UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

ANÁLISIS DISCRIMINANTE PARA DATOS CIRCULARES

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIO

P R E S E N T A :

JORGE EDUARDO ROJAS JIMÉNEZ

DIRECTORA DE TESIS:

MAT. MARGARITA ELVIRA CHÁVEZ CANO

Ciudad Universitaria, CD. MX. 2018

1. Datos del alumno
Rojas
Jiménez
Jorge Eduardo
55 30 77 77 66
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
308234190

2. Datos del tutor
Mat.
Margarita Elvira
Chávez
Cano

3. Datos del sinodal 1
Dra.
Ruth Selene
Fuentes
García

4. Datos del sinodal 2
Act.
Jaime
Vázquez
Alamilla

5. Datos del sinodal 3
Act.
Ángel Manuel
Godoy
Aguilar

6. Datos del sinodal 4
Act.
Francisco
Sánchez
Villareal

7. Datos del trabajo escrito
Análisis discriminante para datos circulares
100 p
2018

INTRODUCCIÓN

En múltiples trabajos científicos, se obtienen observaciones que pueden considerarse como valores sobre una circunferencia o en la superficie de una esfera. Estas observaciones son llamadas datos direccionales, han surgido de manera natural en la geología, meteorología, astronomía, biología, medicina y en otros campos. La literatura sobre la estadística para datos direccionales ha aumentado considerablemente, sin embargo, sobre el análisis discriminante con este enfoque poco se ha escrito. El propósito de este trabajo es estudiar el problema de análisis discriminante para datos circulares, en el que desarrolla el caso de dos grupos solamente debido a la complejidad de los desarrollos.

En el primer capítulo se presenta el análisis discriminante usual para dos grupos, se ilustra con un ejemplo de mosquitos y la formulación de reglas para la clasificación de futuras observaciones. Se define una medida de distancia entre dos poblaciones multivariadas, primero llamada distancia estándar multivariada y después se ve que se trata de la distancia de Mahalanobis cuando se eleva al cuadrado. La distancia está estrechamente relacionada con la función discriminante lineal, que es una combinación lineal de las variables tal que separa lo más posible a las dos distribuciones. En la sección 1.2 se desarrollan e ilustran estas nociones con detalle. En la sección 1.3, se aplican los conceptos de distancia estándar y de función discriminante con el apoyo de ejemplos para mostrar los alcances de esta metodología. Las secciones 1.2 y 1.3 se basan en conceptos heurísticos, sin hacer ninguna suposición de distribución específica. En la

sección 1.4 se retoma la función lineal discriminante, pero basándose en un desarrollo teórico con la distribución normal, utilizando el principio de clasificación basado en la probabilidad máxima posterior. Se observa que la suposición de distribuciones normales multivariadas en dos grupos con matrices de varianzas y covarianzas iguales conduce al mismo método de reducir los datos p-dimensionales a una sola variable como en el enfoque del principio heurístico de la sección 1.2. En la sección 1.5 se estudian varios tipos de tasas de error para medir (o estimar) el éxito de una regla de clasificación. Finamente en la sección 1.6 se observa una conexión relativamente poco conocida pero muy interesante entre las funciones discriminantes lineales y las diferencias de medias condicionales.

En el segundo capítulo se presenta la estadística circular introduciendo a los datos circulares con un ejemplo, se continua con la representación gráfica y la estadística descriptiva de forma general para este tipo de datos. En la sección 2.3 se define la distancia circular, que es una forma intuitiva para llevar a cabo el análisis discriminante para datos circulares, en la sección 2.4 se mencionan algunos métodos para obtener distribuciones circulares y se muestra principalmente, la distribución von Mises con sus propiedades. En la sección 2.5 se presentan los datos direccionales multidimensionales, donde se define la dirección de un vector y se estudia la distribución von Mises-Fisher como una generalización adecuada de la distribución von Mises. En la sección 2.6 se presentan los métodos de clasificación para datos circulares, principalmente se desarrolla el enfoque del análisis discriminante bajo el supuesto de la distribución von Mises, definiendo una función discriminante direccional que permite desarrollar un algoritmo de clasificación con el discriminante circular.

En el tercer capítulo se muestran las aplicaciones de los capítulos anteriores, utilizando el análisis discriminante para datos esféricos y circulares basados en la distribución von Mises-Fisher, con apoyo de funciones codificadas en el programa estadístico R y también, se ejemplifica la predicción de nuevas observaciones utilizando este análisis.

ÍNDICE GENERAL

CAPÍTULO 1

ANÁLISIS DISCRIMINANTE

1.1. Discriminación y clasificación

Ejemplo 1.1.1. Clasificación de mosquitos

Los biólogos Grogan y Wirth (1981) describen dos especies recién descubiertas de mosquitos depredadores, Fasciata Amerohelea (AF) y A. Pseudofasciata (APF). Debido a que las dos especies son similares en apariencia, es útil para los biólogos clasificar a un espécimen por las características que son fáciles de medir. Entre las muchas características que distinguen a AF de APF, Grogan y Wirth recolectaron mediciones de longitudes de antenas y de alas en milímetros, de nueve AF y seis APF. Los datos se muestran en la Tabla 1.1.

¿Se puede encontrar una regla que permita clasificar un mosquito dado como AF o APF, respectivamente, basándose únicamente en las mediciones de las longitudes de antena y de ala?, tal regla podría ser de importancia en la práctica, por ejemplo, una especie puede ser un polinizador valioso, otra especie portadora de alguna enfermedad.

Para comenzar, se utilizará un enfoque descriptivo con el fin de atacar el problema de clasificación para este ejemplo. La mejor manera de visualizar dos conjuntos pequeños de datos es con un diagrama de dispersión y con esto desarrollar un análisis, como en la Figura 1.1. Los elementos de las dos especies mencionadas parecen estar claramente separadas en esta

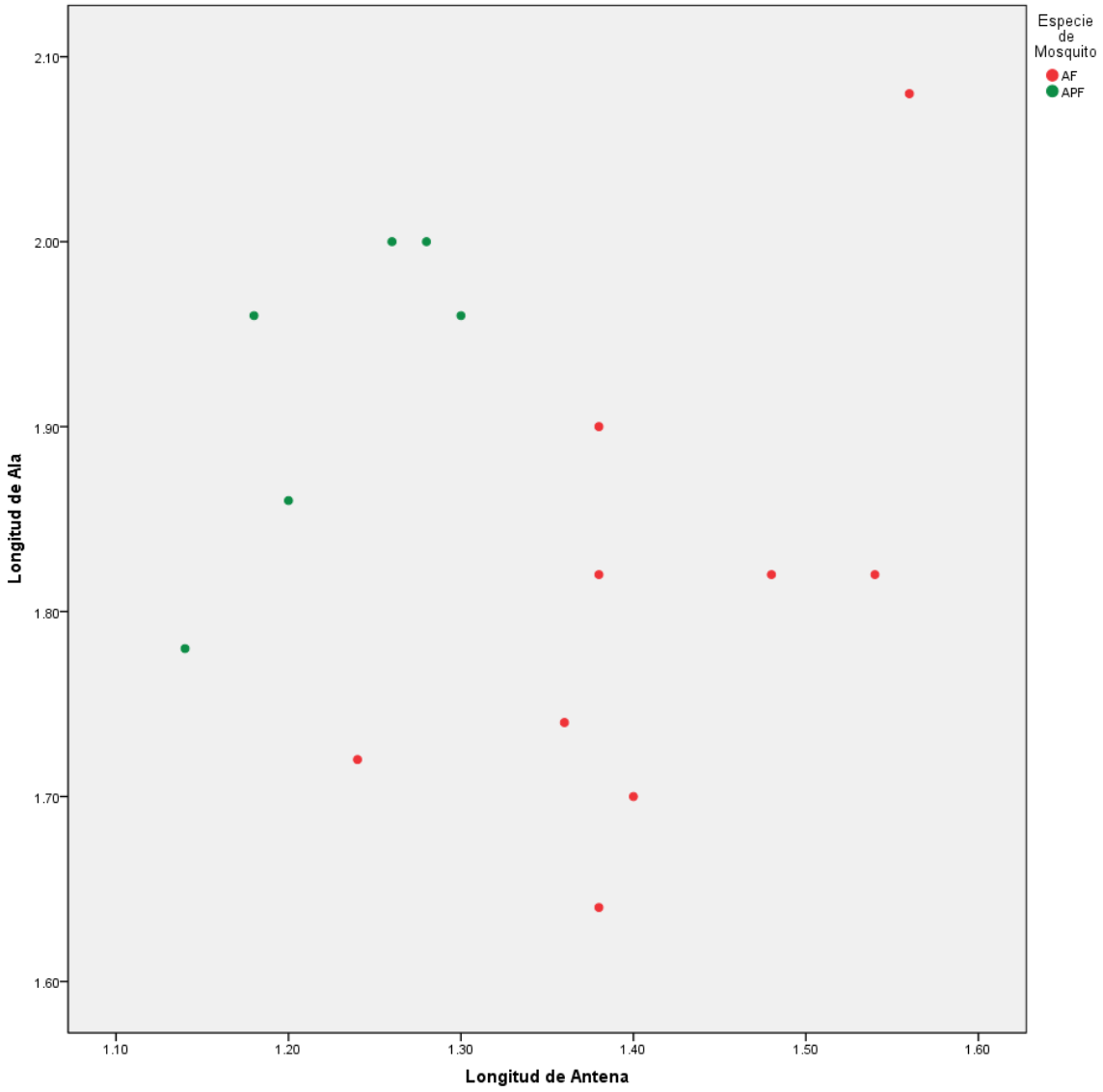


Figura 1.1: Diagrama de dispersión del Ejemplo 1.1.1

Especie	Longitud de Antena (x) (mm)	Longitud de Ala (y) (mm)
AF	1.38	1.64
AF	1.40	1.70
AF	1.24	1.72
AF	1.36	1.74
AF	1.38	1.82
AF	1.48	1.82
AF	1.54	1.82
AF	1.38	1.90
AF	1.56	2.08
APF	1.14	1.78
APF	1.20	1.86
APF	1.18	1.96
APF	1.30	1.96
APF	1.26	2.00
APF	1.28	2.00

Tabla 1.1: Datos de mosquitos. Fuente: Grogan y Wirth, 1981

gráfica. Incluso sin conocimientos de estadística, se podría formular un método de clasificación observando la ubicación de los puntos en la gráfica.

Este método puede ser más preciso trazando una línea que separe a los dos conjuntos de puntos (La Figura 1.2 muestra el mismo diagrama de dispersión, pero con esta línea delimitante).

$$\text{Longitud de ala} - \text{Longitud de antena} = 0.58$$

La cual es bastante arbitraria, pero es útil hasta este punto. La línea delimitante es $y - x = 0.58$, y las dos regiones de clasificación se puede definir como:

$$C_1 = (x, y) : y - x < 0.58 \quad \text{(para AF)}$$

$$C_2 = (x, y) : y - x > 0.58 \quad \text{(para APF)}$$

Por lo tanto, se puede tomar una decisión para clasificar a los mosquitos únicamente basándose en el valor de la diferencia de $y - x$.

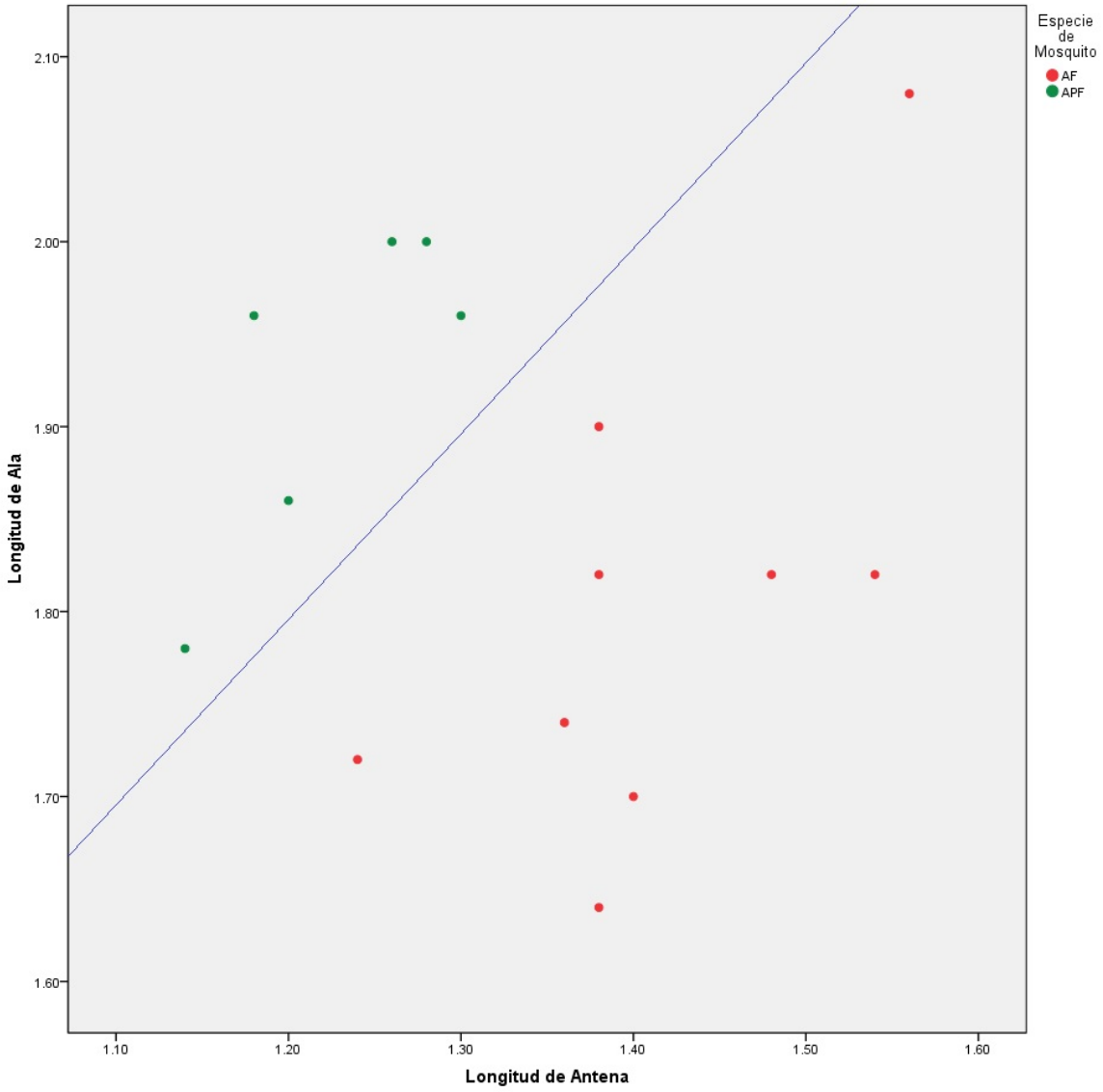


Figura 1.2: Diagrama de dispersión del Ejemplo 1.1.1 con línea delimitante $y - x = 0.58$

¿Qué pasa si se necesitan analizar tres variables (o más)? Claramente, el desarrollo será más complicado, ya que construir y visualizar diagramas de dispersión tridimensionales es complicado. Sin embargo, con el ejemplo anterior (Observando la Figura 1.1) es claro que ninguna de las dos variables por sí misma permitiría hacer un desarrollo para una buena regla de clasificación, ya que no habría un "traslape" considerable entre los dos grupos.

La idea de mirar distribuciones unidimensionales es prometedora si se permite, para una mayor flexibilidad, cambiar a un nuevo sistema de coordenadas. Pero primero, se va a ilustran gráficamente cómo una distribución unidimensional se puede obtener a partir de una distribución bidimensional (se entenderá el propósito en el desarrollo este trabajo). La Figura 1.3 muestra el mismo diagrama de dispersión, esta vez con distribuciones univariadas trazadas a lo largo de los ejes de coordenadas, como si fuera una lluvia de puntos sobre los dos nuevos ejes trazados, que cae horizontal y verticalmente según la localización de cada uno en la gráfica original.

La idea clave, ahora, es jugar con el sistema de coordenadas, rotarlo tal vez 45 grados. Esto introducirá a formular una buena regla de clasificación, ya que con las gráficas rotadas se observa que hay "traslape" en el conjunto de puntos colocados en las nuevas coordenadas y es difícil tomar una decisión de clasificación a primera vista.

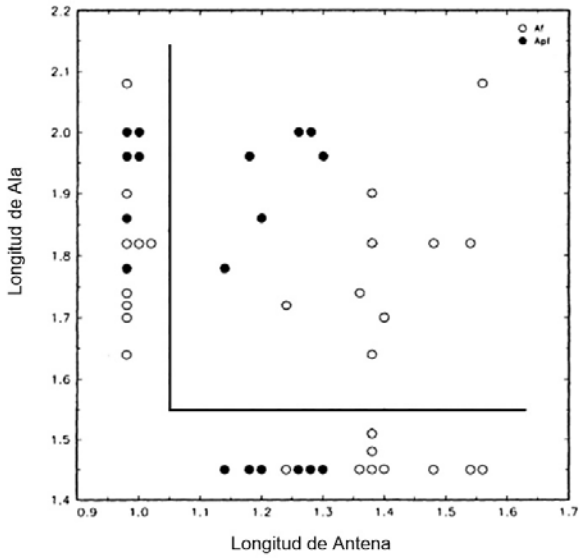


Figura 1.3: Diagrama de dispersión del Ejemplo 1.1.1 con una proyección de los datos en los ejes

El ejemplo anterior es bastante común para el análisis discriminante. Las dos muestras de mosquitos con la pertenencia a grupos conocidos por lo general hacen referencia a las muestras de entrenamiento.

Se definirá una medida de distancia entre dos distribuciones multivariadas, llamada "distancia estándar multivarianda". La distancia estándar está estrechamente relacionada con la función discriminante lineal, que se puede considerar como una combinación lineal de las variables que separan las dos distribuciones de la manera más óptima posible.

1.2. Distancia estándar y la función discriminante lineal

En el análisis de regresión, las funciones de regresión lineal son introducidas como combinaciones lineales de p variables regresoras $X_1, \ldots, X_p$ tal que la variable dependiente Y se aproxima óptimamente utilizando el método de los mínimos cuadrados. Un enfoque similar da lugar a la definición de la función discriminante lineal.

Definición 1.2.1. *Sea X una variable aleatoria con media μ y varianza $\sigma^2 > 0$. La distancia estándar entre dos números x_1 y x_2 con respecto a la variable aleatoria X está dada por:*

$$\Delta_X(x_1, x_2) = \frac{|x_1 - x_2|}{\sigma} \tag{1.1}$$

Observaciones:

1. Si $\sigma = 1$, entonces la distancia estándar es igual a la distancia Euclidiana.

2. La distancia estándar es invariante bajo transformaciones lineales no degeneradas: sea $Y = aX + b$, donde $a \neq 0$ y b son constantes fijas. Similarmente, transformando x_1 y x_2 a $y_i = ax_i + b$, $i = 1, 2$. Entonces:

$$\begin{aligned} \Delta_Y(y_1, y_2) &= \frac{|y_1 - y_2|}{\sqrt{Var[Y]}} \\ &= \frac{|a(x_1 - x_2)|}{\sqrt{a^2\sigma^2}} \\ &= \Delta_X(x_1, x_2). \end{aligned} \tag{1.2}$$

3. La distancia estándar es más usada para distribuciones simétricas. Si la distribución de X es simétrica con media μ, entonces la $f.d.p$ de X depende del argumento x sólo a través de $\Delta_X(x,\mu)$.

La observación anterior se ilustrará con un ejemplo. Por comodidad, se denotará $\Delta(x)$ a cambio de $\Delta_X(x,\mu)$.

Ejemplo 1.2.1. Suponga que X se distribuye normal con media μ y varianza σ^2. Entonces, la función de densidad de X está dada por:

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\Delta^2(x)\right], x \in \Re$$

Se hará referencia a una situación directamente relacionada con el análisis discriminante. Suponga que una variable X es medida en términos de dos poblaciones, con medias diferentes μ_j pero con varianza común σ^2. Cuando se habla de esperanzas, a veces se agrega un índice al operador para indicar la distribución en la que hace referecnia el cálculo, esto es:

$$E_1[X] = \mu_1, \quad Var[X] = \sigma^2 \quad \text{para la población 1}$$

$$E_2[X] = \mu_2, \quad Var[X] = \sigma^2 \quad \text{para la población 2} \tag{1.3}$$

Se puede medir la distancia entre dos medias utilizando:

$$\Delta_X(\mu_1,\mu_2) = \frac{|\mu_1 - \mu_2|}{\sigma} \tag{1.4}$$

Si la distribución de X es simétrica en ambas poblaciones, entonces cualquier valor dado de $\Delta = \Delta(\mu_1,\mu_2)$ está relacionado de forma única con una cierta cantidad de traslape. Es importante, sobre todo para el propósito del análisis de clasificación, desarrollar intuitivamente el concepto de distancia estándar, el siguiente ejemplo muestra esta idea.

Ejemplo 1.2.2. Este ejemplo lleva directamente a la teoría clásica de la distribución normal de clasificación. Suponga que $X \sim \mathcal{N}\left(\mu_j, \sigma^2\right)$ en la población j, $j = 1,2$. Sea $\Delta = \Delta(\mu_1,\mu_2)$ la distancia estándar entre las dos medias, y sea $\gamma = (\mu_1 + \mu_2)/2$ el punto medio entre las dos

medias. La Figura 1.4 muestra las curvas de densidad para $\Delta = 1, 2, 3, 4, 5$. El traslape entre las dos curvas de densidad es considerable para $\Delta = 1$ y casi cero para $\Delta = 5$, pero, ¿Cómo se calcula el traslape?, haciendo referencia a las tasas de error, que se estudiaran en un capitulo posterior, se consideran las áreas bajo la curva de densidad a la derecha y a la izquierda del punto medio γ. Suponga por simplicidad, que $\mu_1 > \mu_2$. Se puede requerir la probabilidad de que X sea inferior a γ, dado que X es de la población I, es decir, se va a evaluar la integral dada por el área sombreada en la Figura 1.4. Para $Z \sim \mathcal{N}(0, 1)$, se tiene que:

$$\Phi(z) = Pr[Z \leq z] \tag{1.5}$$

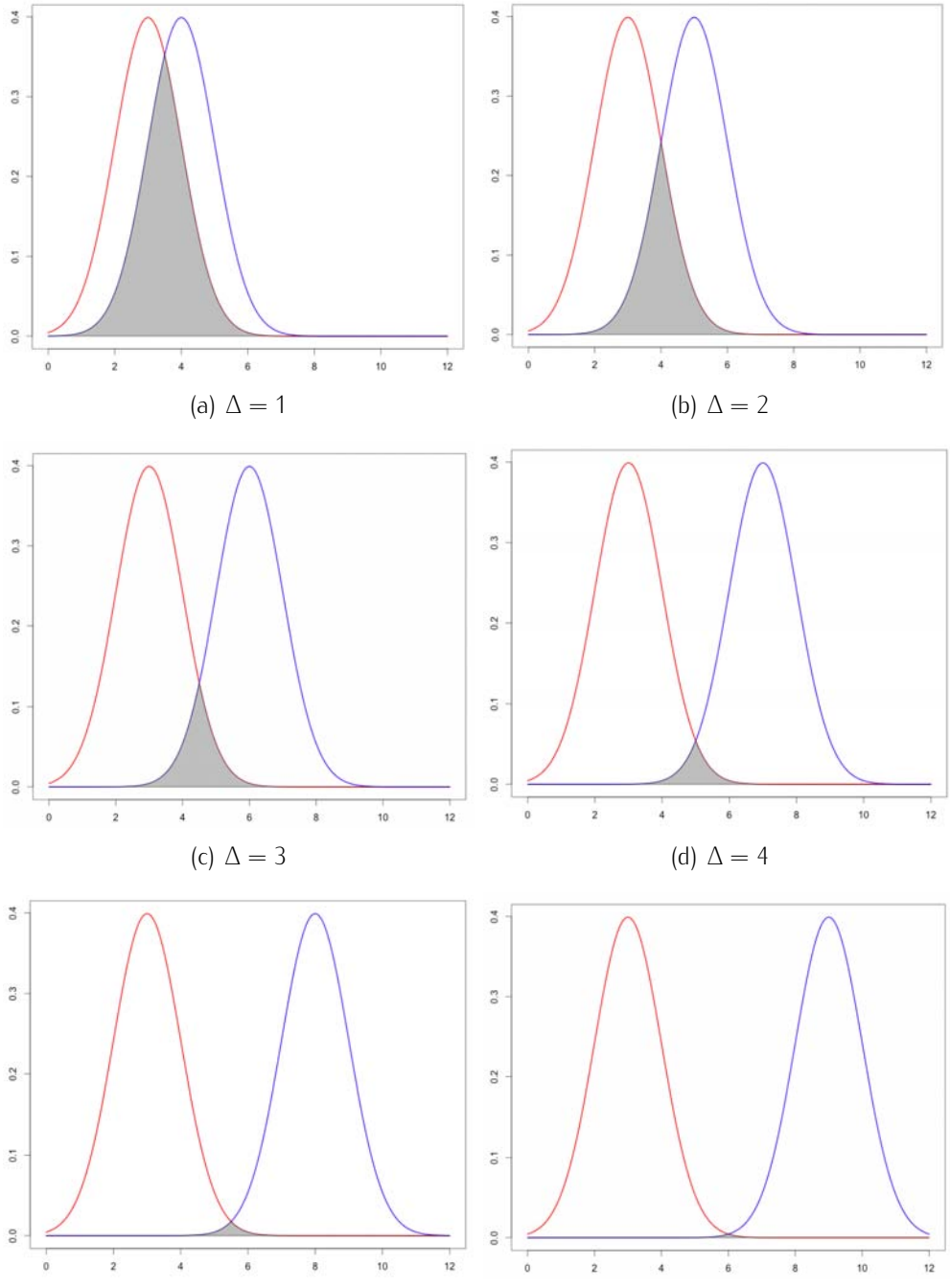
y $Pr[X \leq \gamma \mid P_1]$ para calcular la integral deseada, donde "P_1" indica que el cálculo está basado en la distribución de la población 1. Entonces:

$$\begin{aligned} Pr[X \leq \gamma \mid P_1] &= Pr\left[\frac{X - \mu_1}{\sigma} \leq \frac{\gamma - \mu_1}{\sigma} \mid P_1\right] \\ &= Pr\left[Z \leq \frac{\gamma - \mu_1}{\sigma}\right] \\ &= \Phi\left[\frac{\frac{1}{2}(\mu_1 + \mu_2) - \mu_1}{\sigma}\right] \\ &= \Phi[-\Delta/2] \end{aligned} \tag{1.6}$$

Por el mismo argumento, $Pr[X \geq \gamma \mid P_2] = \Phi[-\Delta/2]$. Así, se puede definir $\Phi[-\Delta/2]$ como una medida de traslape entre los dos grupos. El traslape decrece monótonamente a medida que Δ va de cero a infinito.

