = \frac{1}{2} \left(\bar{X}_1 + \bar{X}_2 \right)$.

Demostración. Se sigue de aplicar el Teorema 3.2, sustituyendo los parámetros por los estimadores máximo verosímiles. □

Por lo tanto las regiones de clasificación pueden ser escritas como:

$$\begin{aligned} \mathbf{R}_i &= \left\{ X \in \Re^p \mid a^t \left(X - \frac{1}{2} \bar{X} \right) > 0,\, i \neq k \right\}, \\ &= \left\{ X \in \Re^p \mid a^t \left(X - \frac{1}{2} \left(\bar{X}_i + \bar{X}_k \right) \right) > 0,\, i \neq k \right\}. \end{aligned}$$

Discriminación Cuadrática

Ahora se considera el caso en que medias y las matrices de covarianza son distintas para cada grupo, y que los estimadores máximo verosímiles están dados como en (3.33) y (3.34) Estos estimadores pueden sustituirse simplemente en el Teorema 3.3 para obtener el siguiente resultado

Teorema 3.6 *En el caso en el que* Π_i *tiene densidad asociada* $N_p(\mu_i, \Sigma)$ $i = 1, ..., g$ *de parámetros desconocidos, la r.d.m.m.v asigna* X *a* Π_i *si* $L_i(x) \geq L_k(x)$ $\forall$ $k \neq i$, *lo que ocurre si y sólo si*

$$\begin{aligned} \mathbf{R}_i &= \{X \in \Re^p \mid \left(X - \bar{X}_i\right)^t S_i^{-1} \left(X - \bar{X}_i\right) < \left(X - \bar{X}_k\right)^t S_k^{-1} \left(X - \bar{X}_k\right), \\ k &= 1, ..., g\}, \text{ para } i = 1, ..., g. \end{aligned}$$

Demostración. Se sigue del Teorema 3.3 reemplazando los parámetros por los estimadores máximo verosímiles.

3.1.3 Regla Discriminante de la Razón de Verosimilitudes

Una alternativa a la regla de asignación máximo verosímil es utilizar el criterio de la razón de verosimilitud, dado por Anderson (1958), p.141. Considerando el caso en que si X es un individuo que pertenece a Π_r, entonces X se distribuye como $N_p(\mu_r, \Sigma)$. El criterio consiste en calcular las verosimilitudes de las siguientes hipótesis:

$$\begin{aligned} H_r &: X \text{ y los renglones de } \mathbf{X}_r \text{ pertenecen a } \Pi_r, \\ &\quad \text{y las filas de } \mathbf{X}_k \text{ pertenecen al grupo } \Pi_k,\ k \neq r. \end{aligned}$$

La regla consiste en asignar X a la población cuya hipótesis H_r tiene la mayor verosimilitud; esta verosimilitud bajo H_r está dada por:

$$\begin{aligned} L_r(\mu_1, ..., \mu_g, \Sigma) &= \prod_{i=1}^{g} \prod_{j=1}^{n_j} |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(X_{ij} - \mu_i)^t \Sigma^{-1}(X_{ij} - \mu_i)\right\} \quad (3.36) \\ &\times \quad |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(X - \mu_r)^t \Sigma^{-1}(X - \mu_r)\right\}. \end{aligned}$$

Si los parámetros son conocidos, la región de clasificación en Π_r, es simplemente:

$$\mathrm{R}_r = \{X \in \Re^p \mid L_r(x) \geq L_k(x) \ \forall\, r \neq k\}\,, \text{ para } r = 1, ..., g,$$

donde $L_r(x)$ es la función de verosimilitud de la $r-$ *ésima* población definida en (3.36).

Si los parámetros son desconocidos, estos se reemplazan por sus correspondientes estimadores máximo verosímiles. Sean $\hat{\mu}_k^{(r)}$ y $\hat{\Sigma}^{(r)}$ los estimadores para la media y de la matriz de covarianzas asociados al $k-$ *ésimo* grupo según la hipótesis H_r, y que toman las siguientes expresiones:

$$\hat{\mu}_k^{(r)} = \begin{cases} \bar{X}_k & \text{si } k \neq r, \\ \bar{X}_k^{(r)} = \frac{n_r \bar{X}_r + X}{n_r + 1} & \text{si } k = r. \end{cases}$$

y

$$\begin{aligned} \hat{\Sigma}^{(r)} &= W^{(r)}, \\ &= \frac{1}{n+1}\left[\sum_{k=1}^{g}\left(n_k + I_{\{r\}}(k)\right) S_k^{(r)}\right], \end{aligned}$$

donde

$$S_k^{(r)} = \begin{cases} S_k = \frac{1}{n_k}\sum_{j=1}^{n_k}\left(X_{kj} - \bar{X}_k\right)\left(X_{kj} - \bar{X}_k\right)^t & \text{si } k \neq r \\ S_r^{(r)} = \frac{1}{n_r+1}\left[\sum_{j=1}^{n_r}\left(X_{rj} - \bar{X}_r^{(r)}\right)\left(X_{rj} - \bar{X}_r^{(r)}\right)^t + \left(X - \bar{X}_r^{(r)}\right)\left(X - \bar{X}_r^{(r)}\right)^t\right] & \text{si } r = k \end{cases}$$

Por lo que las regiones de clasificación están dadas por:

$$\mathrm{R}_i = \left\{X \in \Re^p \mid \hat{L}_r(X) \geq \hat{L}_k(X) \ \forall\, r \neq k\right\},\ r = 1, ..., g.$$

y donde

$$\begin{aligned} \hat{L}_r(X) &= \prod_{i=1}^{g}\prod_{j=1}^{n_j} |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(X_{ij} - \bar{X}^{(r)}\right)^t W^{(r)-1}\left(X_{ij} - \bar{X}^{(r)}\right)\right\} \qquad (3.37) \\ &\quad \times |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(X - \bar{X}^{(r)}\right)^t W^{(r)-1}\left(X - \bar{X}^{(r)}\right)\right\},\ r = 1, ..., g. \end{aligned}$$

De la ecuación anterior puede observarse que las regiones de clasificación están en términos de la matriz $W^{(r)}$ que puede simplificarse desarrollando la siguiente igualdad

$$
\begin{aligned}
(n_r+1)\,S_r^{(r)} &= \sum_{j=1}^{n_r}\left(X_{rj}-\bar{X}_r^{(r)}\right)\left(X_{rj}-\bar{X}_r^{(r)}\right)^t+\left(X-\bar{X}_r^{(r)}\right)\left(X-\bar{X}_r^{(r)}\right)^t,\\
&= \sum_{j=1}^{n_r}\left(X_{rj}-\bar{X}_r+\bar{X}_r-\bar{X}_r^{(r)}\right)\left(X_{rj}-\bar{X}_r+\bar{X}_r-\bar{X}_r^{(r)}\right)^t\\
&+ \left(X-\bar{X}_r^{(r)}\right)\left(X-\bar{X}_r^{(r)}\right)^t,\\
&= \sum_{j=1}^{n_r}\left(X_{rj}-\bar{X}_r\right)\left(X_{rj}-\bar{X}_r\right)^t+\sum_{j=1}^{n_r}\left(\bar{X}_r-\bar{X}_r^{(r)}\right)\left(\bar{X}_r-\bar{X}_r^{(r)}\right)^t\\
&+ \left(X-\bar{X}_r^{(r)}\right)\left(X-\bar{X}_r^{(r)}\right)^t,\\
&= n_rS_r+n_r\left(\bar{X}_r-\bar{X}_r^{(r)}\right)\left(\bar{X}_r-\bar{X}_r^{(r)}\right)^t+\left(X-\bar{X}_r^{(r)}\right)\left(X-\bar{X}_r^{(r)}\right)^t,\\
&= n_rS_r+n_r\left(\bar{X}_r-\frac{n_r\,\bar{X}_r+X}{n_r+1}\right)\left(\bar{X}_r-\frac{n_r\,\bar{X}_r+X}{n_r+1}\right)^t\\
&\quad+\left(X-\frac{n_r\,\bar{X}_r+X}{n_r+1}\right)\left(X-\frac{n_r\,\bar{X}_r+X}{n_r+1}\right)^t,\\
&= n_rS_r+\frac{n_r}{(n_r+1)^2}\left(\bar{X}_r-X\right)\left(\bar{X}_r-X\right)^t+\frac{n_r^2}{(n_r+1)^2}\left(X-\bar{X}_r\right)\left(X-\bar{X}_r\right)^t,\\
&= n_rS_r+\frac{n_r\,(n_r+1)}{(n_r+1)^2}\left(\bar{X}_r-X\right)\left(\bar{X}_r-X\right)^t,\\
&= n_rS_r+\frac{n_r}{(n_r+1)}\left(\bar{X}_r-X\right)\left(\bar{X}_r-X\right)^t,
\end{aligned}
$$

Entonces

$$
\begin{aligned}
W^{(r)} &= \frac{1}{n+1}\left[\sum_{k=1}^{g}\left(n_k+I_{\{r\}}(k)\right)S_k^{(r)}\right],\\
&= \frac{1}{n+1}\left[\sum_{k\neq r}n_kS_k+S_r^{(r)}\right],\\
&= \frac{1}{n+1}\left[\sum_{k\neq r}n_kS_k+n_rS_r+\frac{n_r}{n_r+1}\left(X-\bar{X}_r\right)\left(X-\bar{X}_r\right)^t\right],\\
&= \frac{n}{n+1}W+\frac{n_r}{(n+1)\,(n_r+1)}\left(X-\bar{X}_r\right)\left(X-\bar{X}_r\right)^t
\end{aligned}
$$

y la función de verosimilitud de la muestra $\mathbf{X}_1, ..., \mathbf{X}_g$ bajo la hipótesis H_r dada en (3.37) puede simplificarse como sigue:

$$\begin{aligned}
\hat{L}_r(X) &= |2\pi W^{(r)}|^{-\frac{(n+1)}{2}} \exp\left\{-\frac{1}{2} tr W^{(r)} \left[\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i^{(r)}\right)\left(X_{ij}-\bar{X}_i^{(r)}\right)^t \right.\right. \\
&\quad \left.\left. + \left(X-\bar{X}_r^{(r)}\right)^t W^{(r)-1}\left(X-\bar{X}_r^{(r)}\right)\right]\right\}, \\
&= |2\pi W^{(r)}|^{-\frac{(n+1)}{2}} \exp\left\{-\frac{1}{2} tr\left(W^{(r)-1}(n+1)W^{(r)}\right)\right\}, \\
&= |2\pi W^{(r)}|^{-\frac{(n+1)}{2}} \exp\left\{-\frac{n+1}{2} p\right\}.
\end{aligned}$$

Por lo que ahora se debe encontrar

$$\begin{aligned}
\min_r |W^{(r)}| &= \min_r \left|\frac{n}{n+1} W^{(r)} + \frac{n_r}{(n+1)(n_r+1)}\left(X-\bar{X}_r^{(r)}\right)\left(X-\bar{X}_r^{(r)}\right)^t\right|, \\
&= \min_r |nW^{(r)}|\left[1+\frac{n_r}{n_r+1}\left(X-\bar{X}_r^{(r)}\right)^t (nW^{(r)})^{-1}\left(X-\bar{X}_r^{(r)}\right)\right], \\
&= \min_r \frac{n_r}{n_r+1}\left(X-\bar{X}_r^{(r)}\right)^t (nW^{(r)})^{-1}\left(X-\bar{X}_r^{(r)}\right),
\end{aligned}$$

desarrollando la forma cuadrática en la ecuación anterior se obtiene que la regla asigna X a $\mathbf{\Pi}_r$ en

$$\max_r \frac{n_r}{n_r+1} a_r^t \left(X - \frac{1}{2}\bar{X}_r\right),$$

donde se define $a_r = W^{-1}\bar{X}_r$.

En particular, si se considera el caso de dos poblaciones normales con matriz de covarianzas común, para las cuales un nuevo individuo X debe discriminarse, el juego de hipótesis está dado como sigue:

H_1 : X y los renglones de $\mathbf{X}_1$ pertenecen a $\mathbf{\Pi}_1$, y las filas de $\mathbf{X}_2$ provienen de $\mathbf{\Pi}_2$.

v.s

H_2 : Los renglones de $\mathbf{X}_1$ provienen de $\mathbf{\Pi}_1$ y X y las filas de $\mathbf{X}_2$ pertenecen a $\mathbf{\Pi}_2$.

Si los parámetros de $\bar{X}_1$ y $\bar{X}_2$ son desconocidos se deben reemplazar por sus estimadores máximo verosímiles. Los estimadores para μ_1, μ_2 y Σ bajo la hipótesis H_1 están dadas por:

$$\hat{\mu}_i = \begin{cases} \bar{X}_1^{(1)} = \frac{n_1\bar{X}_1+X}{n_1+1} & \text{si } i=1, \\ \bar{X}_2 & \text{si } i=2, \end{cases}$$

y

$$\hat{\Sigma}^{(1)} = \frac{n}{n_1+n_2+1}\left\{W + \frac{n_1}{n_1+1}\left(X-\bar{X}_1\right)\left(X-\bar{X}_1\right)^t\right\},$$

donde $nW = n_1S_1 + n_2S_2$.

Bajo la hipótesis H_2 los estimadores máximo verosímiles para μ_1, μ_2 y Σ están definidos por:

$$\hat{\mu}_i = \begin{cases} \bar{X}_1 & \text{si } i=1, \\ \bar{X}_2^{(2)} = \frac{n_2\bar{X}_2+X}{n_2+1} & \text{si } i=2, \end{cases}$$

y

$$\hat{\Sigma}^{(2)} = \frac{n}{n_1+n_2+1}\left\{W + \frac{n_2}{n_2+1}\left(X-\bar{X}_2\right)\left(X-\bar{X}_2\right)^t\right\}.$$

Entonces el cociente de la verosimilitud es la proporción:

$$\left|\frac{\hat{\Sigma}^{(2)}}{\hat{\Sigma}^{(1)}}\right| = \left|\frac{\frac{n}{n_1+n_2+1}\left\{W + \frac{n_2}{n_2+1}\left(X-\bar{X}_2\right)\left(X-\bar{X}_2\right)^t\right\}}{\frac{n}{n_1+n_2+1}\left\{W + \frac{n_1}{n_1+1}\left(X-\bar{X}_1\right)\left(X-\bar{X}_1\right)^t\right\}}\right|,$$

$$\left|\frac{\hat{\Sigma}^{(2)}}{\hat{\Sigma}^{(1)}}\right| = \frac{1+\frac{n_2}{n_2+1}\left(X-\bar{X}_2\right)^t W^{-1}\left(X-\bar{X}_2\right)}{1+\frac{n_1}{n_2+1}\left(X-\bar{X}_1\right)^t W^{-1}\left(X-\bar{X}_1\right)}.$$

La prueba acepta la hipótesis H_0, es decir, se asigna X a Π_1 si y sólo si

$$1+\frac{n_2}{n_2+1}\left(X-\bar{X}_2\right)W^{-1}\left(X-\bar{X}_2\right)^t > 1+\frac{n_1}{n_2+1}\left(X-\bar{X}_1\right)W^{-1}\left(X-\bar{X}_1\right)^t.$$

Si los tamaños de muestra son iguales, en otras palabras, si $n_1 = n_2$ este criterio es equivalente a la regla de máxima verosimilitud definida en la sección anterior. Pero si los tamaños de muestra son diferentes este método tiende a clasificar a X a la población que tiene el tamaño de muestra mas grande.

3.1.4 Regla Discriminante de Bayes

En algunas ocasiones es conveniente suponer que varias poblaciones tengan asignadas probabilidades *a priori*, por ejemplo, en los diagnósticos médicos se puede pensar que un paciente que padece de presión arterial alta es más propenso a ciertas enfermedades cardiacas que con respecto a algún otro paciente cuya presión arterial es normal.

La regla discriminante de Bayes, utiliza precisamente estas probabilidades a priori para efectuar la asignación de un individuo X a la población con mayor probabilidad *posterior*, esto es,a la cual el producto $\pi_i L_i(x)$ sea máxima.

Definición 3.4 *Si las poblaciones* $\Pi_1, ..., \Pi_g$ *tienen probabilidades a priori* $\pi' = (\pi_1, ..., \pi_g)$, *entonces la regla discriminante de Bayes (con respecto a* π*) asigna una observación* X *a la población para la cual*

$$\pi_i L_i(x),$$

sea maximizada.

Nótese que la regla de máxima verosimilitud es un caso especial del criterio de Bayes cuando todas las probabilidades iniciales son iguales.

En el caso de discriminación entre $g = 2$ poblaciones, el efecto que causa introducir probabilidades *a priori* es simplemente aumentar el valor crítico de la función de discriminación por la cantidad $\ln\left(\frac{\pi_2}{\pi_1}\right)$. En este caso la regla de asignación según el criterio de Bayes está dada por:

$$\text{asignar } X \text{ a } \begin{cases} \Pi_1 & \text{si } h(x) > \ln\left(\frac{\pi_2}{\pi_1}\right), \\ \Pi_2 & \text{en otro caso,} \end{cases}$$

donde

$$h(x) = \ln\left(\frac{L_1(x)}{L_2(x)}\right). \tag{3.38}$$

Desarrollando la ecuación (3.38) bajo el supuesto adicional de que las dos poblaciones tienen una matriz de covarianzas común conocida se obtiene

$$\ln\left(\frac{\left[|2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(X-\mu_1)' \Sigma^{-1}(X-\mu_1)\right\}\right]}{\left[|2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(X-\mu_2)' \Sigma^{-1}(X-\mu_2)\right\}\right]}\right) > \ln\left(\frac{\pi_2}{\pi_1}\right),$$

si y sólo si

$$-\frac{1}{2}(X-\mu_1)' \Sigma^{-1}(X-\mu_1) + \frac{1}{2}(X-\mu_2)' \Sigma^{-1}(X-\mu_2) > \ln\left(\frac{\pi_2}{\pi_1}\right),$$

lo cual es equivalente a

$$\mu_1^t \Sigma^{-1} X - \frac{1}{2}\mu_1^t \Sigma^{-1}\mu_1 - \mu_2^t \Sigma^{-1} X + \frac{1}{2}\mu_2 \Sigma^{-1}\mu_2 > \ln\left(\frac{\pi_2}{\pi_1}\right)$$

$\Leftrightarrow$

$$(\mu_1 - \mu_2)^t \Sigma^{-1}\left\{X - \frac{1}{2}(\mu_1 + \mu_2)\right\} > \ln\left(\frac{\pi_2}{\pi_1}\right). \tag{3.39}$$

Haciendo $\delta = \Sigma^{-1}(\mu_1 - \mu_2)$ y $\mu = (\mu_1 + \mu_2)$ la ecuación (3.39) se puede escribir como:

$$h(x) = \delta^t\left(X - \frac{1}{2}\mu\right) > \ln\left(\frac{\pi_2}{\pi_1}\right).$$

Propiedades Optimas

Las reglas discriminantes de Bayes previamente descritas (incluyendo la *r.d.m.v.*) cumplen con ciertas propiedades que son óptimas. Nótese en primer lugar que estas reglas son determinísticas en el sentido de que si $\bar{X}_1 = \bar{X}_2$ entonces X_1 y X_2 siempre se asignarán a la misma población. Sin embargo para propósitos matemáticos es conveniente considerar una clase más general de reglas discriminantes, mismas que describiremos sin profundizar los desarrollos matemáticos.

Una regla discriminante aleatoria d, asigna una observación X a una población i con probabilidad $\phi_i(x)$, en donde $\phi_1, ..., \phi_g$, son funciones no negativas definidas cada una en $\Re^p$, mismas que satisfacen

$$\sum_{i=1}^{g} \phi_i(x) = 1, \ \forall\, x \in \Re^p.$$

Es claro que una regla de asignación determinística es un caso particular de una regla de asignación aleatoria, tomando $\phi_i(x) = 1$ para $X \in \mathbf{R}_i$ y $\phi_i(x) = 0$ en cualquier otro caso.

Por ejemplo, la regla de Bayes con respecto a probabilidades a priori $\pi_1, ..., \pi_g$ está definida por:

$$\phi_i(x) = \begin{cases} 1 \text{ si } \pi_i L_i(x) = \max_k \{\pi_k L_k(x)\}, \\ 0 \text{ en otro caso.} \end{cases} \tag{3.40}$$

Definición 3.5 *La probabilidad de asignar un individuo a la población Π_i, cuando proviene de la población Π_k, está dada por:*

$$p_{ik} = \int \phi_i(x)\, L_k(x)\, dx. \tag{3.41}$$

En particular, si este individuo en $\Re^p$ que proviene de Π_i la probabilidad de asignarlo correctamente está dada por la siguiente expresión:

$$p_{ii} = \int \phi_i(x)\, L_i(x)\, dx.$$

La siguiente definición permite ordenar precisamente las reglas discriminantes

Definición 3.6 *Una regla discriminante d con probabilidad de asignación correcta $\{p_{ii}\}$ es tan buena como cualquier otra regla d^* con probabilidades $\{p_{ii}^*\}$ si*

$$p_{ii} \geq p_{ii}^*, \qquad \forall\ i$$

Se dice que d es mejor que d^* si al menos una de las desigualdades es estricta.

Definición 3.7 *Si d es una regla para la cual no existe otra regla mejor, se dice entonces que d es una regla admisible.*

Teorema 3.7 *Todas las reglas discriminantes de Bayes (incluyendo la r.d.m.v.) son admisibles.*

Demostración.

Sea d^* una regla de Bayes con probabilidades a priori $\pi_1, ..., \pi_g$. Supóngase que existe otra regla d que es mejor que la regla anterior. Sean $\{p_{ii}^*\}$ y $\{p_{ii}\}$ las probabilidades de clasificación correcta para la regla d^* y d respectivamente.

Como d es mejor que d^* y $\pi_i > 0$ para toda i, se sigue que

$$\sum_{i=1}^{g} \pi_i p_{ii} > \sum_{i=1}^{g} \pi_i p_{ii}^*. \tag{3.42}$$

Utilizando las Definiciones 3.40 y 3.5 se obtiene

$$\sum_{i=1}^{g} \pi_i p_{ii} = \sum_{i=1}^{g} \int \phi_i(x)\, \pi_i L_i(x)\, dx,$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{g} \int \phi_i(x) \max_j \pi_j L(x) \pi_j dx, \\
&= \int \left[\sum_{i=1}^{g} \phi_i(x) \right] \max_j \pi_j L_j(x) \, dx \\
&= \int \max_j \pi_j L(x) \, dx, \\
&= \int \sum_{i=1}^{g} \phi_i^*(x) \pi_i L_i(x) \, dx, \\
&= \sum_{i=1}^{g} \pi_i p_{ii}^*,
\end{aligned}
$$

lo cual contradice la ecuación (3.42). □

Teorema 3.8 *Si las poblaciones $\Pi_1, ..., \Pi_g$ tienen asociadas las probabilidades a priori $\pi_1, ..., \pi_g$, entonces ninguna otra regla discriminante tiene mejores probabilidades de asignación correcta que la regla de Bayes con respecto a esas probabilidades a priori.*

Demostración.

Sea d^* la regla de Bayes con probabilidades de clasificación correcta $\{p_{ii}^*\}$ y sea d cualquiera otra regla discriminante con probabilidades de clasificación correcta $\{p_{ii}\}$. Ambas reglas con probabilidades a priori $\pi_1, ..., \pi_g$.

Se sigue del Teorema 3.7, que $\sum_{i=1}^{g} \pi_i p_{ii}^* \geq \sum_{i=1}^{g} \pi_i p_{ii}$, es decir, la regla d tiene probabilidades posteriores de asignación correcta a lo más tan grandes que la regla de Bayes. □

3.1.5 Función Lineal Discriminante de Fisher

Otra aproximación al problema de discriminación basado en la matriz de datos $\mathbf{X}$ puede hacerse sin suponer alguna forma particular de distribución paramétrica en las poblaciones $\Pi_1, ..., \Pi_g$. Fisher sugirió una función lineal $a^t X$ la cual deba maximizar la proporción entre la suma de cuadrados entre grupos (B) y la suma de cuadrados dentro de los grupos (W), definiendo a dicha combinación lineal como:

$$Y_{n \times 1} = Xa,$$

$$= \begin{pmatrix} X_1 a \\ \vdots \\ X_g a \end{pmatrix},$$

$$= \begin{pmatrix} Y_1 \\ \vdots \\ Y_g \end{pmatrix}.$$

con a un vector en $\Re^p$.

La matriz de covarianza total para Y esta definida como:

$$T_Y = \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}\right) \left(Y_{ij} - \bar{Y}\right)^t,$$

donde

$$\begin{aligned} \bar{Y} &= \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{n_i} Y_{ij}, \\ &= \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{n_i} a^t X_{ij}, \\ &= a^t \sum_{i=1}^{g} \sum_{j=1}^{n_i} \frac{X_{ij}}{n}, \\ &= a^t \bar{X}. \end{aligned} \tag{3.43}$$

Utilizando la ecuación (3.43), T_Y toma la siguiente expresión:

$$\begin{aligned} T_Y &= \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(a^t X_{ij} - a^t \bar{X}\right) \left(a^t X_{ij} - a^t \bar{X}\right)^t \\ &= a^t \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}\right) \left(X_{ij} - \bar{X}\right)^t a, \\ &= a^t T_X a. \end{aligned}$$

De forma similar la matriz de covarianzas dentro (W) de grupos para Y está dada por:

$$\begin{aligned} W_Y &= \frac{1}{n} \sum_{i=1}^{g} n_i \left(Y_{ij} - \bar{Y}\right) \left(Y_{ij} - \bar{Y}\right)^t, \\ &= a^t \frac{1}{n} \sum_{i=1}^{g} n_i \left(X_{ij} - \bar{X}_i\right) \left(X_{ij} - \bar{X}_i\right)^t a, \\ &= a^t W_X a. \end{aligned}$$

Análogamente la matriz de covarianzas entre grupos (B) para Y toma la expresión:

$$\begin{aligned} B_Y &= \frac{1}{n}\sum_{i=1}^{g} n_i \left(\bar{Y}_i - \bar{Y}\right)\left(\bar{Y}_i - \bar{Y}\right)^t, \\ &= a^t \frac{1}{n}\sum_{i=1}^{g} n_i \left(\bar{X}_i - \bar{X}\right)\left(\bar{X}_i - \bar{X}\right)^t a, \\ &= a^t B_X a. \end{aligned}$$

El criterio de Fisher es intuitivamente atractivo porque es fácil de separar los grupos si la matriz B_Y es relativamente más grande a la matriz W_Y. La proporción que sugiere Fisher está dada por:

$$\frac{a^t B_X a}{a^t W_X a}. \tag{3.44}$$

Si a es el vector que maximiza la ecuación (3.44) se puede llamar a la combinación lineal $a^t X$, la función lineal discriminante de Fisher o primer variable canónica.

Teorema 3.9 *El vector a que maximiza la proporción $\frac{a^t B_X a}{a^t W_X a}$ es el primer vector propio de $W_X^{-1} B_X$ (asociado al mayor valor propio).*

Demostración

Obsérvese que maximizar el cociente $\frac{a^t B_X a}{a^t W_X a}$, es resolver

$$\max_X a^t B_X a \ s.a.\ a^t W_X a = 1,$$

dado que el cociente en la ecuación (3.44) es invariante a cambios de escala en el vector a. El problema puede plantearse equivalentemente como:

$$\begin{aligned} \max_X a^t B_X a &= \max_Y Y^t W_X^{-\frac{1}{2}} B_X W_X^{-\frac{1}{2}} Y \\ s.a.\ Y^t Y &= 1 \end{aligned}$$

donde $Y = W_X^{\frac{1}{2}} a$. Como la matriz W_X es definida semi positiva se sigue que las matrices W_X^{-1}, y $W_X^{\frac{1}{2}}$ lo son también, análogamente la matriz B_X, es definida positiva y por lo tanto $B_X^{\frac{1}{2}}$ es definida semipositiva, entonces

$$W_X^{-\frac{1}{2}} B_X W_X^{-\frac{1}{2}} = W_X^{-\frac{1}{2}} B_X^{\frac{1}{2}} B_X^{\frac{1}{2}} W_X^{-\frac{1}{2}},$$

$$= R^t R,$$
$$\geq 0,$$

donde la matriz R *se define como* $R = B_X^{\frac{1}{2}} W_X^{-\frac{1}{2}}$.

Ya que la matriz $W_X^{-\frac{1}{2}} B_X W_X^{-\frac{1}{2}}$ *es una matriz simétrica y definida positiva se puede descomponer en sus vectores y valores propios como sigue:*

$$W_X^{-\frac{1}{2}} B_X W_X^{-\frac{1}{2}} = U \Lambda U^t,$$

entonces la expresión $Y^t W_X^{-\frac{1}{2}} B_X W_X^{-\frac{1}{2}} Y$ *puede escribirse como:*

$$Y^t W_X^{-\frac{1}{2}} B_X W_X^{-\frac{1}{2}} Y = Y^t U \Lambda U^t Y,$$

y el problema de maximización puede plantearse de la siguiente manera:

$$\max z^t \Lambda z \quad s.a. \quad z^t z = 1$$

con $z = U^t Y$. *Observando que*

$$\sum_{j=1}^{p} \lambda_j z_j^2 \leq \lambda_i \sum_{i=1}^{p} z_i^2$$
$$= \lambda_i,$$

en donde el máximo se alcanza si $z^t = (1, 0, ..., 0)$ *con* $z \in \Re^p$.□

Una vez que la función lineal discriminante ha sido obtenida, una observación X puede ser asignada a una de las g poblaciones (ó grupos), con base en este puntaje discriminante $a^t X$. La media muestral $\bar{X}_i$ tiene un puntaje

$$a^t \bar{X}_i = \bar{Y}_i .$$

Entonces X es asignado a la población Π_i si:

$$\left| a^t X - a^t \bar{X}_k \right| \leq \left| a^t X - a^t \bar{X}_i \right| \ \forall\, i = 1, ..., g.$$

La función lineal discriminante de Fisher es más utilizada cuando existen $g = 2$ grupos, entonces si B tiene rango uno, puede ser escrita como:

$$B = \left(\frac{n_1 n_2}{n}\right) dd^t.$$

donde $d = \left(\bar{X}_1 - \bar{X}_2\right)$. Entonces, $W_X^{-1}B$ tiene sólo un valor propio distinto de cero el cual puede ser obtenido explícitamente como:

$$\begin{aligned} tr\left(W^{-1}B\right) &= tr\left(W^{-1}\left(\frac{n_1 n_2}{n}\right) dd^t\right), \\ &= \frac{n_1 n_2}{n} tr\left(W^{-1}dd^t\right), \\ &= \frac{n_1 n_2}{n} d^t W^{-1} d. \end{aligned}$$

El vector propio correspondiente está dado por:

$$a = W^{-1}d.$$

Entonces la regla discriminante está dada como:

$$\text{asignar } X \text{ a } \begin{cases} \Pi_1 & d^t W^{-1}\left(X - \frac{1}{2}\bar{X}\right) > 0. \\ \Pi_2 & \text{en otro caso.} \end{cases} \tag{3.45}$$

Obsérvese que la expresión (3.45) es exactamente la misma que la obtenida mediante la *r.d.m.v.* para 2 grupos que provienen de una población normal con la misma matriz de covarianza. Sin embargo en la regla de asignación (3.35) existe explícitamente el supuesto de normalidad, mientras que en la regla de asignación (3.45) se tiene simplemente una regla sensible basada en una función lineal de X. Entonces se espera que ésta regla sea apropiada para poblaciones en las que la hipótesis de normalidad no se satisface exactamente.