Δ	1	2	3	4	5	6
$\Phi(-\Delta/2)$	0.308	0.159	0.067	0.023	0.006	0.001

La distancia estándar entre dos medias se menciona frecuentemente en la inferencia estadística. Se observará por medio de un ejemplo, que esta medida predomina en la prueba t de Student para una y dos muestras.



(a) $\Delta = 1$ (b) $\Delta = 2$

(c) $\Delta = 3$ (d) $\Delta = 4$

(e) $\Delta = 5$ (f) $\Delta = 6$

Figura 1.4: Curvas de densidad normal con distancias estandar desde $\Delta = 1$ hasta $\Delta = 6$

- Caso de una muestra: Suponga que se tiene una muestra $x_1, x_2, \ldots, x_N$ de una distribución con media μ y varianza σ^2, donde ambos parámetros son desconocidos. La t para la prueba de hipótesis $H_0 : \mu = \mu_0$, donde μ_0 es una constante fija, está basada en la estadística:

$$t = \sqrt{N}\frac{\bar{x} - \mu_o}{s} \tag{1.7}$$

Donde, $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ y $s = \left[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{1/2}$ denotan la media muestral y la desviación estándar, respectivamente. Se define la distancia estándar muestral entre dos números u_1 y u_2 con respecto a la muestra dada, como:

$$D(u_1, u_2) = |u_1 - u_2| / s \tag{1.8}$$

de donde se obtiene $D(\bar{x}, \mu_0) = |\bar{x} - \mu_0| / s$, y por lo tanto,

$$|t| = \sqrt{N} D(\bar{x}, \mu_0) \tag{1.9}$$

Dado que existe una relación tan estrecha, ¿Por qué utilizar la distancia estándar?, la respuesta es que la distancia estándar y la estadística t se utilizan para diferentes propósitos. La distancia estándar

$$D(\bar{x}, \mu_0) = \frac{|\mu_0 - \bar{x}|}{s} \tag{1.10}$$

es una medida descriptiva, hace referencia a qué tan lejos están la media de la muestra $\bar{x}$ y la media hipotética μ_0 en términos de la desviación estandar; también, puede ser considerada como una estimación de la distancia estándar teórica

$$\Delta(\mu_0, \mu) = \frac{|\mu_0 - \mu|}{\sigma} \tag{1.11}$$

La estadística t, por otro lado, es más adecuada para probar la hipótesis $H_0 : \mu = \mu_0$. Por sí misma no dice mucho acerca de qué tan distantes son μ_0 y $\bar{x}$,de hecho, un valor grande de $|t|$ puede resultar de una distancia estándar grande o de un tamaño de muestra

grande N .

- Caso de dos muestras: Considere muestras independientes $x_{11}, \ldots, x_{1N_1}$ y $x_{21}, \ldots, x_{2N_2}$ de distribuciónes con medias $\mu_j\,(j = 1, 2)$, y varianza común σ^2. Sean:

$$\bar{x_j} = \frac{1}{N_j} \sum_{i=1}^{N_j} x_{ij} \quad j = 1, 2$$

y

$$s_j = \left[\frac{1}{N_j - 1} \sum_{i=1}^{N_j} \left(x_{ij} - \bar{x_j} \right)^2 \right]^{\frac{1}{2}}, \quad j = 1, 2$$

las medias y desviaciones estándar muestrales usuales, respectivamente, y se denota por:

$$s^2 = \frac{\left[(N_1 - 1)\, s_1^2 + (N_2 - 1)\, s_2^2 \right]}{(N_1 + N_2 - 2)}$$

al estimador de la varianza común σ^2, también llamada la varianza combinada. La prueba t de Student para dos muestras para probar la hipótesis $H_0 : \mu_1 = \mu_2$ se basa en la estadística:

$$t = \frac{\bar{x_1} - \bar{x_2}}{s} \sqrt{\frac{N_1 N_2}{N_1 + N_2}} \tag{1.12}$$

Se define la versión muestral de $\Delta(\mu_1, \mu_2)$ como:

$$D(\bar{x_1}, \bar{x_2}) = \frac{\left| \bar{x_1} - \bar{x_2} \right|}{s} \tag{1.13}$$

que puede ser considerada una estimación de $\Delta(\mu_1, \mu_2)$. Entonces

$$|t| = \sqrt{\frac{N_1 N_2}{N_1 + N_2}} \cdot D(\bar{x_1}, \bar{x_2}) \tag{1.14}$$

La distancia estándar $D(\bar{x_1}, \bar{x_2})$ tiene el objetivo de ser una medida descriptiva de la separación entre los dos grupos.

Suponga que $\boldsymbol{X} = \left(X_1, ..., X_p\right)^{'}$ es un vector aleatorio de dimensión p con:

$$E[\boldsymbol{X}] = \mu$$

$$Var[\boldsymbol{X}] = \Sigma$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

Sea $\boldsymbol{\Sigma_{pxp}}$ una matriz de varianzas y covarianzas, se toman dos puntos fijos $\boldsymbol{x_1} \in \Re^p$ y $\boldsymbol{x_2} \in \Re^p$, ¿cómo se puede medir su distancia con respecto a $\boldsymbol{X}$?. La idea principal es reducir el problema de dimensión p a un problema univariado haciendo combinaciones lineales. Sea $\boldsymbol{a} = \left(a_1, ...a_p\right)^{'} \in \Re^p$ un vector de coeficientes de una combinación lineal y sea $Y = \boldsymbol{a'X}$, por el Teorema A.1.3 del Apéndice, se puede obtener $E[Y] = \boldsymbol{\mu_Y} = \boldsymbol{a}^{'}\boldsymbol{\mu}$ y $Var[Y] = \sigma_Y^2 = \boldsymbol{a'\Sigma a}$. Aplicando la misma transformación lineal a los puntos $\boldsymbol{x_1}$ y $\boldsymbol{x_2}$ se obtiene $y_1 = \boldsymbol{a}^{'}\boldsymbol{x_1}$ y $y_2 = \boldsymbol{a}^{'}\boldsymbol{x_2}$, que son números fijos en $\Re$, por lo tanto, es válida la idea unidimensional de la distancia estándar para la combinación lineal $Y = \boldsymbol{a'X}$ y obtener:

$$\begin{aligned} \Delta_Y(y_1, y_2) &= \Delta_{\boldsymbol{a'X}}\left(\boldsymbol{a}^{'}\boldsymbol{x_1}, \boldsymbol{a}^{'}\boldsymbol{x_2}\right) \\ &= \frac{|\boldsymbol{a'}(\boldsymbol{x_1} - \boldsymbol{x_2})|}{(\boldsymbol{a'\Sigma a})^{\frac{1}{2}}} \end{aligned} \tag{1.15}$$

En este punto, se tiene que asegurar que la expresión anterior esté bien definida, es decir, que el denominador no sea cero. Dado que $\boldsymbol{a'\Sigma a} = Var[\boldsymbol{a'X}]$ es la varianza de una variable aleatoria, se observa que $\boldsymbol{a'\Sigma a} \geq 0$, se tendrá la igualdad si $\boldsymbol{a} = \boldsymbol{0}$, así que se excluye la combinación lineal degenerada dada por el vector donde todos sus coeficientes son cero. Por lo tanto, se requiere que:

$$\boldsymbol{a'\Sigma a} > 0 \quad \text{para toda} \quad \boldsymbol{a} \in \Re^p \quad (\boldsymbol{a} \neq \boldsymbol{0}) \tag{1.16}$$

Esta condición es conocida frecuentemente como una matriz $\boldsymbol{\Sigma}$ de varianzas y covarianzas definida positiva. Esto significa que ninguna combinación lineal no trivial de $\boldsymbol{X}$ puede tener varianza igual a cero. La matriz definida positiva $\boldsymbol{\Sigma}$ es una generalización de la varianza positiva en el caso univariado.

Las combinaciones lineales "interesantes" son aquellas que generan una distancia estándar grande, esto conduce a la definición de distancia estándar multivariada.

Definición 1.2.2. *Sea $\boldsymbol{X}$ un vector aleatorio de dimensión p con media $\boldsymbol{\mu} = E[\boldsymbol{X}]$ y matriz de varianzas y covarianzas $Var[\boldsymbol{X}] = \boldsymbol{\Sigma}$, con el supuesto de que es definida positiva. La distancia estándar de dimensión p entre dos vectores $\boldsymbol{x}_1 \in \Re^p$ y $\boldsymbol{x}_2 \in \Re^p$, con respecto a $\boldsymbol{X}$, es la distancia estándar máxima univariada entre $\boldsymbol{a}'\boldsymbol{x}_1$ y $\boldsymbol{a}'\boldsymbol{x}_2$ con respecto a la variable aleatoria $\boldsymbol{a}'\boldsymbol{X}$, se toma el máximo sobre todas las combinaciones lineales, esto es*

$$\begin{aligned} \Delta_X(\boldsymbol{x}_1, \boldsymbol{x}_2) &= \max_{\substack{\boldsymbol{a}\in\Re^p \\ \boldsymbol{a}\neq 0}} \Delta_{\boldsymbol{a}'\boldsymbol{X}}\left(\boldsymbol{a}'\boldsymbol{x}_1, \boldsymbol{a}'\boldsymbol{x}_2\right) \\ &= \max_{\substack{\boldsymbol{a}\in\Re^p \\ \boldsymbol{a}\neq 0}} \frac{|\boldsymbol{a}'(\boldsymbol{x}_1 - \boldsymbol{x}_2)|}{(\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})^{\frac{1}{2}}} \end{aligned} \tag{1.17}$$

La definición muestra que se deben considerar todas las posibles combinaciones lineales $\boldsymbol{a}'\boldsymbol{X}$ de $\boldsymbol{X}$, excepto la trivial dada por $\boldsymbol{a} = \boldsymbol{0}$, y encontrar la o las combinaciones para las cuales $\boldsymbol{a}'\boldsymbol{x}_1$ y $\boldsymbol{a}'\boldsymbol{x}_2$ estén tan distantes entre sí como sea posible 'tan distantes como sea posible' en términos de la desviación estándar de $\boldsymbol{a}'\boldsymbol{X}$. La máxima distancia estándar univariada, suponiendo que exista (un punto aún por demostrar), se pude llamar distancia estándar multivariada entre $\boldsymbol{x}_1$ y $\boldsymbol{x}_2$ con respecto al vector aleatorio $\boldsymbol{X}$.

Antes de continuar con el desarrollo de la distancia estándar multivariada, se mostrarán algunas propiedades.

Definición 1.2.3. *Dos combinaciones lineales $\boldsymbol{a}'\boldsymbol{X}$ y $\boldsymbol{b}'\boldsymbol{X}$ de un vector aleatorio $\boldsymbol{X}$ se dicen equivalentes si son proporcionales entre sí, es decir, si $\boldsymbol{b} = \boldsymbol{a} \cdot c$ para alguna $c \in \Re$ distinta de cero.*

La importancia del concepto de combinaciones lineales equivalentes se desprende del siguiente resultado:

Lema 1.2.1. *Las combinaciones lineales equivalentes llevan al mismo valor de la distancia estándar (univariada) entre dos puntos.*

Demostración: Sea $Y = \boldsymbol{a}'\boldsymbol{X}$ una combinación lineal de $\boldsymbol{X}$, y sea Y^* una combinación lineal equivalente, es decir, $Y^* = c\boldsymbol{a}'\boldsymbol{X}$ para cualquier $c \neq 0$. Para dos puntos $\boldsymbol{x}_1$ y $\boldsymbol{x}_2 \in \Re^p$,

$$\begin{aligned}\Delta_{Y^*}\left(c\boldsymbol{a}'\boldsymbol{x}_1, c\boldsymbol{a}'\boldsymbol{x}_2\right) &= \frac{|c\boldsymbol{a}'(\boldsymbol{x}_1 - \boldsymbol{x}_2)|}{(c^2\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})^{1/2}} \\ &= \frac{|\boldsymbol{a}'(\boldsymbol{x}_1 - \boldsymbol{x}_2)|}{(\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})^{1/2}} \\ &= \Delta_Y\left(\boldsymbol{a}'\boldsymbol{x}_1, \boldsymbol{a}'\boldsymbol{x}_2\right)\end{aligned}$$

El Lema 1.2.1 dice que, en el desarrollo para obtener la máxima distancia estándar, puede haber una restricción para un solo miembro de cada clase de combinaciones lineales equivalentes. Por ejemplo, se puede requerir que todas las combinaciones lineales consideradas estén normalizadas, es decir, establecer la restricción $\boldsymbol{a}'\boldsymbol{a} = 1$ en el vector de coeficientes. Para cualquier combinación lineal dada $Y = \boldsymbol{a}'\boldsymbol{X}$, la combinación lineal normalizada equivalente está dada por:

$$Y^* = \frac{1}{(\boldsymbol{a}'\boldsymbol{a})^{1/2}} Y = \frac{1}{(\boldsymbol{a}'\boldsymbol{a})^{1/2}} \boldsymbol{a}'\boldsymbol{X}$$

Con el propósito de desarrollar alguna intuición geométrica, la idea de buscar sólo en combinaciones lineales normalizadas es útil.

Por costumbre, el caso $p = 2$ es el más sencillo de tratar. Sea $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ un vector aleatorio bivariado y $Y = \boldsymbol{a}'\boldsymbol{X} = a_1X_1 + a_2X_2$ una combinación lineal de $\boldsymbol{X}$, una combinación lineal normalizada equivalente es $Y^* = \boldsymbol{b}'\boldsymbol{X}$, donde $\boldsymbol{b} = \boldsymbol{a}/(\boldsymbol{a}'\boldsymbol{a})^{1/2}$, es decir, $b_1 = a_1/\sqrt{a_1^2 + a_2^2}$ y $b_2 = a_2/\sqrt{a_1^2 + a_2^2}$. Cuando $\boldsymbol{b}'\boldsymbol{b} = b_1^2 + b_2^2 = 1$, se puede escribir $b_1 = \cos\phi$ y $b_2 = \sin\phi$ para un cierto ángulo $\phi \in [0, 2\pi]$. La combinación lineal normalizada Y^* está dada por:

$$Y^* = (\cos\phi)\, X_1 + (\sin\phi)\, X_2$$

Por lo tanto, observando todas las combinaciones lineales normalizadas de las cantidades de $\boldsymbol{X}$ para hacer variaciones en el ángulo ϕ de 0 a 2π (en concreto, de 0 a π es suficiente) y determinar la distribución de Y^* para cada ϕ.

Regresando al desarrollo de la distancia estándar, para encontrar una expresión explícita para la distancia estándar multivariada, se necesitará usar un resultado conocido como la desigualdad de Cauchy-Schwartz extendida, que es una generalización de la desigualdad de Cauchy Schwartz ordinaria, dada en el siguiente lema.

Lema 1.2.2. *(Desigualdad de Cauchy-Schwartz) Para cualesquiera dos vectores distintos de cero* $\boldsymbol{x} \in \Re^p$ *y* $\boldsymbol{y} \in \Re^p$,

$$\left(x'y\right)^2 \leq \left(x'x\right)\left(y'y\right) \tag{1.18}$$

donde la igualdad se da si $\boldsymbol{y} = c\boldsymbol{x}$ para cualquier $c \in \Re$

Lema 1.2.3. *(Desigualdad de Chauchy Schwartz extendida) Para cualesquiera dos vectores distintos de cero* $\boldsymbol{u} \in \Re^p$ *y* $\boldsymbol{v} \in \Re^p$ *y cualquier matriz* $\boldsymbol{M_{pxp}}$ *simétrica definida positiva,*

$$\left(u'v\right)^2 \leq \left(u'Mu\right)\left(v'M^{-1}v\right) \tag{1.19}$$

con la igualdad exacta si $\boldsymbol{v} = c\boldsymbol{M}\boldsymbol{u}$ para algún $c \in \Re$ (o, equivalentemente, $\boldsymbol{u} = c^*\boldsymbol{M}^{-1}\boldsymbol{v}$ para cualquier $c^* \in \Re$).

Demostración: Dado que $\boldsymbol{M}$ es definida positiva y simétrica, existe una matriz simétrica, no singular $M^{1/2}$ tal que $\left(M^{1/2}\right)^2 = M$, con inversa $M^{-1/2} = \left(M^{1/2}\right)^{-1}$. Sea $\boldsymbol{x} = M^{1/2}\boldsymbol{u}$ y $\boldsymbol{y} = M^{-1/2}\boldsymbol{v}$, entonces del Lema 1.2.2 se tiene que:

$$\begin{aligned}\left(u'v\right)^2 &= \left(x'M^{-1/2}M^{1/2}y\right)^2 \\ &= \left(x'y\right)^2 \\ &\leq \left(x'x\right)\left(y'y\right) \\ &= \left(u'Mu\right)\left(v'M^{-1}v\right)\end{aligned} \tag{1.20}$$

La igualdad en (1.20) se da si $\boldsymbol{y} = c\boldsymbol{x}$ para cualquier $c \in \Re$, es decir, si $\boldsymbol{M}^{-1/2}\boldsymbol{v} = c\boldsymbol{M}^{1/2}\boldsymbol{u}$ o $\boldsymbol{v} = c\boldsymbol{M}\boldsymbol{u}$ para $c \in \Re$.

Como una consecuencia del Lema 1.2.3 se tiene que para un vector $\boldsymbol{v} \in \Re^p$ y una matriz $\boldsymbol{M}$ definida positiva y simétrica,

$$\underset{u \in \Re^p}{max} \frac{(\boldsymbol{u}'\boldsymbol{v})^2}{\boldsymbol{u}'\boldsymbol{M}\boldsymbol{u}} = \boldsymbol{v}'\boldsymbol{M}^{-1}\boldsymbol{v} \tag{1.21}$$

se alcanza el máximo para cualquier vector $\boldsymbol{u}$ proporcional a $\boldsymbol{M}^{-1}\boldsymbol{v}$.

Ahora se tiene el resultado principal de este apartado.

Teorema 1.2.1. *Sea $\boldsymbol{X}$ un vector aleatorio de dimensión p con una matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$ definida positiva. La distancia estándar multivariada entre dos puntos $\boldsymbol{x}_1 \in \Re^p$ y $\boldsymbol{x}_2 \in \Re^p$ está dada por:*

$$\Delta_X(\boldsymbol{x}_1, \boldsymbol{x}_2) = \left[(\boldsymbol{x}_1 - \boldsymbol{x}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_1 - \boldsymbol{x}_2)\right]^{1/2} \tag{1.22}$$

Demostración: En lugar de maximizar $\Delta_{a'X}(\boldsymbol{a}'\boldsymbol{x}_1, \boldsymbol{a}'\boldsymbol{x}_2)$, se puede maximizar equivalentemente su cuadrado $\Delta^2_{a'X}(\boldsymbol{a}'\boldsymbol{x}_1, \boldsymbol{a}'\boldsymbol{x}_2) = [\boldsymbol{a}'(\boldsymbol{x}_1 - \boldsymbol{x}_2)]^2 / (\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})$ con $\boldsymbol{a} \in \Re^p, \boldsymbol{a} \neq \boldsymbol{0}$.

Usando la ecuación (1.21) con $\boldsymbol{u} = \boldsymbol{a}$, $\boldsymbol{v} = \boldsymbol{x}_1 - \boldsymbol{x}_2$, y $\boldsymbol{M} = \boldsymbol{\Sigma}$, alcanza su valor máximo si se elige una $\boldsymbol{a}$ proporcional a $\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_1 - \boldsymbol{x}_2)$, entonces:

$$\underset{\substack{a \in \Re^p \\ a \neq 0}}{\boldsymbol{max}} \frac{|\boldsymbol{a}'(\boldsymbol{x}_1 - \boldsymbol{x}_2)|^2}{(\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})^{1/2}} = (\boldsymbol{x}_1 - \boldsymbol{x}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_1 - \boldsymbol{x}_2) \tag{1.23}$$

El resultado se sigue tomando la raíz cuadrada en (1.23).

Dada la importancia del Teorema 1.2.1, se harán algunas observaciones.

- Si el vector aleatorio X se compone de variables aleatorias por parejas, todas con la misma varianza $\sigma^2 > 0$, entonces la distancia estándar entre dos puntos $\boldsymbol{x}_1$ y $\boldsymbol{x}_2$ está dada por:

$$\begin{aligned} \Delta_X(\boldsymbol{x}_1, \boldsymbol{x}_2) &= \left[(\boldsymbol{x}_1 - \boldsymbol{x}_2)' \left(\sigma^2 \boldsymbol{I}_p\right)^{-1} (\boldsymbol{x}_1 - \boldsymbol{x}_2)\right]^{1/2} \\ &= \frac{1}{\sigma}\sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)' (\boldsymbol{x}_1 - \boldsymbol{x}_2)} \end{aligned} \tag{1.24}$$

Con esto, la distancia estándar es igual a la distancia Euclidiana en términos de la desviación estándar común.

- En muchos libros de estadística multivariada se da un nombre que difiere a la distancia estándar, por ejemplo, distancia estadística, distancia elíptica y la distancia de Mahalanobis son términos comúnmente utilizados. Por error, Δ^2 con frecuencia es denominada distancia de Mahalanobis, aunque no satisface los axiomas usuales de una medida de distancia.

- El término "distancia elíptica" se refiere a que, para una matriz $\boldsymbol{A}$ dada definida positiva y simétrica y un vector $\boldsymbol{m} \in \Re^p$ el conjunto de puntos

$$\varepsilon_c = \{\boldsymbol{x} \in \Re^p : (\boldsymbol{x} - \boldsymbol{m})' \boldsymbol{A}^{-1} (\boldsymbol{x} - \boldsymbol{m}) = \boldsymbol{c}^2\} \tag{1.25}$$

es una elipse ($p = 2$), elipsoide ($p = 3$) ó hiperelipsoide ($p > 3$) para cualquier $c > 0$. La distribución normal multivariada y las distribuciones multivariadas elípticas, en general, tienen la propiedad de que la densidad es constante en los conjuntos ε_c, con $\boldsymbol{m}$ y $\boldsymbol{A}$ como el vector de medias y la matriz de varianzas y covarianzas múltiple respectivamente.

A continuación se abordará el caso de dos grupos, que llevará directamente a un análisis discriminante lineal.

Suponga que un vector aleatorio X de dimensión p es medido en dos poblaciones, y

$$\begin{aligned} E_1[\boldsymbol{X}] = \mu_1, \quad Cov[\boldsymbol{X}] = \Sigma \quad \text{para la población 1} \\ E_2[\boldsymbol{X}] = \mu_2, \quad Cov[\boldsymbol{X}] = \Sigma \quad \text{para la población 2} \end{aligned} \tag{1.26}$$

donde $\boldsymbol{\Sigma}$ es definida positiva.

Definición 1.2.4. *La distancia estándar multivariada entre $\boldsymbol{\mu_1}$ y $\boldsymbol{\mu_2}$ para un vector aleatorio X como en (1.26) está dada por:*

$$\begin{aligned}\Delta_X(\mu_1,\mu_2) &= \max_{\substack{a\in\Re^p\\ a\neq 0}} \Delta_{a'X}\left(a'\mu_1, a'\mu_2\right)\\ &= \max_{\substack{a\in\Re^p\\ a\neq 0}} \frac{|a'(\mu_1-\mu_2)|}{(a'\Sigma a)^{1/2}}\end{aligned} \tag{1.27}$$

Ahora se puede definir intuitivamente una función discriminante.

Definición 1.2.5. *Sea $\boldsymbol{X}$ un vector aleatorio de dimensión p con $E_1[\boldsymbol{X}] = \mu_1$ en la población 1, $E_2[\boldsymbol{X}] = \mu_2$ en la población 2, y $Var[\boldsymbol{X}] = \boldsymbol{\Sigma}$ en ambas poblaciones, con $\boldsymbol{\Sigma}$ definida positiva. Sea $Y = \beta'\boldsymbol{X}$ una combinación lineal de $\boldsymbol{X}$. Si*

$$\Delta_Y\left(\beta'\mu_1, \beta'\mu_2\right) = \Delta_X(\mu_1,\mu_2)$$

Entonces $Y = \beta'\boldsymbol{X}$ es llamada una función discriminante lineal para las dos poblaciones.

Las definiciones 1.2.4 y 1.2.5 son importantes para comprender el análisis discriminante lineal. La distancia estándar multivariada entre dos poblaciones con la misma matriz de varianzas y covarianzas es la máxima distancia estándar univariada sobre todas las combinaciones lineales. Cualquier combinación lineal que alcanza el máximo es una función discriminante lineal para las dos poblaciones.

El siguiente Teorema es una consecuencia directa del Teorema 1.2.1 y su demostración.

Teorema 1.2.2. *La distancia estándar multivariada entre dos poblaciones con vectores de medias $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$ y matriz de varianzas y covarianzas común $\boldsymbol{\Sigma}$ está dada por:*

$$\Delta(\mu_1-\mu_2) = \left[(\mu_1-\mu_2)'\,\Sigma^{-1}(\mu_1-\mu_2)\right]^{1/2} \tag{1.28}$$

Cualquier combinación $Y = \beta'\boldsymbol{X}$, con

$$\beta = c\cdot\Sigma^{-1}(\mu_1-\mu_2)$$

Donde $c \neq 0$ es una constante arbitraria y una función discriminante lineal para las dos poblaciones.

Ejemplo 1.2.3. Suponga que $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ es un vector aleatorio bivariado con

$$E_1[\boldsymbol{X}] = \mu_1 = \begin{pmatrix} 2 \\ 5 \end{pmatrix} \quad \text{para la población 1}$$

$$E_2[\boldsymbol{X}] = \mu_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{para la población 2}$$

y

$$Var[\boldsymbol{X}] = \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix} \quad \text{para ambas poblaciones}$$

Las distancias estándar univariadas son $|2-0|/1 = 2$ para la variable X_1 y $|5-0|/\sqrt{5} = \sqrt{5} \approx 2.24$ para la variable X_2. La distancia estándar bivariada, por la ecuación (1.28), está dada por:

$$\Delta(\mu_1, \mu_2) = \left[(2,5) \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{5} \end{pmatrix} \begin{pmatrix} 2 \\ 5 \end{pmatrix} \right]^{1/2} = 3$$

y el vector de coeficientes de la función discriminante está dado por

$$\beta = c \cdot \Sigma^{-1}(\mu_1 - \mu_2) = c \cdot \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{5} \end{pmatrix} \begin{pmatrix} 2 \\ 5 \end{pmatrix} = c \cdot \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

Donde $c \neq 0$ es una constante arbitraria. Si se toma $c = 1$, entonces la función discriminante está dada por $Y = 2X_1 + X_2$.

Por ahora se hará referencia a la función discriminante lineal como si estuviera definida de forma única, se va a establecer $c = 1$, a menos que se indique de otra forma. Esto es, el vector de coeficientes será:

$$\beta = \Sigma^{-1}(\mu_1 - \mu_2) \tag{1.29}$$

Esta elección de la constante c típicamente produce una función discriminante no normalizada,

pero ésta tiene la siguiente propiedad:

$$Var[Y] = Var\left[\beta' \boldsymbol{X}\right] = \beta' \Sigma \beta = (\mu_1 - \mu_2)\, \Sigma^{-1} (\mu_1 - \mu_2) = \Delta_X^2 (\mu_1, \mu_2) \tag{1.30}$$

Adicionalmente, para $c = 1$,

$$\Delta_X^2 (\mu_1, \mu_2) = \beta' (\mu_1 - \mu_2)\,. \tag{1.31}$$

Pero $\beta' (\mu_1 - \mu_2)$ es la diferencia de medias entre los dos grupos para la función discriminante lineal. Así, para $c = 1$, la varianza y la diferencia de medias de la función discriminante son idénticas al cuadrado de la distancia estándar multivariada, nóte que esto es cierto para $c = 1$, pero no para una elección arbitraria de c.

Dependiendo de la convención particular utilizada, en el conocimiento del investigador, o en los estándares utilizados en los programas para el cálculo numérico de una función discriminante, se puede obtener una solución diferente β^* para los coeficientes de la función discriminante lineal. Utilizando una idea que puede ser negativa en el desarrollo, a veces una constante de intercepción se añade, lo que lleva a la función discriminante siguiente

$$Y^* = \beta_0 + c \cdot Y = \beta_0 + c \cdot \boldsymbol{\beta' X} \tag{1.32}$$

Para el propósito de discriminación , cualquier Y^* de este tipo resulta cómoda como $Y = (\mu_1 - \mu_2)' \Sigma^{-1} X$, porque lleva a la misma distancia estándar, siempre que $c \neq 0$. Aunque esto suele ser confuso (si no se tienen los conocimientos), se tiene la libertad que proporciona la elección de los dos parámetros β_0 y c para transformar la función discriminante a una forma conveniente para fines de clasificación. Por ejemplo, si se escribe $\Delta = \Delta_X (\mu_1, \mu_2)$ y se toma

$$c = \frac{1}{\Delta}$$

y

$$\beta_0 = -\frac{1}{2\Delta} (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 + \mu_2) \tag{1.33}$$

entonces, la función discriminante transformada es

$$Y^* = \beta_0 + c \cdot \boldsymbol{\beta}' \boldsymbol{X}$$

$$= -\frac{1}{2\Delta} \cdot (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 + \mu_2) + \frac{1}{\Delta} (\mu_1 - \mu_2)' \Sigma^{-1} X \tag{1.34}$$

La variable Y^* tiene la siguiente propiedad

$$var[Y^*] = \frac{1}{\Delta^2} (\mu_1 - \mu_2)' \Sigma^{-1} \Sigma \Sigma^{-1} (\mu_1 - \mu_2) = 1 \tag{1.35}$$

en ambas poblaciones. Adicionalmente, si $\gamma_1 = E[Y^* \mid P_1]$ y $\gamma_2 = E[Y^* \mid P_2]$ denotan las medias de Y^* en las dos poblaciones, se obtiene

$$\begin{aligned} \gamma &= -\frac{1}{2\Delta} (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 + \mu_2) + \frac{1}{\Delta} (\mu_1 - \mu_2) \Sigma^{-1} \mu_1 \\ &= \frac{1}{\Delta} (\mu_1 - \mu_2)' \Sigma^{-1} \left[-\frac{1}{2} (\mu_1 + \mu_2) + \mu_1 \right] \\ &= \frac{1}{\Delta} (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) /2 \\ &= \frac{1}{2} \Delta \end{aligned} \tag{1.36}$$

y de forma similar,

$$\gamma_2 = -\frac{1}{2} \Delta \tag{1.37}$$

Por lo tanto, el punto medio entre las medias de Y^* es cero, y la varianza de Y^* es uno. La distancia estándar entonces es idéntica a la distancia Euclidiana en la escala Y^*.

Ejemplo 1.2.4. Suponga que el vector aleatorio $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ tiene distribución normal con vector de medias $\mu_1 = \begin{pmatrix} 5 \\ 1 \end{pmatrix}$ en la población 1, $\mu_2 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ en la población 2, y matriz de varianzas y covarianzas $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ en ambas poblaciones. Las distancias estándar univariadas son

$|5-2|/\sqrt{2} \approx 2.121$ para X_1 y 0 para X_2, los coeficientes de la función discriminante lineal son

$$\beta = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 3 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ 0 \end{pmatrix} = \begin{pmatrix} 3 \\ -3 \end{pmatrix}$$

por lo tanto, la función discriminante lineal es $Y = 3X_1 - 3X_2$ o cualquier combinación lineal equivalente a $X_1 - X_2$. La distancia estándar bivariada se calcula como:

$$\Delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) = \beta' (\mu_1 - \mu_2) = (3, -3) \begin{pmatrix} 3 \\ 0 \end{pmatrix} = 9$$

y, $\Delta = 3$. Esta es más grande que la distancia estándar univariada de X_1. Tenga en cuenta que la función discriminante lineal depende de X_2, si bien es cierto que la diferencia de medias en X_2 es cero.

La Figura 1.5 muestra una grafica de contorno para este ejemplo, con la distribución de la función discriminante normalizada,

$$Y' = \frac{1}{\sqrt{2}} (X_1 - X_2) = \frac{1}{3\sqrt{2}} Y$$

En la línea correspondiente a Y', las dos medias son $\Delta = 3$ desviaciones estándar de separación. Por último, se observa la transformación que se sugiere en (1.34). Nóte que:

$$E[Y] = E[3X_1 - 3X_2] = (3, -3)\, \mu_i$$

$$= \begin{cases} 12 \quad en \quad el \quad grupo \quad 1 \\ 3 \quad en \quad el \quad grupo \quad 2 \end{cases}$$

y $var[Y] = (3, -3) \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 3 \\ -3 \end{pmatrix} = 9$ en ambos grupos. Se elige la transformación

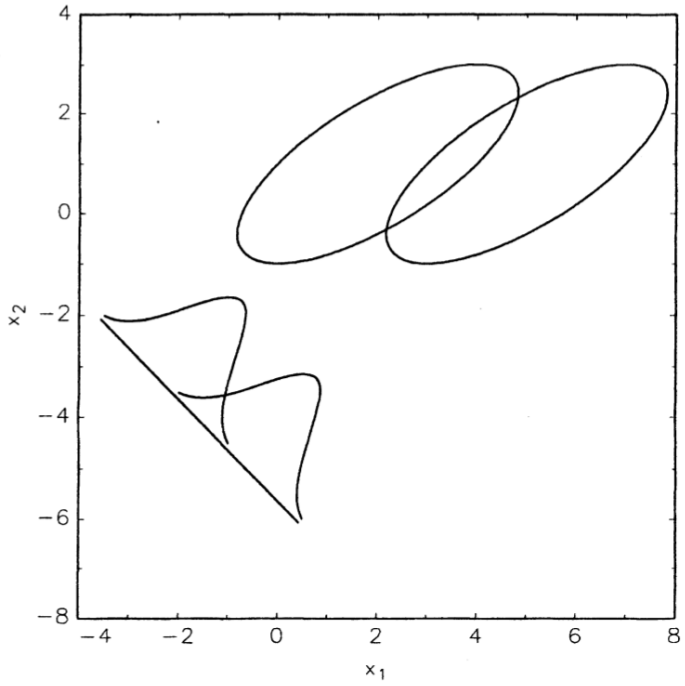


Figura 1.5: La distribución de la función discriminante lineal en el ejemplo 1.2.4

$Y^* = \beta_0 + c \cdot Y$, con $c = 1/3$ y $\beta_0 = -2.5$, entonces:

$$Y^* = -2.5 + \frac{1}{3}Y = X_1 - X_2 - 2.5$$

y $E_1[Y^*] = 1.5$ en el grupo 1, $E_2[Y^*] = -1.5$ en el grupo 2, y $var[Y^*] = 1$ en ambos grupos. Por lo tanto, en una gráfica de las densidades de Y^* en ambos grupos, se intersectan exactamente en 0, y las medias difieren por $\Delta = 3$. Los resultados numéricos no dependen de ninguna manera de la normalidad de X.

1.3. Utilizando la función discriminante lineal

En esta sección, se aplicará la teoría desarrollada hasta ahora a ejemplos prácticos, mediante la sustitución de las estadísticas de la muestra para los parámetros del modelo. Se trata de un enfoque heurístico y principalmente descriptivo, en el que no se consideran los problemas de inferencia estadística. No hay supuestos de distribución.

Sean $\boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \ldots, \boldsymbol{x}_{1N_1}$ los vectores de datos observados del grupo 1 y $\boldsymbol{x}_{21}, \boldsymbol{x}_{22}, \ldots, \boldsymbol{x}_{2N_2}$ los vectores de datos observados del grupo 2. Estas N_1+N_2 observaciones constituyen las muestras de entrenamiento. Sean

$$\bar{x}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \boldsymbol{x}_{ji} \quad j = 1, 2 \tag{1.38}$$

los vectores de medias muestrales, y

$$\boldsymbol{S}_j = \frac{1}{N_j - 1} \sum_{i=1}^{N_j} \left(\boldsymbol{x}_{ji} - \bar{\boldsymbol{x}}_j\right) \left(\boldsymbol{x}_{ji} - \bar{\boldsymbol{x}}_j\right)' \quad j = 1, 2 \tag{1.39}$$

las matrices de varianzas y covarianzas muestrales usuales, con denominadores $N_j - 1$. Cuando se usan los denominadores $N_j - 1$ en vez de N_j, en realidad se violan los supuestos iniciales, pero se sigue con la convención aceptada por la mayoría de los estadísticos.

La única dificultad en la aplicación de las nociones de la distancia estándar multivariada y la función discriminante lineal a los datos observados es que se tienen dos matrices de varianzas y covarianzas $\boldsymbol{S}_1$ y $\boldsymbol{S}_2$, en lugar de una única matriz de varianzas y covarianzas común, esta dificultad se soluciona mediante el uso de un promedio ponderado de $\boldsymbol{S}_1$ y $\boldsymbol{S}_2$, la matriz de varianzas y covarianzas combinada muestral

$$\boldsymbol{S} = \frac{1}{N_1 + N_2 - 2} \left[\left(N_1 - 1\right) \boldsymbol{S}_1 + \left(N_2 - 1\right) \boldsymbol{S}_2\right] \tag{1.40}$$

Únicamente con los supuestos iniciales de estimación, sería difícil justificar el uso de (1.40), pero la matriz de varianzas y covarianzas combinada muestral es un estimador insesgado de la matriz de varianzas y covarianzas común de dos poblaciones , independientemente de la distribución exacta, esto puede ser útil como justificación.

De forma análoga a las definiciones de la sección anterior, se define la distancia estándar multivariada muestral y la función discriminante lineal muestral.

Definición 1.3.1. *Con el planteamiento de las ecuaciones (1.38) a (1.40), la distancia estándar multivariada entre $\bar{\boldsymbol{x}}_1$ y $\bar{\boldsymbol{x}}_2$ está dada por:*

$$D(\bar{x}_1, \bar{x}_2) = \max_{\substack{a \in \Re^p \\ a \neq 0}} \frac{|a'(\bar{x}_1 - \bar{x}_2)|}{(a'Sa)^{1/2}} \tag{1.41}$$

Es la máxima distancia estándar multivariada sobre todas las combinaciones lineales, siempre que el máximo exista. Cualquier combinación lineal para la cual se alcanza el máximo se llama una función discriminante lineal para las muestras dadas.

Este problema de maximización es matemáticamente idéntico al que se observó en la Definición 1.2.4 y el Teorema 1.2.5, por lo tanto, se obtiene inmediatamente el siguiente Teorema:

Teorema 1.3.1. *Con el planteamiento de las ecuaciones (1.38) a (1.40), suponiendo que la matriz de varianzas y covarianzas combinada muestral* $\boldsymbol{S}$ *es no singular, la distancia estándar multivariada entre* $\bar{x}_1$ *y* $\bar{x}_2$ *está dada por:*

$$D(\bar{x}_1, \bar{x}_2) = \left[(\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2)\right]^{1/2} \tag{1.42}$$

y el vector de coeficientes de la función discriminante lineal es:

$$b = S^{-1}(\bar{x}_1 - \bar{x}_2) \tag{1.43}$$

o cualquier vector proporcional a (1.43)

Se observa que si $N_1 + N_2 < p + 2$, entonces $\boldsymbol{S}$ siempre es sigular. Estas restricciones del uso del análisis discriminante lineal para muestras suficientemente grandes se usan para que la matriz de varianzas y covarianzas combinada sea definida positiva. Note, además que, en la ecuación (1.41), el término $\boldsymbol{a'Sa}$ es la varianza muestral combinada de la combinacion lineal $Y = \boldsymbol{a'X}$. Se puede decir que la distancia estándar muestral tiene las mismas propiedades que su análogo teórico; éstas satisfacen los axiomas de una medida de distancia. Como en la sección anterior, se mencionará la ecuación (1.43) como la definición de vector de coeficientes de una función discriminante lineal, es decir, se elegirá la constante de proporcionalidad $c = 1$. En el desarrollo descriptivo actual, no se va a hacer ninguna suposición sobre los datos más allá que $\boldsymbol{S}$ es definida positiva, en particular, no hay supuestos de que los datos constituyen muestras aleatorias de una familia particular de distribuciones o que las matrices de varianzas

y covarianzas en las poblaciones son idénticas, sin embargo, se discutirá el uso de la matriz de varianzas y covarianzas combinada de la muestra para tener claro bajo qué circunstancias esto es razonable.

Ejemplo 1.3.1. Las matrices de varianzas y covarianzas (con X_1 =longitud de antena, X_2 =longitud de ala) están dadas por

$$\boldsymbol{S}_1 = \begin{pmatrix} 98.00 & 80.83 \\ 80.83 & 168.78 \end{pmatrix} \text{ Para AF}$$

y

$$\boldsymbol{S}_2 = \begin{pmatrix} 39.47 & 43.47 \\ 43.47 & 77.87 \end{pmatrix} \text{ Para APF}$$

Con tamaños de muestra $N_1 = 9$, $N_2 = 6$, la matriz de varianzas y covarianzas combinada es

$$\boldsymbol{S} = \frac{1}{13}[8\boldsymbol{S}_1 + 5\boldsymbol{S}_2] = \begin{pmatrix} 75.49 & 66.46 \\ 66.46 & 133.81 \end{pmatrix}$$

y su inversa es

$$\boldsymbol{S}^{-1} = \begin{pmatrix} 23.54 & -11.69 \\ -11.69 & 13.28 \end{pmatrix} \cdot 10^{-3}$$

El vector de diferencia de medias está dado por

$$\boldsymbol{d} = \bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2 = \begin{pmatrix} 18.67 \\ -12.22 \end{pmatrix}$$

Finalmente, se obtienen los coeficientes de la función discriminante lineal

$$\boldsymbol{b} = \boldsymbol{S}^{-1}\boldsymbol{d} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 0.582 \\ -0.381 \end{pmatrix}$$

y la distancia estándar bivariada

$$D = \left[\boldsymbol{d}'\boldsymbol{S}^{-1}\boldsymbol{d}\right]^{1/2} = \left[\boldsymbol{d}'\boldsymbol{b}\right]^{1/2} = 3.94$$

En la combinación lineal óptima de longuitud de antena y de longitud de ala, las dos medias son $D = 3.94$ desviaciones estándar de separación. Ahora, se puede calcular el valor de

$$V = b_1 X_1 + b_2 X_2 = 0.582 X_1 - 0.381 X_2$$

para cada una de las 15 observaciones. Se pretende hacer la clasificación, de nuevo, usando un enfoque heurístico.

Se obtienen las medias muestrales de la función discriminante en ambos grupos por los datos en la tabla 1.2 o por los cálculos directos: $\bar{v}_1 = 13.633$ para el grupo 1 (AF), y $\bar{v}_2 = -1.890$ para el grupo 2 (APF). El punto medio entre $\bar{v}_1$ y $\bar{v}_2$ es $m = (\bar{v}_1 + \bar{v}_2)/2 = 5.872$, por lo tanto, una regla simple de clasificación puede ser:

$$Asignar \quad a \quad \begin{cases} AF \quad si \quad V > 5.87 \\ APF \quad si \quad V < 5.87 \end{cases}$$

Para clasificar a un mosquito como perteneciente a un grupo desconocido pero con valores conocidos x_1 y x_2, se calcula $v = 0.582 x_1 - 0.381 x_2$ y se verifica si $v < m$ o $v > m$.

Visto en el espacio de las variables originales, el punto de corte para la clasificación corresponde a una línea recta. Para ver esto, se observa que $v = m$ corresponde a la ecuación

$$b_1 x_1 + b_2 x_2 = m$$

o (suponiendo $b_2 \neq 0$)

$$x_2 = \frac{m}{b_2} - \frac{b_1}{b_2} x_1.$$

Finalmente, se muestra una versión equivalente de la función discriminante, sugerida por la transformación (1.34) en la seccion anterior. La varianza combinada de V es $D^2 = 15.52$ y el

Especie	Longitud de Antena(X_1)	Longitud de Ala (X_2)	V	V*
AF	138	164	17.95	3.07
AF	140	170	16.83	2.78
AF	124	172	6.75	0.22
AF	136	174	12.98	1.80
AF	138	182	11.10	1.33
AF	148	182	16.92	2.81
AF	154	182	20.42	3.69
AF	138	190	8.05	0.55
AF	156	208	11.69	1.48
APF	114	178	-1.36	-1.83
APF	120	186	-0.91	-1.72
APF	118	196	-5.88	-2.98
APF	130	196	1.11	-1.21
APF	126	200	-2.74	-2.19
APF	128	200	-1.58	1.89

Tabla 1.2: Datos de mosquitos en mm/100 y las dos versiones de la función discriminante

punto medio entre $\bar{v}_1$ y $\bar{v}_2$ es $m = 5.87$, es preferible usar

$$V^* = \frac{1}{D}(V - m) = (V - 5.87)/3.94$$

$$= 1.49 + 0.1478X_1 - 0.0966X_2$$

Hay situaciones en las que incluso las diferencias considerables entre las matrices de varianzas y covarianzas parecen irrelevantes en el siguiente sentido. Suponga que el vector aleatorio $\boldsymbol{X}$ tiene vectores de medias μ_1 y μ_2 en dos poblaciones y matrices de varianzas y covarianzas $\Sigma_1 \neq \Sigma_2$. Dado que las matrices de varianzas y covarianzas no son iguales, se utiliza una combinación convexa de ellas, es decir $\Sigma_\alpha = \alpha\Sigma_1 + (1-\alpha)\Sigma_2$, donde $\alpha \in [0,1]$, y se define el vector de coeficientes de la función discriminante lineal como $\beta = \Sigma_\alpha^{-1}(\mu_1 - \mu_2)$, que depende de la elección particular de α. Note que la matriz de varianzas y covarianzas combinada en la situación muestral es análoga a esta. Se podría argumentar que se debe elegir Σ_1 o Σ_2 en la ecuación de los coeficientes de la función discriminante lineal, es decir, se necesitan calcular dos funciones discriminantes lineales $Y_1 = \beta_1'\boldsymbol{X}$ y $Y_2 = \beta_2'\boldsymbol{X}$ donde $\beta_1 = \Sigma_1^{-1}(\mu_1 - \mu_2)$

y $\beta_2 = \Sigma_2^{-1}(\mu_1 - \mu_2)$, y luego decidir cuál es mejor.