En general, $W^{-1}B$ tiene un $\min\{p, g-1\}$ valores propios distintos de cero. Sus correspondientes valores propios definen la segunda, tercera, etc. variables canónicas. Las primeras k variables son utilizadas cuando se espera que la diferencia entre los grupos esté concentrada en k dimensiones.

3.1.6 Probabilidades de Mala Clasificación

Probabilidades cuando los parámetros son estimados

Formalmente, las probabilidades de mala clasificación p_{ik} están dadas en la ecuación (3.41). Si los parámetros de las distribuciones en cuestión son estimados de los datos, entonces se obtienen las probabilidades estimadas $\hat{p}_{ik}$.

Considérese el caso de dos poblaciones normales $N_p(\mu_1, \Sigma)$ y $N_p(\mu_2, \Sigma)$. Si X es un individuo que proviene de Π_1, debe cumplirse que la función lineal discriminante

$$h(X) = \alpha^t \left(X - \frac{1}{2}\mu \right) > 0,$$

donde $\alpha = \Sigma^{-1}(X - \mu)$ y $\mu = (\mu_1 + \mu_2)$, tiene densidad normal con media y varianza dados por:

$$\begin{aligned} E(h(X)) &= E\left(\alpha^t \left(X - \frac{1}{2}\mu \right) \right), \\ &= \alpha^t E\left(X - \frac{1}{2}\mu \right), \\ &= \frac{1}{2}\alpha^t (\mu_1 - \mu_2). \end{aligned}$$

y varianza

$$\begin{aligned} V(h(X)) &= V\left(\alpha^t \left(X - \frac{1}{2}\mu \right) \right), \\ &= \alpha^t V(X) \alpha, \\ &= \alpha^t \Sigma \alpha. \end{aligned}$$

Entonces puede verse que si X proviene de Π_1

$$h(X) \sim N_p\left(-\frac{1}{2}\Delta^2, \Delta^2 \right),$$

donde

$$\Delta^2 = (\mu_1 - \mu_2)^t \Sigma (\mu_1 - \mu_2),$$

la cual es conocida como la distancia de Mahalanobis entre ambas poblaciones. Análogamente puede demostrarse que si X proviene de Π_2

$$h(X) \sim N_p\left(-\frac{1}{2}\Delta^2, \Delta^2 \right).$$

Entonces las probabilidades de mala clasificación están dadas por:

$$p_{12} = P\left(h\left(X\right) > 0 \mid \Pi_2\right). \tag{3.46}$$

Aclarando que el símbolo 1 denota la población (Π_1) a la cual fue asignada la observación X condicionado a que proviene de la población 2 (Π_2). Conociendo la distribución de $h\left(X\right)$, la ecuación (3.46) puede reescribirse de la siguiente manera:

$$\begin{aligned} p_{12} &= P\left(\frac{h\left(X\right) - E\left(h\left(X\right)\right)}{\sqrt{V\left(h\left(X\right)\right)}} > -\frac{E\left(h\left(X\right)\right)}{\sqrt{V\left(h\left(X\right)\right)}} \mid \Pi_2\right), \\ &= P\left(\frac{h\left(X\right) - E\left(h\left(X\right)\right)}{\sqrt{V\left(h\left(X\right)\right)}} > \frac{-\frac{1}{2}\Delta^2}{\Delta} \mid \Pi_2\right), \\ &= \Phi\left(-\frac{1}{2}\Delta\right). \end{aligned}$$

Donde Φ denota a la función de distribución normal estándar. De forma similar si X proviene de 2 y fue asignada a 1, la probabilidad de mala clasificación está definida por:

$$p_{21} = \Phi\left(-\frac{1}{2}\Delta\right).$$

Sí los parámetros son estimados de los datos, entonces un estimador de Δ^2 es naturalmente

$$D^2 = \left(\bar{X}_1 - \bar{X}_2\right)' S^{-1} \left(\bar{X}_1 - \bar{X}_2\right),$$

y las probabilidades de mala clasificación según el desarrollo anterior toman la expresión:

$$\begin{aligned} \hat{p}_{12} &= \hat{p}_{21}, \\ &= \Phi\left(-\frac{1}{2}D\right). \end{aligned}$$

Desafortunadamente esta aproximación tiende a ser bondadosa, i.e., la aproximación tiende a subestimar las verdaderas probabilidades de mala clasificación cuando el tamaño de muestra n es pequeño. Aitchison et al. (1977) argumentan que las probabilidades basadas en la regla discriminante de Bayes son más realistas.

Método de Resustitución

Supóngase que la discriminación se basa en una matriz de datos $\mathbf{X}_{n\times p}$ que está definida como en (3.27), donde n_k individuos provienen de la población k. Si la regla discriminante

define las regiones de clasificación $\mathbf{R}_i$ sean n_{ik} el números de individuos de $\mathbf{\Pi}_k$ que cayeron en la región $\mathbf{R}_i$, es decir, n_{ik} individuos están mal clasificados si $i \neq k$ ($\sum_i n_{ik} = n_k$). La proporción:

$$\hat{p}_{ik} = \frac{n_{ik}}{n_k},$$

es un estimador de p_{ik}, desafortunadamente este método también tiende a ser optimista en las probabilidades de mala clasificación.

Método U de Jack-Knifing

Sea X_r, $r = 1, ..., n_1$ los primeros n_1 renglones de la matriz $\mathbf{X}_{n \times p}$ que representan a los individuos de $\mathbf{\Pi}_1$. Para cada r, sean $\mathbf{R}_1^{(r)}, ..., \mathbf{R}_g^{(r)}$ las regiones de asignación de la regla discriminante basada en una matriz de datos con dimensión $(n-1) \times p$ que se obtiene eliminando el $r-ésimo$ renglón de $\mathbf{X}$. Entonces el punto X_r puede ser utilizado para juzgar la calidad de esta regla, ya que fue derivada sin utilizar X_r. Si n_{k1}^* denota el número de los primeros n_1 individuos para los cuales $X_r \in \mathbf{R}_i^{(r)}$, entonces las proporciones:

$$p_{k1}^* = \frac{n_{k1}^*}{n_1}, \ k = 1, ..., g.$$

son un estimador de las tasas de error. Repitiendo el procedimiento para cada una de las otras poblaciones $s = 2, ..., g$ se obtiene el resto de los estimadores p_{ij}^*.

3.1.7 Selección de Variables

Uno de los problemas que con frecuencia ocurren dentro del Análisis Discriminante, es elegir únicamente aquellas variables que tienen poder de discriminación y eliminar las variables que simplemente no contribuyan a mejorar las probabilidades de clasificación correcta. En esta Sección, se analizará una prueba para aquella o aquel conjunto de variables para las cuales se deseé probar su poder discriminatorio.

Para fijar ideas denótese con el símbolo 2 la (las) variable(s) para la(s) que se desea probar su capacidad discriminatoria (variables fuera del modelo discriminante); y con el símbolo 1 la(s) variable(s) en la(s) que está condicionada la prueba (variables dentro del modelo discriminante).

Prueba de igualdad de medias condicional para poblaciones normales

Considérese el problema de discriminación entre g poblaciones normales multivariadas con matriz de covarianzas común. Sea una muestra $\mathbf{X}_{ij} \sim N_p(\mu_i, \Sigma)$, $i = 1, ..., g$, $j = 1, ..., n_i$ y $\sum_{i=1}^{g} n_i = n$; en donde se cumple que los vectores X_{ij} son independientes, los cuales están particionados como:

$$\mathbf{X}_{ij(p\times 1)} = \begin{pmatrix} \mathbf{X}_{ij1(q\times 1)} \\ \mathbf{X}_{ij2(p-q\times 1)} \end{pmatrix}.$$

donde $\mathbf{X}_{ij1}$ denota el conjunto de variables que están incluidasdas en la discriminación y $\mathbf{X}_{ij2}$ es el conjunto de variables cuya utilidad en la discriminación será probada. La matriz de covarianzas Σ y el vector de medias μ_i se particionan de la siguiente manera:

$$\Sigma = \begin{pmatrix} \Sigma_{11(q\times q)} & \Sigma_{12(q\times p-q)} \\ \Sigma_{21(p-q\times q)} & \Sigma_{22(p-q\times p-q)} \end{pmatrix} \quad \text{y} \quad \mu_i = \begin{pmatrix} \mu_{i1(q\times 1)} \\ \mu_{i2((p-q)\times 1)} \end{pmatrix}.$$

Si X es un nuevo individuo el cual proviene de Π_i, esto implica que $X \sim N_p(\mu_i, \Sigma)$ y la f.d. condicional de $X_2 \mid X_1$ está dada como una $N_p(\mu_{2.1}, \Sigma_{2.1})$, donde

$$\begin{aligned} \mu_{2.1} &= \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \mu_1), \\ \Sigma_{2.1} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}. \end{aligned}$$

El juego de hipótesis a probar está definido como:

$$\begin{aligned} H_0 &: \mu_{i2} + \Sigma_{21}\Sigma_{11}^{-1}(X_j - \mu_{i1}) = \mu_{j2} + \Sigma_{21}\Sigma_{11}^{-1}(X_i - \mu_{j1}) \quad \forall\, i \neq j. \qquad (3.47) \\ &\quad vs \\ H_a &: \text{no } H_0. \end{aligned}$$

Obsérvese que si H_0 es cierta, la discriminación incluyendo las variables fuera del modelo no tiene sentido ya que se concluye que las medias son iguales dado que las matrices de escala son iguales.

La hipótesis (3.47) es equivalente a probar:

$$\begin{aligned} H_0 &: \mu_{i2} - \Sigma_{21}\Sigma_{11}^{-1}\mu_{i1} = \mu_{j2} - \Sigma_{21}\Sigma_{11}^{-1}\mu_{j1} \ \forall i \neq j. \\ &\quad vs \\ H_a &: \text{no } H_0. \end{aligned}$$

La verosimilitud bajo $H_0 \cup H_a$ está dada como:

$$\begin{aligned} L(\mu_1, ..., \mu_g, \Sigma) &= \prod_{i=1}^{g} \prod_{j=1}^{n_i} f_{X_{ij}}(x_{ij}) \\ &= |2\pi\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (X_{ij} - \mu_i)^t \Sigma^{-1} (X_{ij} - \mu_i)\right\} \end{aligned} \tag{3.48}$$

cuyos estimadores máximo verosímiles pueden obtenerse maximizando la expresión:

$$-\frac{1}{2} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (X_{ij} - \mu_i)^t \Sigma^{-1} (X_{ij} - \mu_i),$$

con respecto de μ_i y Σ. Dichos estimadores están dados como:

$$\begin{aligned} \hat{\mu}_i &= \bar{X}_{i\cdot} \\ &= \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \, i = 1, ..., g. \\ \hat{\Sigma} &= W, \\ &= \frac{1}{n} \sum_{i=1}^{g} n_i S_i. \end{aligned}$$

Entonces el supremo de la función de verosimilitud 3.48 se alcanza en

$$\begin{aligned} \sup_{H_0 \cup H_a} L(\mu_1, ..., \mu_g, \Sigma) &= L\left(\hat{\mu}_1, ..., \hat{\mu}_g, \hat{\Sigma}\right), \\ &= |2\pi W|^{-\frac{n}{2}} \exp\left\{-\frac{np}{2}\right\}. \end{aligned}$$

La verosimilitud bajo la hipótesis H_0 toma la expresión:

$$L(\mu_1, ..., \mu_g, \Sigma) = \prod_{i=1}^{g} \prod_{j=1}^{n_i} \left[f_{X_{ij2}|X_{ij1}}(x_{ij1}, x_{ij2}) \, f_{ij1}(x_{ij1})\right],$$

Recordando que las X_{ij} son variables independientes tales que, $X_{ij} \sim N_p(\mu_i, \Sigma)$, además $\mathbf{X}_{ij1}$ se distribuye como una $N_p(\mu_{i1}, \Sigma_{11})$ y que la f.d. condicional de $f_{X_{ij2}|X_{ij1}}$ toma la expresión:

$$f_{X_{ij2}|X_{ij1}}(x_{ij1}, x_{ij2}) = |2\pi\Sigma_{2.1}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(X_{ij2} - \mu_{2.1})^t (\Sigma_{2.1})^{-1} (X_{ij2} - \mu_{2.1})\right\}, \tag{3.49}$$

$$= |2\pi\Sigma_{2.1}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\left[X_{ij2} - (\mu_{i2} + \Sigma_{21}\Sigma_{11}^{-1}(X_{ij1} - \mu_{i1}))\right]^t (\Sigma_{2.1})^{-1} \quad (3.50)$$
$$\times \left[X_{ij2} - (\mu_{i2} + \Sigma_{21}\Sigma_{11}^{-1}(X_{ij1} - \mu_{i1}))\right]\},$$

utilizando (3.50) se puede expresar la función de densidad conjunta para obtener

$$f_{X_{ij2},X_{ij1}}(x_{ij1}, x_{ij2}) = (2\pi)^{-\frac{np}{2}} |\Sigma_{11}|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(X_{ij1} - \mu_{i1})^t \Sigma_{11}^{-1}(X_{ij1} - \mu_{i1})\right\} \times$$
$$|\Sigma_{2.1}|^{-\frac{n}{2}} \exp\{-\frac{1}{2}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(X_{ij2} - M_{21}X_{ij1} - \theta_i)^t (\Sigma_{2.1})^{-1}$$
$$\times (X_{ij2} - M_{21}X_{ij1} - \theta_i)\},$$

donde $\theta_i = \mu_{i2} - \Sigma_{21}\Sigma_{11}^{-1}\mu_{i1}$ y $M_{21} = \Sigma_{21}\Sigma_{11}^{-1}$. Maximizando primero respecto a los vectores de medias y considerando la hipótesis $H_0 : \theta_1 = \cdots = \theta_g = \theta$, los estimadores están dados por:

$$\hat{\mu}_{i1} = \bar{X}_{i1}, \; i = 1, \ldots, g$$

y

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(X_{ij2} - M_{21}X_{ij1}) \quad (3.51)$$
$$= \bar{X}_2 - M_{21}\bar{X}_1 .$$

en donde, $\bar{X}_i$ está particionado como:

$$\bar{X}_i = \begin{pmatrix} \bar{X}_{i1} \\ \bar{X}_{i2} \end{pmatrix}, i = 1, .., g.$$

Con las expresiones anteriores y aplicando propiedades de la traza la función de verosimilitud se reduce a

$$L\left(\hat{\mu}_1, \ldots, \hat{\mu}_g, \Sigma\right) = (2\pi)^{-\frac{np}{2}} |\Sigma_{11}|^{-\frac{n}{2}} \exp\left\{-\frac{n}{2} tr\left(\Sigma_{11}^{-1} W_{11}\right)\right\} \quad (3.52)$$
$$\times |\Sigma_{22.1}|^{-\frac{n}{2}} \exp\{-\frac{1}{2} tr\left(\Sigma_{22.1}^{-1}\right) \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(X_{ij2} - M_{21}X_{ij1} - \hat{\theta}\right)$$
$$\times \left(X_{ij2} - M_{21}X_{ij1} - \hat{\theta}\right)^t\},$$

La f.d. condicional (3.50) de acuerdo a (3.51) se escribe de la siguiente manera:

$$f_{X_{ij2}|X_{ij1}} = |\Sigma_{22.1}|^{-\frac{n}{2}} \exp\{-\frac{1}{2}tr\left(\Sigma_{22.1}^{-1}\right)\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left[X_{ij2}-\bar{X}_2-M_{21}\left(X_{ij1}-\bar{X}_1\right)\right] \times \left[X_{ij2}-\bar{X}_2-M_{21}\left(X_{ij1}-\bar{X}_1\right)\right]^t.$$

Para simplificar la expresión anterior debe desarrollarse la doble suma que aparece dentro del exponente como:

$$\begin{aligned} SS &= \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left[\left(X_{ij2}-\bar{X}_2\right)-M_{21}\left(X_{ij1}-\bar{X}_1\right)\right]\left[\left(X_{ij2}-\bar{X}_2\right)-M_{21}\left(X_{ij1}-\bar{X}_1\right)\right]^t, \\ &= \sum_{i=1}^{g}\sum_{j=1}^{n_i}[\left(X_{ij2}-\bar{X}_2\right)\left(X_{ij2}-\bar{X}_2\right)^t - M_{21}\left(X_{ij1}-\bar{X}_1\right)\left(X_{ij2}-\bar{X}_2\right)^t \\ &\quad -\left(X_{ij2}-\bar{X}_2\right)\left(X_{ij1}-\bar{X}_1\right)^t M_{21}^t + M_{21}\left(X_{ij1}-\bar{X}_1\right)\left(X_{ij1}-\bar{X}_1\right)^t M_{21}^t], \end{aligned}$$

de donde

$$\begin{aligned} SS &= \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(X_{ij2}-\bar{X}_2\right)\left(X_{ij2}-\bar{X}_2\right)^t - \sum_{i=1}^{g}\sum_{j=1}^{n_i}M_{21}\left(X_{ij1}-\bar{X}_1\right)\left(X_{ij2}-\bar{X}_2\right)^t \\ &\quad -\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(X_{ij2}-\bar{X}_2\right)\left(X_{ij1}-\bar{X}_1\right)^t M_{21}^t + \sum_{i=1}^{g}\sum_{j=1}^{n_i}M_{21}\left(X_{ij1}-\bar{X}_1\right)\left(X_{ij1}-\bar{X}_1\right)^t M_{21}^t. \end{aligned} \tag{3.53}$$

Sea

$$T_{fd} = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(X_{ijf}-\bar{X}_f\right)\left(X_{ijd}-\bar{X}_d\right)^t, \tag{3.54}$$

obsérvese que la dimensión de T_{12} es $(q \times p-q)$, T_{11} es de $(q \times q)$ y que T_{22} es de $(p-q \times p-q)$. La ecuación (3.53) puede simplificarse de acuerdo a la definición (3.54) como:

$$SS = nT_{22} - nM_{21}T_{12} - nT_{21}M_{21}^t + nM_{21}T_{11}M_{21}^t.$$

Por lo que la expresión de la ecuación (3.52) toma la forma:

$$\begin{aligned} L\left(\hat{\mu}_1,...,\hat{\mu}_g,\Sigma\right) &= (2\pi)^{-\frac{np}{2}}|\Sigma_{11}|^{-\frac{n}{2}}\exp\left\{-\frac{n}{2}tr\left(\Sigma_{11}^{-1}W_{11}\right)\right\} \\ &\quad \times|\Sigma_{2.1}|^{-\frac{n}{2}}\exp\{-\frac{n}{2}tr\left(\Sigma_{2.1}^{-1}\right)[T_{22}-M_{21}T_{12} \\ &\quad -T_{21}M_{21}^t + M_{21}T_{11}M_{21}^t]\}. \end{aligned} \tag{3.55}$$

Para obtener el estimador máximo verosímil de $M_{21} = \Sigma_{21}\Sigma_{11}^{-1}$ se debe maximizar la expresión:

$$\exp\{-\frac{n}{2}tr\left(\Sigma_{2.1}^{-1}\right)\left[T_{22} - M_{21}T_{12} - T_{21}M_{21}^t + M_{21}T_{11}M_{21}^t\right]\},$$

lo cual es equivalente a minimizar

$$tr\Sigma_{2.1}^{-1}(T_{22} - M_{21}T_{12} - T_{21}M_{21}^t + M_{21}T_{11}M_{21}^t),$$

pero esta última expresión se puede escribir como

$$\begin{aligned} E &= tr\Sigma_{2.1}^{-1}\left(T_{22} - T_{21}M_{21}^t - M_{21}T_{12} + M_{21}T_{11}M_{21}^t\right), \\ &= tr\Sigma_{2.1}^{-1}\left[T_{22} + \left(M_{21} - T_{21}T_{11}^{-1}\right)T_{11}\left(M_{21} - T_{21}T_{11}^{-1}\right)^t - T_{21}T_{11}^{-1}T_{12}\right], \end{aligned}$$

reacomodando se obtiene

$$E = tr\Sigma_{2.1}^{-1}T_{2.1} + tr\left(\Sigma_{2.1}^{-\frac{1}{2}}\left(M_{21} - T_{21}T_{11}^{-1}\right)^t T_{11}\left(M_{21} - T_{21}T_{11}^{-1}\right)\Sigma_{2.1}^{-\frac{1}{2}}\right), \tag{3.56}$$

donde $T_{2.1} = T_{22} - T_{21}T_{11}^{-1}T_{12}$. Si se define $D = T_{11}^{\frac{1}{2}}\left(M_{21} - T_{21}T_{11}^{-1}\right)\Sigma_{2.1}^{-\frac{1}{2}}$, entonces la ecuación (3.56) puede reescribirse como:

$$E = tr\Sigma_{2.1}^{-1}T_{2.1} + trD^tD, \tag{3.57}$$

donde se sabe que D^tD es una forma cuadrática definida semipositiva, ya que si se toma Z un vector no nulo en $\Re^p$ se tiene

$$\begin{aligned} Z^tD^tDZ &= r^tr, r = DZ \\ &\geq 0. \end{aligned}$$

Nótese que $\Sigma_{22.1}^{-1} \geq 0$ y que la matriz $T_{22.1}$ es constante dado que no depende de M_{21}, entonces el mínimo con respecto a la matriz M_{21} se alcanza igualando a cero la matriz

$$\left(M_{21} - T_{21}T_{11}^{-1}\right)^t T_{11}\left(M_{21} - T_{21}T_{11}^{-1}\right),$$

la cual se satisface si y sólo si $M_{21} = T_{21}T_{11}^{-1}$. Por lo tanto, el estimador máximo verosímil para M_{21} es

$$\hat{M}_{21} = T_{21}T_{11}^{-1}.$$

De esta forma la verosimilitud dada en la ecuación (3.55) bajo H_0, de acuerdo al desarrollo anterior queda expresada como:

$$L\left(\hat{\mu}_{11},...,\hat{\mu}_{g1},\Sigma_{11},\Sigma_{22.1},\hat{M}_{21}\right) = (2\pi)^{-\frac{np}{2}}|\Sigma_{11}|^{-\frac{n}{2}}\exp\left\{-\frac{n}{2}tr\left(\Sigma_{11}^{-1}W_{11}\right)\right\}\cdot |\Sigma_{22.1}|^{-\frac{n}{2}}\exp\left\{-\frac{n}{2}tr\Sigma_{22.1}^{-1}T_{22.1}\right\}.$$

Utilizando el Teorema de Estimación de parámetros dado en el Capítulo 1 (ver Teorema 1.23), los estimadores para Σ_{11} y $\Sigma_{22.1}$ están dados por:

$$\begin{aligned}\hat{\Sigma}_{11} &= W_{11(q\times q)},\\ \hat{\Sigma}_{22.1} &= T_{22.1(p-q\times p-q)}.\end{aligned}$$

Entonces el cociente de las verosimilitudes se escribe como:

$$\begin{aligned}\lambda &= \frac{\sup_{H_0}L\left(\mu_{11},...,\mu_{g1},\Sigma\right)}{\sup_{H_0\cup H_a}L\left(\mu_{1,..}\mu_g,\Sigma\right)},\\ &= \frac{(2\pi)^{-\frac{np}{2}}|W_{11}|^{-\frac{n}{2}}|T_{22.1}|^{-\frac{n}{2}}\exp\left\{-\frac{ng}{2}-\frac{n}{2}(p-q)\right\}}{(2\pi)^{-\frac{np}{2}}|W|^{-\frac{n}{2}}\exp\left\{-\frac{np}{2}\right\}},\\ &= \left(\frac{|W_{11}||T_{22.1}|}{|W|}\right)^{-\frac{n}{2}},\\ &= \left(\frac{|W_{11}||T_{22.1}|}{|W_{11}||W_{22.1}|}\right)^{-\frac{n}{2}}.\end{aligned}$$

Finalmente la región de rechazo está dada por:

$$R=\left\{X_{ij}\ i=1,...,g\ \text{y}\ j=1,..,n_i\ \mid \lambda_1=\frac{|W_{22.1}|}{|T_{22.1}|}\leq k\right\}.$$

Puede demostrarse (ver por ejemplo Mendoza 1987), que las matrices $W_{22.1}$ y $D_{22.1}=T_{22.1}-W_{22.1}$ son independientes y que se distribuyen Wishart, es decir

$$\begin{aligned}W_{22.1} &\sim W_{p-q}\left(\Sigma_{22.1},n-g-q\right).\\ D_{22.1} &\sim W_{p-q}\left(\Sigma_{22.1},g-1\right).\end{aligned}$$

Entonces el cociente

$$L=\frac{|W_{22.1}|}{|T_{22.1}|}\sim\Lambda\left(p-q,n-g-q,g-1\right). \tag{3.58}$$

Cabe señalar, que la variable para la cual el cociente sea más grande tiende a salir del modelo, concluyendo que ésta variable tiene un poder discriminante pobre o nulo.

Si se prueba el poder discriminante para una variable a la vez, es decir, si $(p-q)=1$, se puede relacionar la densidad Λ con una distribución F, siguiendo las propiedades dadas en la Sección (1.5) del Capítulo 1. La transformación que lleva a asociar estas dos densidades es la siguiente:

$$\frac{1-L}{L} \sim \frac{g-1}{n-g-q} F_{(g-1,n-g-q)},$$

$\Leftrightarrow$

$$\frac{(1-L)}{L}\frac{(n-g-q)}{(g-1)} \sim F_{(g-1,n-g-q)}. \tag{3.59}$$

Un caso particular de esta prueba, es cuando se considera el caso de discriminación entre dos normales multivariadas, $N_p(\mu_1,\Sigma)$ y $N_p(\mu_2,\Sigma)$. La *r.d.m.v.* asigna un individuo X a Π_1 si:

$$\alpha^t\left(X-\frac{1}{2}\mu\right)>0.$$

donde $\alpha=\Sigma^{-1}(\mu_1-\mu_2)$ y $\mu=(\mu_1+\mu_2)$.

Sí los parámetros fueron estimados de los datos, los estimadores máximo verosímiles están dados por:

$$\begin{aligned}\hat{\mu}_1 &= \bar{X}_1,\\ \hat{\mu}_2 &= \bar{X}_2,\\ \hat{\Sigma} &= \frac{1}{n}(n_1S_1+n_2S_2).\end{aligned}$$

Entonces la *r.d.m.v.* está dada como:

$$X \text{ se asigna a } \begin{cases}\Pi_1 & \text{si } a^t\left(X-\frac{1}{2}\bar{X}\right)>0,\\ \Pi_2 & \text{en otro caso,}\end{cases}$$

donde $a=W^{-1}\left(\bar{X}_1-\bar{X}_2\right)$ y $\bar{X}=\left(\bar{X}_1+\bar{X}_2\right)$.

Si se tiene la finalidad de encontrar aquellas variables cuyo poder discriminatorio sea significativo, se deben particionar los coeficientes de la función de verosimilitud, este vector se denotará por α y los coeficiente del vector de medias que estará denotado con la letra δ.

Dichas particiones están dadas como:

$$a = \begin{pmatrix} a_{1(q\times 1)} \\ a_{2(p-q)\times 1} \end{pmatrix} \qquad \delta = \begin{pmatrix} \delta_{1(q\times 1)} \\ \delta_{2(p-q)\times 1} \end{pmatrix}.$$

Supóngase que $a_2 = 0$, lo cual significa que las últimas $(p-q)$ componentes no tienen poder discriminatorio, dado que no contribuyen en la evaluación del puntaje discriminante. Para probar la significancia de las variables se consideran tres hipótesis equivalentes (la demostración puede encontrarse por ejemplo en Alvarez 1997), que están dadas como:

$$H_{01} : a_2 = 0.$$

es decir, los coeficientes de $X_{q+1}, ..., X_p$ son cero.

La hipótesis de igualdad de medias condicional

$$H_{02} : \delta_{2.1} = 0$$

con $\delta_{2.1} = \delta_2 - \Sigma_{21}\Sigma_{11}^{-1}\delta_1$, y la hipótesis:

$$H_{03} : \Delta_p^2 = \Delta_q^2.$$

donde

$$\begin{aligned} \Delta_p^2 &= \delta^t\Sigma^{-1}\delta, \\ \Delta_q^2 &= \delta_1^t\Sigma_{11}^{-1}\delta_1, \end{aligned}$$

es decir, la distancia de Mahalanobis entre ambas poblaciones es la misma, si el análisis se hace con las primeras p variables o solamente con las q primeras.

El estadístico de prueba, para la hipótesis $H_{01} : a_2 = 0$ utilizando las distancias de Mahalanobis

$$\begin{aligned} D_p^2 &= md^tW^{-1}d, \\ D_q^2 &= md_1^tW_{11}^{-1}d_1, \end{aligned}$$

en donde $m = n_1 + n_2 - 2$ y $d = \left(\bar{X}_1 - \bar{X}_2\right)$, ha sido propuesta por Rao (1973, p.568), y está definido como:

$$l = \frac{(m-p+1)}{(p-q)} \frac{c^2\left(D_p^2 - D_q^2\right)}{\left(m + c^2 D_q^2\right)},$$

con $c^2 = \frac{n_1 n_2}{n}$, bajo la hipótesis H_{01} la estadística de prueba tiene asociada una distribución F con $(p-q, m-p+1)$ grados de libertad, rechazándose H_{01} para valores grandes de esta estadística.

3.2 Discriminante no Paramétrico

Esta técnica tiene un uso más amplio ya que ningún supuesto distribucional se hace sobre la muestra. La técnica consisten en estimar la función de densidad de la muestra, con base en los estimadores Kernel y una vez que la función de densidad de la muestra ha sido estimada, la clasificación de un cierto individuo se hace maximizando las funciones de verosimilitud como en el caso de discriminación normal.

3.2.1 Criterio de Estimación No Paramétrico

El objetivo de la estimación no paramétrica es diferente al de la estimación paramétrica. En la estimación paramétrica, dada una familia de la densidad $f(\cdot \mid \theta)$, como por ejemplo la familia de la densidad normal con parámetros $\theta = (\mu, \sigma^2)$, la finalidad es obtener el mejor estimador $\hat{\theta}$ de θ. Para el caso no paramétrico, el objetivo primordial es obtener el mejor estimador $\hat{f}(\cdot)$ de la función de densidad $f(\cdot)$ sin suponer ninguna forma funcional o forma paramétrica para $f(\cdot)$.

Estimación de la Función de Densidad Acumulativa

Una función simple para estimar no paramétricamente, es la función de densidad acumulativa ($F.d.a$) correspondiente a alguna $v.a.$ X, que está definida como:

$$F(x) = P[X \leq x].$$

El estimador obvio que proporciona la teoría de la probabilidad elemental, es la función de distribución acumulativa empírica ($F.d.a.e$), que se define por:

$$\begin{aligned} F_n(x) &= \frac{\#\{X_i \leq x\}}{n}, \\ &= \frac{\#X_i \in (-\infty, x]}{n}, \\ &= \frac{1}{n}\sum_{i=1}^{n} I(X_i), \end{aligned} \tag{3.60}$$

donde $\{X_1, ..., X_n\}$ es una *m.a.* de F y

$$I_A(x) = \begin{cases} 1 & \text{si } X \in A, \\ 0 & \text{si } X \notin A. \end{cases}$$

La función $F_n(x)$ definida en (3.60) tiene la forma de una escalera

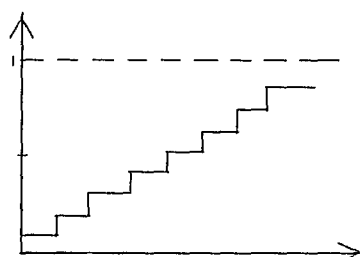


Figura 1: Gráfica de la función de distribución acumulativa.

Es fácil observar que $nF(x)$ tiene excelentes propiedades matemáticas para estimar la función $F(x)$ para todo valor fijo de x,

$$\begin{aligned} E[F_n(x)] &= E\left[I_{(-\infty,x]}(X)\right], \\ &= P[X \in (-\infty, x]], \\ &= F(x), \end{aligned}$$

por lo que $F_n(x)$ es un estimador insesgado de $F(x)$. De hecho, $nF_n(x)$ es una variable Binomial de parámetros n y $p = F(x)$ para cada x fijo, por lo que la $V[F_n(x)] = \frac{p(1-p)}{n}$. Observando que $\sum_{i=1}^{n} I_{(-\infty,x]}(X_i)$ es una estadística suficiente y completa se dice que no existe otro estimador insesgado con varianza mínima.

La definición de la *F.d.a.e* para el caso multivariado es simplemente:

$$F_n(x) = \frac{\#\{X_j \leq x\}}{n} \text{ para } X \in \Re^p,$$

donde la desigualdad $\{X_j \leq x\}$ se aplica a cada una de las entradas del vector x.

3.2.2 Criterio del Error para Densidades Estimadas

Con el deseo de comparar diferentes estimadores e identificar el mejor de ellos se debe elegir un método que proporcione el estimador óptimo. En la estadística paramétrica, un estimador óptimo es aquel que es óptimo para cualquier propósito. Mientras que en el mundo no paramétrico, un estimador puede ser óptimo para algunos propósitos e ineficiente para otros, pero este es lo que ocurre por trabajar con una clase más general de estimadores.

Cuando los parámetros se aproximan mediante estimadores sesgados, el criterio de la varianza comúnmente se reemplaza por el Error Cuadrático Medio (mean squared error MSE), el cual está formado por la suma de la varianza y del sesgo al cuadrado. Sea $X_1, ..., X_n$ una muestra aleatoria de dimensión p de la $f.d.p$ desconocida $f(x)$ y $\hat{f}(x)$ su estimador con base en la muestra. Entonces el MSE se define como:

$$\begin{aligned}
MSE\left[\hat{f}_n(x)\right] &= E\left[\left(\hat{f}_n(x) - f(x)\right)^2\right], \\
&= E\left[\left\{\left(\hat{f}_n(x) - E\left[\hat{f}_n(x)\right]\right) + \left(E\left[\hat{f}_n(x)\right] - f(x)\right)\right\}^2\right], \\
&= E\left[\left(\hat{f}_n(x) - E\left[\hat{f}_n(x)\right]\right)\right]^2 + E\left[\left(E\left[\hat{f}_n(x)\right] - f(x)\right)^2\right] \\
&+ \quad 2E\left\{\left(\hat{f}_n(x) - E\left[\hat{f}_n(x)\right]\right) \times \left(E\left[\hat{f}_n(x)\right] - f(x)\right)\right\} \\
&= \left(E\left[\hat{f}_n^2(x)\right] - E^2\left[\hat{f}_n(x)\right]\right) + \left(E\left[\hat{f}_n(x)\right] - f(x)\right)^2 \\
&+ \quad 2\left(E\left[\hat{f}_n(x)\right] - E\left(\hat{f}_n(x)\right)\right)\left(E\left[\hat{f}_n(x)\right] - f(x)\right) \\
&= Var\left[\hat{f}_n(x)\right] + B^2\left[\hat{f}_n(x), f(x)\right],
\end{aligned}$$

donde $B\left[\hat{f}_n(x), f(x)\right]$ denota el sesgo del estimador $\hat{f}_n(x)$ de $f(x)$. Esta ecuación trata el problema de la estimación no paramétrica de la densidad como un problema de estimación puntual estándar con el parámetro desconocido $\theta = f(x)$. Pero es conveniente considerar

una medida global que proporcione la diferencia entre $\hat{f}(x)$ y $f(x)$. Dicha medida está dada por el *Error Cuadrático Integrado* (*integrated squared error ISE*) y está definida como:

$$ISE = \int_{\Re^p} \left(\hat{f}(x) - f(x) \right)^2 dx.$$

Ya que, el ISE es una *v.a.* complicada que depende de la verdadera función de densidad, del estimador y del tamaño de la muestra, se pueden mantener estas 3 cantidades fijas y el ISE es una función particular de la observación de esos n puntos. Para algunos propósitos será suficiente examinar el promedio del ISE, es decir, la media de la *v.a* ISE conocido como el *Error Cuadrático Integrado Promedio* (*mean integrated squared error*, $MISE$)

$$\begin{aligned} MISE\left[\hat{f}(x)\right] &= E[ISE], \\ &= E\left[\int_{\Re^p} \left(\hat{f}(x) - f(x) \right)^2 dx\right], \\ &= \int_{\Re^p} E\left(\hat{f}(x) - f(x) \right)^2 dx, \\ &= \int_{\Re^p} MSE\left[\hat{f}(x)\right] dx \\ &\equiv IMSE\left[\hat{f}(x)\right]. \end{aligned}$$

El intercambio de la integral y el operador esperanza se justifica por la aplicación del Teorema de Fubini para funciones positivas. La última cantidad es el *Error Cuadrático Medio Integrado* (*integrated mean squared error*, $IMSE$). Por lo tanto, el criterio del error $MISE$ tiene dos interpretaciones equivalentes: es una medida del promedio global del error y es una medida también del error puntual acumulado.

3.2.3 Histogramas

La metodología clásica de los histogramas es utilizada comúnmente para definir la teoría no paramétrica. Dado que un histograma transmite de forma visual la información tanto de las observaciones, esta técnica ha sido utilizada para captar la esencia de la función de densidad.

Los histogramas clásicos de frecuencia están formados por un conjunto de intervalos que no se translapan, donde estos pueden o no tener el mismo ancho. El caso que se analiza en este trabajo es el de los intervalos de la misma longitud. Entonces los histogramas están completamente determinados por dos parámetros: el ancho de banda h y el origen t_0 del primer intervalo, el cual puede elegirse de manera arbitraria en cualquier punto del intervalo. Algunas veces el origen del intervalo se selecciona en $t_0 = 0$.

Criterio de Error para Estimadores no Paramétricos

En esta sección se presentan las propiedades del $MISE$ de la densidad de un histograma. La diferencia entre un histograma de frecuencias y el histograma de una densidad es que este último está normalizado para que integre a 1. Como se dijo anteriormente, un histograma está completamente determinado por la muestra $\{X_1, ..., X_n\}$ de $f(x)$ y la elección de los intervalos $\{t_k, -\infty < k < \infty\}$. Sea $B_k = (t_k, t_{k+1}]$ que denota el $k - ésimo$ intervalo. Supóngase que la diferencia $t_{k+1} - t_k = h$ para toda elección de k, entonces se dice que los intervalos tienen ancho fijo. El histograma de una densidad se construye haciendo cubos de altura $\frac{1}{nh}$, tal que todo bloque posea una área igual a $\frac{1}{n}$. Defínase v_k el número de puntos que caen en el intervalo B_k, entonces el histograma se puede definir como:

$$\begin{aligned} \hat{f}(x) &= \frac{v_k}{nh}, \\ &= \frac{1}{nh} \sum_{i=1}^{n} I_{(t_k, t_{k+1}]} X_i \qquad x \in B_k. \end{aligned}$$

La variable v_k tiene una distribución binomial es decir

$$v_k \sim B(n, p_k),$$

donde

$$p_k = \int_{B_k} f(t)\, dt.$$

Considérese el $MISE$ de $\hat{f}(x)$ para $x \in B_k$. Obsérvese que $E(v_k) = np_k$ y $Var(v_k) = np_k(1 - p_k)$, entonces:

$$\begin{aligned} Var\left(\hat{f}(x)\right) &= \frac{Var(v_k)}{(nh)^2}, \\ &= \frac{p_k(1 - p_k)}{nh^2}, \end{aligned}$$

y el sesgo

$$\begin{aligned} B\left[\hat{f}(x)\right] &= E\left[\hat{f}(x) - f(x)\right], \\ &= \frac{1}{nh} E[v_k] - f(x), \\ &= \frac{p_k}{h} - f(x). \end{aligned} \tag{3.61}$$

Entonces el MSE se puede obtener utilizando las expresiones de sesgo y varianza dadas anteriormente

$$\begin{aligned} MSE\left[\hat{f}(x)\right] &= V\left[\hat{f}(x)\right] + SB\left[\hat{f}(x)\right], \\ &= \frac{p_k(1-p_k)}{nh^2} + \left[\frac{p_k}{h} - f(x)\right]^2, \end{aligned}$$

y esta última expresión no puede simplificarse ya que depende de $f(x)$. Como MSE proporciona una medida puntual de como es la dispersión entre el estimador $\hat{f}(x)$ y $f(x)$ debe considerarse una medida de carácter general, la cual está dada como:

$$IMSE\left[\hat{f}(x)\right] = IV\left[\hat{f}(x)\right] + IB\left[\hat{f}(x)\right],$$

una expansión en series de Taylor de primer orden alrededor de f se obtiene:

$$\begin{aligned} p_k &= \int_{B_k} f(t)\,dt, \\ &= \int_{B_k} \left[f(x) + (t-x) f'(x) + \frac{(t-x)^2}{2} f''(\xi_k) \right] dt \xi_k \in B_k, \\ &= hf(x) + \frac{(t-x)^2}{2} f'(x) \Big|_{t_k}^{t_{k+1}} + \frac{(t-x)^3}{3} f''(\xi_k) \Big|_{t_k}^{t_{k+1}}, \\ &= hf(x) + \frac{f'(x)}{2}\left[-(t_k - x)^2 + (t_k + h - x)^2\right] + o\left(h^3\right). \end{aligned}$$

Entonces el sesgo de $\hat{f}(x)$ puede escribirse utilizando (3.61) como:

$$\begin{aligned} B\left[\hat{f}(x)\right] &= \frac{f'(x)}{2h}\left[2h(t_k - x) + h^2\right] + o\left(h^2\right), \\ &= f'(x)\left[(t_k - x) + \frac{h}{2}\right] + o\left(h^2\right). \end{aligned} \tag{3.62}$$

La varianza integrada puede ser aproximada observando que

$$
\begin{aligned}
IV\left[\hat{f}(x)\right] &= \int_{-\infty}^{\infty} Var\left[\hat{f}(x)\right] dx, \\
&= \sum_{k=-\infty}^{\infty} \int_{B_k} Var\left[\hat{f}(x)\right] dx, \\
&= \sum_{k=-\infty}^{\infty} \int_{B_k} \frac{p_k(1-p_k)}{nh^2} dx. \\
&= \sum_{k=-\infty}^{\infty} \frac{p_k(1-p_k)}{nh}, \\
&= \sum_{k=-\infty}^{\infty} \frac{p_k}{nh} - \sum_{k=-\infty}^{\infty} \frac{p_k^2}{nh}, \\
&= \frac{1}{nh} - \frac{1}{nh} \sum_{k=-\infty}^{\infty} p_k^2.
\end{aligned}
$$

Utilizando el Teorema del valor medio para integrales se deduce que:

$$
\begin{aligned}
\sum_{k=-\infty}^{\infty} p_k^2 &= \sum_k \left[\int_{B_k} f(x)\, dx\right]^2, \\
&= \sum_k [hf(\xi_k)]^2 \quad \xi_k \in B_k, \\
&= \sum_k h^2 f^2(\xi_k), \\
&= h\left[\int f^2(x)\, dx + o(1)\right], \\
&= h[R(f) + o(1)],
\end{aligned}
$$

donde $R(f) = \int_{-\infty}^{\infty} f^2(x)\, dx$, entonces $IV\left[\hat{f}(x)\right]$ se obtiene como:

$$
\begin{aligned}
IV\left[\hat{f}(x)\right] &= \frac{1}{nh} - \frac{1}{nh} \sum_k p_k^2, \\
&= \frac{1}{nh} - \frac{1}{n} R(f) + o\left(n^{-1}\right).
\end{aligned}
\tag{3.63}
$$

Conociendo la expresión del sesgo se puede obtener una medida global de esta cantidad conocida como Sesgo Cuadrático Integrado (integrated squared bias ISB) que es obtenida

elevando al cuadrado la ecuación (3.62), es decir

$$\begin{aligned}
ISB\left[\hat{f}(x)\right] &= \sum_k \int_{B_k} SB\left[\hat{f}(x)\right] dx, \\
&= \sum_k \left\{ \int_{B_k} \left[f'(x)\right]^2 \left[\frac{h}{2} + t_k - x\right]^2 dx + \int_{B_k} 2o\left(h^2\right) f'(x) \left(t_k - x + \frac{h}{2}\right) \right. \\
&\quad \left. + \int_{B_k} \frac{o^2\left(h^2\right)}{h^2} dx \right\}, \\
&= \sum_{-\infty}^{\infty} \left[f'(\eta_k)\right]^2 \int_{B_k} \left(x - t_k - \frac{h}{2}\right)^2 dx + o\left(h^2\right), \\
&= \sum_{-\infty}^{\infty} \left[f'(\eta_k)\right]^2 \frac{\left(x - t_k - \frac{h}{2}\right)^3}{3} \Big|_{t_k}^{t_{k+1}} + o\left(h^2\right), \\
&= \sum_{-\infty}^{\infty} \left[f'(\eta_k)\right]^2 \frac{1}{3} \left[\left(\frac{h}{2}\right)^3 - \left(\frac{-h}{2}\right)^3\right] + o\left(h^2\right), \\
&= \sum_{-\infty}^{\infty} \left[f'(\eta_k)\right]^2 \frac{h^3}{12} + o\left(h^2\right), \\
&= \frac{h^2}{12} \sum_{-\infty}^{\infty} \left[f'(\eta_k)\right]^2 h + o\left(h^2\right), \\
&= \frac{h^2}{12} \left[\int f'(x)\, dx + o(1)\right]^2 + o\left(h^2\right),
\end{aligned}$$

Como $R\left(f'\right) = \int \left[f'(x)\right]^2 dx$ la ecuación anterior se puede escribir como:

$$ISB\left[\hat{f}(x)\right] = \frac{h^2}{12} R\left(f'\right) + o\left(h^2\right). \tag{3.64}$$

Entonces de (3.63) y (3.64) se obtiene el $IMSE$ como:

$$\begin{aligned}
IMSE\left[\hat{f}(x)\right] &= IV\left[\hat{f}(x)\right] + ISB\left[\hat{f}(x)\right], \\
&= \frac{1}{nh} - \frac{1}{n} R(f) + o\left(n^{-1}\right) + \frac{h^2}{12} R\left(f'\right) + o\left(h^2\right), \\
&= \frac{1}{nh} - \frac{1}{n} R(f) + \frac{h^2}{12} R\left(f'\right) + o\left(h^2\right) + o\left(n^{-1}\right), \\
&= \frac{1}{nh} + \frac{h^2}{12} R\left(f'\right) + o\left(h^2\right) + o\left[(nh)^{-1}\right].
\end{aligned}$$

Asintóticamente se puede decir que el $IMSE\left[\hat{f}(x)\right]$ está dado por:

$$AMISE\left[\hat{f}(x)\right] = \frac{1}{nh} + \frac{h^2}{12}R\left(f'\right). \tag{3.65}$$

Teorema 3.10 *Supóngase que f tiene derivada continua y $\left(f'\right)^2$ es integrable. El valor h que minimiza el $AMISE\left[\hat{f}(x)\right]$ es*

$$h^* = \frac{\left[\frac{6}{R(f')}\right]^{\frac{1}{3}}}{n^{\frac{1}{3}}},$$

y el correspondiente $AMISE$ está dado por:

$$\begin{aligned} AMISE^*\left[\hat{f}(x)\right] &= AMISE\left[\hat{f}_{h^*}(x)\right], \\ &= \left(\frac{3}{4}\right)^{\frac{2}{3}} R\left(f'\right)^{\frac{1}{3}} n^{-\frac{1}{3}}. \end{aligned}$$

Demostración

El valor del parámetro h^* se obtiene calculando

$$\frac{\partial AMISE}{\partial h} = -n^{-1}h^{-2} + \frac{1}{6}hR\left(f'\right),$$

igualando a cero se tiene

$$h^* = \left[\frac{6}{nR\left(f'\right)}\right]^{\frac{1}{3}}.$$

Por lo tanto el $AMISE^*$ óptimo se obtiene sustituyendo h^* en (3.65) obteniéndose

$$\begin{aligned} AMISE^*\left[\hat{f}(x)\right] &= \frac{1}{nh^*} + \frac{1}{12}h^{*2}R\left(f'\right), \\ &= \frac{1}{n\left[\frac{6}{nR(f')}\right]^{\frac{1}{3}}} + \frac{1}{12}\left[\frac{6}{nR\left(f'\right)}\right]^{\frac{2}{3}} R\left(f'\right), \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n\left[\frac{6}{R(f')}\right]^{\frac{1}{3}} n^{-\frac{1}{3}}} + \frac{\left[\frac{6}{R(f')}\right]^{\frac{1}{3}} n^{-\frac{1}{3}}}{12} R\left(f'\right), \\
&= R\left(f'\right)^{-\frac{1}{3}} n^{-\frac{1}{3}} \left[\frac{1}{6^{\frac{1}{3}}} + \frac{1}{12} 6^{\frac{2}{3}}\right], \\
&= \left(\frac{3}{4}\right)^{\frac{2}{3}} R\left(f'\right)^{\frac{1}{3}} n^{-\frac{2}{3}}.
\end{aligned}
$$

La generalización de la teoría presentada anteriormente puede consultarse en Scott (1992) pp (80-82).

3.2.4 Histogramas Promediados Corridos

Debido a que cualquier histograma tiene la influencia del punto inicial t_0, esta selección subjetiva puede ser corregidos promediando los histogramas que son obtenidos al hacer variar el punto de origen sobre algún intervalo de longitud h. Esta técnica es conocida en la literatura como Histogramas Promediados Corridos (Average Shifted Histograms ASH, Scott (1992)) y son derivados con el desarrollo que se presenta a continuación.

Construcción

Particiónese la recta real en intervalos $B_k = [kh, (k+1)h)$ cada uno de ellos de longitud h, como se ilustra en la siguiente Figura:

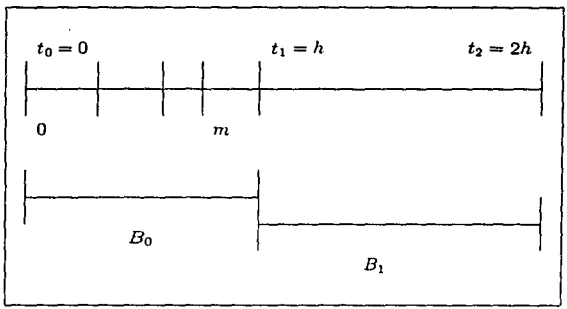


Figura 2. Partición de la recta real en intervalos de longitud h.

Divídase cada intervalo B_k en m intervalos $\widetilde{B}_k$ de longitud $\delta = \frac{h}{m}$, los cuales cumplen la siguiente propiedad:

$$B_k = \cup_{j=1}^{m} \widetilde{B}_k \, .$$

Si se toma, un intervalo cuyo origen esta en $t_0 = 0$ los intervalos $\widetilde{B}_j$están definidos como:

$$\begin{aligned} \widetilde{B}_0 &= [0, \delta) \, , \\ \widetilde{B}_1 &= [\delta, 2\delta) \, , \\ &\vdots \end{aligned}$$

Considérese una colección de m histogramas $\hat{f}_1, ..., \hat{f}_m$ cada uno con intervalo de ancho h, y cuyos orígenes están dados por:

$$h_0 = 0, \frac{h}{m}, \frac{2h}{m}, \cdots, \frac{(m-1)\,h}{m}, \tag{3.66}$$

respectivamente. Esto se puede apreciar en la gráfica siguiente:

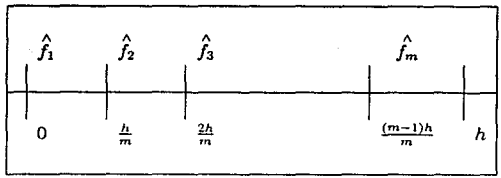


Figura 3. Puntos iniciales en los histogramas promediados corridos.

La función empírica f de los histogramas promediados corridos esta dada como:

$$\begin{aligned} \hat{f}(x) &= \hat{f}_{ASH}(x), \\ &= \frac{1}{m}\sum_{i=1}^{m} \hat{f}_i(x), \end{aligned} \tag{3.67}$$

Se puede observar que el ASH sobre cada intervalo de ancho $\delta_i = \frac{h}{m}$ es constante y que cada origen del intervalo definido como en la ecuación (3.66) difiere precisamente por esta cantidad. Para el caso multivariado la técnica ASH se construye promediando los histogramas corridos, cada uno con intervalo de dimensión $h_1 \times h_2 \times \cdots \times h_p$. Ya que la técnica ASH es constante sobre todos los intervalos $[k\delta, (k+1)\delta)$, es conveniente referirnos a este intervalo reducido como el intervalo $\widetilde{B}_j$, donde

$$\begin{aligned} \nu_k &= \text{la frecuencia en el bin } \widetilde{B}_k, \\ \widetilde{B}_k &= [k\delta, (k+1)\delta). \end{aligned}$$

Considere el ASH estimado en el bin $\widetilde{B}_k$; el peso estimado para este intervalo reducido es el promedio de los pesos de los m histogramas corridos, cada uno con ancho $h = m\delta$. definidos por:

$$\frac{\nu_{k-m+1} + v_{k-m+2} + \cdots + v_k}{nh}, \frac{\nu_{k-m+2} + v_{k-m+3} + \cdots + v_{k+1}}{nh}, \ldots, \frac{\nu_k + v_{k+1} + \cdots + v_{k+m-1}}{nh}.$$

Gráficamente este promedio se puede apreciar en la siguiente Figura:

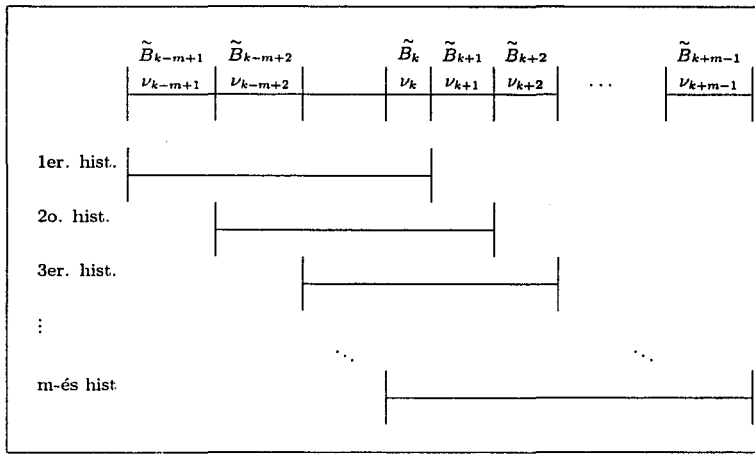


Figura 4. Frecuencias correspondientes al ASH en el intervalo $\widetilde{B}_K$.

Entonces la expresión empírica para el ASH sumando los m pesos de los histogramas es .

$$\begin{aligned}\hat{f}(x,m) &= \frac{\nu_{k-m+1}+v_{k-m+2}+\cdots+v_k}{nh}+\cdots+\frac{\nu_k+v_{k+1}+\cdots+v_{k+m-1}}{nh}, x\in B_k \\ &= \frac{1}{nh}\left(\nu_{k-m+1}+2\nu_{k-m+2}+\cdots+m\nu_k+(m-1)\,\nu_{k+1}+\cdots+2\nu_{k+m-2}+\nu_{k+m-1}\right)\ x\in B_k\end{aligned}$$

Por lo tanto la función empírica de $\hat{f}_h(x;m)$ para el ASH en la ecuación (3.67) está dada por la expresión:

$$\begin{aligned}\hat{f}(x,m) &= \frac{1}{m}\left[\frac{1}{nh}\sum_{i=1-m}^{m-1}(m-|i|)\,\nu_{k+i}\right], \qquad (3.68)\\ &= \frac{1}{nh}\sum_{i=1-m}^{m-1}\left[1-\frac{|i|}{m}\right]\nu_{k+i}\ x\in\widetilde{B}_k\ .\end{aligned}$$

Los pesos en la frecuencia del intervalo en la ecuación (3.68) forman un triángulo isosceles con base en el intervalo $(-1, 1)$.

La frecuencia sobre todos los intervalos está dada por:

$$\hat{f}(x, m) = \frac{1}{nh} \sum_{-\infty}^{\infty} I_{\{1-m, m-1\}}(i) \left[1 - \frac{|i|}{m}\right] \nu_{k+i}, \; x \in \widetilde{B}_k$$

y

$$\lim_{m\to\infty} \hat{f}(x) = \frac{1}{nh} \sum_{i=-\infty}^{\infty} \lim_{m\to\infty} I_{\{1-m, m-1\}}(i) \left[1 - \frac{|i|}{m}\right] \nu_{k+i}.$$

Cuando los parámetros h y n están fijos y m incrementa, es fácil observar que un punto X_j contribuye en la función $\hat{f}(x)$ estimada por el ASH, cuando $x \in B_j$ y $X_j \in B_{j+i}$, si y sólo si

$$(|i| - 1)\,\delta \leq |x - X_j| \leq (|i| + 1)\,\delta \quad \forall\, i$$
$$(|i| - 1) \leq \frac{|x - X_j|}{\delta} \leq (|i| + 1) \quad \forall\, i.$$

Desarrollando la desigualdad anterior se obtiene:

$$\frac{|x - X_j|}{\delta} - 1 \leq |i| \leq \frac{|x - X_j|}{\delta} + 1, \tag{3.69}$$

recordando que $\delta = \frac{h}{m}$ y utilizando las propiedades del valor absoluto se tiene (3.69) ocurre si y sólo si

$$\left| |i| - \frac{|x - X_j|}{\delta} \right| \leq 1,$$

lo cual es equivalente a

$$\left| |i| - \frac{m\,|x - X_j|}{h} \right| \leq 1,$$

equivalentemente se tiene

$$\left| \frac{|i|}{m} - \frac{|x - X_j|}{h} \right| \leq \frac{1}{m}.$$

La igualdad anterior, implica que

$$\lim_{m\to\infty} \frac{|i|}{m} = \frac{|x - X_j|}{h},$$

por lo que

$$\lim_{m\to\infty} \left| \frac{m - |i|}{m} \right| = \left| 1 - \frac{|x - x_j|}{h} \right|.$$

Entonces

$$\begin{aligned}\hat{f}(x) &= \lim_{m\to\infty} \hat{f}(x,m), \\ &= \frac{1}{nh}\sum_{i=1}^{n}\left|1-\frac{|x-X_j|}{h}\right| I_{[-1,1]}\left(\frac{x-X_i}{h}\right).\end{aligned}$$

Finalmente, la estimación puede escribirse como:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_j}{h}\right).$$

donde

$$K_h(\xi) = \frac{1}{h}K\left(\frac{\xi}{h}\right), \tag{3.70}$$

la ecuación (3.70) se conoce en la Literatura como el *Kernel Triangular*. La generalización de la estimación anterior al caso multivariado esta dada por:

$$\hat{f}(X) = \frac{1}{nh_1\cdots h_p}\sum_{i=1}^{n}\left[\prod_{j=1}^{p} K\left(\frac{x_j - x_{ij}}{h_j}\right)\right],$$

donde $X^t = (X_1, ..., X_p) \in \Re^p$ cada h_j $j = 1, ..., p$ es el parámetro de suavizamiento en la $j-$*ésima* dirección y $X_i^t = (X_{i1}, ..., X_{ip})$ define la muestra aleatoria. Este desarrollo puede encontrarse en Scott (1992) p. 123.

Un análisis similar al presentado en la Sección 3.2.3 puede hacerse para estimar el $AMISE$ en la técnica ASH . Scott (1992) presenta el siguiente resultado

Teorema 3.11 *Para un ASH con kernel triángular el $AMISE$ esta dado por:*

$$AMISE = \frac{2}{3nh}\left(1+\frac{1}{2m^2}\right) + \frac{h^2}{12m^2}R\left(f'\right) + \frac{h^4}{144}\left(1-\frac{2}{m^2}+\frac{3}{5m^4}\right)R\left(f''\right).$$

Demostración. Ver Scott (1992) p.119.

El ancho de intervalo óptimo para un ASH está dado por la expresión:

$$h^* = \left(\frac{24}{nR(f'')}\right)^{\frac{1}{5}}, \tag{3.71}$$

cuando el número de histogramas que se promedian tiende a infinito, i.e., siempre que $m \longrightarrow \infty$. En particular cuando se utiliza la densidad $N(\mu, \sigma^2)$ para aproximar el valor h^* en (3.71) el óptimo está dado por:

$$h^* = 2.576\sigma n^{-\frac{1}{5}}. \tag{3.72}$$

Si se desea hacer un análisis mas profundo de esta técnica se sugiere ver Scott (1992).

Ejemplo 3.1 *Para ilustrar la técnica ASH (Histogramas promediados corridos) se utilizó la muestra de los Irises de Fisher (ver Tabla B1 del Apéndice) y se consideró únicamente la variable Sepallen de la especie Setosa, la cual consta de 50 observaciones. El análisis fue realizado en el paquete S-Plus para obtener las gráficas del promedio de los histogramas y se utilizó (3.72) para aproximar el ancho de banda óptimo. En este caso,* $\hat{\sigma} = s = 0.3525$ *y* $n = 50$, *obteniéndose de este modo*

$$h^* = (2.576)\,(0.3525)\,[(50)]^{-\frac{1}{5}} = 0.41.$$

El origen para todos los intervalos se dio en $t_0 = 3.8432$, *y el intervalo final en* $t_m = 6.2568$

Los histogramas para $m = 1, 5, 10, 50$ *se presentan en las gráficas 1 a la 4.*

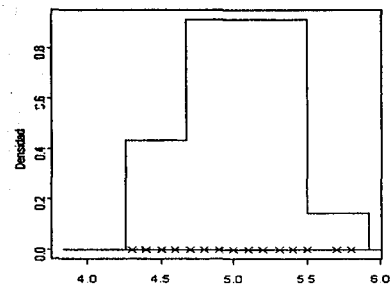


Figura 1. Histograma ASH con $m = 1$ de la variable Sepallen de la especie Setosa de los Irises de Fisher.

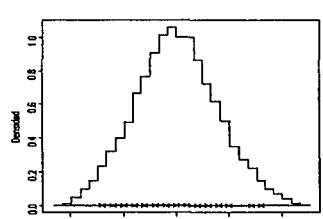


Figura 2. Histograma ASH con $m = 5$ de la variable Sepallen de la especie Setosa para los Irises de Fisher.

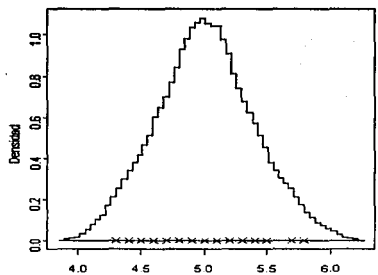


Figura 3. Histograma ASH con $m = 10$ de la variable Sepallen de la especie Setosa de los Irises de Fisher.