Ejemplo 1.3.2. Sea $\boldsymbol{X}$ un vector aleatorio bivariado con $E[\boldsymbol{X}] = \mu_1 = \begin{pmatrix} 7 \\ 7 \end{pmatrix}$, $Var[\boldsymbol{X}] = \Sigma_1 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ para la población 1 y $E[\boldsymbol{X}] = \mu_2 = \begin{pmatrix} 4 \\ 4 \end{pmatrix}$, $Var[\boldsymbol{X}] = \Sigma_2 = \begin{pmatrix} 5 & -3 \\ -3 & 5 \end{pmatrix}$ en la población 2. Si se usa Σ_1 para calcular los coeficientes de la función discriminante, se obtiene

$$\beta_1 = \Sigma_1^{-1}(\mu_1 - \mu_2) = \frac{1}{2} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 3 \end{pmatrix} = \frac{3}{2} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

y si se elige Σ_2, entonces

$$\beta_2 = \Sigma_2^{-1}(\mu_1 - \mu_2) = \frac{1}{16} \cdot \begin{pmatrix} 5 & 3 \\ 3 & 5 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 3 \end{pmatrix} = \frac{3}{2} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

si se utiliza Σ_1 o Σ_2, en ambos casos se obtiene

$$V = \frac{3}{2}(X_1 + X_2)$$

es la función discriminante, o cualquier combinación equivalente a V. Las distancias estandar bivariadas son idénticas, usando Σ_1 o Σ_2 y esto a pesar del hecho de que las distancias estándar univariantes no dependen de la elección de Σ_1 o Σ_2.

En general, los vectores $\beta_1 = \Sigma_1^{-1}(\mu_1 - \mu_2)$ y $\beta_2 = \Sigma_2^{-1}(\mu_1 - \mu_2)$ no son ni idénticos ni proporcionales, pero una función discriminante lineal puede ser definida de forma única, multiplicando por una constante aún en los casos donde las matrices de varianzas y covarianzas son diferentes.

En las aplicaciones, se deben calcular dos vectores de coeficientes de la función discriminante

$$\boldsymbol{b}_1 = \boldsymbol{S}_1^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \quad \text{y} \quad \boldsymbol{b}_2 = \boldsymbol{S}_2^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \tag{1.44}$$

para evaluar el efecto de las diferencias entre las matrices de varianzas y covarianzas en la función discriminante lineal. Se pueden calcular los valores de $V_1 = \boldsymbol{b}_1'\boldsymbol{X}$ y $V_2 = \boldsymbol{b}_2'\boldsymbol{X}$ para todas las observaciones y estudiar la distribución conjunta de V_1 y V_2 en un diagrama de dispersión. Idealmente, si las diferencias entre $\boldsymbol{S}_1$ y $\boldsymbol{S}_2$ no afectan a la función discriminante lineal en absoluto, la correlación entre la V_1 y V_2 sería 1, aunque este no es un procedimiento común en el análisis discriminante.

1.4. Teoría normal en la discriminación lineal

En esta sección se retoma a la función discriminante lineal usando la distribución normal multivariada, concretamente, se especifican las distribuciones normales multivariadas para un vector aleatorio en dos grupos, formulando un principio de clasificación en términos de probabilidades a posteriori, y, se mostrará que esa clasificación lleva al uso de la función discriminante lineal derivada de los principios heurísticos.

Se cambiará la notación utilizada en la secciones anteriores, X denotará una variable aleatoria discreta que indica la pertenencia a un grupo. Suponga que X toma valores 1 y 2 con probabilidades π_1 y π_2, $\pi_1 + \pi_2 = 1$, además que, condicionando a $X = 1$, $\boldsymbol{Y}$ sigue alguna distribución p-variada con f.d.p. $f_1(\boldsymbol{y})$. Del mismo modo, condicionando a $X = 2$, $\boldsymbol{Y}$ tiene f.d.p. $f_2(\boldsymbol{y})$. Las π_j son las probabilidades a priori de pertenecer a un grupo j, $j = 1, 2$. Se Denota por $f_j(\boldsymbol{y})$ a la f.d.p. condicional de $\boldsymbol{Y}$, dada $X = j$, la f.d.p marginal de $\boldsymbol{Y}$ es la densidad

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \pi_1 f_1(\boldsymbol{y}) + \pi_2 f_2(\boldsymbol{y}) \tag{1.45}$$

Finalmente, la probabilidad condicional de pertenecer al grupo j, dado que $\boldsymbol{Y} = \boldsymbol{y}$, está dada

por

$$\pi_{jy} = Pr[X = j \mid \boldsymbol{Y} = \boldsymbol{y}] = \frac{\pi_j f_j(\boldsymbol{y})}{\pi_1 f_1(\boldsymbol{y}) + \pi_2 f_2(\boldsymbol{y})} = \frac{\pi_j f_j(\boldsymbol{y})}{f_Y(\boldsymbol{y})} \quad j = 1, 2. \tag{1.46}$$

Las $\pi_{j\boldsymbol{y}}$ son llamadas usualmente probabilidades a posteriori.

Con esta notación, se puede formular un principio intuitivamente razonable de clasificación como sigue: clasificar una observación $\boldsymbol{y} \in \Re^p$ en el grupo 1 si $\pi_{1\boldsymbol{y}} > \pi_{2\boldsymbol{y}}$, y clasificarla en el grupo 2 si $\pi_{2\boldsymbol{y}} > \pi_{1\boldsymbol{y}}$. Por el momento, no hay preocupación de la posibilidad de que $\pi_{1\boldsymbol{y}} = \pi_{2\boldsymbol{y}}$. Como $\pi_1 + \pi_2 = 1$ para toda $\boldsymbol{y} \in \Re^p$, esta regla es equivalente a lo siguiente: clasificar en el grupo 1 si $\pi_{1\boldsymbol{y}} > \frac{1}{2}$, y clasificar en el grupo 2 si $\pi_{2\boldsymbol{y}} > \frac{1}{2}$, esto define dos regiones de clasificación $C_j = \left\{\boldsymbol{y} \in \Re^p : \pi_{jy} > \frac{1}{2}\right\} \quad j = 1, 2$, con una frontera de clasificación dada por el conjunto $\left\{\boldsymbol{y} \in \Re^p : \pi_{1y} = \frac{1}{2}\right\}$.

El propósito principal de esta sección es mostrar que, si las distribuciones condicionales de $\boldsymbol{Y}$ son normales multivariadas con matrices de varianzas y covarianzas iguales, entonces esta regla establecerá que la clasificación debe basarse en la función discriminante lineal.
Continuando con la situación de dos distribuciones normales con matriz de varianzas y covarianzas común, es decir,

$$\boldsymbol{Y} \sim \mathcal{N}_p(\mu_1, \Sigma) \quad \text{en el grupo 1}$$

$$\boldsymbol{Y} \sim \mathcal{N}_p(\mu_2, \Sigma) \quad \text{en el grupo 2}$$

Donde Σ es definida positiva. Entonces la f.d.p. de $\boldsymbol{Y}$ en el grupo j está dada por:

$$f_j(\boldsymbol{y}) = (2\pi)^{-p/2} (det\Sigma)^{-1/2} exp\left[-\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{\mu_j}\right)' \Sigma^{-1} \left(\boldsymbol{y} - \boldsymbol{\mu_j}\right)\right], \quad y \in \Re^p, \quad j = 1, 2. \tag{1.47}$$

La condición $\pi_{1\boldsymbol{y}} > \pi_{2\boldsymbol{y}}$ es equivalente a $\pi_1 f_1(\boldsymbol{y}) > \pi_2 f_2(\boldsymbol{y})$, que a su vez es equivalente a:

$$\log \frac{\pi_1 f_1(\boldsymbol{y})}{\pi_2 f_2(\boldsymbol{y})} > 0. \tag{1.48}$$

pero

$$\begin{aligned}\log \frac{f_1(\boldsymbol{y})}{f_2(\boldsymbol{y})} &= \frac{1}{2}\Big[(\boldsymbol{y}-\boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_2) - (\boldsymbol{y}-\boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_1)\Big] \\ &= (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}\boldsymbol{y} + \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) \\ &= \boldsymbol{\beta}'\left[\boldsymbol{y} - \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right]\end{aligned} \tag{1.49}$$

donde

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \tag{1.50}$$

es el vector de coeficientes de la función discriminante lineal. Así, el principio de asignar $\boldsymbol{y} \in \Re^p$ para el grupo con la mayor probabilidad a posteriori conduce a una regla de clasificación basada en el valor de $\boldsymbol{\beta}'\boldsymbol{y}$:

$$\text{Clasificar en el grupo} \begin{cases} 1 & si \quad \beta' y - \frac{1}{2}\beta'(\mu_1+\mu_2) > \log(\pi_2/\pi_1) \\ 2 & si \quad \beta' y - \frac{1}{2}\beta'(\mu_1+\mu_2) < \log(\pi_2/\pi_1) \end{cases} \tag{1.51}$$

Si $\pi_1 = \pi_2 = \frac{1}{2}$, entonces esta regla se simplifica a un punto de corte en 0, sin embargo, incluso para probabilidades a priori diferentes de $\frac{1}{2}$, la regla establece que la clasificación debe depender del vector $\boldsymbol{y}$ de datos observados sólo a través del valor de la función discriminante lineal $\boldsymbol{\beta}'\boldsymbol{y}$. Este es un resultado notable y altamente deseable, ya que implica que los datos multivariados pueden reducirse sin pérdida de información (al menos en la disposición de dos distribuciones normales multivariadas con matrices de varianzas y covarianzas idénticas) a los valores de la función discriminante lineal, independientemente de las probabilidades a priori. En otras palabras, la búsqueda de los límites de clasificación es equivalente a encontrar un solo punto de corte para la distribución univariada de la función discriminante.

En el desarollo de la teoría normal, la función discriminante lineal $V = \boldsymbol{\beta}'Y$ tiene una distribución normal univariada con medias $v_1 = \boldsymbol{\beta}'\boldsymbol{\mu}_1 = (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1$ y $v_2 = \boldsymbol{\beta}'\boldsymbol{\mu}_2 =$

$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2$, y con varianza común $\sigma^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, que es idéntica a la distancia estándar cuadrada Δ^2, esto implica que el cálculo de las probabilidades a posteriori puede basarse en la variable discriminante V.

Ejemplo 1.4.1. Continuando con el ejemplo de clasificación de los mosquitos AF y APF. Sea $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$ un vector aleatorio normal bivariado con medias $\hat{\mu}_1 = \begin{pmatrix} 141.33 \\ 180.44 \end{pmatrix}$ para el grupo 1, $\hat{\mu}_2 = \begin{pmatrix} 122.67 \\ 192.67 \end{pmatrix}$ para el grupo 2 y matriz de varianzas y covarianzas común $\hat{\Sigma} = \begin{pmatrix} 75.49 & 66.46 \\ 66.46 & 133.81 \end{pmatrix}$. Sustituyendo las cantidades de la muestra por los parámetros del modelo. Suponga que las probabilidades a priori son $\pi_1 = \pi_2 = \frac{1}{2}$. La Figura 1.6 muestra una gráfica de contorno de $\pi_1 f_1(\boldsymbol{y})$ y $\pi_2 f_2(\boldsymbol{y})$, donde f_1 y f_2 son densidades normales. El límite entre las dos regiones de clasificación es el conjunto de todos los puntos donde las curvas de igual altitud de las dos "montañas normales" se intersectan.

La región límite esta definida por las ecuaciones (1.49) y (1.50).

Las probabilidades a posteriori de pertenencia en el grupo j pueden ser calculadas como:

$$\begin{aligned} \pi_{jv} &= Pr[X = j | V = v] \\ &= Pr\left[X = j | \beta' Y = v\right] \\ &= \frac{\pi_j h_j(v)}{\pi_1 h_1(v) + \pi_2 h_2(v)}, \quad j = 1, 2 \end{aligned} \tag{1.52}$$

Donde $h_j(v)$ es la f.d.p. de una variable aleatoria normal con media $v_j = \beta' \mu_j$ y varianza Δ^2. La línea delimitante para la clasificación corresponde al valor único de z tal que $\pi_{1v} = \pi_{2v} = \frac{1}{2}$.

Una cuestión importante en las aplicaciones prácticas es cómo elegir los valores adecuados para π_1 y π_2 en la determinación de las probabilidades a posteriori, y por lo tanto, identificar las regiones de clasificación. A menudo, en situaciones donde se utiliza el análisis discriminante, las muestras de entrenamiento se obtienen de las distribuciones condicionales de $\boldsymbol{Y}$, dada $X = j$, y los tamaños de muestra N_1 y N_2 son fijos, a veces, las probabilidades a priori iguales se utilizan para expresar la ignorancia del investigador acerca de la verosimilitud que las observaciones futuras serán de los grupos 1 y 2, respectivamente.

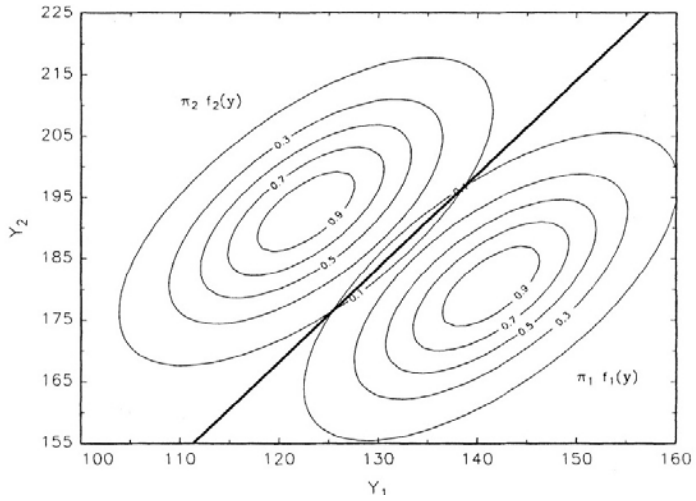


Figura 1.6: Gráfica de contorno utlizando la teoría normal en el ejemplo de la clasificación de mosquitos

En otros casos, el muestreo puede ser de la distribución conjunta de X y $\boldsymbol{Y}$, es decir, tanto el código del grupo X y las variables medidas $\boldsymbol{Y}$ son aleatorias. Los tamaños de la muestra N_j, también son aleatorios, y tiene sentido intuitivo para estimar π_j por $\hat{\pi}_j = N_j/(N_1 + N_2)$, $j = 1, 2$, en realidad estas $\hat{\pi}_j$ también son estimaciones de máxima verosimilitud.

El análisis discriminante involucra a la distribución condicional de $\boldsymbol{Y}$, dado X, en particular, en esta sección se ha discutido el caso en el que la distribución condicional de $\boldsymbol{Y}$, dada X, es normal multivariada, tanto para $X = 1$ como para $X = 2$. Al clasificar un nuevo vector de datos $\boldsymbol{y}$ en uno de los dos grupos, se cambió la perspectiva y se estudió la distribución condicional de X, dado $\boldsymbol{Y} = \boldsymbol{y}$, lo que produjo las probabilidades a posteriori π_{jy}, sin embargo, en algunos casos puede ser más razonable para modelar la distribución condicional de X, dado $\boldsymbol{Y} = \boldsymbol{y}$, directamente, sin realizar algún supuesto distribucional de $\boldsymbol{Y}$. De esta forma se desea calcular la probabilidad de éxito como una función de $\boldsymbol{Y}$. En una situación más general, sería deseable tratar a X como la variable dependiente y a $\boldsymbol{Y}$ como un vector de dimensión p "independiente", o mejor , como las variables predictoras, la técnica estadística más popular basada en este enfoque es la regresión logística.

1.5. Tasas de error

En esta sección se estudiarán las tasas de error frecuentemente utilizadas para evaluar el desempeño de los procedimientos de clasificación.

Una variable discreta X toma valores 1 y 2, indicando la pertenencia a un grupo, con probabilidades π_1 y π_2, y un vector aleatorio $\boldsymbol{Y}$ de dimensión p con f.d.p. $f_j(\boldsymbol{y})$ condicionado a $X = j$, $j = 1, 2$. En ocaciones, se hará referencia a las dos distribuciones condicionales como las poblaciones, o simplemente como grupos.

Un procedimiento de clasificación sirve para el propósito de identificar la pertenencia al grupo de un nuevo vector observado $\boldsymbol{y} \in \Re^p$, sin conocer el valor de X. Por lo general, los errores no pueden evitarse, y, por lo tanto, sería conveniente desarrollar un procedimiento que haga que, en cierto sentido, disminuyan los errores.

Se mencionará un método de clasificación en particiones de dos grupos del espacio muestral de $\boldsymbol{Y}$ en dos regiones de clasificación $\mathcal{C}_1$ y $\mathcal{C}_2$. Por comodidad, en este desarrollo se supone que el espacio muestral es $\Re^p$, así $\mathcal{C}_1 \cup \mathcal{C}_2 = \Re^p$. Se define una función $x^* : \Re^p \to \{1, 2\}$ tal que

$$x^*(\boldsymbol{y}) = \begin{cases} 1 & si \quad y \in C_1 \\ 2 & si \quad y \in C_2 \end{cases} \tag{1.53}$$

Es decir, $x^*(\boldsymbol{y})$ es la función que asigna a $\boldsymbol{y}$ la "pertenencia a un grupo pre-establecido".

Con lo anterior, se puede mostrar una definición formal de una regla de clasificación.

Definición 1.5.1. *Una regla de clasificación es una variable aleatoria*

$$X^* = x^*(\boldsymbol{Y}), \tag{1.54}$$

que toma el valor 1 si $\boldsymbol{Y} \in \mathcal{C}_1$, y el valor 2 si $\boldsymbol{Y} \in \mathcal{C}_2$.

Siguiendo esta definición, X^* puede ser considerada como la pertenencia al grupo pre-establecido, mientras que X representa la pertenencia al grupo verdadero. Idealmente, se tendría que $X = X^*$ con probabilidad uno, es decir, la clasificación correcta en todos los casos.

El desempeño de una regla de clasificación puede ser evaluado por una tabla de la distribución conjunta de X y X^* como sigue:

	$X = 1$	$X = 2$	Marginal
$X^* = 1$	$q_{11}\pi_1$	$q_{12}\pi_2$	$q_{11}\pi_1 + q_{12}\pi_2$
$X^* = 2$	$q_{21}\pi_1$	$q_{22}\pi_2$	$q_{21}\pi_1 + q_{22}\pi_2$
Marginal	π_1	π_2	

Con la tabla anterior, las $\pi_j = Pr[X = j]$ y q_{ij} representan las probabilidades condicionales de X^* dada X, es decir, $q_{ij} = Pr[X^* = i \mid X = j]$, con $q_{11} + q_{21} = q_{12} + q_{22} = 1$. Las dos entradas, tales que $X^* = X$ corresponde a la clasificación correcta, mientras que las entradas $(X^* = 1, X = 2)$ y $(X^* = 2, X = 1)$ representan asignaciones incorrectas. Una buena regla de clasificación intentará hacer pequeñas las probabilidades de asignaciones incorrectas.

En algunos casos, se pueden observar las probabilidades condicionales de clasificación errónea q_{12} y q_{21}, haciéndolas simultáneamente tan pequeñas como sea posible o fijando una de ellas, independientemente de las probabilidades a priori π_j, en otros casos, se desea encontrar una regla de clasificación que minimice una evaluación general de error, como se muestra en la siguiente definición.

Definición 1.5.2. *La probabilidad general de clasificación errónea, o tasa de error, de una regla de clasificación X^* está dada por:*

$$\gamma(X^*) := Pr[X^* \neq X] = q_{21}\pi_1 + q_{12}\pi_2 \tag{1.55}$$

$$\textit{donde} \quad \pi_j = Pr[x = j] \quad \textit{y} \quad q_{ij} = Pr[X^* = i \mid X = j] \quad \textit{para} \quad i, j = 1, 2$$

Ejemplo 1.5.1. Este es un ejemplo práctico del caso univariado para ilustrar las definiciones anteriores. Suponga que la distribución condicional de $\boldsymbol{Y}$, dada $X = j$, es $\mathcal{N}\left(\mu_j, \sigma^2\right)$, $j = 1, 2$,

adicionalmente que $\mu_1 > \mu_2$. Se define una regla de clasificación como

$$X^* = \begin{cases} 1 & si \quad Y > c \\ 2 & si \quad Y \leq c \end{cases}$$

Aquí, c es una constante real que aún no se ha especificado, entonces

$$q_{21} = Pr[X^* = 2 \mid X = 1] = \Phi\left(\frac{c - \mu_1}{\sigma}\right)$$

y

$$q_{12} = Pr[X^* = 1 \mid X = 2] = 1 - \Phi\left(\frac{c - \mu_2}{\sigma}\right)$$

Donde Φ es la función de distribución de una variable aleatoria normal estándar. Para cualquier $c \in \Re$, la tasa de error $\gamma(X^*)$ puede ser calculada ahora con la ecuación (1.55), siempre que las probabilidades a priori sean conocidas.

El desarrollo del ejemplo anterior podría representar una situación en la que los dos grupos corresponden a las personas que son inmunes (grupo 1) o susceptibles (grupo 2) a una cierta enfermedad , y Y es el resultado de una prueba de detección para la inmunidad. En tal caso, se podría querer mantener la probabilidad q_{12}=Pr[Una persona se declara inmune | la persona es suspectible] pequeña, independientemente de la probabilidad de error q_{21} o de la tasa de error.

Si las probabilidades a priori son conocidas o estimadas, se desearía encontrar una regla de clasificación X^* que minimice $\gamma(X^*)$.

Teorema 1.5.1. *Una regla de clasificación X^* que minimice la tasa de error $\gamma(X^*)$ está dada por:*

$$X^*_{opt} \begin{cases} 1 & si \quad y \in \mathcal{C}_1 \\ 2 & si \quad y \in \mathcal{C}_2 \end{cases} \tag{1.56}$$

donde

$$\mathcal{C}_1 = \{y \in \Re^p : \pi f_1(\boldsymbol{y}) \geq \pi_2 f_2(\boldsymbol{y})\}$$

y

$$\mathcal{C}_2 = \{y \in \Re^p : \pi f_1(\boldsymbol{y}) < \pi_2 f_2(\boldsymbol{y})\} \tag{1.57}$$

El teorema muestra que el mejor procedimiento de clasificación, en términos de las tasas de error, es el mismo que la regla de clasificación basada en la máxima probabilidad a posteriori. Se refiere a menudo como la regla de Bayes. Note que en (1.57), el conjunto de fronteras $\{y \in \Re^p : \pi_1 f_1(y) = \pi_2 f_2(y)\}$ está asociado arbitrariamente con la región $\mathcal{C}_1$; también se podría asociar con $\mathcal{C}_2$ o dividirlo entre $\mathcal{C}_1$ y $\mathcal{C}_2$, así X^*_{opt} no se define de forma única , pero si el conjunto de fronteras tiene probabilidad cero en ambos grupos, entonces esta no unicidad es irrelevante. De ahora en adelante, se hará referencia a una regla de clasificación X^*_{opt} que minimiza la tasa de error como una regla de clasificación óptima y a la tasa de error asociada $\gamma_{opt} = \gamma\left(X^*_{opt}\right)$ como la tasa de error óptima para un problema de clasificación dado.

Ejemplo 1.5.2. En el desarrollo de la teoría normal con matrices de covarianzas idénticas, la regla de clasificación óptima es

$$X^*_{opt} \begin{cases} 1 & si \quad \beta'\left[Y - \frac{1}{2}(\mu_1 - \mu_2)\right] \geq \log(\pi_2/\pi_1) \\ 2 & \text{en otro caso} \end{cases} \tag{1.58}$$

donde $\beta' = \psi^{-1}(\mu_1 - \mu_2)$. Esto se deduce de las ecuaciones (1.48) y (1.49) de la sección anterior y el teorema 1.6.1. Con $\Delta = \left[(\mu_1 - \mu_2)' \psi^{-1}(\mu_1 - \mu_2)\right]^{1/2}$ como la distancia estándar , la tasa de error óptima está dada por

$$\gamma_{opt} = \pi_1 \Phi\left(-\frac{1}{2}\Delta + \frac{1}{\Delta}\log\frac{\pi_2}{\pi_1}\right) + \pi_2 \Phi\left(-\frac{1}{2}\Delta - \frac{1}{\Delta}\log\frac{\pi_2}{\pi_1}\right) \tag{1.59}$$

Para $\pi_1 = \pi_2 = \frac{1}{2}$, se reduce a

$$\gamma_{opt} = \Phi\left(-\frac{1}{2}\Delta\right) \tag{1.60}$$

que se definió como "la función de traslape" con anterioridad.

Una regla de clasificación obtenida a partir de muestras de entrenamiento, en lugar de modelos conocidos , se denota por $\hat{X}^*$, por ejemplo, la regla de clasificación basada en la función discriminante lineal muestral puede ser descrita como:

$$\hat{X}^* = \begin{cases} 1 & si \quad b'\left[y - \frac{1}{2}\left(\bar{y}_1 - \bar{y}_2\right)\right] \geq c \\ 2 & \text{en otro caso} \end{cases}$$

Donde $\boldsymbol{b} = \boldsymbol{S}^{-1}\left(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2\right)$. De acuerdo con el principio de optimalidad, elegiríamos $c = \log\left(\pi_2/\pi_1\right)$, si las probabilidades a priori se suponen fijas y conocidas, y $c = \log\left(\hat{\pi}_2/\hat{\pi}_1\right) = \log\left(N_2/N_1\right)$, si las probabilidades a priori son estimadas de acuerdo a los tamaños de muestra.

Ahora, se estudiarán varios métodos para evaluar el desempeño de reglas de clasificación estimadas a partir de las muestras. La más simple y popular medida, es llamada tasa de error "plug-in", que es la proporción de observaciones clasificadas erróneamente cuando se aplica la regla de clasificación $\hat{X}^*$ para los datos de las muestras de entrenamiento.

Formalmente, para cada una de las observaciones $N_1 + N_2$ en las muestras de entrenamiento se tiene la pertenencia a un grupo observado x_i y la pertenencia a un grupo pre-establecido $\hat{x}_i^*$. Sea $e_i = 0$, si $x_i = \hat{x}_i^*$, y $e_i = 1$ si $x_i \neq \hat{x}_i^*$, entonces la tasa de error plug-in está dada por

$$\hat{\gamma}_{plug-in} = \frac{1}{N_1 + N_2} \sum_{i=1}^{N} e_i \tag{1.61}$$

Si la clasificación se hace utilizando la funcion discriminante lineal y si el investigador está dispuesto a suponer que la función discriminante lineal es aproximadamente normal con varianza igual en ambos grupos, entonces se puede usar una estimación con apoyo de la teoría normal de la tasas de error siguiendo la ecuación (1.59)

$$\hat{\gamma}_{normal} = \pi_1 \Phi\left(-\frac{1}{2}D + \frac{1}{D}\log\frac{\pi_2}{\pi_1}\right) + \pi_2 \Phi\left(-\frac{1}{2}D - \frac{1}{D}\log\frac{\pi_2}{\pi_1}\right) \tag{1.62}$$

D es la distancia estándar muestral, y las probabilidades a priori π_j pueden ser sustituidas por estimaciones, dependiendo de la situación. En particular, para $\pi_1 = \pi_2 = \frac{1}{2}$, se obtiene

$\hat{\gamma}_{normal} = \Phi\left(-\frac{1}{2}D\right)$.

1.6. Funciones discriminantes lineales y medias condicionales

En esta sección se estudiaran las propiedades de la función discriminante lineal con más detalle, en particular se investigará el problema de redundancia, es decir, ¿bajo qué circunstancias una variable o un conjunto de variables pueden ser excluidas del análisis sin pérdida de información?. Los resultados presentados en este apartado son de valor considerable para una comprensión completa del análisis discriminante lineal.

A lo largo de esta sección se asume que $\boldsymbol{Y} = \left(Y_1, \ldots, Y_p\right)'$ es un vector aleatorio de dimensión p con $E_1[\boldsymbol{Y}] = \mu_1$ en el grupo 1, $E_2[\boldsymbol{Y}] = \mu_2$ en el grupo 2, y $Var[\boldsymbol{Y}] = \Sigma$ en ambos grupos, Σ es definida positiva, y sea $\delta = \mu_1 - \mu_2$ el vector de la diferencia de medias y $\beta = \Sigma^{-1}\delta$ el vector de coeficientes de la función discriminante.

El primer resultado indica cómo los coeficientes de la función discriminante lineal cambian si las variables se transforman linealmente.

Lema 1.6.1. *Sea $\boldsymbol{A}$ una matriz no singular de orden $p \times p$, y $\boldsymbol{b} \in \Re^p$ un vector fijo. Se define $\boldsymbol{Z} = \boldsymbol{AY} + \boldsymbol{b}$, y sean $\boldsymbol{\beta_Y}$ y $\boldsymbol{\beta_Z}$ los vectores de coeficientes de la función discriminante de $\boldsymbol{Y}$ y $\boldsymbol{Z}$, respectivamente, entonces*

$$\boldsymbol{\beta_Y} = \boldsymbol{A'\beta_Z} \tag{1.63}$$

o, equivalentemente,

$$\boldsymbol{\beta_Z} = \left(\boldsymbol{A'}\right)^{-1}\boldsymbol{\beta_Y} \tag{1.64}$$

Demostración: Para todos los parámetros involucrados, se usarán subíndices *"Y"* y *"Z"* para indicar que se hará referencia a los vectores aleatorios $\boldsymbol{Y}$ y $\boldsymbol{Z}$, respectivamente. Dada $\boldsymbol{Z} = \boldsymbol{Ay} + \boldsymbol{b}$

$$E_1[\boldsymbol{Z}] = \boldsymbol{A\mu_1} + \boldsymbol{b} \quad \text{en el grupo 1}$$

$$E_2[\boldsymbol{Z}] = \boldsymbol{A\mu_2} + \boldsymbol{b} \quad \text{en el grupo 2}$$

y

$$Var[Z] = \boldsymbol{\Sigma}_Z = A\boldsymbol{\Sigma}_Y A' \quad \text{en ambos grupos}$$

Entonces, el vector de diferencia de medias en Z está dado por $\boldsymbol{\delta}_Z = A(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = A\boldsymbol{\delta}_Y$, y el vector de coeficientes de la función discriminante lineal en Z es:

$$\begin{aligned}\boldsymbol{\beta}_Z &= \boldsymbol{\Sigma}_Z^{-1}\boldsymbol{\delta}_Z \\ &= \left(A\boldsymbol{\Sigma}_Y A'\right)^{-1} A\boldsymbol{\delta}_Y \\ &= \left(A'\right)^{-1}\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\delta}_Y \\ &= \left(A'\right)^{-1}\boldsymbol{\beta}_Y\end{aligned}$$

Se estudió anteriormente que la distancia estándar multivariada es invariante bajo transformaciones lineales no singulares y que la distancia estándar cuadrada entre los dos vectores de medias se puede escribir como

$$\Delta_Y^2 = \boldsymbol{\beta}_Y'\boldsymbol{\delta}_Y;$$

Con apoyo de la ecuación (1.31). Para un vector aleatorio Z (como en el lema 1.6.1),

$$\begin{aligned}\Delta_Z^2 &= \boldsymbol{\beta}_Z'\boldsymbol{\delta}_Z \\ &= \left(\boldsymbol{\beta}_Y' A^{-1}\right)(A\boldsymbol{\delta}_Y) \\ &= \Delta_Y^2\end{aligned}$$

que confirma la invarianza de la distancia estándar.

A continuación, se divide el vector aleatorio Y en subvectores de dimensión q y $p - q$, respectivamente:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \tag{1.65}$$

donde

$$Y_1 = \begin{pmatrix} Y_1 \\ \vdots \\ Y_q \end{pmatrix}$$

y

$$Y_2 = \begin{pmatrix} Y_{q+1} \\ \vdots \\ Y_p \end{pmatrix}$$

Contienen la primera q y las últimas $(p-q)$ variables. La partición del vector de diferencia de medias δ y la matriz de varianzas y covarianzas Σ es análoga, es decir,

$$\boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \end{pmatrix}$$

y

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \tag{1.66}$$

además, sea

$$\Delta^2_{Y_1} = \boldsymbol{\delta}'_1 \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\delta}_1$$

y

$$\Delta^2_{Y_2} = \boldsymbol{\delta}'_2 \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\delta}_2 \tag{1.67}$$

que denotan las distancias estándar multivariadas cuadradas de $\boldsymbol{Y}_1$ y $\boldsymbol{Y}_2$, respectivamente. Entonces se puede establecer el siguiente resultado:

Lema 1.6.2. *En el desarrollo de las ecuaciones (1.65) a (1.67), si* $\boldsymbol{\Sigma}_{12} = \mathbf{0}$*, es decir, si todas las covarianzas entre* $\boldsymbol{Y}_1$ *y* $\boldsymbol{Y}_2$ *son cero, entonces:*

- *El vector de coeficientes de la función discriminate lineal sobre todas las p variables es*

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\delta}_1 \\ \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\delta}_2 \end{pmatrix} \tag{1.68}$$

- *La distancia estándar cuadrada de dimensión p es la suma de las distancias estándar*

cuadradas en $\boldsymbol{Y}_1$ y $\boldsymbol{Y}_1$, respectivamente, es decir,

$$\Delta^2(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = \boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} = \Delta^2_{Y_1} + \Delta^2_{Y_2} \tag{1.69}$$

Note que la ecuación (1.69) no es cierta en general, es decir, si $\boldsymbol{\Sigma}_{12} \neq \mathbf{0}$, entonces por lo general

$$\Delta^2(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) \neq \Delta^2_{Y_1} + \Delta^2_{Y_2}$$

Se puede intuir el resultado principal de esta sección, una caracterización de condiciones en las que una variable o un conjunto de variables pueden ser omitidas del análisis sin pérdida de información . En primer lugar, se definirá el concepto de redundancia.

Definición 1.6.1. *Sea $\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix}$ particionada en q y $(p-q)$ componentes como en la ecuación (1.65), y sea $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$ el vector de coeficientes de la función discriminante, particionado análogamente como Y. Entonces las variables $\boldsymbol{Y}_2$ son redundantes si $\boldsymbol{\beta}_2 = \mathbf{0}$. Equivalentemente, se dice que el subconjunto de variables contenidas en $\boldsymbol{Y}_1$ es suficiente.*

Una variable es redundante si no aparece en la función discriminante, o, en otras palabras, si su omisión no afecta a la función discriminante. Por simplicidad en la notación, la definición es en términos de redundancia de las últimas $(p - q)$ variables, pero se aplica a cualquier subconjunto de variables propiamente ordenadas.

Ejemplo 1.6.1. Suponga que $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$ es un vector aleatorio bivariado con matriz de varianzas y covarianzas $\Sigma = \begin{pmatrix} 2 & 2 \\ 2 & 3 \end{pmatrix}$ y los vectores de medias $\mu_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$ y $\mu_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. Entonces $\beta = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, es decir, la función discriminante lineal está dada $Z = Y_1$. Por lo tanto, la variable Y_2 es redundante en el conjunto de variables $\{Y_1, Y_2\}$. Por otro lado, si el conjunto de variables

consiste únicamente de Y_2, entonces Y_2 no es redundante debido a que su diferencia de medias no es cero.

El caso contrario también puede suceder, como lo ilustra el caso $\Sigma = \begin{pmatrix} 2 & 2 \\ 2 & 3 \end{pmatrix}$, $\delta = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$. Entonces $\beta = \begin{pmatrix} 3 \\ -2 \end{pmatrix}$, es decir, la función discriminante lineal está dada por $Z = 3Y_1 - 2Y_2$. Para el conjunto de variables $\{Y_1, Y_2\}$, ninguna de las variables es redundante, aunque la variable Y_2 por sí sola no proporciona ninguna información acerca de las diferencias en la ubicación.

Por el teorema principal de esta sección, se requiere notación adicional. Sea $\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix}$ particionado en q y $(p-q)$ variables, sean

$$\Delta_q^2 = \boldsymbol{\delta}'_1 \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\delta}_1$$

y

$$\Delta_p^2 = \boldsymbol{\delta}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta} \tag{1.70}$$

las distancias estándar cuadradas de las primeras q variables $\left(\Delta_q^2\right)$ y de todas las p variables $\left(\Delta_p^2\right)$, respectivamente. Sea la partición del vector de coeficientes de la función discriminante $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$, el vector de diferencia de medias $\boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \end{pmatrix}$ y

$$\delta_{2.1} = \delta_2 - \Sigma_{21} \Sigma_{11}^{-1} \delta_1 \tag{1.71}$$

Teorema 1.6.1. *Las siguientes tres condiciones son equivalentes*

(a) $\beta_2 = \mathbf{0}$, *es decir, las ultimas (p-q) variables son redundantes*

(b) $\Delta_p^2 = \Delta_q^2$

(c) $\delta_{2.1} = \mathbf{0}$

Demostración: Considere la transformación lineal

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} I_q & 0 \\ -\Sigma_{21}\Sigma_{11}^{-1} & I_{p-q} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \tag{1.72}$$

corresponde a $Z_1 = Y_1$ y $Z_2 = Y_2 - \Sigma_{21}\Sigma_{-1}^{-1}Y_1$. El vector de la diferencia de medias de Z, dado en la forma particionada usual, es:

$$\delta_Z = \begin{pmatrix} \delta_1 \\ \delta_{2.1} \end{pmatrix} \tag{1.73}$$

Y la matriz de varianzas y covarianzas de Z es:

$$\Sigma_Z = Var[Z] = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22.1} \end{pmatrix} \tag{1.74}$$

Donde

$$\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \tag{1.75}$$

Dado que la matriz de la transformación en la ecuación (1.72) es no singular, se puede aplicar el lema 1.6.1 y 1.6.2. Se deonta el vector de coeficientes de la función discriminante de Z como $\beta^* = \begin{pmatrix} \beta_1^* \\ \beta_2^* \end{pmatrix}$

$$\begin{aligned} \begin{pmatrix} \beta_1^* \\ \beta_2^* \end{pmatrix} &= \begin{pmatrix} I_q & \Sigma_{11}^{-1}\Sigma_{12} \\ 0 & \Sigma_{p-q} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \\ &= \begin{pmatrix} \beta_1 + \Sigma_{11}^{-1}\Sigma_{12}\beta_2 \\ \beta_2 \end{pmatrix} \\ &= \begin{pmatrix} \Sigma_{11}^{-1}\delta_1 \\ \Sigma_{22.1}^{-1}\delta_{2.1} \end{pmatrix} \end{aligned} \tag{1.76}$$

y

$$\Delta_p^2 = \delta'_1\Sigma_{11}^{-1}\delta_1 + \delta'_{2.1}\Sigma_{22.1}^{-1}\delta_{2.1}$$

$$= \Delta_q^2 + \boldsymbol{\delta}'_{2.1}\boldsymbol{\Sigma}_{22.1}^{-1}\boldsymbol{\delta}_{2.1} \tag{1.77}$$

De la ecuación (1.77) y porque $\Sigma_{22.1}$ es definida positiva, $\Delta_p^2 = \Delta_q^2$ implica $\delta_{2.1} = \mathbf{0}$. A continuación, de la ecuación (1.76), $\delta_{2.1} = \mathbf{0}$ implica $\beta_2 = \mathbf{0}$. Finalmente, $\beta_2 = \mathbf{0}$ implica

$$\Delta_p^2 = \boldsymbol{\beta}'\boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\beta}_1' & \mathbf{0}' \end{pmatrix} \begin{pmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \end{pmatrix} = \boldsymbol{\beta}'_1\boldsymbol{\delta}_1 = \Delta_q^2$$

Note que la ecuación (1.76) implica

$$\boldsymbol{\beta}_2 = \boldsymbol{\Sigma}_{22.1}^{-1}\boldsymbol{\delta}_{2.1} \tag{1.78}$$

y, análogamente

$$\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_{11.2}^{-1}\boldsymbol{\delta}_{1.2} \tag{1.79}$$

Donde $\delta_{1.2}$ y $\Sigma_{11.2}^{-1}$ son definidas análogamente como en (1.71) y (1.75). La ecuación (1.77) permite el cálculo simplificado de Δ_q^2.

Se relacionarán los coeficientes de la función discriminante con medias condicionales, y por lo tanto, el análisis discriminante lineal con el análisis de regresión. Con apoyo del Teorema A.2.3 del Apéndice, se observa que, para dos variables aleatorias distribuidas conjuntamente Y y Z, si $E[Z \mid Y = y] = \alpha + \beta y$ para cualquier $\alpha \in \Re$ y $\beta \in \Re$,

$$\beta = \frac{cov[Y, Z]}{var[Y]}, \quad \alpha = E[Z] - \beta E[Y] \tag{1.80}$$

Se expresará una generalización multivariada de este resultado.

Teorema 1.6.2. *Sea* $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$ *un vector aleatorio de dimensión* p *con media* $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ *y matriz de varianzas y covarianzas* $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, *particionado en* q *y* $(p-q)$ *componentes. Se asume que la media condicional de* $\boldsymbol{Y}_2$, *dada* $\boldsymbol{Y}_1 = \boldsymbol{y}_1$, *es una función lineal de* $\boldsymbol{y}_1$, *es decir,* $E[\boldsymbol{Y}_2 \mid \boldsymbol{Y}_1 = \boldsymbol{y}_1] = \boldsymbol{c} + \boldsymbol{D}\boldsymbol{y}_1$, *donde* $\boldsymbol{c} \in \Re^{p-q}$ *y* $\boldsymbol{D}$ *es una matriz de dimensión* $(p-q) \times q$.

Entonces

$$\boldsymbol{D} = \Sigma_{21}\Sigma_{11}^{-1}$$

y

$$\boldsymbol{c} = \mu_2 - D\mu_1 \tag{1.81}$$

Se aplicará el Teorema 1.6.2 para el desarrollo de dos grupos que difieren en localización pero no en su matriz de varianzas y covarianzas Σ. Sea

$$\mu^{(j)} = \begin{pmatrix} \mu_1^{(j)} \\ \mu_2^{(j)} \end{pmatrix}, \quad j = 1, 2 \tag{1.82}$$

el vector de medias en el grupo $j - esimo$, asumiendo linealidad en las medias condicionales, el Teorema 1.6.2 dice que

$$E_1[Y_2 \mid Y_1 = y_1] = \mu_2^{(1)} + \Sigma_{21}\Sigma_{11}^{-1}\left(y_1 - \mu_1^{(1)}\right) \quad \text{en el grupo 1}$$

y

$$E_2[Y_2 \mid Y_1 = y_1] = \mu_2^{(2)} + \Sigma_{21}\Sigma_{11}^{-1}\left(y_1 - \mu_1^{(2)}\right) \quad \text{en el grupo 2} \tag{1.83}$$

Condicionando a $Y_1 = y_1$, la diferencia de medias entre los dos grupos para la variable $\boldsymbol{Y_2}$ es

$$\begin{aligned} \delta_{Y_2}(\boldsymbol{y_1}) &= E_2[\boldsymbol{Y_2} \mid \boldsymbol{Y_1} = \boldsymbol{y_1}] - E[\boldsymbol{Y_2} \mid \boldsymbol{Y_1} = \boldsymbol{y_1}] \\ &= \left(\mu_2^{(1)} - \mu_2^{(2)}\right) - \Sigma_{21}\Sigma_{11}^{-1}\left(\mu_2^{(1)} - \mu_2^{(2)}\right) \\ &= \delta_2 - \Sigma_{21}\Sigma_{11}^{-1}\delta_1 \\ &= \delta_{2.1} \end{aligned} \tag{1.84}$$

Note que $\delta_{2.1}$ es un vector de constantes que no dependen de $\boldsymbol{y_1}$, demostrando que las líneas de regresión o planos son paralelos, es el vector de las diferencias de medias condicional de Y_2, dado Y_1, bajo el supuesto de linealidead de la esperanza condicional.

El siguiente corolario muestra que la interpretación de $\Sigma_{22.1}$ como la matriz de covarianzas de una distribución condicional no se limita al caso de normalidad multivariada

Corolario 1.6.1. *Bajo los supuestos del Teorema 1.6.2, suponga que* $Cov[\boldsymbol{Y}_2 \mid \boldsymbol{Y}_1 = \boldsymbol{y}_1]$ *no depende de* $\boldsymbol{y}_1$*, entonces*

$$Cov[\boldsymbol{Y}_2 \mid \boldsymbol{Y}_1 = \boldsymbol{y}_1] = \Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \tag{1.85}$$

En particular, para cualquier variable única Y_h , su coeficiente en la función discriminante lineal es el mismo que la diferencia media dividida entre la varianza de la distribución condicional de Y_h, dadas todas las otras variables. Esto demuestra, una vez más, que los coeficientes de la función discriminante dependen del conjunto completo de variables consideradas. Si el conjunto de variables cambia , entonces los coeficientes de la función discriminante también pueden cambiar. Una variable es redundante exactamente si su diferencia de medias condicional, dadas todas las demás variables, es cero. Si se añaden o eliminan las variables del conjunto utilizado entonces las variables anteriormente importantes pueden ser redundantes , o viceversa.

CAPÍTULO 2

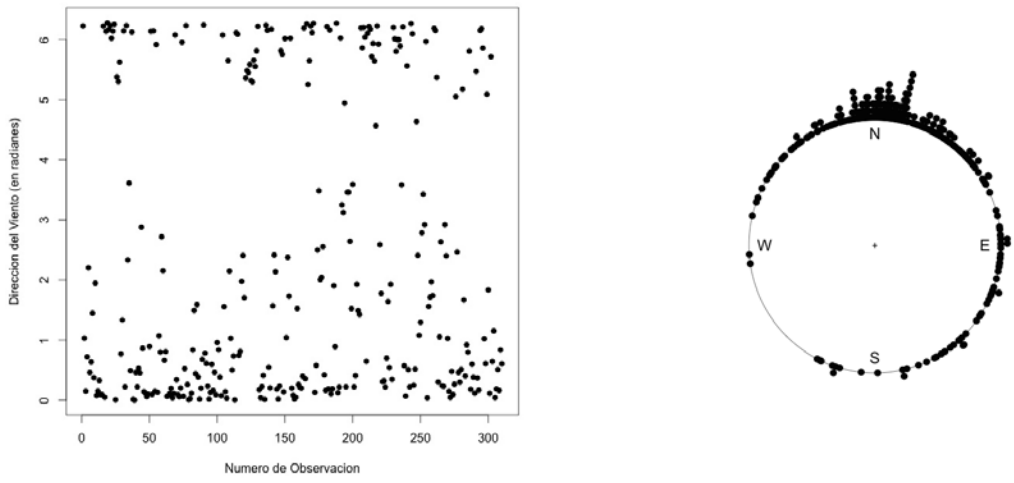
ESTADÍSTICA CIRCULAR

2.1. Datos circulares

En muchos campos científicos diversos, las medidas son direcciones. Por ejemplo, una persona que estudia biología puede medir la dirección de vuelo de un ave o la orientación de un animal, mientras que una persona que se dedica a la geología puede estar interesado(a) en la dirección de los polos magnéticos de la Tierra. Estas direcciones pueden ser en dos dimensiones como en los primeros dos ejemplos o en tres dimensiones como en el último ejemplo. También, cualquier fenómeno periódico con un periodo conocido, digamos un día, un mes, un año, puede ser representado en un círculo donde la circunferencia corresponde a dicho periodo. Por ejemplo, tiempos de llegada a un hospital de pacientes que dicen sufrir ataques al corazón durante el día.