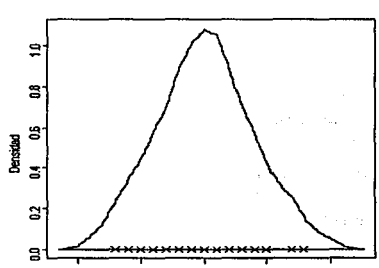


Figura 4. Histograma ASH con $m = 50$ de la variable Sepallen de la especie Setosa de los Irises de Fisher.

Como puede observarse de estas gráficas, el incremento en el suavizamiento de la estimación es notorio conforme m crece, donde la estimación con $m = 50$ produce una curva bastante suave.

3.2.5 Densidad Estimada del Kernel

Es conveniente destacar que la investigación relacionada con la estimación no paramétrica, fue desarrollada hasta 1950, cuando la teoría paramétrica de estimación de densidades ya era amplia. Fix y Hodges (1951) introdujeron el algoritmo básico para estimar no paramétricamente una densidad y ellos fueron quienes encaminaron el problema de la discriminación estadística cuando la forma de la densidad de la muestra no era conocida. Durante la siguiente década, aparecieron varios algoritmos generales y teorías alternativas, las cuales fueron introducidas por Rosenblatt (1956), Parzen (1962) y Cencov (1962). Las investigaciones posteriores fueron realizadas por Watson y Leadbetter (1963), Loftsgaarden y Queseberry (1965), Schwartz (1967), Epanechnikov (1969), Tarter y Kronmal (1970) y Wahaba (1971). La generalización multivariada fue introducida por Cacoullos (1966). Finalmente en 1970, llegaron los primeros escritos sobre aplicaciones prácticas de estos métodos y que fueron realizados por Scott et al. (1978) y Silverman (1978).

La estimación por medio del Kernel promediado es flexible en cuanto a la forma y al desarrollo matemático. Como analogía al caso de los histogramas, se tiene un parámetro de suavizamiento el cual regula el grado de ajuste en estos suavizadores del Kernel, llamados anchos de banda h. El Kernel básico estimado puede ser escrito en forma compacta como

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right),$$

con $K(\cdot)$ la función de densidad. Promediando con respecto a estas funciones de Kernel, se obtiene

$$\begin{aligned} \hat{f}(x) &= \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i), \\ &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \end{aligned} \tag{3.73}$$

Como la función $K(\cdot)$ es una densidad se cumple:

$$\int_{-\infty}^{\infty} \hat{f}(x)\,dx = \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K\left(\frac{x-X_i}{h}\right)dx,$$

haciendo $y = \frac{x-X_i}{h}$ se sigue que:

$$\begin{aligned}\int_{-\infty}^{\infty} \hat{f}(x)\,dx &= \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K(y)\,h dy,\\ &= 1.\end{aligned}$$

Criterio del Error en las Densidades Estimadas

El análisis estadístico de los estimadores *Kernel* es mucho mas sencillo que el de los histogramas, ya que el estimador *Kernel* en la ecuación (3.73) es la media aritmética de las n *v.a.i.i.d*, $X_1, ..., X_n$ con densidad común $f(x)$. Definiendo

$$K_h(x, X_i) = \frac{1}{h}K\left(\frac{x-X_i}{h}\right).$$

Entonces

$$E\left[\hat{f}(x)\right] = E\left[K_h(x, X)\right],$$

y

$$Var\left[\hat{f}(x)\right] = \frac{1}{n}Var\left[K_h(x, X)\right],$$

donde $X \sim f(x)$. El valor esperado está dado por:

$$E\left[\hat{f}(x)\right] = \int_{-\infty}^{\infty}\frac{1}{h}K_h(x, t)\,f(t)\,dt,$$

tomando $w = \frac{x-t}{h}$ se tiene que $t = x - hw$ y se sigue

$$\begin{aligned}E\left[\hat{f}(x)\right] &= \int_{-\infty}^{\infty}\frac{1}{h}K(w)\,f(x-hw)\,h dw, \qquad (3.74)\\ &= \int_{-\infty}^{\infty}K(w)\,f(x-hw)\,dw.\end{aligned}$$

Expandiendo en series de Taylor $f(x-hw)$ alrededor de x se obtiene

$$f(x-hw) = f(x) + f'(x)(-wh) + \frac{f''(x)}{2}(-wh)^2 + \frac{f'''(x)}{3}(-wh)^3 + \cdots \qquad (3.75)$$

Entonces la ecuación (3.74) se puede simplificar como:

$$E\left[\hat{f}(x)\right] = \int_{-\infty}^{\infty} K(w)\left[f(x) + f'(x)(-wh) + \frac{f''(x)}{2}(-wh)^2 + \frac{f'''(x)}{3}(-wh)^3 + \cdots\right] dw \tag{3.76}$$

Si $\int K(w) = 1$, $\int wK(w) = 0$, $\int w^2 K(w) = \sigma_k^2 > 0$ y $\left|\int w^k K(w)\, dw\right| < \infty \ \forall k \in \mathbf{N}$, entonces la expresión (3.76) se simplifica como

$$E\left[\hat{f}(x)\right] = f(x) + \frac{h^2}{2} f''(x)\,\sigma_k^2 + o\left(h^2\right), \tag{3.77}$$

y por lo tanto, el sesgo se puede expresar como:

$$\begin{aligned} B\left[\hat{f}(x)\right] &= E\left[\hat{f}(x)\right] - f(x), \\ &= \frac{h^2}{2} f''(x)\,\sigma_k^2 + o\left(h^2\right). \end{aligned}$$

Por otro lado el sesgo cuadrático integrado puede calcularse como:

$$\begin{aligned} ISB\left[\hat{f}(x)\right] &= \int_{-\infty}^{\infty} B^2\left[\hat{f}(x)\right] dx, \\ &= \int_{-\infty}^{\infty} \left[\frac{h^2}{2} f''(x)\,\sigma_k^2 + o\left(h^2\right)\right]^2 dx, \\ &= \int_{-\infty}^{\infty} \left[\frac{h^4}{4}\left[f''(x)\right]^2 \sigma_k^4 + o\left(h^2\right) h^2 f''(x)\,\sigma_k^2 + o^2\left(h^2\right)\right] dx, \\ &= \frac{h^4\sigma_k^4}{4}\int_{-\infty}^{\infty}\left[f''(x)\right]^2 dx + h^2\sigma_k^2\int_{-\infty}^{\infty} o\left(h^2\right) f''(x)\, dx \\ &\quad + \int_{-\infty}^{\infty} \sigma_k^2\left(h^2\right) dx, \end{aligned}$$

de donde finalmente

$$ISB\left[\hat{f}(x)\right] = \frac{h^4}{4}\sigma_k^4 R\left(f''\right) + o\left(h^4\right). \tag{3.78}$$

Para esta muestra aleatoria la varianza se calcula como:

$$\begin{aligned} Var\left[\hat{f}(x)\right] &= \frac{1}{n} Var\left[K_h(x, X)\right] X \sim f(x), \\ &= \frac{1}{n}\left\{E\left[\frac{1}{h}K\left(\frac{x-t}{h}\right)\right]^2 - E^2\left[\frac{1}{h}K\left(\frac{x-t}{h}\right)\right]\right\}, \\ &= \frac{1}{n}\left\{\frac{1}{h^2}K^2\left(\frac{x-t}{h}\right) f(t)\, dt - \left[f(x) - o(h)\right]^2\right\}, \end{aligned}$$

donde esta igualdad se satisface utilizando (3.77). Haciendo el cambio de variable $w = \frac{x-t}{h}$ se tiene $t = x - hw$ la ecuación anterior se puede escribir como:

$$Var\left[\hat{f}(x)\right] = \frac{1}{n}\left\{\frac{1}{h^2}\int_{-\infty}^{\infty} K^2(w) f(x-hw)\, h dw - [f(x) + o(h)]^2\right\},$$

y expandiendo $f(x-hw)$ como en (3.75) se tiene

$$\begin{aligned}
Var\left[\hat{f}(x)\right] &= \frac{1}{n}\left\{\frac{1}{h}\int_{-\infty}^{\infty} K^2(w)\left[f(x) - hwf'(x) + \frac{h^2w^2}{2}f''(x)\right]dw - [f(x) - o(h)]^2\right\}, \\
&= \frac{1}{n}n\left\{\frac{f(x)}{h}R(K) + f'(x)\int_{-\infty}^{\infty} wK^2(w)\,dw + o(1) - [f(x) - o(h)]^2\right\}, \\
&= \frac{1}{nh}f(x)\,R(K) + \frac{1}{n}f'(x)\int_{-\infty}^{\infty} wK^2(w)\,dw + \frac{1}{n}o(1) - \frac{1}{n}[f(x) - o(h)]^2, \\
&= \frac{1}{nh}f(x)\,R(K) + o\left([nh]^{-1}\right).
\end{aligned}$$

Entonces se puede calcular el $IV\left[\hat{f}_h(x)\right]$ con base en la expresión anterior como:

$$\begin{aligned}
IV\left[\hat{f}(x)\right] &= \int_{-\infty}^{\infty}\left\{\frac{1}{nh}f(x)\,R(K) + o\left([nh]^{-1}\right)\right\}dx, \qquad (3.79) \\
&= \frac{R(K)}{nh} + o\left([nh]^{-1}\right).
\end{aligned}$$

De las ecuaciones (3.78) y (3.79) se puede obtener el $IMSE\left[\hat{f}_h(x)\right]$ de la siguiente manera:

$$\begin{aligned}
IMSE\left[\hat{f}_h(x)\right] &= ISB\left[\hat{f}_h(x)\right] + IV\left[\hat{f}_h(x)\right], \\
&= \frac{h^4}{4}\sigma_k^4 R\left(f''\right) + o\left(h^4\right) + \frac{R(K)}{nh} + o\left([nh]^{-1}\right).
\end{aligned}$$

Asintóticamente se puede decir que el $IMSE\left[\hat{f}_h(x)\right] = MISE\left[\hat{f}_h(x)\right]$ está dado por:

$$AMISE\left[\hat{f}_h(x)\right] = \frac{R(K)}{nh} + \frac{h^4}{4}\sigma_k^4 R\left(f''\right),$$

el resultado que se establece en el siguiente Teorema.

Teorema 3.12 *Para un estimador de Kernel*

$$AMISE\left[\hat{f}_h(x)\right] = \frac{R(K)}{nh} + \frac{h^4}{4}\sigma_k^4 R\left(f''\right),$$

y el parámetro de suavizamiento óptimo

$$h^* = \left[\frac{R(K)}{\sigma_k^4 R(f'')}\right]^{\frac{1}{5}} n^{-\frac{1}{5}},$$

proporciona

$$AMISE_{h^*}\left[\hat{f}(x)\right] = \frac{5}{4}\left[\sigma_k R(K)\right]^{\frac{4}{5}} R\left(f''\right)^{\frac{1}{5}} n^{-\frac{4}{5}}.$$

Los resultados sobre estimación de densidades para el caso multivariado de los estimadores de Kernel pueden consultarse en Scott (1992).

3.3 Mapas Territoriales.

Como una opción a las técnicas multivariadas en donde lo que se desea es asignar un nuevo individuo a alguna de las g poblaciones que se tienen como alternativa, se presenta en esta sección un procedimiento gráfico conocido en la literatura como Mapas Territoriales. Estos mapas tienen como finalidad facilitar la asignación de nuevas observaciones evitando los cálculos laboriosos basados en la evaluación de las funciones de clasificación, particularmente cuando el número de variables es elevado.

Dado que la asignación basada en las g funciones de clasificación (lineales o cuadráticas) es similar a la obtenida utilizando las funciones de clasificación canónica significativas, resulta conveniente elaborar los mapas territoriales con base en estas últimas dado que los cálculos resultan más simples debido a la reducción de dimensión de las variables de clasificación.

La idea consiste en que dado un nuevo individuo con vector de atributos X, se obtiene el vector de puntajes canónicos asociados y el individuo se asigna a la población para la cual este vector está más cerca del correspondiente vector de medias canónicas. La utilidad de los mapas es que dado el vector de puntajes canónicos sólo debe localizarse en el mapa el símbolo de la población en la que debe clasificarse el individuo. Estos mapas

se construyen haciendo una retícula en dos o tres dimensiones (que generalmente es el número de funciones canónicas significativas en la práctica) y clasificando cada punto de la retícula en la población correspondiente y graficando el símbolo en el mapa territorial.

Ejemplo 3.2 *Utilizando la muestra de los Irises de Fisher (ver Tabla B1 del Apéndice B) se hizo el análisis discriminante para obtener las funciones de clasificación canónicas no estandarizadas para ejemplificar el uso de los mapas territoriales.*

La muestra consiste en tres especies de flor de Iris: Iris Setosa, Iris Versicolor e Iris Virgínica las cuales se denotarán en el mapa como:

$$\begin{aligned} Setosa &\rightarrow 1 \\ Versicolor &\rightarrow 2 \\ Virgínica &\rightarrow 3 \end{aligned}$$

Y se observó 4 características (variables) en la muestra:

Petallen,

Petalwid,

Sepallen,

Petalwid.

Las funciones de clasificación canónicas no estandarizada asociadas a la muestra están dadas en la siguiente Tabla:

Varible	*Función 1*	*Función 2*
Petallen	*2.2012*	*-.9319*
Petalwid	*2.8104*	*2.8391*
Sepallen	*-.8294*	*0.0241*
Sepalwid	*-1.5344*	*2.1645*
Constante	*-2.1051*	*-6.6614*

Tabla 1. Funciones de clasificación canónicas no estandarizadas

Las combinaciones lineales mediante las que se obtienen las nuevas variables están dadas por:

$$\begin{aligned} Y_1 &= 2.2012X_1 + 2.8104X_2 - 0.8294X_3 - 1.5344X_4 - 2.1051, \\ Y_2 &= -0.9319X_1 + 2.8391X_2 + 0.0241X_3 + 2.1645X_4 - 6.6614. \end{aligned}$$

Entonces para un individuo con vector de atributos $X = (6.4, 2.8, 5.6, 2.2)$, *las funciones canónicas no estandarizadas transforman este individuo en el punto*

$$Y = (11.8326, 0.2209).$$

Este punto se localiza en el Mapa Territorial y se observa que ha sido clasificado en la población 3, la cual corresponde a la Especie Virgínica. Es claro, que el orden en el que se introducen las variables debe coincidir con el orden que establece el paquete sobre las variables.

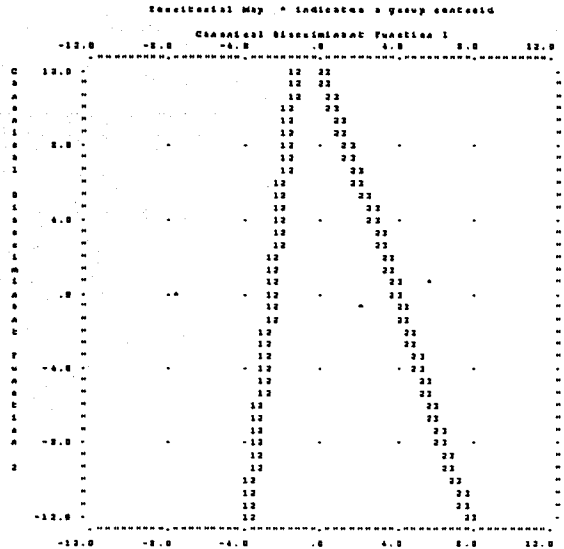


Figura 8. Mapas Territoriales para la muestra de las Irises.

Es fácil observar que las posibles malas clasificaciones están dadas en las fronteras de cada región.

Capítulo 4

DISCRIMINANTE PARAMETRICO vs NO PARAMETRICO

Este capítulo tiene la finalidad de hacer una comparación entre los métodos de discriminación normal y no paramétrico. La comparación se lleva a cabo con el fin de mostrar que los métodos no paramétricos producen resultados similares a los obtenidos en el caso de discriminar poblaciones normales, mientras que la estimación no paramétrica puede producir resultados muy superiores a los que se obtienen con base en la clasificación normal cuando se supone que los datos no siguen una distribución normal. El análisis utiliza cuatro muestras en donde tres de ellas fueron simuladas mediante el paquete estadístico S-Plus y fueron analizadas en el paquete SAS. En la última comparación se utiliza la muestra de los Irises de Fisher y de manera adicional se presentan las gráficas de las probabilidades a posteriori para esta muestra, las cuales servirán para la clasificación de un nuevo individuo.

En esta parte se reproducen algunos de las resultados que son obtenidos en el paquete SAS como son: las tablas de mala clasificación y gráficas de las densidades correspondientes a cada una de las muestras.

4.1 Simulaciones

En este apartado se analizan las muestra que fueron simuladas mediante el paquete S-Plus y que corresponden a los archivos Datosn1.txt, Datosn2.txt, Datosn3.txt y adicionalmente se contempla la muestra de los Irises de Fisher. Para los cuatro casos se hace un análisis de discriminación normal, con base en las funciones de discriminación lineal y las funciones de discriminación cuadráticas, además se obtiene el análisis utilizando los estimadores Kernel, con base en la distribución Normal. Se debe hacer notar que en la discriminación de las observaciones se supone que inicialmente son igualmente probables.

En la muestra de los Irises de Fisher se presentan las gráficas de las probabilidades posteriores que determinan las regiones de clasificación de las especies de Iris.

4.1.1 Simulación 1

En este ejemplo se consideran dos poblaciones bimodales univariadas con 200 observaciones cada una, teniendo una muestra total de dimensión 400×1 observaciones. La función de densidad de cada población se obtuvo utilizando las siguientes combinaciones:

Para la población correspondiente al Grupo 1 su densidad es

$$f_X(x) = 0.4N(x_i; 0, 1) + 0.6N(x_i; 8, 1),$$

donde $N(x; \mu, \sigma^2)$ representa la densidad de una variable $N(\mu, \sigma^2)$ evaluada en x. La densidad de las muestras se presenta en la gráfica siguiente:

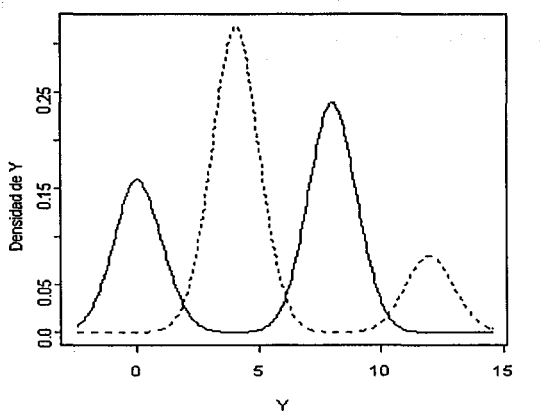


Figura 1. Gráfica de las densidades para la muestra simulada 1.

En la siguiente gráfica puede observarse también el comportamiento de la muestra, mediante un histograma de frecuencias:

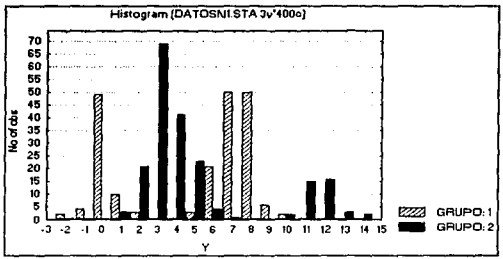


Figura 2. Histograma de frecuencias de los grupos de la Simulación 1.

Bajo el supuesto de normalidad en la población y considerando las funciones de discriminación lineal, se obtiene el siguiente cuadro de clasificación

Grupo	1	2	Total
1	69	131	200
	34.50	65.50	100.00
2	146	54	200
	73.00	27.00	100.00
Total	215	185	400
	53.75	46.25	100.00

Tabla 1. Porcentajes de clasificación por Grupo para la simulación 1

De la Tabla anterior puede observarse que de los 400 individuos en muestra 277 de ellos se encuentran mal clasificados, por lo que es de suponerse que la tasa de clasificación incorrecta para este análisis tenderá a ser muy alta. Estas tasas de error pueden observarse en el cuadro siguiente:

	1	2	Total
Proporción	0.6550	0.7300	0.6925
Prob. apriori	0.5000	0.5000	

Tabla 2. Error Estimado por Grupo

Obsérvese de la Tabla 2, que la proporción total del error total es de 0.6925, lo cual significa que casi el 70% de las observaciones están mal clasificadas. Por lo tanto, dadas las características de la muestra, la discriminación bajo el supuesto de normalidad en las poblaciones y considerando una matriz de covarianzas común no proporciona resultados satisfactorios, ya que el número de traslapes es considerable.

La discriminación cuadrática es obtenida al suponer normalidad en la muestra y considerando que los grupos tienen matriz de covarianzas distinta. Con base en el método anterior se obtiene el siguiente cuadro de clasificación:

Grupo	1	2	Total
1	69	131	200
	34.50	65.50	100.00
2	57	143	200
	28.50	71.50	100.00
Total	126	274	400
	31.50	68.50	100.00

Tabla 3. Porcentajes de clasificación por Grupo para la simulación 1

La cual manifiesta que 131 individuos del grupo 1 se traslaparon en el grupo 2 y 57 individuos pertenecientes al grupo 2 fueron clasificados en el grupo 1, obteniendo así 188 observaciones mal clasificadas de un total de 400 individuos, lo cual implica que el porcentaje de error nuevamente es alto. En la Tabla que se presenta a continuación se puede observar la proporción de error global y por grupo:

	1	2	Total
Proporción	0.6550	0.2850	0.4700
Prob. apriori	0.5000	0.5000	

Tabla 4. Error Estimado por Grupo y Error Total

En la Tabla 4 se observa que el error total es del 47%, lo cual quiere decir que casi la mitad de las observaciones se encuentran mal clasificados.

Adicionalmente se obtiene el análisis de acuerdo a la teoría de los estimadores Kernel, el cual utiliza la densidad Normal para aproximar la densidad de la muestra simulada. Dado que las dos poblaciones son bimodales y se encuentran traslapadas se espera que el

método Kernel estime adecuadamente la densidad de las poblaciones y se obtengan mejores resultados en la clasificación, en el sentido de que las tasas de error sean mínimas.

En la siguiente Tabla se pueden apreciar los resultados de la clasificación obtenida mediante el Kernel Normal,

Grupo	1	2	Total
1	189	11	200
	94.50	5.50	100.00
2	5	195	200
	2.50	97.50	100.00
Total	194	206	400
	48.50	51.50	100.00

Tabla 5. Porcentajes de clasificación por grupos en la simulación 1

Nótese que muy pocas observaciones en ambos grupos fueron clasificados incorrectamente, teniendo 16 individuos mal clasificados de un total de 400, lo cual no induce a pensar que los porcentajes de error en este caso deben ser pequeños. Los resultados de las tasas de error para este método están contenidas en la siguiente Tabla:

	1	2	Total
Proporción	0.0550	0.0250	0.0400
Prob. apriori	0.5000	0.5000	

Tabla 6. Error Estimado por Grupo y Error Total

Como puede verse de la Tabla 6, con el Kernel Normal se tiene un error total del 4%, lo cual significa que sólo una proporción pequeña de individuos se encuentran mal clasificados.

Finalmente con base en el análisis anterior se puede concluir que cuando la muestra observada no cumple con los supuestos de normalidad, la discriminación lineal o cuadrática proporciona tasas de error altas mientras que cuando se estima la densidad de la muestra mediante métodos no paramétricos los resultados obtenidos son superiores en el sentido de que son más precisos en la clasificación.

4.1.2 Simulación 2

Para esta parte se simularon dos muestras Normales bivariadas, independientes con matriz de covarianzas común a ambas muestras, en donde cada muestra está asociada con una matriz de dimensión 150×2. El objetivo en este punto es que dadas las características de la muestra, se obtengan los mejores resultados utilizando la discriminación lineal, y poderla comparar con la discriminación cuadrática y los estimadores Kernel.

La población correspondiente al grupo 1 tiene distribución Normal de parámetros:

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{y} \quad \Sigma = \begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}$$

y la población asociada al grupo 2 es Normal de parámetros:

$$\mu = \begin{pmatrix} 1.7 \\ 1.7 \end{pmatrix} \quad \text{y} \quad \Sigma = \begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}.$$

La siguiente gráfica proporciona la dispersión que siguen las dos poblaciones:

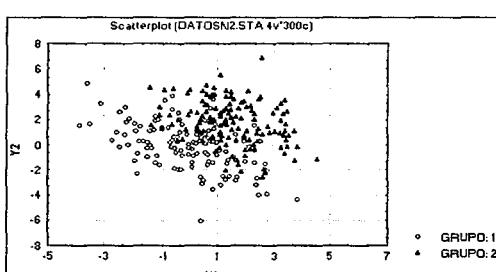


Figura 3. Dipersión de los dos grupos correspondientes a la Simulación 2.

Bajo normalidad en la muestra y usando el puntaje lineal se obtiene la siguiente Tabla:

Grupo	1	2	Total
1	130	20	150
	86.67	13.33	100.00
2	29	121	150
	19.33	80.67	100.00
Total	159	141	300
	53.00	47.00	100.00

Tabla 7. Porcentaje de clasificación por grupos para la simulación 2

Puede observarse de la Tabla 7 que 49 individuos se encuentran mal clasificados de un total de 300, aunque los traslapes no son numerosos, se esperaba obtener mejores resultados dadas las condiciones en las que la muestra fue obtenida. Los errores de clasificación se pueden visualizar de la siguiente Tabla:

	1	2	Total
Proporción	0.1333	0.1933	0.1633
Prob. apriori	0.5000	0.5000	

Tabla 8. Error Estimado por Grupo y Error Total

De esta tabla se puede observar que el porcentaje de clasificación incorrecta bajo el método de discriminación lineal es del 16.33%. Debido a que las poblaciones comparten la misma matriz de covarianzas se puede pensar que la tasa de error no es la satisfactoria.

Si las poblaciones no comparten la misma matriz de covarianzas, se utiliza la discriminación cuadrática. Los resultados se dan a continuación:

Grupo	1	2	Total
1	132	18	150
	88.00	12.00	100.00
2	25	125	150
	16.67	83.33	100.00
Total	157	143	300
	52.33	47.67	100.00

Tabla 9. Porcentaje de clasificación por grupos para la simulación 2

Obsérvese de la Tabla 9 que el 18 observaciones pertenecientes al grupo 1 fueron asignadas a la población 2 y 25 elementos del grupo 2 se traslaparon en el grupo 1, resultando 43 individuos mal clasificados de un total de 300. Nótese que esta estimación es similar a la obtenida con base en la discriminación lineal. Las tasas de error están dadas a continuación

	1	2	Total
Proporción	0.1200	0.1667	0.1433
Prob. apriori	0.5000	0.5000	

Tabla 10. Error Estimado por Grupo y Error Total

La Tabla 10 muestra un error total del 14.33%, el cual es muy parecido al obtenido en la discriminación lineal, sin embargo las diferencias en las tasas de error en la clasificación a simple vista parecen no ser significativas.

En suma a los métodos utilizados anteriormente se analiza esta muestra por medio de los estimadores Kernel ajustando la densidad de la muestra mediante la densidad Normal, suponiéndose un ancho de intervalo igual en las dos direcciones del plano (ya que la muestra contempla dos variables), y considerando también el caso en el que los anchos de los intervalos son diferentes en las dos direcciones.

De acuerdo al método de Kernel Normal cuando el ancho del intervalo es el mismo en ambas direcciones se producen los siguientes resultados en la clasificación que están concentrados en la Tabla que a continuación se cita:

Grupo	1	2	Total
1	133	17	150
	88.67	11.33	100.00
2	29	121	150
	19.33	80.67	100.00
Total	162	138	300
	54.00	46.00	100.00

Tabla 11. Porcentajes de clasificación por Grupos para la simulación 2

Como se observa de la Tabla 11, 47 individuos se encuentran mal clasificados, de los cuales 17 pertenecen al grupo 1 y 29 al grupo 2. Nótese también que estos resultados son muy parecidos a los obtenidos por los métodos anteriores.

Las porcentajes de error en la clasificación que fueron obtenidos mediante el método de Kernel Normal están dados en la Tabla que a continuación se presenta:

	1	2	Total
Proporción	0.1133	0.1933	0.1533
Prob. apriori	0.5000	0.5000	

Tabla 12. Error Estimado por Grupo y Error Total

De esta Tabla se observa que la proporción del error total es del 0.1533, lo cual significa que el 15.33% de los individuos de la muestra fueron clasificados incorrectamente.

Se debe aclarar que el análisis de la muestra simulada con base en el Método de Kernel Normal utilizando distintos anchos de intervalo queda omitido, ya que los resultados obtenidos en la clasificación y en los porcentajes de error son los mismos a los obtenidos por el método del Kernel Normal con el mismo ancho de intervalo.

Con base en los métodos utilizados anteriormente se puede concluir que los resultados en la clasificación son muy parecidos y que las tasas de error en cada uno de ellos no manifiestan grandes diferencias. Dado que el Kernel Normal hace una buena aproximación de la densidad asociada a las muestras sus resultados pueden ser comparables con los obtenidos al utilizar los métodos paramétricos.

4.1.3 Simulación 3

Para este caso se simularon dos muestras normales bivariadas cada una de dimensión 150×2, formando una muestra total de dimensión 300×2.

La población correspondiente al grupo 1 es Normal de parámetros:

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{y} \quad \Sigma = \begin{pmatrix} 3 & -3.7 \\ -3.7 & 5 \end{pmatrix}.$$

y la población asociada al grupo 2 es normal con parámetros:

$$\mu = \begin{pmatrix} 0.25 \\ 0.25 \end{pmatrix} \quad \text{y} \quad \Sigma = \begin{pmatrix} 3 & 3.7 \\ 3.7 & 5 \end{pmatrix}.$$

La dispersión de la muestra simulada puede apreciarse en la siguiente figura:

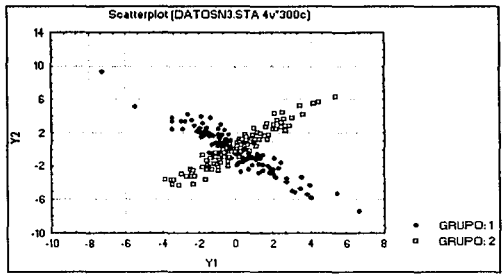


Figura 4. Dispersión de la muestra de la Simulación 3.

El objetivo en esta sección dado que las poblaciones son normales y tienen distinta matriz de covarianzas es comparar los resultados que proporciona la discriminación cuadrática con los obtenidos mediante los otros métodos.

Considerando las funciones de discriminación lineal, los resultados obtenidos en la clasificación de la muestra simulada son los siguientes:

Grupo	1	2	Total
1	82	68	150
	54.67	45.33	100.00
2	74	76	150
	49.33	50.67	100.00
Total	156	144	300
	52.00	48.00	100.00

Tabla 13. Porcentajes de clasificación por Grupo en la simulación 3

De la Tabla anterior, se puede observar que existen 142 individuos de un total de 300 que se encuentran mal clasificados, por lo que se puede inferir que este método no produce resultados satisfactorios en la clasificación de los individuos y se puede pensar que la estimación de la tasa de error tendrá un valor alto. La proporción del error estimado para este método

se puede observar de la siguiente Tabla:

	1	2	Total
Proporción	0.4533	0.4933	0.4733
Prob. apriori	0.5000	0.5000	
Tabla 14. Error Estimado por Grupo y Error Total			

Claramente de la Tabla anterior se observa que el porcentaje de error total es muy alto, y se obtiene que el 47.33% de lo individuos se encuentran mal clasificados.

Con base en las funciones de discriminación cuadráticas, se espera que este análisis proporcione mejores resultados en la clasificación, porque las poblaciones simuladas son normales con matriz de covarianzas distinta. Los porcentajes de la clasificación obtenidos mediante la discriminación cuadrática están dados en la siguiente Tabla:

Grupo	1	2	Total
1	128	22	150
	85.33	14.67	100.00
2	6	144	150
	4.00	96.00	100.00
Total	134	166	300
	44.67	55.33	100.00

Tabla 15. Porcentajes de clasificación para la simulación 3

De la Tabla anterior, se puede observar que el número de individuos mal clasificados es muy pequeño, ya que solamente 28 observaciones se encuentran traslapadas. Los porcentajes de el error estimado global y por grupos está dado en la siguiente Tabla:

	1	2	Total
Proporción	0.1467	0.0400	0.0933
Prob. apriori	0.5000	0.5000	
Tabla 16. Error Estimado por Grupo y Error Total			

En la Tabla anterior se aprecia que la proporción del error estimado es considerablemente más pequeño al obtenido por el método de discriminación lineal, ya que se tiene que el 9% de los individuos están mal clasificados.

Comparativamente a lo que se obtuvo con las funciones de discriminación lineal, la discriminación cuadrática dadas las características de la muestra proporcionó mejores resultados, en el sentido de que es más precisa en la clasificación de los individuos.

Adicionalmente se utilizan los estimadores Kernel con base en la discriminación Normal para estimar la densidad de la muestra simulada y se considera que el ancho de los intervalos en las dos direcciones es la misma. Los porcentajes de clasificación de los individuos en cada grupo se puede apreciar de la siguiente Tabla:

Grupo	1	2	Total
1	136	14	150
	90.67	9.33	100.00
2	20	130	150
	13.33	86.67	100.00
Total	156	144	300
	52.00	48.00	100.00

Tabla 17. Porcentajes de clasificación por grupos en la simulación 3

Como se aprecia de la Tabla 17, se tienen 34 individuos mal asignados, en donde 14 de ellos pertenecen al grupo 1 y fueron clasificados en el grupo 2, y 20 observaciones que son elementos del grupo 2 y se traslaparon en el 3. Sin embargo el número de individuos mal clasificado mediante la estimación Kernel es pequeño.

Las tasas de error con base en la estimación no paramétrica se puede consultar de la siguiente Tabla:

	1	2	Total
Proporción	0.0933	0.1333	0.1133
Prob. apriori	0.5000	0.5000	

Tabla 18. Error Estimado por Grupo y Error Total

De la tabla anterior se puede apreciar que el error global es del 11.33% el cual no difiere en mucho al obtenido mediante las funciones de discriminación cuadráticas. Por lo tanto los resultados en la clasificación de los individuos mediante la estimación no paramétrica son similares a los obtenidos por medio de la discriminación cuadrática.

Por ultimo se hace el análisis de la muestra con base en el Kernel Normal considerando que los anchos de los intervalos son diferentes en las dos direcciones. Los resultados que se obtienen utilizando el método mencionado anteriormente están concentrados en la Tabla que se cita a continuación:

Grupo	1	2	Total
1	129	21	150
	86.00	14.00	100.00
2	7	143	150
	4.67	95.33	100.00
Total	136	164	300
	45.33	54.67	100.00

Tabla 19. Porcentajes de clasificación por grupos para la simulación 3

Se observa de esta Tabla que el número de individuos que se encuentran mal asignados es muy pequeño, lo cual implica que las tasas de error no serán altas. Estas tasas de error se pueden apreciar de la siguiente Tabla:

	1	2	Total
Proporción	0.1400	0.0467	0.0933
Prob. apriori	0.5000	0.5000	

Tabla 20. Error Estimado por Grupo y Error Total

La Tabla 20, produce en error total estimado del 9.33%, la cual coincide con la tasa obtenida mediante las funciones de discriminación cuadráticas (ver Tabla 16 de este Capítulo) y se puede concluir que las clasificaciones con base en estos dos métodos son equivalente.

Por lo tanto, se concluye que cuando la muestra tiene asociada una distribución Normal y los grupos poseen matrices de covarianzas distinta, los métodos de discriminación cuadráticos proporcionan resultados muy parecidos a los obtenidos mediante los métodos discriminantes no paramétricos.

4.1.4 Irises de Fisher (1936)

La muestra de los Irises de Fisher es uno de los ejemplos que se consideran clásicos dentro del Análisis Multivariado y que sirve de apoyo para ilustrar algunas de sus técnicas. La

muestra consiste de 50 observaciones para cada una de las tres especies de Iris: Setosa(1), Versicolor(2) y Virgínica(3). Tomando como variables de interés para analizar la muestra el largo del pétalo (PETALLEN), el ancho del pétalo (PETALWID), el largo del sépalo (SEPALLEN) y el ancho del sépalo (SEPALWID), en donde estas variables están escaladas en milímetros (mm). En la siguiente gráfica se aprecia la dispersión de la muestra:

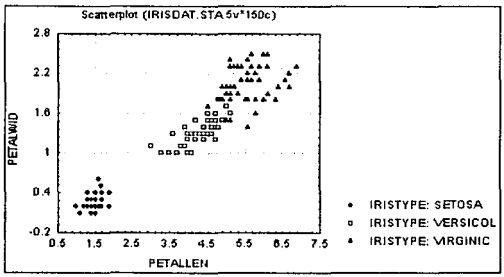


Figura 5. Gráfica de la dispersión de los Irises de Fisher.

El ejemplo se encuentra dividido básicamente en dos partes: En la primera se hace el supuesto de normalidad en la muestra y se obtienen las funciones de clasificación lineal y cuadrática. En la segunda parte no se hace ningún supuesto distribucional y se utiliza el Kernel Normal para estimar la densidad de la muestra. Cualquiera que sea el método utilizado se supone inicialmente que las poblaciones (especies) son igualmente probables.

Bajo el supuesto de normalidad en la muestra y considerando que las matrices de covarianza para los tres grupos son iguales se considera la clasificación de los individuos con base en las funciones de discriminación lineal. El siguiente cuadro proporciona los porcentajes

de la clasificación por grupo:

Especie	Setosa	Versicolor	Virginica	Total
Setosa	50	0	0	150
	100.00	0.00	0.00	100.00
Versicolor	0	48	2	150
	0.00	96.00	4.00	100.00
Virginica	0	4	46	150
	0.00	8.00	92.00	100.00
Total	50	52	48	150
	33.33	34.67	32.00	100.00

Tabla 21. Porcentajes de clasificación por especie para los Irises de Fisher

Obsérvese de la Tabla 21 que existen muy pocas observaciones mal clasificadas y que el grupo que no presenta ningún traslape es la especie Setosa, lo cual nos hace pensar que esta especie presenta diferencias relevantes en comparación con las otras dos especies. Los resultados del error estimado por grupos y el error total se encuentra dados en la Tabla que a continuación se presenta:

	Setosa	Versicolor	Virginica	Total
Proporción	0.0000	0.0400	0.0800	0.0400
Prob. apriori	0.3333	0.3333	0.3333	

Tabla 22. Error Estimado por Grupo y Error Total

Como puede observarse el error total que se presenta al utilizar las funciones de discriminación lineal es muy pequeño, ya que su estimación es del 4%, lo cual significa que sólo 6 observaciones se encuentran mal clasificadas. Los resultados de la validación cruzada proporciona el total de casos que se encuentran mal clasificados así como el grupo del cual proviene la observación en cuestión y el grupo al que fue asignado, estos casos se pueden

observar de la siguiente Tabla:

Obs	Especie	Asignada en	Setosa	Versicolor	Virgínica
8	Virginica	Versicolor*	0.0000	0.8453	0.1547
9	Versicolor	Virginica*	0.0000	0.2130	0.7870
25	Virginica	Versicolor*	0.0000	0.8322	0.1678
57	Virginica	Versicolor*	0.0000	0.8057	0.1943
91	Virginica	Versicolor*	0.0000	0.8903	0.1097
148	Versicolor	Virginica*	0.0000	0.3118	0.6882

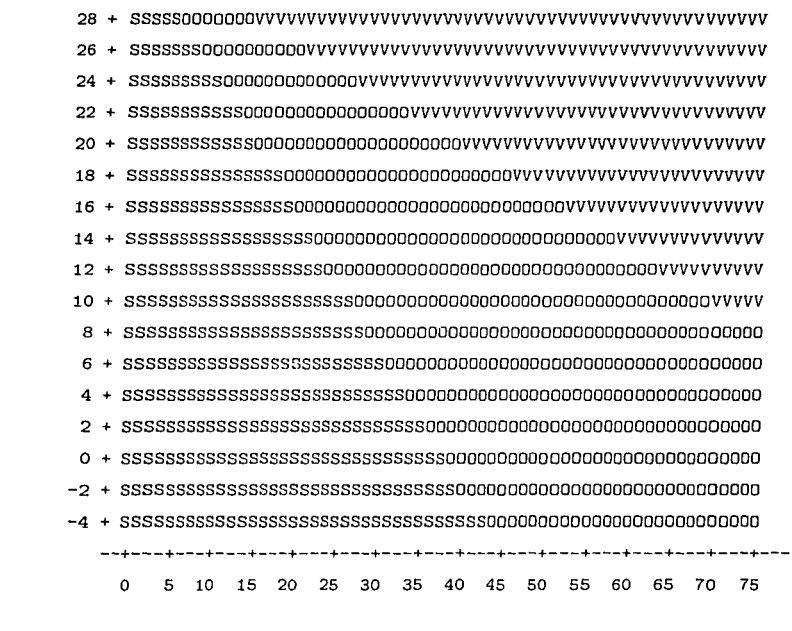
Tabla 23. Resultados de la validación cruzada para los Irises.

Esta clasificación se hace estimando la probabilidad posterior de cada individuo de la muestra y se asigna la observación, en el grupo que tiene la probabilidad más alta. De la Tabla 23 se puede observar que los grupos que se traslapan son los correspondientes a Iris Versicolor e Iris Virgínica.

Las regiones de clasificación en las que esta dividido el plano pueden apreciarse de la siguiente gráfica:

Discriminant Analysis of Fisher (1936) Iris Data 25
Using Normal Density Estimates with Equal Variance
Plot of Classification Results

Plot of PETALWID*PETALLEN. Symbol is value of _INTO_.



Donde la región asociada a la muestra de la especie Setosa esta representada por una S, la región correspondiente a la especies Versicolor con la letra O y mapea la letra V para identificar la muestra correspondiente a los Irises de la especie Virgínica.. Como puede verse de la gráfica, la especie Setosa tiene dimensiones pequeñas en las variables PETALLEN y PETALWID ya que su población esta situada en la parte inferior izquierda de la gráfica. Las mediciones de las individuos pertenecientes a las especies Versicolor y Virgínica tienen un comportamiento parecido, lo cual influye en los resultados de la clasificación, puesto que es en estas especies en donde se encuentran los traslapes. De la gráfica también puede apreciarse que los individuos que pueden ser mal clasificados se encuentran en las fronteras de cada región.

Si se considera ahora que las matrices de covarianzas asociada a cada especie de Iris son distintas, la discriminación se hace con base en las funciones de discriminación cuadráticas. Utilizando este supuesto los porcentajes de clasificación en cada grupo son los siguientes:

Especie	Setosa	Versicolor	Virginica	Total
Setosa	50	0	0	150
	100.00	0.00	0.00	100.00
Versicolor	0	48	2	150
	0.00	96.00	4.00	100.00
Virginica	0	3	47	150
	0.00	6.00	94.00	100.00
Total	50	51	49	150
	33.33	34.00	32.67	100.00

Tabla 24. Porcentaje de clasificación por especie para los Irises

De la tabla 24, se puede observar que los resultados obtenidos en la clasificación son parecidos a los que se obtuvieron por medio de las funciones de discriminación lineal, ya que la especie Setosa sigue clasificándose correctamente y las otras dos especies manifiestan traslapes de una población a otra. El error estimado por especie y total se encuentra en la

siguiente Tabla:

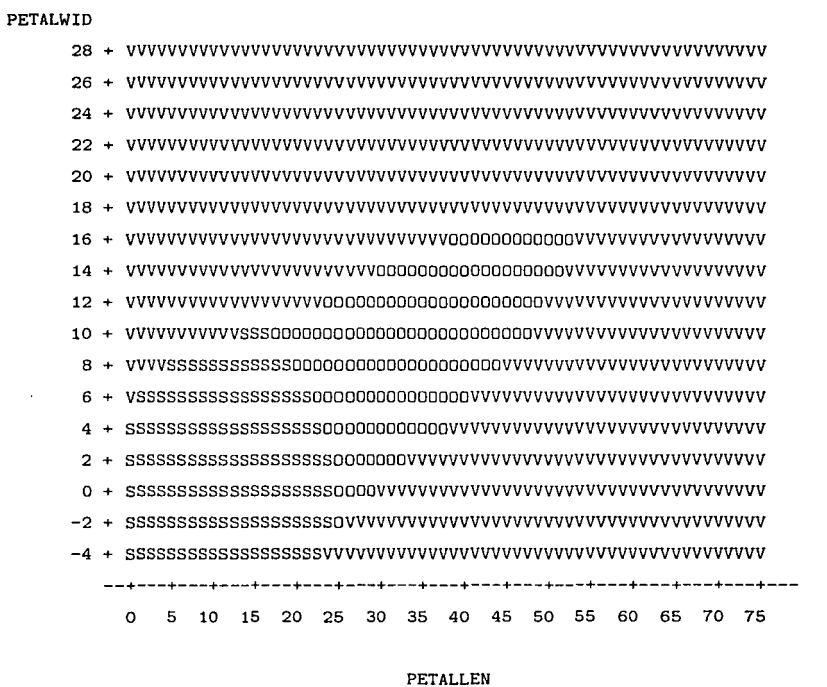
	Setosa	Versicolor	Virginica	Total
Proporción	0.0000	0.0400	0.0600	0.0333
Prob. apriori	0.3333	0.3333	0.3333	

Tabla 25. Error Estimado por Grupo y Error Total

La tasa del error total con base en las funciones de discriminación cuadráticas es del 3.33%, la cual no difiere mucho a la obtenida con las funciones de discriminación lineal.

Análogamente al caso de discriminación lineal, el paquete SAS elabora con base en las probabilidades posteriores las regiones de clasificación con las que puede separarse la muestra de los Irises. Como puede apreciarse de la gráfica que a continuación se presenta, la especie Setosa se sigue acumulando en la parte inferior izquierda de la gráfica, mientras que la especie Versicolor ocupa la región central y la especie Virgínica se localiza por debajo y por encima de la especie Versicolor originando que las especies se traslapen.

Discriminant Analysis of Fisher (1936) Iris Data 48
Using Normal Density Estimates with Unequal Variance
Plot of Classification Results



Si se supone que la muestra no proviene de ninguna distribución paramétrica conocida, se pueden utilizar los estimadores Kernel para aproximar la densidad de la muestra. En este caso se utiliza el Kernel Normal para estimar la densidad de los Irises de Fisher.

Con base en el Kernel Normal y suponiendo que las anchos de los intervalos son iguales se obtienen los siguientes resultados en la clasificación:

Especie	Setosa	Versicolor	Virginica	Total
Setosa	50	0	0	150
	100.00	0.00	0.00	100.00
Versicolor	0	48	2	150
	0.00	96.00	4.00	100.00
Virginica	0	3	47	150
	0.00	6.00	94.00	100.00
Total	50	51	49	150
	33.33	34.00	32.67	100.00

Tabla 26. Porcentaje de clasificación por especie para los Irises

De esta Tabla se aprecia que 5 individuos fueron traslapados de una especie a otra, lo cual significa que existen individuos cuyas características son muy parecidas a las que tiene la otra especie. También obsérvese que el Kernel Normal con igual ancho de banda proporciona los mismos porcentajes de clasificación que lo obtenidos utilizando la discriminación cuadrática (ver Tabla 24). Las regiones de clasificación utilizando las probabilidades posteriores pueden apreciarse en la gráfica siguiente:

Discriminant Analysis of Fisher (1936) Iris Data 74
Using Kernel Density Estimates with Equal Bandwidth
Plot of Classification Results

Plot of PETALWID*PETALLEN. Symbol is value of _INTO_.

En esta gráfica se puede observar que las tres especies siguen el comportamiento que habían manifestado anteriormente, es decir, la especie Setosa se agrupa en la parte inferior izquierda, en la parte central se dispersa la especie Versicolor y por encima de ésta queda agrupada la especie Virgínica, siendo estas dos últimas especies las que se traslapan.

Siguiendo el método del Kernel Normal, cuando se elige un determinado ancho para cada dirección del espacio, se obtienen los siguientes resultados en los porcentajes de la clasificación:

Especie	Setosa	Versicolor	Virginica	Total
Setosa	50	0	0	150
	100.00	0.00	0.00	100.00
Versicolor	0	48	2	150
	0.00	96.00	4.00	100.00
Virginica	0	2	48	150
	0.00	4.00	96.00	100.00
Total	50	50	50	150
	33.33	33.33	33.33	100.00

Tabla 27. Porcentaje de clasificación por especie para los Irises

Obsérvese de la Tabla 27 que sólo se registraron 4 observaciones mal clasificadas, cuyos traslapes ocurren en las especies Versicolor y Virgínica como anteriormente se había notado. El error global estimado está dado en la siguiente Tabla:

	Setosa	Versicolor	Virginica	Total
Proporción	0.0000	0.0400	0.0400	0.0267
Prob. apriori	0.3333	0.3333	0.3333	

Tabla 28. Error Estimado por Grupo y Error Total

Este método proporciona la tasa de error más pequeña, sin embargo los métodos anteriores proporcionaron tasas no mayores a la anterior.

Las regiones en la que queda dividido el plano cuando se utiliza el Kernel Normal con un ancho de intervalo para cada dirección, se pueden apreciar de la gráfica en la siguiente gráfica:

Discriminant Analysis of Fisher (1936) Iris Data

Using Kernel Density Estimates with Unequal Bandwidth

Plot of Classification Results

```
          Plot of PETALWID*PETALLEN.  Symbol is value of _INTO_.
PETALWID
 28 + VVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVV
 26 + VVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVV
 24 + VVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVV
 22 + VVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVV
 20 + VVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVV
 18 + VVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVV
 16 + VVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVOOOOOOOOOOVVVVVVVVVVVVVVVVVVVV
 14 + VVVVVVVVVVVVVVVVVVVVVVVOOOOOOOOOOOOOOOOOOVVVVVVVVVVVVVVVVVVVV
 12 + VVVVVVVVVVVVVVVOOOOOOOOOOOOOOOOOOOOOOOOOOVVVVVVVVVVVVVVVVVVVV
 10 + VVVVVVVVSSSSSSSOOOOOOOOOOOOOOOOOOOOOOOOVVVVVVVVVVVVVVVVVVVVVV
  8 + VVVOOOOSSSSSSSSOOOOOOOOOOOOOOOOOOOOOVVVVVVVVVVVVVVVVVVVVVVVVV
  6 + OOOSSSSSSSSSSSSSSSSOOOOOOOOOOOOOOOVVVVVVVVVVVVVVVVVVVVVVVVVVV
  4 + OSSSSSSSSSSSSSSSSSSSOOOOOOOOOOOVVVVVVVVVVVVVVVVVVVVVVVVVVVVVV
  2 + SSSSSSSSSSSSSSSSSSSSSOOOOOOOVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVV
  0 + SSSSSSSSSSSSSSSSSSSSSOOOOVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVV
 -2 + SSSSSSSSSSSSSSSSSSSSSOVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVV
 -4 + SSSSSSSSSSSSSSSSSSSVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVV
    --+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---
      0   5  10  15  20  25  30  35  40  45  50  55  60  65  70  75

                                PETALLEN
```

En esta gráfica se observa que el comportamiento de las tres especies es estable, lo cual significa que la especies Setosa se agrupa en la parte inferior izquierda mientras que Versicolor ocupa la parte central de la gráfica y la especie Virgínica esta dispersa alrededor de ésta.

Por lo tanto se puede concluir que los resultados obtenidos utilizando el supuesto de normalidad y cuando se supone que la densidad de la muestra no sigue ninguna forma paramétrica conocida son similares.

Capítulo 5

Conclusiones

Debido a que las técnicas multivariadas permiten analizar un número considerable de variables que son observadas sobre individuos que provienen de una o más poblaciones, estas técnicas han sido utilizadas en diversas áreas de la Ciencia como lo son: La Medicina para clasificar a los pacientes que padecen de una cierta enfermedad y aquellos que no la padecen, en La Biología se hacen estudios para analizar la relación entre el crecimiento del tallo de una planta con el tipo y la cantidad de fertilizante que es mezclado en la tierra, en la Arqueología para hacer la clasificación de los cráneos que son encontrados en una cierta región y ubicarlos en el periódo histórico al que pertenecen, etc. Dado que existe interés por conocer cuales son las bases matemáticas en las que descansan estos métodos, el presente trabajo cumplió con el objetivo inicial de hacer una recopilación de los resultados que son necesarios para la comprensión de dos de las técnicas multivariadas más utilizadas: La Técnica de Componentes Principales y El Análisis Discriminante.

Dado la importancia que tiene dentro del análisis multivariado el concepto de la función de densidad de probabilidad multivariada, el Capítulo 1 de esta Tesis extiende este concepto al caso multivariado y hace una recopilación general de las propiedades que caracterizan a las densidades multivariadas donde en algunos casos se presentan ejemplos para ilustrar su uso.

En el Capítulo de Componentes Principales, se presentan los resultados típicos que cita la Literatura especializada en este tópico y se presentan ejemplos ilustrativos de esta técnica

utilizando el paquete STATISTICA debido a su fácil manejo, a la calidad de los resultados que pueden obtenerse y la resolución de sus gráficas.

En la parte correspondiente al Análisis Discriminante, se obtuvieron resultados mucho más ricos, se introdujo la teoría de los estimadores Kernel para estimar la densidad de la muestra cuando no se hace el supuesto de alguna densidad paramétrica, como el modelo normal y se discute la construcción de los mapas territoriales que aparecen en los paquetes estadísticos pero que sin embargo no se encuentra su discusión en libros correspondientes a este tópico.

En el Capítulo 4, se hace una comparación de los resultados que son obtenidos con base en la discriminación normal y aquellos que se derivan de los estimadores Kernel. En esta parte se pudo comprobar (aunque no se elaboraron un número considerable de simulaciones), que cuando una muestra cumple con el supuesto de normalidad, los resultados que se obtienen mediante la discriminación normal son muy parecidos a los obtenidos por medio de los estimadores Kernel, pero cuando la muestra no cumple con este supuesto, los resultados de la clasificación si se utilizan las funciones de discriminación bajo el supuesto de normalidad son pobres, mientras que los estimadores Kernel proporcionan resultados muy superiores, en el sentido de la precisión en la clasificación. Esto conduce a considerar la discriminación no paramétrica basada en la estimación por medio del Kernel, ya que los resultados que de ella se obtienen son tan eficientes como los que se desprenden de la discriminación normal, cuando el supuesto de normalidad se cumple y son superiores cuando la densidad de la muestra no proporciona alguna forma paramétrica conocida.

Finalmente, en el Apéndice A se recopilan los resultados de apoyo en las áreas de Algebra, Cálculo, Estadística y Probabilidad que se utilizaron a lo largo de la presente. El Apéndice B, contiene los resultados de los ejemplos y las muestras asociadas a la técnica de Componentes Principales. En el Apéndice C, se concentran los programas elaborados en el paquete SAS para las simulaciones correspondientes al Capítulo 4.

Apéndice A

Resultados Preliminares

El objetivo de este Apéndice es hacer una recopilación de algunos de los resultados básicos que pueden ser importantes para la comprensión de la teoría matemática que utilizan los métodos estadísticos multivariados. Por esta razón se enuncian algunas definiciones correspondientes a las áreas de Algebra, Cálculo, Estadística y Probabilidad.

A.1 Algebra

A.1.1 Matrices

Definición A.1 *Una matriz A de tamaño $m \times n$, es un conjunto de números con m renglones y n columnas con la siguiente forma*

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix},$$

y se escribe $A_{m\times n} = \{a_{ij}\}$ para denotar una matriz A de dimensión $m \times n$, en donde a_{ij} es el elemento que está en el i-ésimo renglón y en la j-ésima columna. En particular, sí

$m = n$ se dice que A es una matriz cuadrada.

Definición A.2 *Si $n = 1$, entonces se dice que A es un vector columna y dicho vector se escribe:*

$$A = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}.$$

Definición A.3 *Si $m = 1$ entonces se dice que A es un vector renglón el cual tiene la siguiente forma:*

$$A = \begin{pmatrix} a_1 & \dots & a_n \end{pmatrix}.$$

Definición A.4 *A es una matriz nula si todos sus elementos son cero.*

Definición A.5 *Sea A es una matriz cuadrada en donde todos sus elementos arriba de la diagonal son cero,entonces A es llamada una matriz triangular inferior . La matriz tiene la forma:*

$$A = \begin{pmatrix} a_{11} & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \dots & a_{nn} \end{pmatrix}.$$

Definición A.6 *Sea A es una matriz cuadrada, en donde todos sus elementos debajo de la diagonal son cero entonces se dice que es una matriz triangular superior misma que tiene la forma*

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix}.$$

Definición A.7 *Sea $A_{n\times n}$ una matriz cuadrada en donde todos sus elementos fuera de la diagonal son cero, es llamada una matriz diagonal, esta matriz se denota por diag $(a_{11},...,a_{nn})$.*

Definición A.8 *Una matriz cuadrada de dimensión $n\times n$, que se denota por I_n y se define por*

$$I_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

Definición A.9 *Si los renglones y las columnas de una matriz A de $m\times n$ son intercambiados se obtiene como resultado la matriz transpuesta de A, y que se denota por A^t o A'. De este modo si $A=\{a_{ij}\}$ entonces $A^t=\{a_{ji}\}$ de dimensión $n\times m$. Sí en particularcular $A=A^t$, se dice que A es una matriz simétrica.*

Operaciones con Matrices Sean A, B, C matrices de $m\times n$, con elementos arbitrarios a_{ij}, b_{ij}, c_{ij} con $i=1,...,m$ $j=1,...,n$ respectivamente.

Suma La suma de las matrices A y B es una matriz C y que puede ser escrita como:

$$C=A+B,$$

tal que un elemento arbitrario de C, digamos c_{ij} está dado por

$$c_{ij}=a_{ij}+b_{ij}.$$

Multiplicación El producto AB entre una matriz A de $m\times n$ y una matriz B de $n\times k$, es una matriz C de $m\times k$ cuyo elemento c_{ij} está dado por

$$c_{ij}=\sum_{r=1}^{n} a_{ir}b_{rj},$$

con $i=1,...,m$ y $j=1,...,k$. Nótese que el producto AB esta definido únicamente si la dimensión de las columnas de A es igual a la dimensión de los renglones de B.

Definición A.10 *Sean A una matriz de dimensión $m \times n$ y α un escalar. El producto αA del escalar α y de la matriz A es una matriz B de $m \times n$, tal que un elemento arbitrario de B, por decir b_{ij} está dado por la expresión:*

$$b_{ij} = \alpha a_{ij},$$

en donde $i = 1, ..., m$ y $j = 1, ..., n$.

Propiedades de las Operaciones con Matrices

1. $A + B = B + A$ Conmutatividad de la suma.
2. $(A + B) + C = A + (B + C)$ Asociatividad de la suma.
3. $A + 0 = 0 + A = A$ Identidad de la suma.
4. $\alpha (A + B) = \alpha A + \alpha B$ Ley distributiva izquierda.
5. $(\alpha + \beta) A = \alpha A + \beta A$ Ley distributiva derecha.
6. $(\alpha\beta) A = \alpha (\beta A)$ Asociatividad de la multiplicación de escalares.
7. $A (BC) = (AB) C$ Asociatividad de la multiplicación de matrices.
8. $IA = A$ y BI Identidad para la multiplicación.
9. $A (B + C) = AB + AC$ Ley distributiva izquierda.
10. $(A + B) C = AC + BC$ Ley distributiva derecha.
11. $(A^t)^t = A$ Transpuesta de la transpuesta.
12. $(A + B)^t = A^t + B^t$ Transpuesta de la suma.
13. $(AB)^t = B^t A^t$ Transpuesta de un producto.

A.1.2 Propiedades de la Traza, el Determinante y la Matriz Inversa

La función traza que se denota y se define por $tr\,(A_{p\times p}) = \sum_{i=1}^{p} a_{pp}$ satisface las siguientes propiedades.

Sean las matrices $A_{p\times p}, B_{p\times p}, C_{p\times n}, D_{n\times p}$ y un escalar α

1. $tr\,(\alpha) = \alpha$.
2. $tr\,(A \pm B) = trA \pm trB$.
3. $tr\,(\alpha A) = \alpha tr\,(A)$.
4. $tr\,(CD) = tr\,(DC)$.
5. $\sum_{i=1}^{p} x_i^t A x_i = tr\,(A \sum_{i=1}^{p} x_i x_i^t)$ donde x_i $i = 1, ..., p$ son las componentes de un vector X.

Definición A.11 *Determinante de primer y segundo orden.*

El determinante de primer orden asociado a una matriz $A = (a_{11})$ *de* 1×1 *que se denota por* $|A|$ *se define como* $|A| = a_{11}$.

El determinante de segundo orden de una matriz $A_{2\times 2}$ *definida como:*

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$

que se denota por $|A|$ *está dado por*

$$\begin{aligned} |A| &= \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \\ &= (a_{11}a_{22} - a_{21}a_{12}). \end{aligned}$$

Definición A.12 *Cofactores y determinante.*

Sea $A_{n\times n} = \{a_{ij}\}$ una matriz y supóngase que los determinantes de menor orden que n están determinados. El cofactor de a_{ij} en A es

$$a_{ij} = (-1)^{i+j} |A_{ij}|,$$

donde A_{ij} es la matriz menor de A, que resulta de eliminar el $i-$ésimo renglón y la $j-$ésima columna de la matriz A. El determinante de A es

$$\begin{aligned} |A| &= \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} \\ &= a_{11}a'_{11} + a_{12}a'_{12} + \cdots + a_{1n}a'_{1n}. \end{aligned}$$

El determinante de una matriz cumple con las siguientes propiedades:

1. Si A es una matriz cuadrada, se cumple que

$$|A| = |A^t|.$$

2. Sea B la matriz que resulta de intercambiar dos renglones o columnas diferentes de una matriz cuadrada A, entonces la relación que guardan dichas matrices es la siguiente:

$$|B| = -|A|.$$

3. Si dos renglones de una matriz A son iguales, entonces $|A| = 0$.

4. Si A tiene un renglón cero entonces $|A| = 0$.

5. Si A es una matriz triangular de $n \times n$, entonces

$$|A| = \prod_{i=1}^{n} a_{ii},$$

en particularcular si la matriz A tiene unos en la diagonal

$$|A| = 1.$$

6. Sea $B_{n\times n}$ la matriz que resulta de multiplicar un sólo renglón de la matriz $A_{n\times n}$ por un escalar α, la relación entre estás dos matrices es la siguiente:

$$|B| = \alpha^n |A| .$$

7. La operación de sustraer de un renglón el múltiplo de otro no altera el determinante.

8. Si A y B son matrices de $n \times n$ se satisface

$$|AB| = |A||B| .$$

Definición A.13 *Sea $A_{n\times n}$ una matriz cuadrada. Si se puede encontrar una matriz A^{-1} tal que*

$$AA^{-1} = I_n \qquad y \qquad A^{-1}A = I_n,$$

entonces A^{-1} es llamada la matriz inversa de A.