Por ejemplo, se consideran los datos "wind" disponibles en la biblioteca *circular* en el programa **R**. Este objeto contiene 310 direcciones de viento, medidas en sentido horario (como las manecillas del reloj) desde el azimuth en radianes, registrados en una estación meteorológica en los Alpes italianos, cada 15 minutos de las 3:00 a.m. a las 4:00 a.m. del 29 de Enero de 2001 al 31 de Marzo de 2001. Estos datos fueron introducidos en la literatura de la estadística circular por Agostinelli (2007). Aunque los datos contenidos en "wind" son ángulos, no es un objeto

de datos circular; es un objeto de datos estándar que contiene 310 números sin información adicional para que **R** les de alguna interpretación, con la biblioteca circular se pueden trabajar como se muestra en la Figura 2.1.



(a) Datos observados (b) Representación circular

Figura 2.1: Direcciones de viento

2.2. Representación gráfica

2.2.1. Datos sin agrupar

La representación más simple de los datos circulares es una gráfica de datos circulares agrupados, en la que cada observación se representa como un punto en el circulo unitario. La Figura 2.1 (b) ilustra este método.

2.2.2. Datos agrupados

Histogramas circulares

Los datos circulares agrupados pueden ser representados por histogramas circulares, que son análogos a los histogramas en la recta real. Cada barra en un histograma circular está

centrada en el punto medio del grupo correspondiente de ángulos, y el área de la barra es proporcional a la frecuencia en ese grupo.

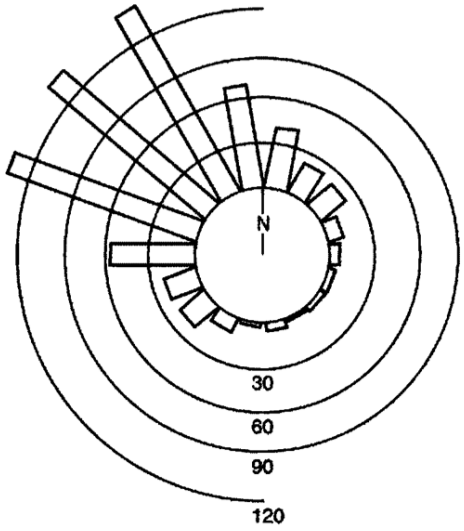


Figura 2.2: Histograma circular

Histogramas lineales

Debido a que los estadísticos han adquirido experiencia en la interpretación de histogramas en la recta, puede ser útil transformar un histograma circular en un histograma lineal. Esto se hace cortando el histograma circular en un punto elegido adecuadamente en el círculo y luego desenrollando el histograma circular a un histograma lineal en un intervalo de longitud de 360°. Si los datos tienen una moda (dirección preferida), entonces es aconsejable utilizar un corte casi enfrente de esta moda. Entonces, el centro del histograma lineal estará cerca de la moda. Un corte cerca de la moda daría la impresión errónea de que los datos son bimodales. Para conjuntos de datos que no tienen una única moda pronunciada, es útil para modificar el histograma circular mediante la repetición de un ciclo completo de los datos, para dar un histograma lineal en un intervalo de longitud 720°.

Diagrama de rosa

Una variante útil del histograma circular es el diagrama de rosa, en el que las barras del histograma circular se sustituyen por sectores. El área de cada sector es proporcional a

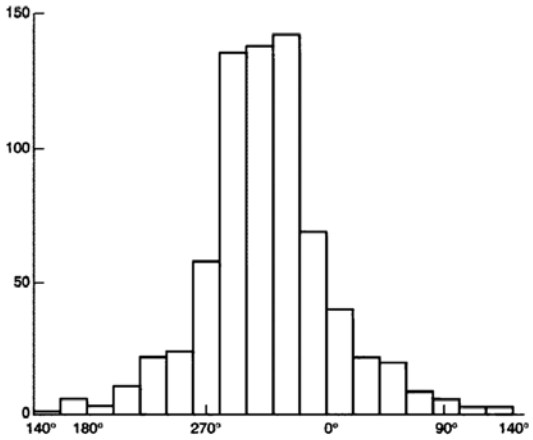


Figura 2.3: Histograma circular

la frecuencia en el grupo correspondiente. Para lograr esto, cuando los grupos son de igual longitud, el radio de cada sector debe ser proporcional a la raíz cuadrada de la frecuencia correspondiente, pero no todos los autores siguen esta convención.

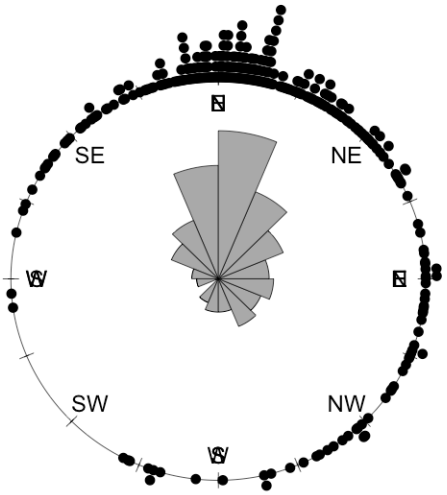


Figura 2.4: Diagrama de rosa

2.3. Estadística descriptiva para datos circulares

Los datos circulares pueden ser representados como ángulos o como puntos sobre una circunferencia en un circulo unitario. La posición direccional tiene una representación única en un sistema coordenado de dos dimensiones, es decir, cuarlquier punto P sobre el plano puede ser representado en términos de coordenadas rectangulares (X, Y) o como cordenadas polares (r, α), donde r es la distancia del punto al origen y α es el ángulo. Para el punto de origen, tenemos que $r = 0$, este no tiene dirección, es decir, α no está definida.

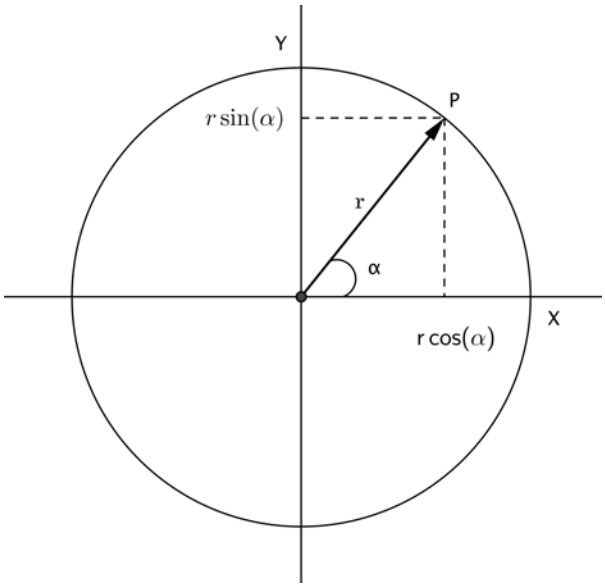


Figura 2.5: Transformada polar

La Figura 2.5 representa las coordenadas rectangulares y polares del punto P. Dadas las coordenadas rectangulares de un punto $P = (x, y)$, se tiene que:

$$x = r\cos(\alpha) \quad \text{y} \quad y = r\,\text{sen}(\alpha) \tag{2.1}$$

En el análisis direccional la dirección es de interés y no la magnitud del vector, por lo que se toma $r = 1$, por conveniencia. Así, cada dirección corresponde a un punto P sobre la circunferencia. Del mismo modo, cada punto en la circunferencia puede ser representado por un ángulo.

2.3.1. Medida de centralidad

Para definir una direccion media, se podría pensar en calcular la media aritmética de los ángulos, pero no tendría sentido, ya que si se toman dos ángulos de 20° y 340°, su promedio aritmético sería 180°, pero como se observa en la Figura 2.6, esto no tiene sentido.

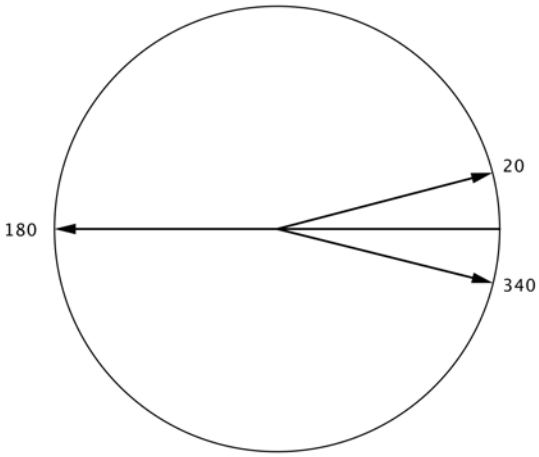


Figura 2.6: Transformada polar

Una medida adecuada para la media direccional, para el caso de distribuciones circulares unimodales, es el vector donde se concentran las direcciones. El cálculo de este vector es el siguiente:

Sea $\alpha_1, \alpha_2, \ldots, \alpha_n$ un conjunto de observaciones circulares dadas en ángulos. Se considera la transformación polar a rectangular de cada observación dada en la ecuación (2.1), es decir $(\cos(\alpha_i), \operatorname{sen}(\alpha_i))$, $i = 1, \ldots, n$. La suma de los n términos sumados componte a componente, da como resultado el siguiente vector:

$$R = \left(\sum_{i=1}^{n} \cos(\alpha_i) , \sum_{i=1}^{n} \operatorname{sen}(\alpha_i) \right) = (C, S) \tag{2.2}$$

Entonces $R = \boldsymbol{R} = \sqrt{C^2 + S^2}$ es la longitud del vector $\boldsymbol{R}$. La dirección del vector anterior,

denotado por $\bar{\alpha}_0$, se obtiene de las siguientes ecuaciones

$$\cos(\bar{\alpha}_0) = \frac{C}{R} \quad \text{y} \quad \operatorname{sen}(\bar{\alpha}_0) = \frac{S}{R} \tag{2.3}$$

Una definición explícita de $\bar{\alpha}_0$ está dada como sigue:

$$\overline{\alpha_0} = \begin{cases} \arctan(S/C) & \text{si } C > 0, S \geq 0 \\ \pi/2 & \text{si } C = 0, S > 0 \\ \arctan(S/C) + \pi & \text{si } C < 0 \\ \arctan(S/C) + 2\pi & \text{si } C \geq 0, S < 0 \\ \text{indefinida} & \text{si } C = 0, S = 0 \end{cases} \tag{2.4}$$

Regresando al ejemplo de la Figura 2.6, se tiene de (2.4) que $C = \cos(20°) + \cos(340°) = 1.879385$ y $S = \operatorname{sen}(20°) + \operatorname{sen}(340°) = 0$, por lo tanto, $\hat{\alpha}_0 = \arctan(0) = 0$, siguiendo el sentido natural de una tendencia central (Figura 2.7).

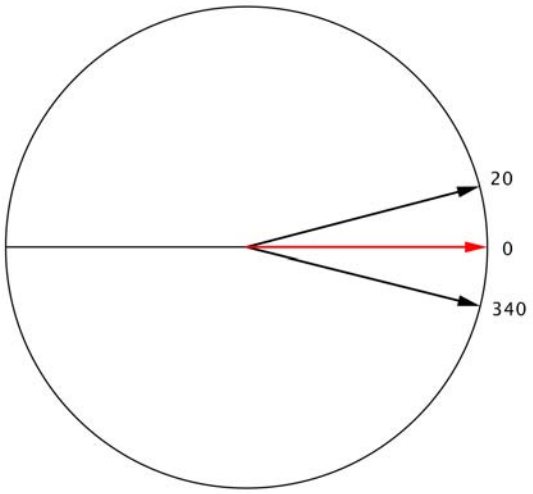


Figura 2.7: Media direccional adecuada

2.3.2. Distancia circular

En [Jammalamadaka y SenGupta, 2001] se define como distancia circular adecuada entre dos puntos a la longitud mínima de los arcos formados entre los dos puntos en la circunferencia, es decir que para cualesquiera dos ángulos α y β se tiene que

$$\begin{aligned}\rho_0(\alpha,\beta) &= min(\alpha-\beta, 2\pi-(\alpha-\beta)) \\ &= \pi - |\pi - |\alpha-\beta||\end{aligned} \tag{2.5}$$

En la Figura 2.8, la distancia entre A y B puede ser la longitud del arco ANB o la del arco ASB. Con (2.5), la distancia sería la longitud de arco ANB. Se puede ver que la distancia circular ρ_0 toma valores entre $[0, \pi]$.

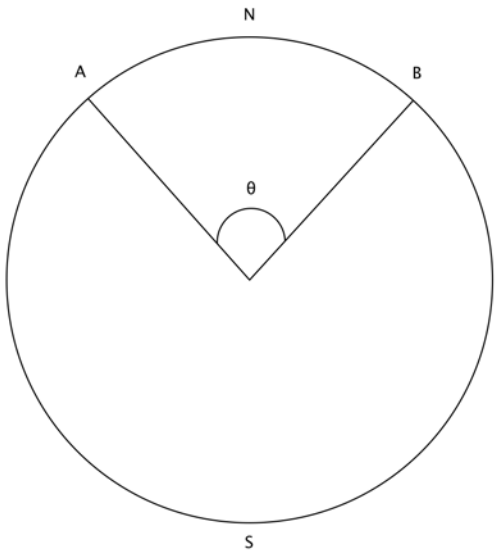


Figura 2.8: La distancia circular ρ_0 es la longitud del arco ANB

En [Jammalamadaka y SenGupta, 2001] se define una segunda distancia circular entre los ángulos α y β como

$$\rho_1(\alpha,\beta) = 1 - \cos(\alpha-\beta) \tag{2.6}$$

donde α y β representan los ángulos correspondientes a A y B respectivamente. Si θ es el ángulo entre los puntos A y B, la función de distancia ρ_1 es monótona creciente con respecto a θ, tomando el valor de 0 cuando $\theta = 0$ y crece hasta 2 si $\theta = \pi$.

En la Figura 2.9 se observa, que salvo unidades, ambas distancias tienen el mismo sentido y su diferencia radica en la mayor curvatura que representa ρ_1.

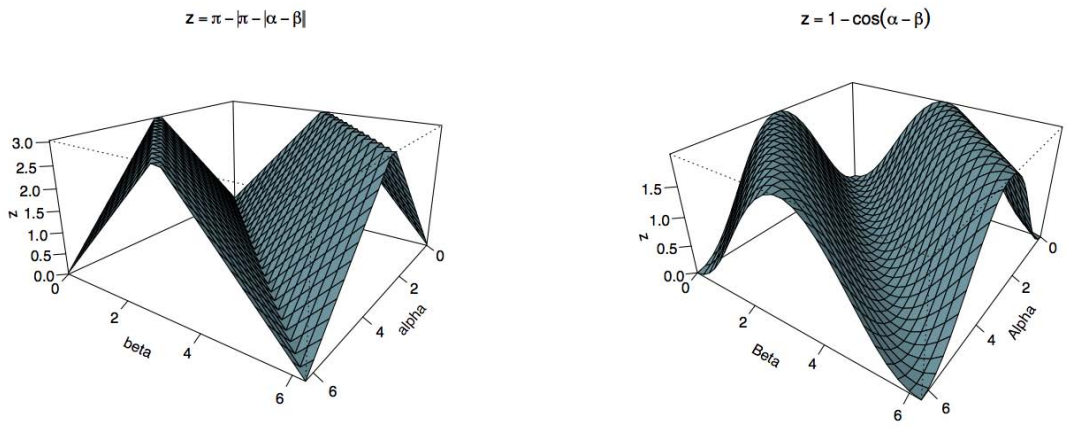


Figura 2.9: Comparación de las medidas de distancias circulares

Para poder utilizar la distancia como instrumento de decisión, es importante determinar si las distancias definidas en (2.5) y (2.6) cumplen las propiedades de medida de disimilaridad. En [Webb, 1999] se indica que una medida ρ entre a y b se dice de disimilaridad si

- $\rho(a, b) \geq 0 \quad \forall a, b$ (Positiva)
- $\rho(a, a) = 0 \quad \forall a$ (Nulidad)
- $\rho(a, b) = \rho(b, a) \quad \forall a, b$ (Simetría)

Proposición 2.3.1. *(Disimilaridad Circular del Coseno). La distancia circular definida en (2.6) es una medida de disimilaridad.*

Demostración. La positividad de la distancia se satisface ya que $\cos(a-b)$ está entre $[-1,1]$ por lo tanto $\rho_1(a,b) \geq 0$. Además, si $\cos(0°) = 1$, se tiene que $1-\cos(0°) = 0$, y para cualquier a se tiene que $\rho(a) = 1-\cos(a-a) = 0$. La simetría de la disimilaridad circular se tiene porque la función coseno es una función par. Es decir $\cos(a) = \cos(-a)$ lleva directamente a $\rho_1(a,b) = 1-\cos(a-b) = 1-\cos(b-a) = \rho_1(b,a)$.

Proposición 2.3.2. *(Disimilaridad Circular del Valor Absoluto). La distancia circular definida en (2.5) es una medida de disimilaridad.*

Demostración. Dado que el máximo valor que toma la diferencia entre dos ángulos en radianes está entre -2π y 2π se tiene que $|\pi - |a-b||$ toma valores entre $-\pi$ y π por lo tanto $\pi - |\pi - |a-b||$ tiene como rango $[0,\pi]$ con lo que se tiene $\rho_0(a,b) \geq 0$ para toda a y b. La nulidad es clara y la simetría se obtiene del valor absoluto, ya que $|a-b| = |b-a|$

2.4. Distribuciones de probabilidad circulares

En [Jammalamadaka y SenGupta, 2001] se define una distribución circular como aquella cuya probabilidad total está concentrada sobre la circunferencia de un círculo unitario. El rango de una variable aleatoria circular θ, medida en radianes, toma valores entre $[0,2\pi)$ o $[-\pi,\pi)$. De la misma forma que las distribuciones clásicas de probabilidad, las distribuciones circulares son dos tipos: Discretas y Continuas. En cualquier caso, una función de densidad de probabilidad $f(\theta)$ debe cumplir con las siguientes propiedades:

- $f(\theta) \geq 0$
- $\int_0^{2\pi} f(\theta)\, d\theta = 1$
- $f(\theta) = f(\theta + k \cdot 2\pi)$ para cualquier entero k (es decir f es periódica)

2.4.1. Algunos métodos para obtener distribuciones circulares

En las secciones anteriores se observó que una variable aleatoria circular puede ser representada en términos del ángulo θ, $(0 \leq \theta \leq 2\pi)$ o como un vector unitario bidimensional

$(X = \cos(\theta), Y = \text{sen}(\theta))'$.

Algunas distribuciones circulares pueden ser generadas a partir de distribuciones de probabilidad conocidas sobre la recta real o sobre el plano, por medio de diferentes desarrollos. Se describirán algunos de ellos:

- Por envoltura (wrapping) de una distribución lineal alrededor del círculo unitario
- Por medio de propiedades características tal como maximizar la entropía
- Transformando una variable aleatoria lineal bivariada en sus componentes direccionales, los cuales son llamadas distribuciones de desplazamiento (offset)
- Iniciando con una distribución sobre la recta real $\Re$, se aplica una proyección estereográfica que indentifica cada punto x en $\Re$ con algún punto sobre la circunferencia del círculo unitario, llamado θ. Esta correspondencia es uno a uno, excepto por el hecho de los valores $-\infty$ y ∞ son identificados con π

2.4.2. Distribución von Mises

Una variable aleatoria θ sigue una distribución von Mises o Normal circular si tiene como función de densidad:

$$f(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta - \mu)}, \quad 0 \leq \theta \leq 2\pi \tag{2.7}$$

donde $0 \leq \mu < 2\pi$ y $\kappa \geq 0$ son parámetros. El término $I_0(\kappa)$ de la constante de normalización es la función modificada de Bessel de primera clase de orden cero [Fisher, 1993] y está dada por

$$\begin{aligned} I_0(\kappa) &= \frac{1}{2\pi} \int_0^{2\pi} exp(\kappa \cos(\theta))\, d\theta \\ &= \sum_{r=0}^{\infty} \left(\frac{\kappa}{2}\right)^{2r} \left(\frac{1}{r!}\right)^2 \end{aligned} \tag{2.8}$$

La función de densidad von Mises tiene las siguientes propiedades:

- **Simetría**: Debido a la simetría de la función coseno, la densidad es simétrica alrededor de la dirección μ.

- **Moda en** μ: Dado que la función coseno tiene máximo en cero, la densidad von Mises tiene máximo en $\theta = \mu$, es decir que μ es la moda direccional cuyo valor máximo es

$$f(\mu) = \frac{e^{\kappa}}{2\pi I_0(\kappa)} \tag{2.9}$$

- **Antimoda en** $\mu \pm \pi$: Ya que $\cos(\pi) = -1$ es el valor mínimo, entonces $\theta = \mu \pm \pi$, da el valor mínimo de densidad en

$$f(\mu \pm \pi) = \frac{e^{-\kappa}}{2\pi I_0(\kappa)} \tag{2.10}$$

  Así, $\mu \pm \pi$ es la dirección antimodal.

- **Parámetro de concentración** κ: De (2.8) y (2.9), se tiene que:

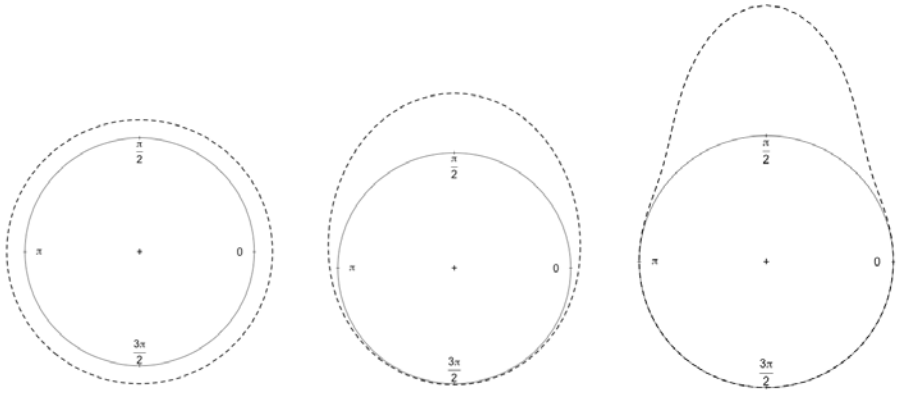
$$\frac{f(\mu)}{f(\mu \pm \pi)} = e^{2\kappa}$$

  A medida que κ aumenta, la razón de $f(\mu)$ y $f(\mu \pm \pi)$ es más grande; indicando mayor concentración alrededor de la media direccional poblacional μ. Por tal motivo, κ es conocido como el parámetro de concentración alrededor de la media direccional.

En la Figura 2.10 se muestra la función de densidad von Mises para diferentes valores de κ y para $\mu = \pi/2$.

2.5. Datos direccionales multidimensionales

Usualmente cuando se tiene un conjunto de datos es de interés estudiar la dirección y la magnitud del vector $x = \left(x_1, \ldots, x_p\right)'$, pero en algunas ocasiones se desea estudiar únicamente la dirección de x. Entonces, dichos puntos pueden ser ubicados en la circunferencia en dos dimensiones o en la superficie de la esfera en tres dimensiones. En general, las direcciones pueden ser visualizadas como puntos en la superficie de la hiperesfera. Sea θ el vector aleatorio



(a) $\kappa = 1/100$ (b) $\kappa = 2$ (c) $\kappa = 7$

Figura 2.10: Densidades von Mises ($\mu = \pi/2$) en los tres casos

direccional de dimensión p, donde $\theta'\theta = 1$. El vector unitario θ toma valores sobre la superficie de la eresfera S_{p-1} de dimensión $p-1$, que tiene radio unitario y centro en origen.

En general, es conveniente considerar la densidad de x en términos de las coordenadas polares esféricas $x = ru(\theta)$ con $\theta = \left(\theta_1, \ldots, \theta_{p-1}\right)'$ donde

$$u_{i(\theta)} = \cos\theta_i \prod_{j=0}^{i-1} \operatorname{sen}\theta_j \quad i = 1, \ldots, p$$

$$\operatorname{sen}\theta_0 = \cos\theta_p = 1 \tag{2.11}$$

y

$$0 \leq \theta_j \leq \pi \qquad 0 \leq \theta_{p-1} < 2\pi \qquad r > 0$$

El jacobiano de la transformación de (r, θ) a x de acuerdo en [Mardia y Jupp, 1999] es

$$J_p = r^{p-1} \prod_{i=1}^{p-1} \operatorname{sen}^{p-i}\theta_{i-1}$$

$$J_2 = r \tag{2.12}$$

Usualmente se utiliza esta transformación sobre la eresfera de radio uno. Es importante mencionar que para $p = 2$, es igual a (2.1).

2.5.1. Distribución von Mises-Fisher

[Jammalamadaka y SenGupta, 2001] sugiere como una generalización adecuada de la distribución von Mises (2.7) sobre esferas de dimensión $p-1$, a distribuciones cuya log-densidad sea lineal en x, es decir, cuya densidad $f(x;\mu,\kappa)$ satisface

$$\log f(x;\mu,\kappa) = \kappa\mu' x + c \tag{2.13}$$

donde c es una constante, éstas son llamadas las distribuciones von Mises-Fisher.

Un vector aleatorio $\boldsymbol{x}$ sigue una distribución de dimensión $(p-1)$ von Mises-Fisher si la función de densidad de probabilidad es

$$f(x;\mu,\kappa) = \left(\frac{\kappa}{2}\right)^{p/2-1} \frac{1}{\Gamma(p/2)\, I_{p/2-1}(\kappa)} exp\{\kappa\mu' x\} \tag{2.14}$$

donde $\kappa \geq 0,\ \mu = 1$ y I_v denota la función modificada de Bessel de primer tipo de orden v. Los parámetros μ y κ son llamados media direccional y parámetro de concentración respectivamente. En el caso particular de $p = 2$, de (2.14) se obtiene

$$f(x;\mu,\kappa) = \frac{1}{I_0(\kappa)} e^{\kappa\mu' u(x)}$$

y usando coordenadas polares

$$u(\theta) = \begin{pmatrix} u_1(\theta_1) \\ u_2(\theta_2) \end{pmatrix} = \begin{pmatrix} \cos\theta_1 \\ \operatorname{sen}\theta_1 \end{pmatrix}$$

se obtiene

$$\begin{aligned} f(\theta;\mu,\kappa) &= \frac{1}{I_0(\kappa)} e^{\kappa(\cos\mu_1\cos\theta_1 + \operatorname{sen}\mu_1 \operatorname{sen}\theta_1)} \\ &= \frac{1}{I_0(\kappa)} e^{\kappa\cos(\theta_1-\mu_1)} \end{aligned}$$

la cual tiene el mismo comportamiento de la distribución circular von-Mises, a excepción del parámetro de normalización $1/(2\pi)$. Para el caso de $p = 3$, la distribución von Mises-Fisher es llamada la distribución Fisher, y está dada por la siguiente expresión:

$$f(x; \mu, \kappa) = \frac{\kappa}{2 \operatorname{senh} \kappa} e^{\mu' x} \tag{2.15}$$

escribiendo a x y a μ en coordenadas polares esféricas se tiene:

$$x = (\cos(\theta), \operatorname{sen}(\theta)\cos(\phi), \operatorname{sen}(\theta)\operatorname{sen}(\phi))', \quad \mu = (\cos(\alpha), \operatorname{sen}(\alpha)\cos(\beta), \operatorname{sen}(\alpha)\operatorname{sen}(\beta))'$$

entonces

$$f(x; \mu, \kappa) = \frac{\kappa}{4\pi \operatorname{senh} \kappa} e^{\kappa[\cos\theta\cos\alpha + \operatorname{sen}\theta\operatorname{sen}\alpha\cos(\phi-\beta)]} \operatorname{sen}\theta \tag{2.16}$$

2.6. Métodos de clasificación

Se tienen J clases de objetos (estados en la naturaleza) de intéres, los cuales utilizan el subíndice j con $j = 1, \ldots, J$, para cada estado respectivamente. La información que poseemos sobre los objetos es resumida en un número finito p, de medidas de valor real denominadas características. Todas juntas forman un vector de características $x \in \Re^p$. En los modelos de incertidumbre que se analizarán a continuación se supone que hay probabilidades a priori $\pi_1, \pi_2, \ldots, \pi_j$ para cada una de las clases. Para modelar la relación entre el vector de características (incluyendo el ruido natural en los procesos de medición y ocasionales de la naturaleza misma), se supone que un objeto de la clase $y \in \{1, 2, \ldots, J\}$ es una realización del vector de variables aleatorias con función de distribución condicional de la clase $F_y(x)$. Vectores de caracerísticas aleatorias X (los observados en el proceso) son generados de acuerdo al siguiente proceso de J estados: la clase aleatoria $Y \in \{1, 2, \ldots, J\}$ es seleccionada de acuerdo a las probabilidades a priori $\{\pi_1, \ldots, \pi_j\}$; el vector de características observadas X, es una selección de la distribución condicional de la clase F_Y. Dada una realización del vector de características X, denotado por x, el problema al que se enfrenta el investigador es decidir a cuál clase pertenece el elemento x. Así, una regla de decisión, es una función $g : \Re^p \to \{1, 2, \ldots, J\}$, donde $g(x)$

denota la clase asignada al vector de características x por el clasificador g, el rendimiento de g puede ser medido por la probabilidad de error, dada por

$$L(g) = P\{g(x) \neq Y\} \tag{2.17}$$

Si las probabilidades a priori y las distribuciones condicionales son conocidas, entonces el problema de clasificar se convierte en un problema de regla óptima de decisión, es decir, minimizar la probabilidad de error. La regla que minimiza esta probabilidad de error es denominada regla de decisión de Bayes, denotada por g^*. Esta regla de decisión usa las distribuciones conocidas y la observación $X = x$ para calcular las probabilidades a posteriori

$$\eta_J(x) = P[Y = J \mid X = x]$$

de las J clases, y se selecciona la clase con mayor probabilidad a posteriori (equivalentemente menor riesgo), es decir

$$g^*(x) = \underset{j \in \{1,,2,\ldots,J\}}{argmax} \quad \eta_j(x) \tag{2.18}$$

La tasa de error de Bayes, está dada por

$$L* = L(g^*) = E[min\{\eta_1(X), \eta_2(X), \ldots, \eta_J(X)\}]$$

Desde luego, en la mayoría de las aplicaciones se desconoce por lo menos parcialmente las distribuciones condicionales, o no se desea realizar supuestos sobre ella. En este caso, generalmente se asume que se tienen realizaciones previas, denominadas muestra de entrenamiento del vector de características X para las dieferenctes clases. Es decir, se tiene

$$D_n = (X_1, Y_1), \ldots, (X_n, Y_n) \tag{2.19}$$

donde $Y_k \in \{1, 2, \ldots, J\}$ corresponde a la clase de los objetos, la cual es asumida idéntica e independientemente distribuida con $\{\Pi_1, \ldots, \Pi_J\}$ para cada una de las clases, y X_k es un vector de características provieniente de la distribución condicional $F_{Y_k}(x)$. Así, la pareja (X_k, Y_k) es

asumida idéntica e independientemente distribuida de acuerdo a las distribuciones (desconocidas) P_y y $F_y(x)$, las cuales caracterizan el problema. Intuitivamente, los datos D_n brindan información parcial de las distribuciones desconocidas, y es de interés usar dichos datos para encontrar buenos clasificadores. Más formalmente, una regla de clasificación o clasificador es una función $g_n(x) = g_n(x, D_n)$, construida a partir del conjunto o de la muestra de entrenamiento D_n, la cual asigna una etiqueta $(1, 2, \ldots, J)$ a cada punto $X \in \Re^p$. Para un conjunto de datos de entrenamiento fijo D_n, la probabilidad condicional del error de una regla de clasificación es

$$L(g_n) = P[g_n(X) \neq Y \mid D_n] \tag{2.20}$$

donde la pareja (X, Y) es independiente de D_n. Es importante notar que este error depende de la muestra de entrenamiento, por lo tanto, la probabilidad de error $L(g_n)$ es una variable aleatoria que depende de D_n, luego se define la probabilidad esperada de error $L(\bar{g}_n) = E[L(g_n)] = P[g_n(X) \neq Y]$ tomando este valor esperado con respecto a la muestra de entrenamiento aleatoria. Teóricamente, cuando se desea evaluar el rendimiento se usa como cota la tasa de error de Bayes L^* la cual es óptima cuando se considera completamente conocidas las distribuciones.

2.6.1. Discriminante direccional y clasificación

Una medida de distancia: A continuación se introduce una regla de discriminación general basada en una distancia circular que puede ser utilizada para cualquier par de distribuciones circulares paramétricas. La idea básica es encontrar qué tan lejos está la nueva observación, de cualquiera de las muestras dadas y asignarla a la población a la que está más cercana.

Para cualesquiera dos puntos $\left(\theta_i, \theta_j\right)$ en el circulo unitario, se define la distancia circular

$$d_{ij} = 1 - \cos\left(\theta_i - \theta_j\right) \tag{2.21}$$

que es no negativa, simétrica en sus indices y es invariante bajo rotación. La distancia promedio

d_i de una nueva observación θ del grupo i, está dada por

$$d_i(\theta) = 1 - \frac{1}{n_j}\sum_j \cos\left(\theta_{ij} - \theta\right) \tag{2.22}$$

Note que esto es lo mismo que la dispersión circular de la muestra i sobre la dirección dada θ. Usando las notaciones estándar.

$$\overline{C}_i = \frac{i}{n_i}\sum_j \cos\left(\theta_{ij}\right), \quad \overline{S}_i = \frac{1}{n_i}\sum_j \operatorname{sen}\left(\theta_{ij}\right), \quad \overline{R}_i = \sqrt{\overline{C}_i^2 + \overline{S}_i^2}, \quad \overline{\theta}_i = \arctan\frac{\overline{S}_i}{\overline{C}_i}$$

La cual puede ser reescrita como sigue

$$d_i(\theta) = 1 - \overline{R}_i \cos\left(\overline{\theta}_i - \theta\right) = 1 - \frac{V_i}{n_i}$$

Donde V_i es la longitud de la proyección de la i-ésima muestra resultante para la dirección θ. La regla es para clasificar a θ como perteneciente a la población 1 si $d_1(\theta) < d_2(\theta)$, es decir, si

$$\frac{V_1}{n_1} > \frac{V_2}{n_2}$$

El caso donde los dos tamaños de muestra son iguales y las concentraciones tambíen son iguales, corresponde a clasificar a θ como perteneciente al grupo cuya dirección media muestral está más cerca.

La regla basada en la cuerda: Sea θ la nueva observación para clasificar, se denota por d_{0i} la distancia de θ a $\hat{\theta}_i$, la media circular para el grupo i, $i = 1, 2$. Se define $D(\theta) = d_{01}(\theta) - d_{02}(\theta)$. Con el supuesto de que las probabilidades a priori para las dos poblaciones son iguales y sea c una constante real, entonces, la regla de clasificación está dada como sigue

$$\begin{aligned} \text{Si} \quad D(\theta) < c \qquad & \text{asignar a } \theta \text{ a la población 1 y} \\ & \text{asignar a } \theta \text{ a la población 2 en otro caso} \end{aligned} \tag{2.23}$$

Ahora

$$D(\theta) = \left(\cos\left(\bar{\theta}_2\right) - \cos\left(\bar{\theta}_1\right)\right)\cos\theta + \left(\operatorname{sen}\left(\bar{\theta}_2\right) - \operatorname{sen}\left(\bar{\theta}_1\right)\right)\operatorname{sen}\theta \tag{2.24}$$

Sea

$$\tan(\theta_0) = \frac{\operatorname{sen}\left(\bar{\theta}_2\right) - \operatorname{sen}\left(\bar{\theta}_1\right)}{\cos\left(\bar{\theta}_2\right) - \cos\left(\bar{\theta}_1\right)} \tag{2.25}$$

Se observa que $P\left(\bar{\theta}_1 = \bar{\theta}_2\right) = 0$, suponiendo que se trata de distribuciones continuas subyacentes y, por tanto, θ_0 está bien definida (con probabilidad 1).

Entonces (2.24) puede escribirse como

$$D(\theta) = \sqrt{2 - 2\cos\left(\bar{\theta}_1 - \bar{\theta}_2\right)}\cos(\theta - \theta_0) \tag{2.26}$$

Note que para (2.25), habrá dos soluciones para θ_0.

La regla de clasificación en (2.23) se reduce para una equivalente, pero simple forma como

$$\text{Si} \quad \cos(\theta - \theta_0) > K \quad \text{asignar } \theta \text{ a la población 1}$$

$$\text{y asignar a } \theta \text{ a la población 2 en otro caso} \tag{2.27}$$

Donde K es una constante apropiada.

Observaciones

- La dirección θ_0 es ortogonal a la bisectriz de $\bar{\theta}_1$ y $\bar{\theta}_2$
- Como ocurre a menudo por la razón de la simplicidad de la regla de clasificación [Ver, e.g. Rao (1973, p.575, Eq. (8e.1.8))], se puede tomar $K = 0$. La regla dada por la ecuación (2.27) particiona el círculo en dos sectores de 180° de amplitud. En este caso, explícitamente, los sectores pueden ser especificados como un semicírculo teniendo a θ_0 como su punto medio, y el arco complementario. Note que si las direcciones medias muestrales son

iguales, las varianzas circulares distintas no tienen efecto en la regla. En este caso θ_0 es la propia dirección media. Además, cuando las direcciones medias muestrales no son iguales, las varianzas circulares afectan a θ_0. La regla puede ser modificada para cubrir el caso de probabilidades a priori diferentes también.

von Mises-Fisher: De forma análoga al desarrollo del análisis discriminante en espacios euclidianos, en el análisis discriminante direccional se supone la distribución von Mises-Fisher de dimensión $p-1$. Es decir, se supone que el vector direccional x sigue la distribución von Mises-Fisher. Se tiene entonces que, si las poblaciones $\Pi_1, \ldots, \Pi_J$ tienen probabilidades a priori $(\pi_1, \ldots, \pi_J) = \pi$, entonces la regla de discriminación de Bayes (con respecto a π) asigna una observación x a la población en la cual $\pi_j L_j(x)$ es máxima. Donde

$$L_j(x) = f_j(x) = \left(\frac{\kappa_j}{2}\right)^{p/2-1} \frac{1}{\Gamma(p/2)\, I_{p/2-1}(\kappa_j)} exp\{\kappa_j \mu_j' x\} \tag{2.28}$$

y κ_j y μ_j son parámetros para cada j con $j = 1, \ldots, J$

Proposición 2.6.1. *(Función discriminante direccional). Bajo el supuesto de la distribución von Mises-Fisher en cada población, la regla de clasificación η para una observación x está dada por*

$$\eta(x) = \underset{j}{\operatorname{argmax}} \left\{ \pi_j \frac{\kappa_j^{p/2-1}}{I_{p/2-1}(\kappa_j)} exp\left\{ \kappa_j \sum_{i=1}^{p} \left(\cos(\mu_{ij}) \cos(\theta_i) \prod_{m=0}^{i-1} \operatorname{sen}(\mu_{mj}) \operatorname{sen}(\theta_m) \right) \right\} \right\} \tag{2.29}$$

Demostración La regla es clasificar a la clase en la que se maximice (2.28):

$$\eta(x) = \underset{j}{\operatorname{argmax}} \{\pi_j f_j(x)\}$$

$$= \underset{j}{\operatorname{argmax}} \left\{ \pi_j \left(\frac{\kappa_j}{2}\right)^{p/2-1} \frac{1}{\Gamma(p/2)\, I_{p/2-1}(\kappa_j)} exp\{\kappa_j \mu_j^T x\} \right\}$$

$$= \underset{j}{\operatorname{argmax}} \left\{ \pi_j \frac{\kappa_j^{p/2-1}}{I_{p/2-1}(\kappa_j)} exp\{\kappa_j \mu_j^T x\} \right\}$$

$$= \underset{j}{\text{argmax}} \left\{ \pi_j \frac{\kappa_j^{p/2-1}}{I_{p/2-1}\left(\kappa_j\right)} exp\{\kappa_j u\left(\mu_j\right)^T u\left(x\right)\} \right\}$$

$$= \underset{j}{\text{argmax}} \left\{ \pi_j \frac{\kappa_j^{p/2-1}}{I_{p/2-1}\left(\kappa_j\right)} exp\left\{ \kappa_j \sum_{i=1}^{p} \left(\cos\left(\mu_{ij}\right) \cos\left(\theta_i\right) \prod_{m=0}^{i-1} \text{sen}\left(\mu_{mj}\right) \text{sen}\left(\theta_m\right) \right) \right\} \right\}$$

donde μ_{mj} y θ_m denotan la m-ésima coordenada de la transformada polar para la media direccional de la clase j y para el vector $\boldsymbol{x}$ respectivamente.

Suponiendo inicialmente que las probabilidades a priori π_j son iguales y el parámetro de concentración κ_j en cada clase (homocedasticidad), se tiene:

$$\eta\left(x\right) = \underset{j}{\text{argmax}} \left\{ \sum_{i=1}^{p} \left(\cos\left(\mu_{ij}\right) \cos\left(\theta_i\right) \right) \prod_{m=0}^{i-1} \text{sen}\left(\mu_{mj}\right) \text{sen}\left(\theta_m\right) \right\} \tag{2.30}$$

denomidada función discriminante direccional en el caso en el cual los parámetros de dispersión son iguales.

Proposición 2.6.2. *(Discriminante circular en el cual κ_j y π_j son iguales). La regla de discriminación suponiendo la distribución von Mises-Fisher circular con parámetro de concentración κ y probabilidades a priori iguales para cada clase es:*

$$\eta\left(x\right) = \underset{j}{\text{argmax}} \left(\cos\left(\theta - \mu_j\right) \right) \tag{2.31}$$

Demostración. Para el caso circular $p = 2$, utilizando el hecho de que en la transformación polar $u\left(\alpha\right)$, $\text{sen}\left(\alpha_0\right) = \cos\left(\alpha_p\right) = 1$ se tiene:

$$\eta\left(x\right) = \underset{j}{\text{argmax}} \left[exp\left\{ \left(\cos\left(\mu_j\right) \cos\left(\theta\right) + \text{sen}\left(\mu_j\right) \text{sen}\left(\theta\right) \right) \right\} \right]$$

$$= \underset{j}{\text{argmax}} \left[exp\left\{ \left(\cos\left(\theta - \mu_j\right) \right) \right\} \right]$$

$$= \underset{j}{\text{argmax}} \left(\cos\left(\theta - \mu_j\right) \right)$$

Esta función discriminante alcanza su valor máximo para la clase j donde θ esté más cercano a μ_j circularmente. Una gráfica de esta clasificación se presenta en la Figura 2.11. Es importante

mencionar que la función discriminante circular (2.31) es similar a la distancia circular (2.5), es decir, mide la menor longitud de arco entre θ y cada μ_j, pero en (2.31) alcanza su valor máximo para la clase donde $\theta = \mu_j$.

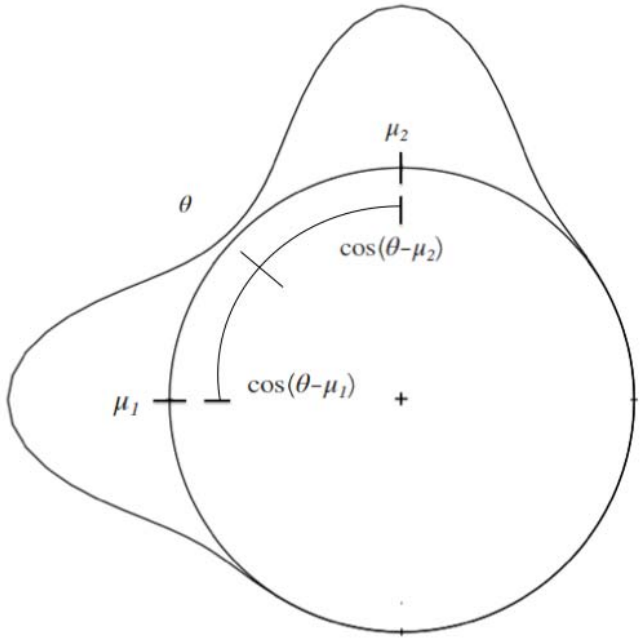


Figura 2.11: Clasificación por discriminante circular

Sin el supuesto de igualdad en el parámetro de concentración se obtiene el siguiente resultado.

Proposición 2.6.3. *(Discriminante Circular). Suponiendo una distribución circular von Mises-Fisher, la función de discriminación es*

$$\eta(x) = \underset{j}{\operatorname{argmax}} \left\{ \ln\left(\Pi_j\right) + \ln\left(\frac{1}{I_0\left(\kappa_j\right)}\right) + \kappa_j \cos\left(\theta - \mu_j\right) \right\} \tag{2.32}$$

Demostración. Para el caso de $p = 2$ sin el supuesto de igualdad del parámetro de dispersión

se tiene

$$\eta(x) = \underset{j}{\text{argmax}} \left\{ \frac{\Pi_j}{I_0(\kappa_j)} exp\left\{ \kappa_j \cos(\mu_j) \cos(\theta) + \text{sen}(\mu_j) \,\text{sen}(\theta) \right\} \right\}$$

$$= \underset{j}{\text{argmax}} \left\{ \frac{\Pi_j}{I_0(\kappa_j)} exp\left\{ \kappa_j \cos(\theta - \mu_j) \right\} \right\}$$

$$= \underset{j}{\text{argmax}} \left\{ \ln(\Pi_j) + \ln\left(\frac{1}{I_0(\kappa_j)} \right) + \kappa_j \cos(\theta - \mu_j) \right\}$$

Este resultado es equivalente al encontrado por [Morris y Laycock, 1974].

La función discriminante tiene una figura de tipo cardioide para cada clase j, centrada en μ_j con aplanamiento para valores pequeños de κ_j y tiene un efecto de picudez para valores grandes de κ_j. La Figura 2.12 muestra las funciones discriminantes en el caso de dos clases, para diferentes valores de κ_j. En la Figura 2.12, el valor de la función de discriminación para cada θ está determinado por la distancia del centro del círculo al punto de la función hacia la dirección θ. Los valores negativos de la función discriminante, se ubican en sentido opuesto a θ.