Definición A.14 *La inversa de una matriz $A_{n\times n}$ en donde $r(A) = n$ es una matriz única que se denota por A^{-1} y que satisface que $AA^{-1} = A^{-1}A = I$ y que cumple con las siguientes propiedades:*

1. $A^{-1} = \frac{1}{|A|} (A_{ij})^t$, donde A_{ij} es una matriz menor de A.
2. $(cA)^{-1} = c^{-1}A^{-1}$.
3. $(AB)^{-1} = B^{-1}A^{-1}$.
4. La única solución de $Ax = b$ es $x = A^{-1}b$.
5. Si existen las matrices de $A_{p\times p}, B_{p\times n}, C_{n\times n}$ y $D_{n\times p}$ entonces

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B \left(C^{-1} + DA^{-1}B\right)^{-1} DA^{-1}.$$

6. Si existen las matrices inversas de la matriz particionada $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ entonces los elementos de A^{-1} están definidos como:

(a) $A_{11}^{-1} = \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1}$.

(b) $A_{12}^{-1} = -A_{11}^{-1}A_{12}A_{22}^{-1}$.

(c) $A_{22}^{-1} = \left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1}$.

(d) $A_{21}^{-1} = -A_{22}^{-1}A_{21}A_{11}^{-1}$.

Definición A.15 *Una matriz cuadrada A se dice que es no singular si $|A| \neq 0$, y cumple con las siguientes propiedades.*

1. Para submatrices cuadradas $A_{p\times p}$ y $B_{q\times q}$

$$\begin{vmatrix} A & 0 \\ 0 & B \end{vmatrix} = |A|\,|B|.$$

2. $\begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix} = |A_{11}|\,|A_{22} - A_{21}A_{11}^{-1}A_{12}| = |A_{22}|\,|A_{11} - A_{12}A_{22}^{-1}A_{21}|.$

3. Para matrices $B_{p\times n}, C_{n\times p}$ y una matriz no singular $A_{p\times p}$, se satisface

$$|A + BC| = |A|\,|I_p - A^{-1}BC| = |A|\,|I_n - CA^{-1}B|.$$

Definición A.16 *Una matriz cuadrada que tiene inversa se llama invertible. Una matriz cuadrada que no tiene inversa se llama singular.*

Definición A.17 *Una matriz A de $n \times n$ es ortogonal si es invertible y $A^{-1} = A^t$.*

Definición A.18 *Una matriz $A_{n\times n}$ cuadrada, es ortogonal si $AA^t = I$ con las siguientes propiedades:*

1. $A^{-1} = A^t$.
2. $A^tA = I$.
3. $|A| = \pm 1$.

4. $A_r^t A_s = \begin{cases} 0 & r \neq s \\ 1 & r = s \end{cases}$, donde A_r es un vector columna de la matriz A.

5. $C = AB$ es ortogonal si A y B son ortogonales.

Definición A.19 *Una base* $\{A_1, A_2, ..., A_n\}$ *para un subespacio* W *de* $\Re^n$ *es ortonormal si los vectores* A_i *son unitarios (es decir, de norma uno) y mutuamente perpendiculares* $A_i^t \cdot A_j = 0$ *para todo* $i \neq j$.

Teorema A.1 *Sea* A *una matriz de* $n \times n$. *Las siguientes condiciones son equivalentes:*

1. *Los renglones de* A *forman una base ortonormal para* $\Re^n$.
2. *Las columnas de* A *forman una base ortonormal para* $\Re^n$.
3. *La matriz* A *es ortogonal, i.e., invertible con* $A^{-1} = A^t$.

Demostración. Ver Fraleigh (1989).

Definición A.20 *Una matriz* A *cuadrada se dice que es idempotente si* $A = A^2$.

Definición A.21 *El rango de una matriz* $A_{n \times p}$, *está definido como el número de renglones (columnas) linealmente independientes en* A. *Si además el rango de la matriz* A *coincide con* n *o* p *se dice que es de rango completo. El rango de la matriz* A *se denota por* $r(A)$ *o* $rango(A)$ *y satisface las siguientes propiedades*

1. $0 \leq r(A) \leq \min(n, p)$.
2. $r(A) = r(A^t)$.
3. $r(A + B) \leq r(A) + r(B)$.
4. $r(AB) \leq \min\{r(A), r(B)\}$.

5. $r(A^tA) = r(AA^t) = r(A)$.

6. Si las matrices $B_{n\times n}$ y $C_{p\times p}$ son no singulares, entonces $r(BAC) = r(A)$.

7. Si $n = p$ entonces $r(A) = p$ si y sólo si A es una matriz no singular.

Definición A.22 *El número λ se dice que es un valor propio de la matriz A, si y sólo si*

$$|A - \lambda I| = 0,$$

es llamada la ecuación característica para la matriz A.

Definición A.23 *Sea A de $n \times n$. El polinomio $P(\lambda) = |A - \lambda I|$ tiene n raíces $\lambda_1, ..., \lambda_n$ y que son llamadas valores propios de la matriz A y cuyos vectores $X_1, ..., X_n$ asociados a estas raíces satisfacen $AX_i = \lambda_i X_i$ para $i = 1, ..., n$ y que se denominan vectores propios de la matriz A.*

Teorema A.2 *Cada una de las siguientes condiciones es necesaria y suficiente para que el número λ sea una valor propio de A :*

1. *Existe un vector $X \neq 0$ tal que*

$$AX = \lambda X.$$

2. *La matriz $(A - \lambda I)$ es singular.*

3. $|A - \lambda I| = 0$.

Demostración. Ver Fraleigh (1989).

Teorema A.3 *La suma de los n valores propios de una matriz A es igual a la suma de las n entradas de su diagonal, es decir*

$$\sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} a_{ii}.$$

Demostración. Ver Strang (1976).

Teorema A.4 *Si la matriz A es triangular (superior, inferior o en particularcular diagonal), entonces los valores propios $\lambda_1, ..., \lambda_n$ son exactamente los mismos que las entradas de la diagonal de A, es decir, $a_{11}, ..., a_{nn}$.*

Demostración. Ver Strang (1976).

Teorema A.5 *Si los vectores propios $X_1, ..., X_k$ corresponden a diferentes valores propios $\lambda_1, ..., \lambda_k$, entonces son ortogonales.*

Demostración. Ver Fraleigh (1989).

Teorema A.6 *Supóngase que A de $n \times n$ tiene asociados n vectores propios linealmente independientes que están dados por $G_1, ..., G_n$, entonces si se eligen estos vectores como las columnas de una matriz G, se sigue que $G^{-1}AG$ es una matriz diagonal definida como:*

$$\begin{aligned} G^{-1}AG &= \Lambda \\ &= \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix}, \end{aligned}$$

en donde $\lambda_1, ..., \lambda_n$ son los valores propios de A.

Demostración. Ver Strang (1976).

Definición A.24 *La matriz A es definida positiva si $X^tAX > 0 \ \forall \ X \neq 0$ dicha propiedad se denota por $A > 0$.*

Teorema A.7 *Cada uno de los siguientes criterios es una condición necesaria y suficiente para que la matriz simétrica $A_{n \times n}$ sea definida positiva:*

1. *$X^tAX > 0$ para todos los vectores $X \neq 0$.*
2. *Todos los valores propios de A satisfacen $\lambda_i > 0 \ \forall i$.*

3. *Todas las submatrices* $A_{k\times k}$ $k = 1, ..., n$ *tienen determinantes positivos.*

Demostración. Ver Strang (1976).

Corolario A.1 *A es definida positiva si y sólo si existe una matriz* W *no singular tal que* $A = W^tW$.

Demostración. Ver Strang (1976).

Definición A.25 *La* $A_{n\times n}$ *es definida semipositiva si* $X^tAX \geq 0 \; \forall \; X \neq 0$ *y se denota por* $A \geq 0$.

Teorema A.8 *Cada uno de los siguientes criterios es una condición necesaria y suficiente para que la matriz A sea definida semipositiva:*

1. $X^tAX \geq 0$ *para todos los vectores* X.
2. *Todos los valores propios de* A *satisfacen* $\lambda_i \geq 0 \; \forall i$.
3. *Todas las submatrices de A tienen determinantes no negativos.*

Demostración. Ver Strang (1976).

Teorema A.9 *Supóngase que* $A > 0$ *y que sus vectores propios unitarios son las columnas de U donde la matriz* $A = U\Lambda U^t$. *Entonces la rotación* $Y = U^tX$ *produce la suma de cuadrados*

$$\begin{aligned} X^tAX &= X^tU\Lambda U^tX \\ &= Y^t\Lambda Y \\ &= \lambda_1 Y_1^2 + \cdots + \lambda_n Y_n^2. \end{aligned}$$

La ecuación $X^tAX = r > 0$ *describe una elipsoidede cuyos ejes están en dirección de los vectores propios.*

Demostración. Ver Mardia (1984).

Teorema A.10 *(Descomposición Espectral) Cualquier matriz simétrica $A_{p\times p}$ puede escribirse como*

$$A = \Gamma\Lambda\Gamma^t = \sum \lambda_i \Gamma_i \Gamma_i^t.$$

donde Λ es una matriz diagonal de los valores propios de A, y Γ es una matriz ortogonal cuyas columnas son los vectores propios estandarizados.

Demostración. Ver Mardia (1984).

Corolario A.2 *Sea $A_{p\times p}$ es una matriz no singular simétrica cuyos vectores propios están dados por $\Gamma = (\Gamma_1, ..., \Gamma_p)$ y con valores propios $\lambda_1, ..., \lambda_p$, para cualquier entero n*

$$\Lambda^n = diag\left(\lambda_i^n\right),$$

y

$$A^n = \Gamma\Lambda^n\Gamma^t.$$

si todos los valores propios de A son positivos entonces se puede definir las fracciones

$$A^{\frac{r}{s}} = \Gamma\Lambda^{\frac{r}{s}}\Gamma^t.$$

donde $\Lambda^{\frac{r}{s}} = diag\left(\lambda_i^{\frac{r}{s}}\right)$, para enteros $s > 0$ y r.

Demostración. Ver Mardia (1984).

Corolario A.3 *$r(A)$ es el número de valores propios distintos de cero en la matriz A.*

Corolario A.4 *Si $A \geq 0$ entonces la matriz A es definida positiva, si y sólo si $r(A) = p$.*

Demostración

$\Rightarrow$] si $A > 0$ entonces $r(A) = p$.

Por el Teorema A.10 se puede descomponer la matriz A en sus vectores y valores propios, como

$$A = \Gamma\Lambda\Gamma^t,$$

con Γ la matriz de vectores propios y Λ es la matriz diagonal de los valores propios de A.

Supóngase que $r(A) < p$, entonces esto implica que al menos un $\lambda_i = 0$. S.P.G. considere que el valor propio nulo es $\lambda_p = 0$, y que $\Gamma_{(p)} \neq 0$ es el vector propio asociado a éste valor, entonces

$$\begin{aligned} 0 &= \Gamma_{(p)}^t A \Gamma_{(p)} \\ &= \Gamma_{(p)}^t \Gamma^t \Gamma \Lambda \Gamma^t \Gamma \Gamma_{(p)} \\ &= \Gamma_{(p)}^t \Gamma^t A \Gamma \Gamma_{(p)}, \end{aligned}$$

con $\Gamma\Gamma_{(p)} \neq 0$, ya que si ocurriera lo contrario, i.e., si $\Gamma\Gamma_{(p)} = 0$ entonces también lo es $\Gamma^t\Gamma\Gamma_{(p)} = 0$ lo cual implica que $\Gamma_{(p)} = 0$ y esto es una contradicción, por lo tanto se sigue que si $A > 0$ entonces $r(A) = p$.

$\Leftarrow$] *Si $r(A) = p$ entonces $A > 0$*

Sea a un vector no nulo en $\Re^p$, entonces

$$\begin{aligned} a^t A a &= a^t \Gamma \Lambda \Gamma^t a \\ &= b^t \Lambda b \text{ con } b = \Gamma^t a \\ &= \sum_{i=1}^{p} b_i^2 \lambda_i > 0, \end{aligned}$$

por hipótesis se sabe que $r(A) = p$, lo cual significa que todos los valores propios de A son distintos de cero, entonces se concluye que A es una matriz definida positiva. □

Corolario A.5 *Si $M_{p\times p}$ es una matriz simétrica, entonces $M = U\Lambda U^t$ en donde $UU^t = I_p$ y $\Lambda_{p\times p}$ es la matriz de los valores propios se cumple que*

1. $tr(M) = tr(\Lambda) = \sum_{i=1}^{p} \lambda_i$.
2. $|M| = \prod_{i=1}^{p} \lambda_i$.

Si en particularcular $M_{p\times p} > 0$ se satisface que $\sum_{i=1}^{p} \lambda_i > 0$ y $\prod_{i=1}^{p} \lambda_i > 0$ y además

$$\frac{1}{p}\sum_{i=1}^{p} \lambda_i \geq \left(\prod_{i=1}^{p} \lambda_i\right)^{\frac{1}{p}}.$$

Demostración. Ver Mardia (1984).

Definición A.26 *Una forma cuadrática en el vector* $X^t = (X_1, ..., X_p)$ *es una función de la forma*

$$Q(X) = X^t A X = \sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij} X_i X_j,$$

donde A *es una matriz simétrica; esto implica que*

$$Q(X) = a_{11} X_1^2 + \cdots + 2a_{12} X_1 X_2 + \cdots + 2a_{p-1p} X_{p-1} X_p.$$

Definición A.27 *1.* $Q(X) = X^t A X$ *es llamada una forma cuadrática definida positiva si* $Q(X) > 0$ *para toda* $X \neq 0$.

2. $Q(X) = X^t A X$ *es llamada una forma cuadrática semidefinida positiva si* $Q(X) \geq 0$ *para toda* $X \neq 0$.

Corolario A.6 *Si* $A > 0$, *entonces* $A^{-1} > 0$.

Corolario A.7 *Descomposición Simétrica. Cualquier matriz* $A \geq 0$ *puede escribirse como*

$$A = B^2.$$

donde B *es una matriz simétrica.*

Demostración. Ver Mardia (1984).

Corolario A.8 *Si* $A \geq 0$ *y* $B > 0$ *matrices de* $p \times p$, *entonces todos los valores propios diferentes de cero de* $B^{-1}A$ *son positivos.*

Demostración. Ver Mardia (1984).

Teorema A.11 *Sean A y B matrices simétricas. El valor máximo (mínimo) de X^tAX s.a $X^tBX = 1$, se alcanza cuando X es el vector propio correspondiente al más grande (más pequeño) valor propio de $B^{-1}A$. De esta forma*

$$\begin{aligned} \max_x X^tAX &= \lambda_i \qquad s.a \qquad X^tBX = 1 \\ (B^{-1}A)\,X &= \lambda_i X. \end{aligned}$$

Demostración. Ver Mardia (1984).

Teorema A.12 *(Descomposición en valores singulares). Si A es una matriz de $(n \times p)$ y de rango r, entonces A puede ser escrita como*

$$A = U\Lambda V^t,$$

donde $U_{(n\times r)}$ y $V_{(p\times r)}$ son las columnas ortonormales de las matrices $U^tU = V^tV = I_r$ y L es una matriz diagonal con elementos positivos.

Demostración. Ver Mardia (1984).

A.2 Cálculo.

Teorema A.13 *(Valor Medio para Integrales). Para una función continuanua $f(x)$ en el intervalo $[a, b]$, existe un valor $\xi \in [a, b]$ tal que*

$$\int_a^b f(x)\,dx = f(\xi)\,(b-a).$$

Demostración. Ver Hasser (1979).

En otras palabras, el Teorema asegura solamente la existencia de por lo menos un ξ en el intervalo, para el cual $f(\xi)$ es igual al valor promedio de f, pero no da ninguna información adicional de la ubicación de ξ.

A.3 Estadística.

A.3.1 Distribuciones Discretas y Continuas

A continuación se enlistan las familias paramétricas de las densidades discretas y continuas que fueron utilizadas a lo largo del presente trabajo.

Distribución Bernoulli y Binomial

Definición A.28 *Se dice que X es una v.a. que tiene una distribución Bernoulli si su función de densidad está dada por la expresión:*

$$f(x) = p^x q^{1-x} I_{\{0,1\}}(x),$$

en donde el parámetro p satisface $0 \leq p \leq 1$ y $q = (1-p)$.

La media, varianza y F.G.M. para una *v.a.* X que se distribuye como una Bernoulli toman las siguientes expresiones

$$\begin{aligned} E(X) &= p, \\ Var(X) &= pq, \\ M_X(t) &= (q + pe^t) \text{ con } t \in \Re. \end{aligned}$$

Definición A.29 *Una v.a. X se distribuye Binomialmente si su función de densidad está dada por:*

$$f(x) = \binom{n}{x} p^x q^{n-x} I_{\{0,1,\ldots,n\}}(x),$$

donde los dos parámetros n y p satisfacen $0 \leq p \leq 1$, $n \in \mathbf{N}$ y $q = (1-p)$.

Cuando una variable X que se distribuye como una binomial ésto se denota frecuentemente como

$$X \sim Bin(n, p).$$

La media, varianza y F.G.M. para X están dadas por

$$\begin{aligned} E(X) &= np, \\ Var(X) &= npq, \\ M_X(t) &= \left(q + pe^t\right)^n \text{ con } t \in \Re. \end{aligned}$$

Distribución Normal

Definición A.30 *Se dice que X es una v.a. que se distribuye normalmente si su función de densidad está dada por*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},$$

en donde los parámetros μ y σ satisfacen $-\infty < \mu < \infty$ y $\sigma > 0$.

Generalmente si la *v.a* X se distribuye normalmente con parámetros μ y σ entonces lo anterior puede denotarse como

$$X \sim N\left(\mu, \sigma^2\right),$$

en donde la media, varianza y F.G.M para la variable X están dadas por

$$\begin{aligned} E(X) &= \mu, \\ Var(X) &= \sigma^2, \\ M_X(t) &= \exp\left(t\mu + \frac{1}{2}\sigma^2 t^2\right) \text{ con } t \in \Re. \end{aligned}$$

Distribución F

Definición A.31 *Si la v.a. X tiene una función de densidad definida por*

$$f(x) = \frac{\Gamma\left[\frac{m+n}{2}\right]}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{\frac{m}{2}} \cdot \frac{x^{\frac{(m-2)}{2}}}{\left[1 + \left(\frac{m}{n}\right)x\right]^{\frac{(m+n)}{2}}} I_{(0,\infty)}(x),$$

donde los parámetros $m,n \in \mathbf{Z}$, entonces se dice que X tiene una distribución F y se puede escribir como

$$X \sim F(m,n).$$

Si la variable $X \sim F(m,n)$ entonces la esperanza y varianza de X toman las expresiones

$$\begin{aligned} E(X) &= \frac{n}{n-2} \text{ con } n > 2, \\ Var(X) &= \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \text{ con } n > 4. \end{aligned}$$

La F.G.M. asociada a esta densidad no existe.

Distribución χ^2

Definición A.32 *Se dice que la v.a. X se distribuye como una χ^2 de parámetro k, si su función de densidad está dada por*

$$f(x) = \frac{1}{\Gamma(\frac{k}{2})} \left(\frac{1}{2}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1} e^{-\frac{1}{2}x} I_{(0,\infty)}(x),$$

en donde el parámetro $k > 0$.

Si X se distribuye como una χ^2 entonces la media, varianza y la F.G.M están definidas como

$$\begin{aligned} E(X) &= k, \\ Var(X) &= 2k, \\ M_X(t) &= (1-2t)^{-\frac{k}{2}} \text{ con } t < \frac{1}{2}. \end{aligned}$$

Teorema A.14 *Una matriz central H de dimensión $n \times n$ definida por $H = I - \frac{1}{n}11^t$ cumple con*

1. $H^t = H$ es simétrica.
2. $H^2 = H$ es idempotente
3. $H1_n = 0$ y $H1_n1_n^t = 1_n1_n^tH = 0$.
4. $HX = X - \bar{X}1_n$, donde $\bar{X} = n^{-1}\sum X_i$.

5. $X^tHX = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^t = nS_X$.

Demostración

1.

$$\begin{aligned} H^t &= \left(I_n - \frac{1}{n}1_n 1_n^t\right)^t \\ &= I_n - \left(\frac{1}{n}1_n 1_n^t\right)^t \\ &= I_n - \frac{1}{n}1_n 1_n^t. \square \end{aligned}$$

$$\begin{aligned} H^2 &= \left(I_n - \frac{1}{n}1_n 1_n^t\right)\left(I_n - \frac{1}{n}1_n 1_n^t\right) \\ &= H - \frac{1}{n}1_n 1_n^t \left(I_n - \frac{1}{n}1_n 1_n^t\right) \\ &= H - \frac{1}{n}1_n \left(1_n^t - \frac{1}{n}1_n^t 1_n 1_n^t\right), \end{aligned}$$

donde $1_n^t 1_n = n$, entonces

$$\begin{aligned} H^2 &= H - \frac{1}{n}1_n \left(1_n^t - 1_n^t\right) \\ &= H. \square \end{aligned}$$

$$\begin{aligned} H1_n &= \left(I_n - \frac{1}{n}1_n 1_n^t\right)1_n \\ &= \left(1_n - \frac{1}{n}1_n 1_n^t 1_n\right), \end{aligned}$$

y del inciso anterior se obtiene

$$\begin{aligned} H1_n &= (1_n - 1_n) \\ &= 0. \square \end{aligned}$$

1.

$$\begin{aligned} HX &= \left(I - \frac{1}{n}1_n 1_n^t\right) X \\ &= X - \frac{1}{n}1_n 1_n^t X \\ &= X - \frac{1}{n}1_n \sum_{i=1}^{n} X_i \\ &= X - \frac{\sum_{i=1}^{n} X_i}{n} 1_n \\ &= X - \bar{X}\, 1_n . \square \end{aligned}$$

Desarróllese nS_X

$$\begin{aligned} nS_X &= \sum_{i=1} n \left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^t \\ &= \sum_{i=1}^{n} \left(X_i X_i^t - X_i\, \bar{X}^t - \bar{X}^t\, X_i + \bar{X}\bar{X}^t\right), \end{aligned}$$

y distribuyendo la suma se obtiene

$$nS_X = \sum_{i=1}^{n} X_i X_i^t - n\, \bar{X}\bar{X}^t,$$

pero

$$X^t X = \sum_{i=1}^{n} X_i X_i^t,$$

y si se define 1_n como un vector columna de dimensión $n \times 1$ entonces la media de la población queda definida como

$$\bar{X}_{n\times 1} = \frac{1}{n}\left(X^t 1_n\right),$$

entonces

$$\begin{aligned} nS_X &= X^t X - n\left(\frac{1}{n}X^t 1_n\right)\left(\frac{1}{n}X^t 1_n\right)^t \\ &= X^t H X . \square \end{aligned}$$

Teorema A.15 *Sea $\Sigma > 0$ la matriz de dispersión de alguna variable aleatoria X, la raíz de Σ puede construirserse utilizando la descomposición espectral $\Sigma = U\Lambda U^t$ como*

$\Sigma^{\frac{1}{2}} = U\Lambda^{\frac{1}{2}}U^t$, *la matriz* $\Sigma^{\frac{1}{2}}$ *es simétrica y definida positiva y* $\Lambda^{\frac{1}{2}}$ *es la matriz diagonal de los valores propios de* $\Sigma^{\frac{1}{2}}$.

Demostración.

$$\begin{aligned}\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}} &= \left(U\Lambda^{\frac{1}{2}}U^t\right)\left(U\Lambda^{\frac{1}{2}}U^t\right)\\ &= U\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}U^t\\ &= U\Lambda U^t\\ &= \Sigma.\end{aligned}$$

$\Sigma^{\frac{1}{2}}$ *es simétrica.*

$$\begin{aligned}\left(\Sigma^{\frac{1}{2}}\right)^t &= \left(U\Lambda^{\frac{1}{2}}U^t\right)^t\\ &= (U^t)^t\left(\Lambda^{\frac{1}{2}}\right)^t U^t\\ &= U\Lambda^{\frac{1}{2}}U^t\\ &= \Sigma^{\frac{1}{2}}.\end{aligned}$$

Por lo tanto $\Sigma^{\frac{1}{2}}$ *es simétrica.*□

$\Sigma^{\frac{1}{2}}$ *es definida positiva. Sea* $z \in \Re^p$, *con* $z \neq 0$, *P.d.* $z^t\Sigma^{\frac{1}{2}}z \geq 0$.

$$\begin{aligned}z^t\Sigma^{\frac{1}{2}}z &= z^t\left(U\Lambda^{\frac{1}{2}}U^t\right)z\\ &= y^t\Lambda^{\frac{1}{2}}y\end{aligned}$$

donde $y = U^t z$,

$$\begin{aligned}z^t\Sigma^{\frac{1}{2}}z &= (y_1,...,y_p)\begin{pmatrix}\sqrt{\lambda_1} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sqrt{\lambda_p}\end{pmatrix}\begin{pmatrix}y_1\\ \vdots\\ y_p\end{pmatrix}\\ &= \sum_{i=1}^{p}\lambda_i^{\frac{1}{2}}y_i^2.\end{aligned}$$

Sabemos que cada $\lambda_i > 0$ *y que no todos los valores* y_i $i = 1,...,p$ *son cero, por lo tanto,* $\Sigma^{\frac{1}{2}}$ *es definida positiva.*□

Teorema A.16 *Sea $\Sigma > 0$ la raíz de Σ^{-1} puede construirse utilizando la descomposición espectral $\Sigma^{-1} = U\Lambda^{-1}U^t$ como $\Sigma^{-\frac{1}{2}} = U\Lambda^{-\frac{1}{2}}U^t$ y además la matriz $\Sigma^{-\frac{1}{2}}$ es simétrica y definida positiva.*

Demostración.

$$\begin{aligned}\Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}} &= \left(U\Lambda^{-\frac{1}{2}}U^t\right)\left(U\Lambda^{-\frac{1}{2}}U^t\right)\\ &= U\Lambda^{-\frac{1}{2}}\Lambda^{-\frac{1}{2}}U^t\\ &= U\Lambda^{-1}U^t\\ &= \Sigma^{-1}.\square\end{aligned}$$

$\Sigma^{-\frac{1}{2}}$ es simétrica.

$$\begin{aligned}\left(\Sigma^{-\frac{1}{2}}\right)^t &= \left(U\Lambda^{-\frac{1}{2}}U^t\right)^t\\ &= \left(U^t\right)^t\left(\Lambda^{-\frac{1}{2}}\right)^t U^t\\ &= U\Lambda^{-\frac{1}{2}}U^t\\ &= \Sigma^{-\frac{1}{2}}.\square\end{aligned}$$

$\Sigma^{-\frac{1}{2}}$ es definida positiva. Sea $s \in \Re^p$, con $z \neq 0$, P.d. $s^t\Sigma^{-\frac{1}{2}}s \geq 0$.

$$\begin{aligned}s^t\Sigma^{-\frac{1}{2}}s &= s^t\left(U\Lambda^{-\frac{1}{2}}U^t\right)s\\ &= t^t\Lambda^{\frac{1}{2}}t \text{ con } t = U^tz\\ &= \sum_{i=1}^{p}\lambda_i t_i^2,\end{aligned}$$

por hipótesis $\lambda_i > 0$ entonces también lo es $\lambda_i^{-\frac{1}{2}}$, luego se sabe que no todas las t_i^2 son cero entonces

$$s^t\Sigma^{-\frac{1}{2}}s > 0,$$

por lo tanto la matriz $\Sigma^{-\frac{1}{2}}$ es definida positiva.$\square$

Teorema A.17 *Sea X un vector aleatorio y sea μ el vector de medias en donde $X, \mu \in \Re^n$ los cuales satisfacen la siguiente igualdad.*

$$\sum_{i=1}^{n}(X_i-\mu)^t(X_i-\mu)=\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^t\left(X_i-\bar{X}\right)+n\left(\bar{X}-\mu\right)^t\left(\bar{X}-\mu\right).$$

Demostración

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)(X_i-\mu)^t &= \sum_{i=1}^{n}\left(X_i-\bar{X}+\bar{X}-\mu\right)^t\left(X_i-\bar{X}+\bar{X}-\mu\right)\\ &= \sum_{i=1}^{n}\left[\left(X_i-\bar{X}\right)+\left(\bar{X}-\mu\right)\right]^t\left[\left(X_i-\bar{X}\right)+\left(\bar{X}-\mu\right)\right].\end{aligned}$$

Desarrollando el producto y distribuyendo la suma se obtiene los productos cruzados se anulan, por lo que

$$\sum_{i=1}^{n}(X_i-\mu)(X_i-\mu)^t=\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^t\left(X_i-\bar{X}\right)+n\left(\bar{X}-\mu\right)^t\left(\bar{X}-\mu\right).\square$$

Teorema A.18 *Sean A, Σ matrices definidas positivas y sea $H(\Sigma)=|\Sigma|^{-\frac{n}{2}}\exp\left\{-\frac{n}{2}tr\Sigma^{-1}A\right\}$ entonces* $\sup_{\Sigma>0}H(\Sigma)=H(A)=|A|^{-\frac{n}{2}}\exp\left\{-\frac{n}{2}trA^{-1}A\right\}$.

Demostración.

Basta demostrar que

$$C=\ln H(A)-\ln H(\Sigma)\geq 0,$$

para toda $\Sigma>0$. Desarrollando la diferencia se tiene

$$\begin{aligned}C &= -\frac{n}{2}\ln|A|-\frac{n}{2}tr\left(A^{-1}A\right)+\frac{n}{2}\ln|\Sigma|+\frac{n}{2}tr\left(\Sigma^{-1}A\right)\\ &= -\frac{n}{2}\left\{\ln|A|-\ln|\Sigma|\right\}-\frac{n}{2}trI_p+\frac{n}{2}tr\left(\Sigma^{-1}A\right)\\ &= -\frac{n}{2}\ln\left|\Sigma^{-1}A\right|-\frac{np}{2}+\frac{n}{2}tr\left(\Sigma^{-1}A\right),\end{aligned}$$

ya que

$$\begin{aligned}\left|\Sigma^{-1}A\right| &= \left|\Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}A\right|\\ &= \left|\Sigma^{-\frac{1}{2}}\right|\left|\Sigma^{-\frac{1}{2}}A\right|\\ &= \left|\Sigma^{-\frac{1}{2}}\right||A|\left|\Sigma^{-\frac{1}{2}}\right|\\ &= \left|\Sigma^{-\frac{1}{2}}A\Sigma^{-\frac{1}{2}}\right|.\end{aligned}$$

Desarrollando también la traza en la matriz $\Sigma^{-1}A$, se obtiene

$$\begin{aligned} tr\left(\Sigma^{-1}A\right) &= tr\left(\Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}A\right) \\ &= tr\left(\Sigma^{-\frac{1}{2}}A\Sigma^{-\frac{1}{2}}\right), \end{aligned}$$

y se deduce que

$$C = \frac{np}{2}\left[\frac{1}{p}tr\left(\Sigma^{-\frac{1}{2}}A\Sigma^{-\frac{1}{2}}\right) - 1 - \ln\left|\Sigma^{-\frac{1}{2}}A\Sigma^{-\frac{1}{2}}\right|^{\frac{1}{p}}\right]. \tag{A.1}$$

Dado que las matrices Σ^{-1} y A son definidas positivas, también lo es la matriz $\Sigma^{-\frac{1}{2}}A\Sigma^{-\frac{1}{2}}$ la cual tiene asociados $\lambda_1, ..., \lambda_p$ valores propios distintos con la propiedad de que cada $\lambda_i > 0$, y según el Corolario A.5 se cumple

$$\begin{aligned} tr\left(\Sigma^{-\frac{1}{2}}A\Sigma^{-\frac{1}{2}}\right) &= \sum_{i=1}^{p}\lambda_i, \\ \left|\Sigma^{-\frac{1}{2}}A\Sigma^{-\frac{1}{2}}\right| &= \prod_{i=1}^{p}\lambda_i, \end{aligned}$$

y además se satisface por el mismo Corolario que

$$\lambda_a = \frac{1}{p}\sum_{i=1}^{p}\lambda_i > \left(\prod_{i=1}^{p}\lambda_i\right)^{\frac{1}{p}} = \lambda_g,$$

entonces la ecuación A.1 puede expresarse como

$$\begin{aligned} C &= \frac{np}{2}\left[\lambda_a - 1 - \ln\lambda_g\right] \\ &\geq \frac{np}{2}\left[\lambda_a - 1 - \ln\lambda_a\right]. \end{aligned}$$

Obsérvese que C tiene la forma del polinomio siguiente

$$F(x) = x - 1 - \ln x.$$

Calculando la primera derivada de $F(x)$ se obtiene

$$\begin{aligned} \frac{dF(x)}{dx} &= 1 - \frac{1}{x} \\ &\geq 0 \text{ si } x \geq 0, \end{aligned}$$

igualándola a cero se obtiene el punto $x = 1$. Si se calcula la segunda derivada y se evalúa en $x = 1$ puede verse que ésta es mayor que cero y se concluye que existe un mínimo local en $x = 1$ y que la función a partir de ese punto crece y de lo cual se deduce que

$$C = \frac{np}{2}\left[\lambda_a - 1 - \ln \lambda_a\right] \geq 0,$$

y por lo tanto

$$\ln H(A) \geq \ln H(\Sigma). \square$$

Teorema A.19 *Sea $X_1, ..., X_n$ una muestra aleatoria de la población $N_p(\mu, \Sigma)$. Si H_0 y H_1 son hipótesis las cuales conducen a estimar Σ como $\hat{\Sigma}$ y S; y $\bar{X}$ es el estimador máximo verosímil de μ bajo ambas hipótesis, entonces el cociente de verosimilitud para probar H_0 vs H_a está dado por:*

$$-2\ln \lambda = np\left\{a - \ln(g) - 1\right\},$$

donde a y g denotan la media aritmética y geométrica respectivamente de los valores propios de $\hat{\Sigma}^{-1} S$.

Demostración. Ver Mardia (1982, p.134).

A.4 Probabilidad

A.4.1 Convergencia en Probabilidad.

Definición A.33 *(Convergencia a cero en probabilidad). Decimos que X_n converge a cero en probabilidad, escribiéndose como $X_n = o_p(1)$ o $X_n \xrightarrow{p} 0$, si para cada $\varepsilon > 0$,*

$$P(|X_n| > \varepsilon) \longrightarrow 0, \text{ cuando } n \longrightarrow \infty.$$

Definición A.34 *(Acotamiento en probabilidad). Se dice que la sucesión $\{X_n\}$ está acotada en probabilidad, denotándose como $X_n = 0_p(1)$, si para cada $\varepsilon > 0$ existe una*

$\delta(\varepsilon) \in (0, \infty)$ *tal que*

$$P(|X_n| > \delta(\varepsilon)) < \varepsilon, \quad \forall n.$$

Teorema A.20 *Serie de Taylor. La serie de Taylor para* $f(x)$ *en* $x = a$ *está definida como*

$$f(x) = f(a) + f^1(a)(x-a) + \frac{1}{2!}f^2(a)(x-a)^2 + ... + \frac{1}{n!}f^n(a)(x-a)^n + R_n.$$

donde

$$f^i(a) = \frac{d^i f(x)}{dx_i}\,|_{x=a},$$

y

$$R_n = \frac{f^{n+1}(c)(x-a)^{n+1}}{(n+1)!} \quad con\ a \leq c \leq x.$$

Demostración. Ver Brockwell & Davis (1991).

Apéndice B

Estadísticas de los Ejemplos de Componentes Principales

Ya que los ejemplos presentados en el Capítulo 2 no contienen las muestras ni los cálculos en las que está basado el análisis, este Apéndice tiene como finalidad hacer una recopilación de las muestras así como también de los resultados numéricos que pueden proporcionar un panorama más amplio de la constitución estadística de las muestras y facilitarle al Lector el seguimiento (si es que lo desea) de los cálculos obtenidos y que pueden ser reproducidos sin la mayor dificultad.

B.1 Ejemplo 1. (Irises de Fisher)

La muestra de los Irises de Fisher es una matriz de datos de dimensión 150×4, que esta formada por 3 especies de flor de Iris: Iris Setosa, Iris Virgínica e Iris Versicolor; para cada una de estas especies se tomaron 50 observaciones y se registró para cada una de estas 50 flores 4 características: Longitud de Sépalo (sepallen), Ancho del Sépalo (sepalwid), Largo de Pétalo (petallen) y Ancho de Pétalo (petalwid).

Variable	Media	Desviación estandar
SP	5.8433	0.8280
SW	3.0573	0.4358
PL	3.7580	1.7652
PW	1.1993	0.7622

Table B.1: Estadísticas básicas para los Irises de Fisher.

	SP	SW	PL	PW
SP	1	−0.0117	0.8717	0.8179
SW	−0.0117	1	−0.4284	−0.0366
PL	0.8717	−0.4284	1	0.9628
PW	0.8179	−0.0366	0.9628	1

Table B.2: Matriz de correlación para los Irises de Fisher.

	SP	SL	PL	PW
SP	0.6856	−0.0422	1.2743	0.5162
SW	−0.0422	0.1899	−0.3296	−0.1216
PL	1.2743	−0.3296	3.1162	1.2956
PW	0.5162	−0.1216	1.2956	0.5810

Table B.3: Matriz de covarianza para los Irises de Fisher.

U_1	U_2	U_3	U_4	λ
0.3050	−0.3947	−1.8783	1.8151	2.9184
−0.1576	−0.9657	0.6379	−0.8581	0.9140
0.3397	−0.0256	0.3710	−5.5684	0.1467
0.3306	−0.0700	1.6556	3.6379	0.0207

Table B.4: Tabla de vectores y valores para los Irises.

Sepal Lenght	Sepal Width	Petal Lenght	Petal Width
6.4	2.8	5.6	2.2
6.7	3.1	5.6	2.4
6.3	2.8	5.1	1.5
6.9	3.1	5.1	2.3
6.5	3.0	5.2	2.0
6.5	3.0	5.5	1.8
5.8	2.7	5.1	1.9
6.8	3.2	5.9	2.3
6.2	3.4	5.4	2.3
7.7	3.8	6.7	2.2
6.7	3.3	5.7	2.5
7.6	3.0	6.6	2.1
4.9	2.5	4.5	1.7
6.7	3.0	5.2	2.3
5.9	3.0	5.1	1.8
6.3	2.5	5.0	1.9
6.4	3.2	5.3	2.3
7.9	3.8	6.4	2.0
6.7	3.3	5.7	2.1
7.7	2.8	6.7	2.0
6.3	2.7	4.9	1.8
7.2	3.2	6.0	1.8
6.1	3.0	4.9	1.8
6.1	2.6	5.6	1.4
6.4	2.8	5.6	2.1
6.2	2.8	4.8	1.8
7.7	3.0	6.1	2.3
6.3	3.4	5.6	2.4
5.8	2.7	5.1	1.9
7.2	3.0	5.8	1.6
7.1	3.0	5.9	2.1
6.4	3.1	5.5	1.8
6.0	3.0	4.8	1.8
6.3	2.9	5.6	1.8
7.7	2.6	6.9	2.3
6.0	2.2	5.0	1.5
6.9	3.2	5.7	2.3
7.4	2.8	6.1	1.9
5.6	2.8	4.9	2.0
7.3	2.9	6.3	1.8
6.7	2.5	5.8	1.8
6.5	3.0	5.8	2.2
6.9	3.1	5.4	2.1
7.2	3.6	6.1	2.5
6.5	3.2	5.1	2.0
6.4	2.7	5.3	1.9
6.8	3.0	5.5	2.1
5.7	2.5	5.0	2.0
5.8	2.8	5.1	2.4
6.3	3.3	6.0	2.5

Table B.5: Datos Iris de Fisher, Especie Virgínica

Sepal Lenght	Sepal Width	Petal Lenght	Petal Width
5.0	3.3	1.4	0.2
4.6	3.4	1.4	0.3
4.6	3.6	1.0	0.2
5.1	3.3	1.7	0.5
5.5	3.5	1.3	0.2
4.8	3.1	1.6	0.2
5.2	3.4	1.4	0.2
4.9	3.6	1.4	0.1
4.4	3.2	1.3	0.2
5.0	3.5	1.6	0.6
4.4	3.0	1.3	0.2
4.7	3.2	1.6	0.2
4.8	3.0	1.4	0.3
5.1	3.8	1.6	0.2
4.8	3.4	1.9	0.2
5.0	3.0	1.6	0.2
5.0	3.2	1.2	0.2
4.3	3.0	1.1	0.1
5.8	4.0	1.2	0.2
5.1	3.8	1.9	0.4
4.9	3.0	1.4	0.2
5.1	3.5	1.4	0.2
5.0	3.4	1.6	0.4
4.6	3.2	1.4	0.2
5.7	4.4	1.5	0.4
5.0	3.6	1.4	0.2
5.4	3.4	1.5	0.4
5.2	4.1	1.5	0.1
5.5	4.2	1.4	0.2
4.9	3.1	1.5	0.2
5.4	3.9	1.7	0.4
5.0	3.4	1.5	0.2
4.4	2.9	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5.1	3.4	1.5	0.2
5.0	3.5	1.3	0.3
4.9	3.1	1.5	0.1
5.4	3.7	1.5	0.2
5.4	3.9	1.3	0.4
5.1	3.5	1.4	0.3
4.8	3.4	1.6	0.2
4.8	3.0	1.4	0.1
4.5	2.3	1.3	0.3
5.7	3.8	1.7	0.3
5.1	3.8	1.5	0.3
5.4	3.4	1.7	0.2
5.1	3.7	1.5	0.4
5.2	3.5	1.5	0.2
5.3	3.7	1.5	0.2

Table B.5: Datos Iris de Fisher, Especie Setosa (continuación)

Sepal Lenght	Sepal Width	Petal Lenght	Petal Width
6.5	2.8	4.6	1.5
6.2	2.2	4.5	1.5
5.9	3.2	4.8	1.8
6.1	3.0	4.6	1.4
6.0	2.7	5.1	1.6
5.6	2.5	3.9	1.1
5.7	2.8	4.5	1.3
6.3	3.3	4.7	1.6
7.0	3.2	4.7	1.4
6.4	3.2	4.5	1.5
6.1	2.8	4.0	1.3
5.5	2.4	3.8	1.1
5.4	3.0	4.5	1.5
5.8	2.6	4.0	1.2
5.5	2.6	4.4	1.2
6.7	3.1	4.4	1.4
5.6	3.0	4.5	1.5
5.8	2.7	4.1	1.0
6.0	2.9	4.5	1.5
5.7	2.6	3.5	1.0
5.7	2.9	4.2	1.3
4.9	2.4	3.3	1.0
5.6	2.7	4.2	1.3
5.7	3.0	4.2	1.2
6.6	2.9	4.6	1.3
5.2	2.7	3.9	1.4
6.0	3.4	4.5	1.6
5.0	2.0	3.5	1.0
5.5	2.4	3.7	1.0
5.8	2.7	3.9	1.2
6.2	2.9	4.3	1.3
5.9	3.0	4.2	1.5
6.0	2.2	4.0	1.0
6.7	3.1	4.7	1.5
6.3	2.3	4.4	1.3
5.6	3.0	4.1	1.3
6.3	2.5	4.9	1.5
6.1	2.8	4.7	1.2
6.4	2.9	4.3	1.3
5.1	2.5	3.0	1.1
5.7	2.8	4.1	1.3
6.1	2.9	4.7	1.4
5.6	2.9	3.6	1.3
6.9	3.1	4.9	1.5
5.5	2.5	4.0	1.3
5.5	2.3	4.0	1.3
6.6	3.0	4.4	1.4
6.8	2.8	4.8	1.4
6.7	3.0	5.0	1.7

Table B.5: Datos Irir de Fisher, EspecieVersicolor (continuación)

B.2 Ejemplo 2 (Datos Simulados)

La muestra consiste de una matriz X de dimensión 100×3 la cual contiene tres variables a las que denotaremos por X_1, X_2 y X_3; y que fueron simuladas haciendo uso del paquete $S - PLUS$ para $WINDOWS$ versión 3.3 (1995). Es conveniente señalar que las dos primeras variables X_1, X_2 fueron simuladas con una distribución normal bivariada y que la variable X_3 fue generada como una combinación lineal de X_1 y X_2 más un error aleatorio ε que fue generado independientemente de X_1 y X_2 y cuya distribución es normal con media 0 y varianza 1.

Variable	Media	Desviación estandar
X_1	5.3490	3.2594
X_2	3.1337	1.3481
X_3	0.3485	8.8620

Table B.6: Estadísticas descriptivas para la tabla B2.

	X_1	X_2	X_3
X_1	1	0.4512	0.7546
X_2	0.4512	1	−0.2360
X_3	0.7546	−0.2360	1

Table B.7: Matriz de correlación para la muestra simulada.

	X_1	X_2	X_3
X_1	10.6234	1.9825	21.7983
X_2	1.9825	1.8172	−2.8196
X_3	21.7983	−2.8196	78.5357

Table B.8: Matriz de covarianza para la muestra simulada.

-0.5510	0.1285	-9.4881	1.7909
-0.1728	0.7892	6.3665	1.2042
-0.4741	-0.4370	8.7052	0.0048

Table B.9: Vectores y valores propios.

	X1	X2	X3
1	3.89433309	4.59996089	-10.19609414
2	2.48170761	1.68741685	-1.63563163
3	3.43892501	2.67324816	-3.87058899
4	5.63577236	4.26865087	-5.49387145
5	3.97327552	1.58407351	5.53148336
6	3.25494783	0.85062199	5.86633236
7	8.42274391	2.11966199	13.22071619
8	9.30728714	3.03255923	13.40887586
9	3.28716312	1.83466632	-1.78027472
10	4.09239414	4.63986041	-11.69785726
11	10.37742194	2.53910658	18.81480135
12	3.87703344	4.33038355	-10.57208539
13	5.54184767	2.36312637	5.31582774
14	2.90430370	3.24861201	-7.58996728
15	5.85039613	5.40027647	-10.01526648
16	0.73508751	0.62693588	-0.68841827
17	12.07055018	2.41021902	24.31033321
18	7.69108887	2.37053679	8.62087245
19	8.61990033	3.97598537	5.99043972
20	10.11287217	3.92794622	11.00374302
21	9.41062699	2.72947920	14.79830056
22	3.61979867	2.50335312	-2.26060069
23	5.17009421	0.13327605	15.43834687
24	-0.50843922	0.80013635	-6.66952193
25	6.42870408	2.07504503	10.09650381
26	5.06241113	1.64676341	5.78794432
27	5.01442194	2.77229214	1.33826430
28	5.59426021	5.00964979	-7.74274145
29	6.56282229	4.78174531	-4.85225317
30	8.59128748	3.96662272	6.88060512
31	5.03421021	4.04482956	-4.05041245
32	9.17732993	3.96676228	7.37555489
33	0.63193616	1.72213835	-6.46428962
34	7.35304982	3.22292573	7.38945939
35	1.63722180	2.96073005	-9.04100471
36	12.65755155	4.11814264	16.35090031
37	5.28596881	3.90602406	-3.84622956
38	2.42157549	3.25904502	-7.85065016
39	3.99936727	3.13276365	-4.57653581
40	4.41452497	3.60031878	-4.71542564

Tabla B10. Muestra simulada en S-PLUS

	X1	X2	X3
41	7.03482417	5.12729976	-3.99361671
42	11.10538631	5.99206855	2.90126373
43	5.20032839	3.35267740	0.57606074
44	10.50517841	2.76806864	15.41024243
45	8.62193910	6.48394737	-5.85258132
46	6.21974103	3.05907842	3.22853846
47	5.12499355	1.36799530	6.68148428
48	3.04946914	3.88470562	-11.32186365
49	-0.85387840	3.48081429	-18.96755220
50	6.78758902	1.68818935	11.02466589
51	3.11529867	2.01834473	-0.33446846
52	4.36079110	1.93007126	2.47200344
53	8.22871960	4.54502774	3.23741733
54	9.04802258	5.15195530	0.47946678
55	1.85957564	1.75283126	-2.19380016
56	0.47541087	0.49430492	-2.64868591
57	2.02981514	2.54399870	-6.70132443
58	4.24467044	3.77053281	-5.11885734
59	11.03310623	1.42478827	25.30540321
60	0.14736975	1.75622576	-8.06774288
61	6.53693431	2.44284000	8.33330182
62	6.90086041	4.48233808	0.03059037
63	6.48635314	4.13161545	-1.34189054
64	-3.00119801	2.80583342	-21.17069948
65	0.47958033	2.25236339	-10.90032165
66	4.22096384	2.38416104	0.22521510
67	-1.90672821	0.85832233	-9.39252212
68	6.88070732	5.85354557	-7.54568455
69	5.78455111	3.99594069	-2.08099345
70	6.02982784	2.09689324	7.38160359
71	6.68906854	2.81121062	5.05492474
72	10.54688257	3.86787675	13.63522346
73	8.70682877	5.35832822	0.92443203
74	6.07071916	3.50325321	0.78651065
75	0.04396223	4.00595409	-19.61583854
76	7.40081457	1.55438586	12.84758283
77	6.92832256	2.92451774	5.26592156
78	6.35721777	4.16976213	-0.36390039
79	2.04636708	3.50669059	-11.10425138
80	4.84017125	3.38675977	-2.36955340

Tabla B10: Muestra simulada en S-Plus (continuación).

	X1	X2	X3
81	3.30414795	4.37767204	-10.90097279
82	4.62602793	2.64549463	1.13080595
83	1.57989036	1.34182774	-5.47360165
84	7.36837016	3.35963790	5.22736213
85	10.72479019	4.70387049	8.22373332
86	5.27953285	2.96457467	0.03402487
87	0.72717712	1.77089162	-7.06591724
88	4.31772182	3.74347575	-6.52520326
89	7.10365401	4.21305650	0.16277413
90	3.95114980	4.12314108	-8.62172670
91	4.96122510	2.77431361	1.84588427
92	4.06397215	2.44218141	1.13072171
93	2.68173291	1.85780157	-3.07242871
94	6.02479649	3.12009812	2.98900232
95	10.15358183	4.77906370	4.93209648
96	5.31285924	3.63312673	-1.99423032
97	4.12547143	4.28232924	-9.83716597
98	2.15648981	1.32186846	1.70495716
99	12.60429259	6.17839108	7.47909756
100	5.33383785	1.91538188	6.73316586

Tabla B10. Muestra simulada en S-Plus (continuación).

Apéndice C

Salidas del Paquete SAS

Este Apéndice concentra las simulaciones obtenidas mediante el paquete estadístico S-Plus con las que se hace la comparación entre los métodos discriminantes paramétrico y no paramétricos analizados en el Capítulo 4 de la presente. En este apartado se encuentra el programa que es típico en las tres primeras muestras simuladas y en la última sección se hace referencia a la muestra de los Irises de Fisher, la cual fue analizada mediante el paquete SAS.

C.1 Programa Elaborado en el Paquete SAS para la Simulaciones 1, 2 y 3

El programa que se presenta a continuación es común para las tres conjuntos de datos simulados en el Capítulo 4, por los cual se consideró pertinente incluirlo sólo una vez.

```
data datos;
   title 'Discriminant Analysis';
```

```
   infile 'c:\datos.txt';
   input IND Y GRUPO;
   drop IND;
run;

proc discrim data=datos method=normal pool=yes short;
   class GRUPO;
   var Y;
   title2 'Using Normal Density Estimates with Equal Variance';
run;

proc discrim data=datos method=normal pool=no short;
   class GRUPO;
   var Y;
   title2 'Using Normal Density Estimates with Unequal Variance';
run;

proc discrim data=datos method=npar kernel=normal r=.4 pool=yes
            short;
   class GRUPO;
   var Y;
   title2 'Using Kernel Density Estimates with Equal Bandwidth';
run;


proc discrim data=datos method=npar kernel=normal r=.4 pool=no
            short;
   class GRUPO;
```

```
   var Y;
   title2 'Using Kernel Density Estimates with Unequal Bandwidth';
run;
```

C.1.1 Muestra Correspondiente a la Simulación 1

IND	Y	GRUPO	IND	Y	GRUPO
1	7.6830	1	41	8.1050	1
2	8.0300	1	42	8.5620	1
3	1.2080	1	43	7.5240	1
4	6.5630	1	44	0.5200	1
5	0.5150	1	45	8.0100	1
6	0.1490	1	46	8.2060	1
7	7.9640	1	47	8.7220	1
8	2.0680	1	48	6.8170	1
9	10.7970	1	49	7.9500	1
10	-1.3720	1	50	8.2890	1
11	8.1130	1	51	8.6980	1
12	7.7720	1	52	7.5770	1
13	6.2620	1	53	-0.7100	1
14	7.6290	1	54	5.3550	1
15	1.2240	1	55	8.4210	1
16	7.1910	1	56	8.8340	1
17	8.7710	1	57	2.6250	1
18	8.1160	1	58	7.4260	1
19	7.8260	1	59	8.0580	1
20	0.1860	1	60	-0.9850	1
21	-0.7750	1	61	6.5640	1
22	8.9000	1	62	1.8620	1
23	7.7580	1	63	7.9210	1
24	8.8160	1	64	7.5460	1
25	8.9580	1	65	-0.2120	1
26	-0.7740	1	66	8.1260	1
27	0.2830	1	67	7.3300	1
28	-0.9350	1	68	7.4940	1
29	0.9390	1	69	7.1430	1
30	8.3900	1	70	7.0880	1
31	0.8760	1	71	2.2530	1
32	6.4830	1	72	7.8200	1
33	7.7480	1	73	7.4690	1
34	0.1430	1	74	6.1980	1
35	7.8960	1	75	7.4970	1
36	8.2010	1	76	-0.4070	1
37	8.8070	1	77	6.4610	1
38	1.2330	1	78	8.6400	1
39	0.1280	1	79	-0.2400	1
40	6.0450	1	80	8.9940	1

Tabla C1. Muestra asociada a la Simulación 1

IND	Y	GRUPO	IND	Y	GRUPO
81	1.8420	1	121	6.8950	1
82	7.4930	1	122	7.6400	1
83	0.0170	1	123	6.8250	1
84	8.2150	1	124	8.4380	1
85	7.1370	1	125	0.7280	1
86	7.2270	1	126	8.4920	1
87	-0.1620	1	127	8.9990	1
88	1.3950	1	128	6.9150	1
89	8.0170	1	129	5.9980	1
90	1.6940	1	130	7.1600	1
91	9.5720	1	131	8.7020	1
92	6.9100	1	132	7.1520	1
93	0.0570	1	133	-1.2730	1
94	9.2590	1	134	-0.5380	1
95	7.8290	1	135	-2.0050	1
96	8.2650	1	136	0.0840	1
97	6.8580	1	137	7.0610	1
98	8.1900	1	138	7.5730	1
99	8.6180	1	139	5.9830	1
100	7.7780	1	140	1.0360	1
101	7.7010	1	141	0.2320	1
102	-1.9160	1	142	1.7250	1
103	7.9170	1	143	6.9770	1
104	0.2410	1	144	7.8640	1
105	8.2050	1	145	5.9370	1
106	6.7630	1	146	9.4400	1
107	-0.4000	1	147	-2.3570	1
108	7.2280	1	148	7.3960	1
109	-1.8530	1	149	-0.5000	1
110	8.9230	1	150	8.3980	1
111	0.2270	1	151	0.5810	1
112	7.5610	1	152	7.4300	1
113	-0.1820	1	153	8.2700	1
114	7.0540	1	154	7.3190	1
115	-0.3280	1	155	6.6310	1
116	8.2600	1	156	-0.8740	1
117	7.9020	1	157	-0.7050	1
118	8.9140	1	158	8.4570	1
119	7.1020	1	159	6.5850	1
120	0.3870	1	160	-0.1560	1

Tabla C1. Muestra correspondiente a la Simulación 1 **(continuación).**

IND	Y	GRUPO	IND	Y	GRUPO
161	0.2070	1	1	4.3260	2
162	0.7330	1	2	3.7400	2
163	8.3810	1	3	12.3840	2
164	7.9370	1	4	3.4200	2
165	7.6070	1	5	4.4900	2
166	6.5640	1	6	11.1590	2
167	-0.3660	1	7	3.8470	2
168	8.0700	1	8	5.0790	2
169	-0.5050	1	9	12.7270	2
170	6.6500	1	10	5.1360	2
171	1.2560	1	11	2.4510	2
172	6.5830	1	12	2.4060	2
173	7.9660	1	13	4.8910	2
174	-0.5730	1	14	2.4950	2
175	0.0410	1	15	3.6700	2
176	8.1270	1	16	4.9300	2
177	8.2640	1	17	1.6150	2
178	9.0590	1	18	5.5340	2
179	7.9130	1	19	3.5120	2
180	8.4040	1	20	13.5490	2
181	8.7110	1	21	3.5200	2
182	8.7100	1	22	3.5520	2
183	0.4870	1	23	6.0390	2
184	6.6500	1	24	4.1750	2
185	8.3580	1	25	4.2510	2
186	0.0060	1	26	3.8060	2
187	9.3130	1	27	4.0730	2
188	7.8760	1	28	4.6370	2
189	9.5040	1	29	3.5060	2
190	-0.5120	1	30	6.1460	2
191	10.2300	1	31	11.6750	2
192	0.1490	1	32	5.5830	2
193	0.7930	1	33	3.8210	2
194	-0.6340	1	34	3.5170	2
195	7.7260	1	35	10.9130	2
196	8.0930	1	36	2.5450	2
197	6.8740	1	37	4.9060	2
198	-0.3480	1	38	3.9580	2
199	0.8650	1	39	3.5940	2
200	7.9350	1	40	3.7030	2

Tabla C1. Muestra correspondiente a la Simulación (continuación).

IND	Y	GRUPO	IND	Y	GRUPO
41	12.9580	2	81	14.680	2
42	4.6820	2	82	12.345	2
43	3.3640	2	83	12.430	2
44	3.4400	2	84	3.521	2
45	5.1900	2	85	12.524	2
46	2.5170	2	86	3.598	2
47	12.1060	2	87	3.358	2
48	11.8220	2	88	3.101	2
49	12.8230	2	89	5.511	2
50	12.5910	2	90	5.405	2
51	4.1370	2	91	3.861	2
52	3.9550	2	92	3.877	2
53	4.0760	2	93	5.395	2
54	2.0700	2	94	4.777	2
55	6.8400	2	95	3.873	2
56	3.1150	2	96	3.352	2
57	4.3560	2	97	4.477	2
58	4.7140	2	98	5.565	2
59	4.5280	2	99	4.570	2
60	3.2890	2	100	4.150	2
61	3.2900	2	101	2.407	2
62	3.6700	2	102	4.243	2
63	4.7150	2	103	3.530	2
64	3.9790	2	104	2.731	2
65	12.8650	2	105	12.667	2
66	3.0460	2	106	4.190	2
67	2.9120	2	107	12.203	2
68	2.9670	2	108	5.233	2
69	11.9700	2	109	2.983	2
70	6.4180	2	110	5.346	2
71	5.9360	2	111	3.464	2
72	5.1090	2	112	12.004	2
73	10.9200	2	113	3.771	2
74	5.6980	2	114	3.729	2
75	3.3140	2	115	5.050	2
76	2.9530	2	116	4.306	2
77	4.2750	2	117	13.071	2
78	3.7690	2	118	11.595	2
79	5.1970	2	119	3.996	2
80	11.4420	2	120	3.391	2

Tabla C1. Muestra asociada a la Simulación 1 (continuación).

IND	Y	GRUPO	IND	Y	GRUPO
121	3.4190	2	161	5.8850	2
122	3.3550	2	162	3.2440	2
123	3.2000	2	163	2.2070	2
124	4.5650	2	164	5.8170	2
125	12.2670	2	165	4.2770	2
126	11.0520	2	166	5.4640	2
127	11.8270	2	167	3.6030	2
128	3.7090	2	168	4.3690	2
129	4.4820	2	169	3.1370	2
130	11.5810	2	170	2.8150	2
131	4.4230	2	171	3.5630	2
132	2.9380	2	172	3.1980	2
133	5.3390	2	173	3.8240	2
134	2.5770	2	174	3.7350	2
135	4.3970	2	175	3.1520	2
136	4.1170	2	176	11.0890	2
137	14.0340	2	177	3.6940	2
138	4.1960	2	178	3.5880	2
139	1.6330	2	179	3.5120	2
140	11.9060	2	180	3.7620	2
141	3.2540	2	181	3.6670	2
142	4.0030	2	182	3.6440	2
143	7.2700	2	183	2.5560	2
144	3.1540	2	184	4.5800	2
145	3.3540	2	185	2.4420	2
146	4.4780	2	186	2.8920	2
147	4.5850	2	187	3.3430	2
148	3.5440	2	188	3.4440	2
149	1.8130	2	189	12.6540	2
150	3.5690	2	190	11.8510	2
151	5.3650	2	191	2.1240	2
152	11.2450	2	192	4.0470	2
153	4.4910	2	193	12.5560	2
154	4.1810	2	194	3.5210	2
155	3.2770	2	195	5.0500	2
156	5.0250	2	196	2.2700	2
157	13.0850	2	197	11.1310	2
158	11.8600	2	198	3.0850	2
159	4.0990	2	199	3.1450	2
160	4.7090	2	200	3.9580	2

Tabla C1. Muestra asociada a la Simulación 1 (continuación).

C.1.2 Muestra Correspondiente a la Simulación 2

IND	Y1	Y2	GRUPO	IND	Y1	Y2	GRUPO
1	-2.0810	2.0540	1	41	0.5950	0.1160	1
2	1.6890	0.8550	1	42	-0.0370	0.0230	1
3	-0.7070	0.0620	1	43	-2.6700	0.3920	1
4	0.7980	0.4850	1	44	1.7230	-2.5670	1
5	1.5800	-1.2080	1	45	-0.5570	0.6630	1
6	-1.4060	1.1100	1	46	1.0050	0.1600	1
7	1.2980	-0.4550	1	47	-0.9780	1.3870	1
8	1.1510	0.7400	1	48	0.3120	1.1030	1
9	0.8940	1.2850	1	49	1.3420	-2.7840	1
10	2.5050	0.3130	1	50	0.6120	3.1460	1
11	2.4240	-2.6640	1	51	1.2490	-2.5450	1
12	-3.4660	1.6560	1	52	-0.4390	0.6000	1
13	-3.0900	3.2800	1	53	-1.3990	-0.0490	1
14	-0.7190	0.1710	1	54	0.6760	-0.8290	1
15	0.1730	0.1910	1	55	-0.4840	-0.6500	1
16	2.4580	-4.0260	1	56	1.0030	-0.4850	1
17	-2.2270	2.9650	1	57	-2.4160	2.5680	1
18	1.4800	-2.5520	1	58	-1.8970	-1.2590	1
19	-1.2090	1.8830	1	59	0.8130	0.3250	1
20	-1.2570	2.0250	1	60	1.5540	2.1440	1
21	0.7740	2.8620	1	61	-1.4720	-1.0260	1
22	0.8840	-3.5790	1	62	0.2190	1.0960	1
23	0.2930	-0.7500	1	63	2.7380	-3.9140	1
24	0.1860	-0.8850	1	64	1.7070	-2.8860	1
25	-2.1990	1.7580	1	65	-1.3330	-0.2880	1
26	0.2920	0.7860	1	66	1.3360	-0.5660	1
27	1.1340	-3.2140	1	67	0.8570	3.5670	1
28	-0.3550	1.9630	1	68	2.3540	-1.4340	1
29	0.9870	-0.8320	1	69	-0.5920	3.8650	1
30	-0.2970	-2.0180	1	70	-0.3800	-1.9590	1
31	-0.2300	-0.8610	1	71	-0.7240	2.6920	1
32	-1.8790	1.0990	1	72	3.8590	-4.3950	1
33	-1.1440	1.9930	1	73	0.1020	-0.3430	1
34	-1.4610	1.0740	1	74	1.6030	-0.9660	1
35	0.4300	0.5830	1	75	0.6970	-0.9250	1
36	-0.5200	-0.2780	1	76	1.3710	-1.5150	1
37	1.3430	-1.6100	1	77	-0.6990	0.2220	1
38	0.1200	-1.4590	1	78	-1.7760	0.3160	1
39	-1.1850	-0.8950	1	79	-0.7390	1.3240	1
40	0.1080	-1.7890	1	80	-0.6350	2.7980	1

Tabla C2. Muestra correspondiente a la Simulación 2.

IND	Y1	Y2	GRUPO	IND	Y1	Y2	GRUPO
81	-2.4130	-0.1660	1	121	-0.6320	0.3800	1
82	-1.6320	0.2930	1	122	-0.2640	1.6100	1
83	-0.5150	-1.8480	1	123	-0.3300	1.0300	1
84	0.6040	-0.6130	1	124	1.6910	0.2690	1
85	-1.2150	2.2590	1	125	0.5700	0.8080	1
86	1.7060	-1.2890	1	126	-2.4960	0.9500	1
87	0.6850	-1.5740	1	127	2.7500	-2.0570	1
88	1.9810	0.2440	1	128	1.0990	-0.4630	1
89	-0.3670	2.5460	1	129	0.5770	1.8460	1
90	0.4960	0.6960	1	130	0.5580	-2.6870	1
91	-2.1240	0.0130	1	131	-0.5440	-0.3710	1
92	0.4540	-2.5650	1	132	-1.2720	1.1660	1
93	0.4650	-6.1330	1	133	-0.8770	3.5000	1
94	-1.2140	-1.4600	1	134	2.3380	-0.4430	1
95	-1.8200	-2.3290	1	135	0.9350	1.5840	1
96	1.3980	-1.2390	1	136	0.0740	-1.8730	1
97	0.7220	1.0930	1	137	0.7360	-1.3400	1
98	-1.5070	0.1670	1	138	0.6830	0.5300	1
99	0.6830	0.1620	1	139	0.7540	0.2550	1
100	-1.0630	-1.6210	1	140	0.3830	0.2290	1
101	0.1070	-1.2530	1	141	2.3400	1.5240	1
102	0.7930	0.1910	1	142	-1.8350	-0.7560	1
103	-2.2110	0.8120	1	143	-0.0060	-0.8370	1
104	-0.6430	-0.1920	1	144	1.3020	-1.5580	1
105	0.2620	0.0510	1	145	0.8700	0.2830	1
106	-0.4850	0.6130	1	146	-3.8280	1.5700	1
107	2.0130	-1.3880	1	147	2.1380	-0.1990	1
108	-0.2890	-0.4360	1	148	-1.7310	1.4550	1
109	-1.4890	-0.2750	1	149	-3.5560	4.8400	1
110	-0.4270	0.2570	1	150	0.4890	-3.1340	1
111	-0.4850	0.1910	1	1	1.9720	3.4490	2
112	-0.0770	0.7760	1	2	1.1610	2.2190	2
113	2.7160	-1.5630	1	3	2.0240	-0.9010	2
114	-1.6880	1.7160	1	4	1.3420	4.1980	2
115	0.2980	1.2750	1	5	0.4880	4.1820	2
116	0.2680	1.1280	1	6	1.3660	1.6050	2
117	0.0060	-0.2320	1	7	0.7970	0.6320	2
118	2.3970	-3.2670	1	8	1.8530	2.2680	2
119	0.6440	-0.4770	1	9	0.6500	3.9460	2
120	0.8520	-1.4710	1	10	3.2840	3.2220	2

Tabla C2. Muestra asociada a la Simulación 2 (continuación).

IND	Y1	Y2	GRUPO	IND	Y1	Y2	GRUPO
11	3.2280	0.8780	2	51	3.1250	2.8410	2
12	3.3140	-0.1110	2	52	2.0530	2.9260	2
13	1.0860	1.5520	2	53	1.3220	3.7970	2
14	1.2030	0.6840	2	54	0.0920	2.6030	2
15	0.8280	3.9420	2	55	3.3870	-0.7050	2
16	2.1000	4.2230	2	56	1.3980	0.5230	2
17	-1.3920	4.5190	2	57	1.6350	2.4200	2
18	2.9800	1.1960	2	58	3.6990	1.0590	2
19	1.7170	-1.2440	2	59	1.1350	1.0020	2
20	-0.3880	1.9950	2	60	3.5010	-0.1370	2
21	2.8250	0.8890	2	61	1.5690	1.1380	2
22	0.9180	1.9390	2	62	3.3740	0.7290	2
23	1.1570	5.5020	2	63	3.0270	0.9520	2
24	1.4370	2.5190	2	64	1.9980	-0.4880	2
25	2.2440	2.6880	2	65	2.6390	-1.9400	2
26	2.9840	2.5230	2	66	1.8940	2.0980	2
27	2.1640	0.3460	2	67	1.4760	3.2880	2
28	2.3600	1.9630	2	68	0.2870	2.5540	2
29	-0.0270	2.3310	2	69	3.6540	-0.1650	2
30	-1.0560	2.3170	2	70	1.8740	3.2940	2
31	0.9560	-0.5740	2	71	1.5360	-0.9910	2
32	0.5850	3.1900	2	72	1.6760	0.3510	2
33	1.7780	-1.6460	2	73	1.9550	0.5790	2
34	1.2950	1.7250	2	74	-0.2490	0.4900	2
35	3.4230	1.1110	2	75	-0.6870	2.3620	2
36	2.2470	3.1320	2	76	2.7830	1.4750	2
37	1.9750	1.1410	2	77	0.4510	2.0160	2
38	1.8950	-2.1240	2	78	-0.4190	0.2900	2
39	1.1110	-0.3030	2	79	1.3270	1.9390	2
40	2.4210	0.8910	2	80	2.3370	2.4190	2
41	1.5210	2.1770	2	81	1.5310	3.7820	2
42	2.4130	0.2820	2	82	0.7480	2.1220	2
43	1.9800	0.6790	2	83	-0.0830	4.0040	2
44	1.8760	2.1900	2	84	3.3620	2.4200	2
45	2.5410	3.7750	2	85	2.3100	-0.0130	2
46	0.5150	0.2890	2	86	3.4040	3.5420	2
47	0.5470	1.1170	2	87	1.6770	2.2430	2
48	1.1220	2.4830	2	88	1.1310	1.5100	2
49	1.5060	2.7520	2	89	0.6570	0.6620	2
50	0.4730	2.2220	2	90	-0.9120	4.3360	2

Tabla C2. Muestra correspondiente a la Simulación 2 (continuación).

IND	Y1	Y2	GRUPO	IND	Y1	Y2	GRUPO
91	3.7810	1.0880	2	121	3.7680	-1.2320	2
92	0.7740	1.8070	2	122	2.4610	3.6240	2
93	0.3940	1.7070	2	123	2.1260	0.4340	2
94	2.2040	3.0390	2	124	0.4760	3.4320	2
95	2.6170	-2.5650	2	125	2.2040	-0.3450	2
96	1.2450	2.7320	2	126	1.6180	1.8190	2
97	-1.3460	1.5570	2	127	3.4600	0.3930	2
98	2.0510	1.0920	2	128	1.3190	-2.0260	2
99	1.4420	3.2500	2	129	3.4330	1.5820	2
100	0.6260	2.6220	2	130	2.0400	3.0110	2
101	1.1180	0.3390	2	131	-0.6150	1.5890	2
102	0.3420	1.1540	2	132	1.2980	3.8470	2
103	3.2840	-0.2900	2	133	3.2020	-0.1190	2
104	0.0300	4.3060	2	134	1.4760	0.7400	2
105	4.5750	-1.1910	2	135	0.9490	2.8650	2
106	2.2890	2.0520	2	136	2.6130	0.8180	2
107	2.4630	1.2590	2	137	2.7280	-2.2030	2
108	2.2610	2.8370	2	138	3.4920	2.5700	2
109	0.6100	4.6910	2	139	0.3220	1.4680	2
110	-0.5590	4.3800	2	140	1.5880	3.6320	2
111	-0.1260	1.1390	2	141	3.4050	0.9580	2
112	0.9200	4.5370	2	142	2.5800	6.8670	2
113	0.5320	2.7340	2	143	1.7570	2.3100	2
114	0.7930	3.7340	2	144	1.5100	4.1080	2
115	3.2690	1.8780	2	145	1.3420	2.6960	2
116	1.7910	2.9230	2	146	-0.9750	1.2610	2
117	1.4030	0.5750	2	147	0.8880	3.9240	2
118	-1.0650	2.9460	2	148	3.5160	0.7640	2
119	1.3230	-0.0100	2	149	1.7650	1.2980	2
120	2.3410	-0.3310	2	150	3.0520	1.1920	2

Tabla C2. Muestra correspondiente a la Simulación 2 (continuación).

C.1.3 Muestra Asociada a la Simulación 3

IND	Y1	Y2	GRUPO	IND	Y1	Y2	GRUPO
1	-0.6770	1.0980	1	41	1.1830	-0.5320	1
2	3.9850	-5.7510	1	42	0.7050	-0.6100	1
3	-1.7680	1.6860	1	43	-0.7910	0.4330	1
4	-0.9640	2.4580	1	44	0.2860	-2.7670	1
5	0.2000	-0.6430	1	45	-1.9380	1.7310	1
6	0.1640	-0.6780	1	46	1.3180	-1.0410	1
7	-3.3910	3.4100	1	47	-0.3770	0.6880	1
8	3.1730	-5.2100	1	48	1.2580	-2.8950	1
9	6.6700	-7.4770	1	49	-0.2400	1.3760	1
10	5.4830	-5.2120	1	50	2.3020	-2.3470	1
11	-1.8460	3.8390	1	51	-1.4360	-0.1240	1
12	-1.5770	2.5210	1	52	2.0630	-2.9930	1
13	-1.6720	1.9700	1	53	3.8110	-5.4390	1
14	-0.4560	0.7950	1	54	1.5190	-0.7450	1
15	1.4850	-1.8680	1	55	-1.9600	2.5180	1
16	2.7800	-3.9360	1	56	-1.9000	2.4640	1
17	2.4630	-1.6190	1	57	2.2140	-3.4030	1
18	-0.9830	0.4420	1	58	-0.1330	-0.7110	1
19	-1.7090	2.7390	1	59	-0.4350	1.7100	1
20	-1.2390	1.5570	1	60	-0.9530	2.0050	1
21	0.8920	-0.9540	1	61	0.9100	-1.8640	1
22	-1.0560	0.3940	1	62	1.2300	-0.9710	1
23	1.4900	-2.5590	1	63	-2.3360	3.5050	1
24	-0.5770	-0.7830	1	64	-2.5650	4.1000	1
25	0.2280	-0.3520	1	65	0.7390	-1.0680	1
26	-1.4740	1.9090	1	66	0.5660	-0.0490	1
27	-0.8580	3.0050	1	67	0.6180	-0.3250	1
28	1.4520	-1.9480	1	68	1.7740	-2.1010	1
29	-2.0580	2.1410	1	69	-2.9150	3.3260	1
30	-1.4610	1.6600	1	70	-0.2010	-1.0400	1
31	-1.3980	0.4820	1	71	1.1130	-0.2780	1
32	0.2030	-1.6360	1	72	-3.4150	2.4320	1
33	-1.9680	2.8280	1	73	-0.6710	0.8560	1
34	0.7300	-0.1420	1	74	-3.4110	3.7200	1
35	1.4150	-0.8690	1	75	2.1420	-1.1340	1
36	0.1000	-1.9450	1	76	-1.8850	1.4390	1
37	-1.8600	1.5680	1	77	0.4350	-1.1210	1
38	-0.5180	-0.7870	1	78	-2.2770	3.7090	1
39	0.6510	-0.3200	1	79	3.0310	-4.9440	1
40	-1.7160	1.7340	1	80	-1.9710	1.9460	1

Tabla C3. Muestra asociada a la Simulación 3.

IND	Y1	Y2	GRUPO	IND	Y1	Y2	GRUPO
81	0.2970	0.1470	1	121	-2.2560	2.7960	1
82	-2.3820	3.8160	1	122	-0.4450	0.4300	1
83	-0.7880	1.7260	1	123	0.1390	-0.1600	1
84	-1.9260	2.8940	1	124	0.4180	-0.1710	1
85	3.5480	-3.4410	1	125	0.2690	-0.1230	1
86	0.5450	-0.1760	1	126	-2.7440	3.4180	1
87	1.2730	-1.1230	1	127	3.5330	-4 4030	1
88	-1.3620	3.7020	1	128	-0.2680	0.2250	1
89	1.0170	0.7670	1	129	-1.7780	1.6670	1
90	0.4140	0.4630	1	130	0.1060	0.6920	1
91	1.0900	-1.3700	1	131	0.6680	-1.2540	1
92	3.9730	-4.2840	1	132	0.6120	0.3600	1
93	-0.7780	0.5210	1	133	-1.2660	-0.5390	1
94	3.4550	-4.7190	1	134	-0.6680	1.2940	1
95	-1.6040	1.6850	1	135	-0.0740	-0.1650	1
96	0 0490	0.0690	1	136	1.9130	-2.5070	1
97	-0.9340	1.6680	1	137	-1.0610	0.9230	1
98	-0.6750	0.6910	1	138	-1.3160	0.4960	1
99	-1.1150	0.7920	1	139	-0.9170	1.2690	1
100	1.8400	-2.8880	1	140	0.7190	-2.5110	1
101	1.2960	-1.8140	1	141	0.2180	-1.3660	1
102	0.0530	0.0310	1	142	-2.8490	2.4320	1
103	2.7270	-3.5260	1	143	0.7840	-0.9950	1
104	-0.2480	1.0650	1	144	2.8510	-3.3100	1
105	-7.2230	9.3090	1	145	-2.1830	2.7120	1
106	-0.6120	2.2180	1	146	0.3960	0.5950	1
107	2.0660	-2.4360	1	147	-0.4780	-0.0670	1
108	-1.6530	2.1670	1	148	-0.9030	-0.3620	1
109	0.0370	-0.5910	1	149	-0.2900	-0.2900	1
110	1.1660	-1.2990	1	150	0.4260	-1.5580	1
111	0.1600	-0.7770	1	1	1.055	1.864	2
112	1.3920	-1.7760	1	2	-2.617	-3.222	2
113	-0.5500	1.1560	1	3	-0.861	-1.412	2
114	-5.4580	5.0560	1	4	1.539	1.132	2
115	-0.8790	0.7830	1	5	5.408	6.256	2
116	-0.8750	1.1710	1	6	1.014	0.695	2
117	1.3460	-1.8110	1	7	-1.12	-0.985	2
118	-1.4330	-0.4140	1	8	0.094	-0.014	2
119	1.2930	-0.9230	1	9	0.108	0.724	2
120	-1.8100	3.0970	1	10	-0.831	-0.671	2

Tabla C3. Muestra correspondiente a la Simulación 3 (continuación).

IND	Y1	Y2	GRUPO	IND	Y1	Y2	GRUPO
11	2.038	2.433	2	51	0.791	-0.178	2
12	0.212	-0.474	2	52	-1.527	-1.863	2
13	-1.705	-2.317	2	53	-0.9	-1.416	2
14	-1.805	-0.643	2	54	-1.826	-2.216	2
15	-1.532	-2.096	2	55	2.146	4.274	2
16	0.853	1.127	2	56	0.069	0.436	2
17	-0.725	-0.717	2	57	2.066	2.914	2
18	-2.253	-3.305	2	58	-1.665	-1.412	2
19	1.006	2.532	2	59	-3.476	-3.606	2
20	0.241	0.878	2	60	0.119	0.796	2
21	0.652	0.488	2	61	-0.434	-2.099	2
22	1.051	1.818	2	62	-1.734	-1.505	2
23	1.264	2.074	2	63	2.53	4.382	2
24	-0.369	-1.449	2	64	2.468	3.164	2
25	0.267	0.538	2	65	-0.322	-0.308	2
26	-0.797	-0.983	2	66	2.264	3.439	2
27	-2.357	-3.29	2	67	1.59	1.593	2
28	0.205	-1.01	2	68	-0.319	-1.073	2
29	1.372	1.85	2	69	1.671	1.22	2
30	-0.258	0.605	2	70	-1.504	-1.404	2
31	0.272	-0.15	2	71	0.608	1.015	2
32	2.604	2.336	2	72	-2.389	-3.03	2
33	-0.829	-1.151	2	73	2.689	2.753	2
34	1.086	0.206	2	74	-1.197	-1.743	2
35	0.519	1.683	2	75	0.532	0.328	2
36	-2.274	-2.561	2	76	1.066	2.168	2
37	1.776	1.79	2	77	-1.594	-2.395	2
38	-3.868	-3.682	2	78	-1.57	-1.241	2
39	-0.977	-2.521	2	79	-2.443	-2.498	2
40	-0.591	-0.945	2	80	0.21	1.131	2
41	4.408	5.635	2	81	-1.375	-1.766	2
42	-1.351	-1.099	2	82	-0.922	-1.148	2
43	3.623	4.096	2	83	1.243	1.184	2
44	-3.331	-3.639	2	84	2.691	2.315	2
45	-1.052	-1.964	2	85	-3.401	-4.18	2
46	0.6	0.658	2	86	-1.279	-2.419	2
47	1.454	1.698	2	87	-0.75	-0.657	2
48	1.26	1.003	2	88	0.293	0.144	2
49	1.167	0.422	2	89	-1.632	-3.632	2
50	-3.085	-4.37	2	90	1.2	1.97	2

Tabla C3. Muestra asociada a la Simulación 3 (continuación).

IND	Y1	Y2	GRUPO	IND	Y1	Y2	GRUPO
91	1.592	2.657	2	121	0.289	-0.162	2
92	-1.567	-1.373	2	122	-0.581	-0.365	2
93	-0.363	-0.934	2	123	0.908	1.778	2
94	-1.667	-1.465	2	124	-1.038	-1.648	2
95	2.636	3.597	2	125	2.599	3.676	2
96	2.458	2.733	2	126	-0.262	-0.158	2
97	2.01	2.639	2	127	-1.234	-1.756	2
98	-0.334	0.677	2	128	2.206	2.461	2
99	-0.702	-0.056	2	129	-0.996	-1.125	2
100	-1.609	-1.402	2	130	1.41	1.987	2
101	2.705	2.276	2	131	3.009	3.756	2
102	0.32	-0.321	2	132	-1.073	-0.472	2
103	-0.875	-0.906	2	133	3.438	5.18	2
104	-0.025	-0.062	2	134	-1.38	-1.762	2
105	-0.01	-0.313	2	135	1.673	3.029	2
106	-0.978	-2.213	2	136	-2.925	-3.013	2
107	1.04	1.302	2	137	-1.059	-2.264	2
108	-1.584	-1.305	2	138	-1.608	-1.456	2
109	1.097	1.431	2	139	0.809	0.895	2
110	-2.415	-2.431	2	140	-0.023	0.712	2
111	0.399	1.216	2	141	2.36	3.015	2
112	-2.516	-3.41	2	142	-2.285	-4.209	2
113	1.069	1.21	2	143	-0.843	-1.69	2
114	0.325	-0.027	2	144	-0.225	-0.263	2
115	0.574	0.364	2	145	-0.915	-1.236	2
116	0.673	1.557	2	146	2.71	2.873	2
117	-0.512	-0.656	2	147	-0.452	0.1	2
118	0.051	-0.079	2	148	0.573	1.501	2
119	4.157	5.53	2	149	-0.004	0.464	2
120	1.103	1.164	2	150	-2.514	-2.995	2

Tabla C3. Muestra correspondiente a la Simulación 3 (continuación).

C.2 Muestra de los Irises de Fisher

C.2.1 Programa en el Paquete SAS para los Irises de Fisher

```
                S A S   S A M P L E   L I B R A R Y

      NAME: DISCEX2
      TITLE: DISCRIM DOCUMENTATION EXAMPLE 2
PRODUCT: SAS/STAT
  SYSTEM: ALL
      KEYS: DISCRIM
     PROCS: DISCRIM FORMAT PLOT
      DATA: FISHER (1936) IRIS DATA

 SUPPORT: WFK
      REF: PROC DISCRIM, EXAMPLE 2
     MISC:



proc format;
   value specname
      1='SETOSA     '
      2='VERSICOLOR'
      3='VIRGINICA ';
   value specchar
      1='S'
      2='O'
```

```
      3='V';
run;

data iris;
   title 'Discriminant Analysis of Fisher (1936) Iris Data';
   input sepallen sepalwid petallen petalwid species @@;
   format species specname.;
   label sepallen='Sepal Length in mm.'
         sepalwid='Sepal Width  in mm.'
          petallen='Petal Length in mm.'
          petalwid='Petal Width  in mm.';
   cards;
50 33 14 02 1 64 28 56 22 3 65 28 46 15 2 67 31 56 24 3
63 28 51 15 3 46 34 14 03 1 69 31 51 23 3 62 22 45 15 2
59 32 48 18 2 46 36 10 02 1 61 30 46 14 2 60 27 51 16 2
65 30 52 20 3 56 25 39 11 2 65 30 55 18 3 58 27 51 19 3
68 32 59 23 3 51 33 17 05 1 57 28 45 13 2 62 34 54 23 3
77 38 67 22 3 63 33 47 16 2 67 33 57 25 3 76 30 66 21 3
49 25 45 17 3 55 35 13 02 1 67 30 52 23 3 70 32 47 14 2
64 32 45 15 2 61 28 40 13 2 48 31 16 02 1 59 30 51 18 3
55 24 38 11 2 63 25 50 19 3 64 32 53 23 3 52 34 14 02 1
49 36 14 01 1 54 30 45 15 2 79 38 64 20 3 44 32 13 02 1
67 33 57 21 3 50 35 16 06 1 58 26 40 12 2 44 30 13 02 1
77 28 67 20 3 63 27 49 18 3 47 32 16 02 1 55 26 44 12 2
50 23 33 10 2 72 32 60 18 3 48 30 14 03 1 51 38 16 02 1
61 30 49 18 3 48 34 19 02 1 50 30 16 02 1 50 32 12 02 1
61 26 56 14 3 64 28 56 21 3 43 30 11 01 1 58 40 12 02 1
51 38 19 04 1 67 31 44 14 2 62 28 48 18 3 49 30 14 02 1
```

```
51 35 14 02 1 56 30 45 15 2 58 27 41 10 2 50 34 16 04 1
46 32 14 02 1 60 29 45 15 2 57 26 35 10 2 57 44 15 04 1
50 36 14 02 1 77 30 61 23 3 63 34 56 24 3 58 27 51 19 3
57 29 42 13 2 72 30 58 16 3 54 34 15 04 1 52 41 15 01 1
71 30 59 21 3 64 31 55 18 3 60 30 48 18 3 63 29 56 18 3
49 24 33 10 2 56 27 42 13 2 57 30 42 12 2 55 42 14 02 1
49 31 15 02 1 77 26 69 23 3 60 22 50 15 3 54 39 17 04 1
66 29 46 13 2 52 27 39 14 2 60 34 45 16 2 50 34 15 02 1
44 29 14 02 1 50 20 35 10 2 55 24 37 10 2 58 27 39 12 2
47 32 13 02 1 46 31 15 02 1 69 32 57 23 3 62 29 43 13 2
74 28 61 19 3 59 30 42 15 2 51 34 15 02 1 50 35 13 03 1
56 28 49 20 3 60 22 40 10 2 73 29 63 18 3 67 25 58 18 3
49 31 15 01 1 67 31 47 15 2 63 23 44 13 2 54 37 15 02 1
56 30 41 13 2 63 25 49 15 2 61 28 47 12 2 64 29 43 13 2
51 25 30 11 2 57 28 41 13 2 65 30 58 22 3 69 31 54 21 3
54 39 13 04 1 51 35 14 03 1 72 36 61 25 3 65 32 51 20 3
61 29 47 14 2 56 29 36 13 2 69 31 49 15 2 64 27 53 19 3
68 30 55 21 3 55 25 40 13 2 48 34 16 02 1 48 30 14 01 1
45 23 13 03 1 57 25 50 20 3 57 38 17 03 1 51 38 15 03 1
55 23 40 13 2 66 30 44 14 2 68 28 48 14 2 54 34 17 02 1
51 37 15 04 1 52 35 15 02 1 58 28 51 24 3 67 30 50 17 2
63 33 60 25 3 53 37 15 02 1
;

proc plot data=iris;
   plot petalwid*petallen=species
        / vpos=17 vaxis=-4 to 28 by 2 haxis=0 to 75 by 5;
   format species specchar.;
```

```
run;

data plotdata;
   do petalwid=-4 to 28 by 2;
   do petallen=0 to 75;
         output;
   end;
  end;
run;

 macro contour;
   proc plot data=plotd;
      plot petalwid*petallen=setosa
            petalwid*petallen=versicol
            petalwid*petallen=virginic
           / contour=6 vpos=17 hpos=76 haxis=0 to 75 by 5;
      title3 'Plot of Estimated Densities';
   run;
   proc plot data=plotp;
      plot petalwid*petallen=setosa
            petalwid*petallen=versicol
            petalwid*petallen=virginic
           / contour=6 vpos=17 hpos=76 haxis=0 to 75 by 5;
      title3 'Plot of Posterior Probabilities';
   run;
   proc plot data=plotp;
      plot petalwid*petallen=_into_
           / vpos=17 hpos=76 haxis=0 to 75 by 5;
```

```
      format _into_ specchar.;
      title3 'Plot of Classification Results';
   run;
mend;

proc discrim data=iris testdata=plotdata testout=plotp testoutd=plotd
         method=normal pool=yes short noclassify crosslisterr;
   class species;
   var petal:;
   title2 'Using Normal Density Estimates with Equal Variance';
run;

 contour

proc discrim data=iris testdata=plotdata testout=plotp testoutd=plotd
         method=normal pool=no short noclassify crosslisterr;
   class species;
   var petal:;
   title2 'Using Normal Density Estimates with Unequal Variance';
run;

 contour

proc discrim data=iris testdata=plotdata testout=plotp testoutd=plotd
         method=npar kernel=normal r=.5 pool=yes
         short noclassify crosslisterr;
   class species;
   var petal:;
```

```
   title2 'Using Kernel Density Estimates with Equal Bandwidth';
run;

 contour

proc discrim data=iris testdata=plotdata testout=plotp testoutd=plotd
          method=npar kernel=normal r=.5 pool=no
          short noclassify crosslisterr;
   class species;
   var petal:;
   title2 'Using Kernel Density Estimates with Unequal Bandwidth';
run;

 contour
```

Bibliografía

Anderson, T. W., (1994), *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

Brokwell, P. J. & Davis, R. A. (1991), *Time Series: Theory and Methods.* Springer-Verlag, New York.

Castrejón, J. L. (1996), *Análisis de Componentes Principales y otras Proyecciones Lineales.* Tesis de Licenciatura en Actuaría, Facultad de Ciencias UNAM, México D.F..

Chatfield, C., & Collins, A. F. (1980), *Introduction to Multivariate Analysis.* Chapman and Hall, London, New York.

Chambers, J. M. & Hastie, T. J. (1993), *Statistical Models in S.* Chapman and Hall, London, New York.

Courant, R. & John F. (1978), *Introducción al Cálculo y al Análisis Matemático. Vol I.* LIMUSA, México D. F..

Dillon, W. R. & Goldstein M. (1984), *Multivariate Analysis Methods and Applcations.* Wiley, New York.

Everitt, B. S. & Dunn, G. (1991), *Applied Multivariate Data Analysis.* Edward Arnold, Great Britain.

Fraleigh, J. B. & Beauregard R. A. (1989), *Linear Algebra.* Addison-Wesley, Wilmington, Delaware.

Friedberg, S. H. & Insel, A. J. (1979), *Linear Algebra.* Prentice Hall.

Hand, D. J. (1981), *Discrimination and Classification.* Wiley, New York.

Harris, R. J. (1975), *A Primer of Multivariate Statistics.* Academic Press, London.

Hasser, N. B., LaSalle, J. P. & Sullivan, J. A. (1979), *Análisis Matemático I, Curso de Introducción.* Trillas, México D. F..

Härdle, W. (1990), *Smoothing Techniques: with Implementation in S.* Springer-Verlang, New York.

Huberty, P. J. (1994), *Applied Discriminant Analysis.* Wiley, New York.

Jackson, J. E. (1991), *A User's Guide to Principal Componentes.* Wiley.

Johnson, R. A. & Wichern, D. W. (1992), *Applied Multivariate Statistical Analysis.* Prentice Hall, New Jersey.

Kshirsagar, A. M., *Multivariate Analysis.* Marcel Dekker, New York.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1984), *Multivariate Analysis.* Academic Press, London.

Mendoza, J. R. (1987), *La Distribución Normal Multivariada y su Relación con otras Distribuciones.* Tesis de Licenciatura en Actuaría, Facultad de Ciencias UNAM, México D.F..

MacLachlan, J. G. (1992), *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York.

Mood, A. M., Graybill, F.A. & Boes, D. C. (1974), *Introduction to the Theory of Statistics.* McGraw-Hill, Singapore.

Press, S. J. (1982), *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference.* Robert E. Krieger Publising Company, Malabar Florida.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice end Visualization.* Wiley, New York.

Seber, G. A. F. (1984), *Multivariate Observations.* Wiley, New York.

Srivastava, M. S. & Carter, E. M. (1983), *An Introduction to Applied Multivariate Statistics.* Elsevier Science Publishing, New York.

Strang, G. (1976), *Linear Algebra and Its Applications.* Academic Press, London.

Venables, W. N. & Ripley, B. D. (1994), *Modern Applied Statistics with S-Plus.* Springer-Verlag, New York.