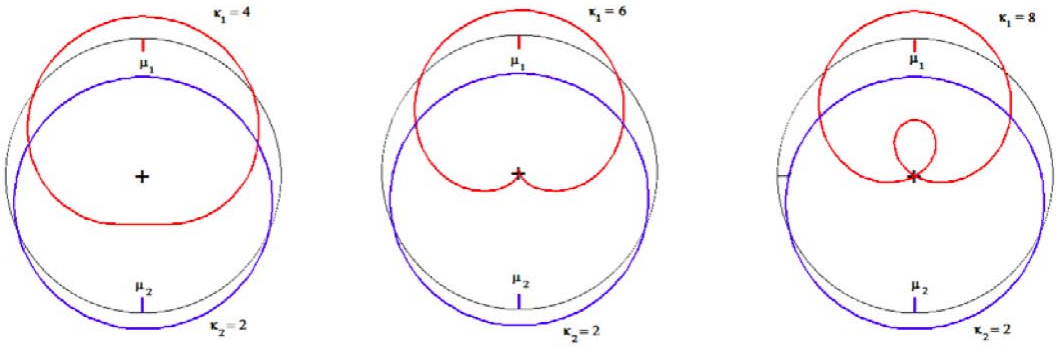


Figura 2.12: Funciones discriminantes para ($\kappa_1 = 4, 6, 8, \mu_1 = \pi/2$) y ($\kappa_2 = 2, \mu_2 = 3\pi/2$)

2.6.2. Algoritmo de clasificación por discriminante direccional

Suponga que se tiene un conjunto de datos $\boldsymbol{X}$ de dimensión nxp y un vector de clases $\boldsymbol{Y}$ de dimensión $nx1$, donde $Y[i,1] = j$ indica que la fila i-ésima de $\boldsymbol{X}$ pertenece a la clase j, con $j = 1, \ldots, J$. Suponga que se tiene $n_1, n_2, \ldots, n_J$ registros de la clase $1, 2, \ldots, J$ respectivamente. El algoritmo de clasificación por discriminante direccional, para un nuevo registro x_{nuevo} de dimensión $1xp$ se describe a continuación:

Conversión a datos direccionales: Suponga que la matriz $\boldsymbol{X}$ está ordenada de acuerdo a la clase que pertenece. Sea $Z = [x_{nuevo}|\boldsymbol{X}]$ con $n+1$ filas y p columnas. Se calcula $\boldsymbol{R} = \rho\left(Z^T\right)$ la matriz de correlación entre individuos y luego la función inversa $\Theta = \arccos(\boldsymbol{R})$. Es importante mencionar que la primera fila Θ_1 contiene el ángulo entre x_{nuevo} y los vectores de la matriz $\boldsymbol{X}$. Específicamente, las columnas $2, 3, \ldots, 1 + n_1$ contienen las correlaciones de x_{nuevo} con los registros de la clase 1. Asimismo entre $1 + n_1$ y $1 + n_1 + n_2$ la correlación con la clase 2, y así hasta las columnas $n - n_J$ y n donde se encuentra la correlación con la clase J. Es decir: $\Theta_1 = (0, \underbrace{\theta_2, \ldots, \theta_{n_1+1}}_{\text{Clase 1}}, \underbrace{\theta_{n_1+2}, \ldots, \theta_{n_1+n_2+1}}_{\text{Clase 2}}, \ldots, \underbrace{\theta_{n-n_J}, \ldots, \theta_n}_{\text{Clase J}})$ y se denota $\Theta_1 = (\Theta_{11}, \Theta_{12}, \ldots, \Theta_{1J})$.

Aplicación de clasificadores direccionales: Se calcula el discriminante direccional para cada clase, es decir:

$$\eta_j\left(\Theta_{1j}\right) = \pi_j \frac{\kappa_j^{n_j/2-1}}{I_{n_j/2-1}\left(\kappa_j\right)} exp\left\{\kappa_j \sum_{i=1}^{n_j}\left(\cos\left(\mu_{ij}\right)\cos\left(\theta_{ij}\right)\prod_{m=0}^{i-1}\text{sen}\left(\mu_{mj}\right)\text{sen}\left(\theta_{mj}\right)\right)\right\}$$

donde θ_i es la i-ésima coordenada de Θ_j, y κ_j y el vector μ_j son calculados a partir de la matriz Θ_{ij} la cual contiene los datos circulares para las i que pertenecen a la clase j.

Selección de la clase: El vector x_{nuevo} es clasificado a la clase donde la función de discriminación sea mayor, es decir:

$$\eta_j\left(\Theta_{1j}\right) = \underset{j}{\text{argmax}}\left\{\pi_j \frac{\kappa_j^{n_j/2-1}}{I_{n_j/2-1}\left(\kappa_j\right)} exp\left\{\kappa_j \sum_{i=1}^{n_j}\left(\cos\left(\mu_{ij}\right)\cos\left(\theta_{ij}\right)\prod_{m=0}^{i-1}\text{sen}\left(\mu_{mj}\right)\text{sen}\left(\theta_{mj}\right)\right)\right\}\right\}$$

CAPÍTULO 3

APLICACIONES

Son pocos los documentos que abordan el análisis discriminante para datos circulares o esféricos. Se utilizará la distribución von Mises-Fisher para realizar este análisis (Morris y Laycock, 1974) que es similar a la distribución normal multivariada (o univariada) en $\Re^p$.

3.1. Análisis discriminante de datos esféricos y circulares utilizando la distribución von Mises-Fisher

Para cada grupo se calcula el vector de medias y el parámetro de concentración, y después se calcula la densidad de una observación para cada grupo. El grupo para el cual la densidad tiene el valor más grande, es el grupo al que se asigna la observación. Para evitar cualquier colapso computacional derivado de la función de Bessel se aplicará el logaritmo de la densidad y será la determinante para la discriminación

$$\delta_i = \frac{p}{2}\log\kappa_i + \kappa_i z^T \mu_i - \frac{1}{2}\log(2\pi) - \log\left[I_{p/2-1}(\kappa_i)\right] \tag{3.1}$$

para $i = 1, \ldots, g$, donde g es el número de grupos, k_i el parámetro de concentración, μ_i es la dirección media del i-esimo grupo y z es una observación en S^{p-1}. Primero se tiene que

observar que tan eficiente es el método, para esto se ha codificado la siguiente función en el programa estadístico R para estimar la tasa de clasificación

```
vmf.da(x, ina, fraction = 0.2, R = 1000, seed = FALSE)
```

donde

- x: es una matriz del conjunto de datos en coordenadas euclidianas, es decir, vectores unitarios
- ina: es una variable que indica la pertenencia de los datos a un grupo
- fraction: indica el porcentaje de la muestra que se utilizará como muestra de ensayo
- R: es el número de repeticiones
- Seed: si seed es TRUE, el resultado siempre será el mismo

Se realiza un procedimiento repetido de validación cruzada para estimar la tasa de clasificación correcta. Ésta consiste básicamente en recoger el doble de datos de los estrictamente necesarios para la estimación de los parámetros y utilizar inicialmente la mitad para, posteriormente, comprobar si los resultados son los mismos con la segunda mitad. Dicho en otras palabras, la validación cruzada consiste en dividir la muestra al azar en dos partes, una para el análisis y otra para la validación. De este modo el investigador o la investigadora evita tener que volver a recoger nuevos datos, con los inconvenientes logísticos y prácticos que ello supone.

Resultados

- Percent: es el porcentaje estimado de clasificación (discriminación) correcta y dos desviaciones estándar estimadas. Uno de los porcentajes es la desviación estándar de las tasas y el otro está suponiendo una distribución binomial
- Ci: arroja tres tipos de intervalos de confianza, el estándar, otro basado en la distribución binomial y el tercero es el empírico, que calcula el 2.5% superior e inferior de las tasas

Ejemplo 3.1.1. Se trabajará con una matriz de datos x que contiene dos poblaciones que se distribuyen von Mises-Fisher, para obtenerlo se hace una simulación con la siguiente función

```
rvmf(n, mu, k)
```

donde

- n: es el tamaño de la muestra
- mu: es la dirección media
- k: es el parámetro de concentración

En la Tabla 3.1 se muestra el vector *ina* y la matriz x que contiene a los grupos $V1$ y $V2$ que servirán para el ejercicio de esta sección

```
library(Directional)
x <- rvmf(50, rnorm(2), 5)
ina <- rep(1:2, each = 25)
y<-vmf.da(x, ina, fraction = 0.15, R = 8, seed = FALSE)

## $percent$
##   percent       sd1       sd2
## 0.7812500 0.1456206 0.1461585
## $ci$
##                2.5%     97.5%
## standard  0.4958336 1.000000
## binomial  0.4947794 1.000000
## empirical 0.6250000 0.978125
## $runtime$
##    user  system elapsed
##       0       0       0
```

ina	V1	V2	ina	V1	V2
1	-0.19397239	-0.9810070	2	0.88546724	0.4647018
1	-0.30856119	0.9512045	2	0.40671528	0.9135550
1	-0.35894265	-0.9333596	2	-0.05101978	0.9986976
1	0.72992581	0.6835264	2	-0.81789468	-0.5753680
1	0.64243103	0.7663435	2	-0.95706853	-0.2898617
1	-0.76373858	-0.6455257	2	-0.25319604	0.9674150
1	0.42283982	0.9062044	2	-0.11099364	0.9938211
1	0.97004581	0.2429221	2	0.46820880	0.8836179
1	-0.07330177	0.9973098	2	0.05985349	0.9982072
1	-0.05074022	-0.9987119	2	-0.24324378	-0.9699652
1	0.82283441	0.5682812	2	0.05071633	-0.9987131
1	-0.89815551	-0.4396779	2	0.87619661	-0.4819538
1	0.24927960	0.9684316	2	-0.07609672	-0.9971004
1	0.16539251	0.9862278	2	-0.26093114	-0.9653574
1	-0.00701150	-0.9999754	2	-0.81022230	-0.5861227
1	0.39064498	0.9205414	2	-0.15995833	-0.9871238
1	-0.91013208	-0.4143182	2	-0.33676357	-0.9415892
1	-0.55085722	-0.8345995	2	-0.58343615	-0.8121590
1	-0.43190345	-0.9019198	2	-0.49760415	-0.8674042
1	0.56052534	0.8281373	2	0.17469833	-0.9846220
1	0.79202648	0.6104867	2	-0.80517279	-0.5930403
1	0.20893073	0.9779304	2	-0.41023125	-0.9119815
1	0.47576078	0.8795747	2	0.10111711	-0.9948745
1	0.58059920	0.8141895	2	-0.04888489	-0.9988044
1	0.39925467	0.9168401	2	0.69846620	0.7156430

Tabla 3.1: Vector ina y matriz x (contiene dos grupos de datos V1 y V2)

Para el vector x se obtuvo un porcentaje de clasificación correcta de 78.12 % por lo que se tienen los elementos para decir que el método utilizado es eficiente.

3.2. Predicción de nuevas observaciones utilizando el análisis discriminante basado en la distribución von Mises-Fisher

Continuado con el ejemplo anterior, ahora se desea predecir la pertenencia de un nuevo conjunto de datos a algún grupo de la matriz x, para esto se utiliza la siguiente función

codificada en el programa estadístico R

```
vmfda.pred(xnew, x, ina)
```

donde

- xnew: es la nueva observación u observaciones (vector(es) unitario(s)) cuyo grupo se va a predecir
- x: una matriz de datos con vectores unitarios, es decir, los datos direccionales
- ina: es el vector que indica la pertenencia a algún grupo de los datos

Ejemplo 3.2.1. Se va a trabajar con una matriz *y* de nuevas observaciones que se les asignará un grupo de pertenencia *id*, con lo anterior se va a predecir utilizando la función de este apartado y la matriz de datos *x* y el vector *ina* del ejemplo anterior

```
m1 <- rnorm(2)
m2 <- rnorm(2)
y <- rbind(rvmf(2, m1, 5), rvmf(2, m2, 5))
y

##             [,1]       [,2]
## [1,]  0.3838680 -0.9233880
## [2,]  0.3408036 -0.9401345
## [3,]  0.9166191 -0.3997618
## [4,]  0.8348666  0.5504523

id <- rep(1:2, each =2)
id

## [1] 1 1 2 2

g <- vmfda.pred(y, x, ina)
table(id, g)
```

```
##    g
## id  1 2
##   1 2 0
##   2 0 2
```

Con la salida de $table(id, g)$ se observa que las nuevas observaciones se clasificaron correctamente según el grupo de pertenencia de la matriz de datos x, basándose en los resultados de la diagonal del cruce de id con la función g.

Se realizaron múltiples simulaciones y los resultados que arrojaban los algoritmos fueron similares a los mostrados en los dos ejemplos anteriores, debido a esto se tomó la decisión de solo presentar un ejemplo.

CAPÍTULO 4

CONCLUSIONES

La literatura sobre el análisis discriminante para datos circulares no es muy abundante, se revisaron principalmente tres publicaciones: Figueiredo A. (2009), Moffit S. D. (1976) y Morris y Laycock (1974). Este último artículo es al que hacen referencia casi todas las publicaciones posteriores. Después de presentar la teoría en los dos primeros capítulos, en el tercer capítulo se muestra una aplicación de la función discriminante circular, en donde se realiza un proceso de validación cruzada para estimar la tasa de clasificación correcta de dos grupos de datos, que resultó en el 78 %. Con los ejemplos prácticos de este capítulo se pueden hacer los siguientes comentarios:

- Para la función discriminante se realizaron ocho repeticiones del proceso, ya que se observó que un número excesivo de repeticiones suele ser muy tardado el proceso y puede presentarse algún problema computacional. La tasa de clasificación correcta depende más de la elección de los datos.

- Se trabajó con un número de 100 observaciones (50 para cada población), esta elección es debido a que se observó que la discriminación de un número excesivo de datos que siguen una distribución von Mises-Fisher arroja una tasa de clasificación correcta en un

intervalo de 53 % a 65 % para las muestras de entrenamiento y esto se puede justificar debido a la complejidad de clasificar una gran cantidad de datos que se traslapen.

- De acuerdo al desarrollo de este trabajo, es importante considerar los parámetros de la distribución von Mises-Fisher. Para el caso del parámetro de concentración κ, como se puede observar en las gráficas 2.10 y 2.12, si disminuye o aumenta su valor los datos se dispersan o concentran respectivamente alrededor de la media o las medias.

- Para predecir la pertenencia a un grupo de nuevas observaciones, se observó con el ejemplo que el algoritmo arrojó los resultados deseados, debido a que se seleccionaron los datos simulados con los mismos parámetros de la distribución de la matriz de datos original.

APÉNDICE A

RESULTADOS

A.1. Matrices

Definición A.1.1. *Un vector de n componentes se define como un conjunto ordenado de n números escritos de la siguiente manera:*

$$(a_1, a_2, \ldots, a_n)$$

Definición A.1.2. *Un vector columna de n componentes es un conjunto ordenado de n números escritos de la siguiente manera:*

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

Definición A.1.3. *Una matriz A de $m \times n$ es un arreglo rectangular de mn números dispuestos en m renglones y n columnas*

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

Definición A.1.4. *El rango de una matriz de $A_{m\times n}$, se define como el máximo número de renglones linealmente independientes; se denota por $ran(A)$*

Definición A.1.5. *se tiene que,*

1. *La matriz $A_{n\times n}$ es definida positiva si $X^TAX > 0$ para $X \neq 0$, y se denota $A > 0$.*
2. *La matriz $A_{n\times n}$ es semidefinida positiva si $X^TAX \geq 0$ para $X \neq 0$, y se denota $A \geq 0$.*

Definición A.1.6. *Sea $A = (a_{ij})$ una matriz de $m \times n$. Entonces la matriz transpuesta de A, que se escribe A^T, es la matriz de $n \times m$ que se obtiene al intercambiar los renglones por las columnas de A. De manera breve, se puede escribir $A^T = (a_{ji})$. En otras palabras*

$$Si \quad A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \quad entonces \quad A^T = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{nm} \end{pmatrix}$$

Definición A.1.7. *Una matriz A es simétrica si $A^T = A$.*

Teorema A.1.1. *La transpuesta de una matriz satisface las siguientes propiedades:*

- $(A^T)^T = A$.
- $(A + B)^T = A^T + B^T$.
- $(AB)^T = B^TA^T$.

Definición A.1.8. *La traza de una matriz se denota por $tr(A)$ y se define como*

$$tr(A) = \sum_{i=1}^{n} a_{ii}$$

Definición A.1.9. *El determinante de una matriz cumple:*

- *Si A es una matriz cuadrada, el determinante se denota por $|A|$.*
- *Una matriz cuadrada es no singular si $|A| \neq 0$; de otra forma se dice que A es una matriz singular.*
- *Si A tiene un renglón de ceros, entonces $|A| = 0$.*
- *Si A es una matriz triangular, entonces $|A| = \prod_{i=1}^{n} a_{ii}$ y en particular si cada $a_{ii} = 1$, entonces $|A| = 1$.*
- $|AB| = |BA|$.

Definición A.1.10. *La inversa de una matriz cuadrada $A_{n\times n}$ es la matriz única denotada A^{-1}, que satisface $AA^{-1} = A^{-1}A = I$, y existe si y solo si A es no singular; por lo que A es invertible.*

A.2. Preliminares para el análisis discriminante

Definición A.2.1. *Suponga que la variable aleatoria Y tiene una distribución que puede ser representada por una $f.d.p.$ de la forma:*

$$f_Y(y) = p_1 f_1(y) + \ldots + p_k f_k(y)$$

Donde todas las p_j son positivas, $p_1 + \ldots + p_k = 1$, y todas las $f_k(y)$ son funciones de densidad de probabilidad. Entonces se dice que Y tiene una distribución mixta finita con k componentes.

Si el estudio se centra en la clasificación, las p_j se conocen como probabilidades a priori. El concepto de *priori* se refiere a que p_j da la probabilidad de que una observación sea del j-ésimo grupo antes de que se tenga cualquier medida. Después de haber calculado el valor de Y, es posible que se deseé volver a evaluar la probabilidad de pertenencia a un grupo. Se tendría entonces que utilizar la función de probabilidad condicional de X, dada $Y = y$, y definir

las probabilidades a posteriori de pertenencia a un grupo j como:

$$p_{jy} = Pr[X = j \mid Y = y] = \frac{p_j f_j(y)}{f_Y(y)}$$

Las probabilidades a posteriori utilizan la información obtenida mediante el calculo de Y y son, por lo tanto, *conjeturas con información*, en contra de las *conjeturas sin información* expresadas por las probabilidades a priori.

Definición A.2.2. *Sea* $\boldsymbol{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_2 \end{pmatrix}$ *un vector aleatorio de dimensión p. Entonces:*

$$E[\boldsymbol{X}] = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_p] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$$

es el vector de medias de $\boldsymbol{X}$.

Con frecuencia se utilizará el símbolo $\boldsymbol{\mu} = \boldsymbol{E}[\boldsymbol{X}]$ para indicar que se esta hablando de un vector de medias. De forma más general, si

$$Y = \begin{pmatrix} Y_{11} & \cdots & Y_{1q} \\ \vdots & & \vdots \\ Y_{p1} & \cdots & Y_{pq} \end{pmatrix}$$

se denota como una matriz de dimensión $p \times q$ de variables aleatorias (o una matriz aleatoria), entonces se define su esperanza como:

$$E[Y] = \begin{pmatrix} E[Y_{11}] & \cdots & E[Y_{1q}] \\ \vdots & & \vdots \\ E[Y_{p1}] & \cdots & E[Y_{pq}] \end{pmatrix}$$

Definición A.2.3. *La matriz de varianzas y covarianzas de dimensión* $p \times p$ *de un vector aleatorio*

de dimensión p es:

$$Var(X) = Cov(X, X)$$

$$= E\left\{\begin{bmatrix} (X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) & \cdots & (X_1-\mu_1)\left(X_p-\mu_p\right) \\ (X_2-\mu_2)(X_1-\mu_1) & (X_2-\mu_2)^2 & \cdots & (X_2-\mu_2)\left(X_p-\mu_p\right) \\ \vdots & \vdots & \ddots & \vdots \\ \left(X_p-\mu_p\right)(X_1-\mu_1) & \left(X_p-\mu_p\right)(X_2-\mu_2) & \cdots & \left(X_p-\mu_p\right)^2 \end{bmatrix}\right\}$$

$$= \begin{pmatrix} var[X_1] & cov[X_1, X_2] & \cdots & cov[X_1, X_p] \\ cov[X_2, X_2] & var[X_2] & \cdots & cov[X_2, X_p] \\ \vdots & \vdots & \ddots & \vdots \\ cov[X_p, X_1] & cov[X_p, X_2] & \cdots & var[X_p] \end{pmatrix}$$

la cual se denota como

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

Considere un vector aleatorio $\boldsymbol{X} = \left(X_1, \ldots, X_p\right)'$ de dimensión p y una matriz

$$A = \left(a_{ij}\right) = \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kp} \end{pmatrix}$$

de orden $k \times p$. Se define un nuevo vector aleatorio

$$\boldsymbol{Y} = \boldsymbol{AX} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_k \end{pmatrix}$$

con

$$Y_j = \sum_{h=1}^{p} a_{jh} X_h$$

como su j-ésimo componente. En este caso, los elementos de la matriz $\boldsymbol{A}$ son números reales conocidos, por lo tanto, cada variable Y_j es una combinación lineal de $X_1, \ldots, X_p$. Ahora, se puede mencionar la relación entre los momentos de $\boldsymbol{X}$ y los de $\boldsymbol{Y}$, suponiendo que todas la medias y varianzas existen.

Teorema A.2.1. *Si $\boldsymbol{Y} = \boldsymbol{AX} + \boldsymbol{c}$, donde $\boldsymbol{Y}$ es un vector aleatorio de dimensión p, $\boldsymbol{A}$ es una matriz conocida de orden $k \times p$, y $\boldsymbol{c} \in \Re^k$ es un vector conocido, entonces*

$$E[\boldsymbol{Y}] = \boldsymbol{A} \cdot \boldsymbol{E}[\boldsymbol{X}] + \boldsymbol{c}$$

y

$$Var[\boldsymbol{Y}] = \boldsymbol{A} \cdot \boldsymbol{Var}[\boldsymbol{X}] \cdot \boldsymbol{A'}$$

Teorema A.2.2. *Si $\boldsymbol{Y} = \boldsymbol{AX} + \boldsymbol{c}$, donde $\boldsymbol{X}$ es un vector aleatorio de dimensión p, $\boldsymbol{A}$ es una matriz fija de dimensión $k \times p$, y $\boldsymbol{c} \in \Re^k$ un vector fijo, entonces*

$$E[\boldsymbol{Y}] = \boldsymbol{A} \cdot \boldsymbol{E}[\boldsymbol{X}] + \boldsymbol{c}, \quad y$$

$$Cov[\boldsymbol{Y}] = \boldsymbol{A} \cdot \boldsymbol{Cov}[\boldsymbol{X}] + \boldsymbol{A'}$$

Teorema A.2.3. *Sea (X, Y) una variable aleatoria bivariada tal que $E[Y|X = x] = \alpha + \beta x$, es decir, la regresión de Y sobre X es una línea recta con alguna intersección α y alguna pendiente β. Entonces*

$$\beta = cov[X, Y]/var[x], \quad y\alpha = E[Y] - \beta \cdot E[X]$$

A.3. Medidas de localización para datos circulares

A.3.1. La dirección media

Suponga que se tienen vectores unitarios $x_1, \ldots, x_n$ con sus correspondientes ángulos θ_i, $i = 1, \ldots, n$. La dirección media $\bar{\theta}$ de $\theta_1, \ldots, \theta_n$ es la dirección resultante $x_1 + \ldots + x_n$ de $x_1, \ldots, x_n$. Esto también es la dirección del centro de masa $\bar{x}$ de $x_1, \ldots, x_n$. Desde las coordenadas

cartesianas de x_j que son $\left(\cos\theta_j, \sin\theta_j\right)$ para $j = 1, \ldots, n$, las coordenadas cartesianas del centro de masa son $\left(\bar{C}, \bar{S}\right)$, donde

$$\bar{C} = \frac{1}{n}\sum_{j=1}^{n}\cos\theta_j \quad \bar{S} = \frac{1}{n}\sum_{j=1}^{n}\sin\theta_j \tag{A.1}$$

Por lo tanto, $\bar{\theta}$ es la solución de las ecuaciones:

$$\bar{C} = \bar{R}\cos\bar{\theta} \quad \bar{S} = \bar{R}\sin\bar{\theta} \tag{A.2}$$

(Siempre que $\bar{R} > 0$), donde la longitud media resultante $\bar{R}$ esta dada por

$$\bar{R} = \left(\bar{C}^2 + \bar{S}^2\right)^{1/2} \tag{A.3}$$

Note que $\bar{\theta}$ no esta definida cuando $\bar{R} = 0$. Si $\bar{R} > 0$, $\bar{\theta}$ está dada explícitamente por

$$\bar{\theta} = \begin{cases} \arctan\left(\bar{S}/\bar{C}\right) & \text{si } \bar{C} \geq 0 \\ \arctan\left(\bar{S}/\bar{C}\right) + \pi & \text{si } \bar{C} < 0 \end{cases} \tag{A.4}$$

Donde $arctan$ toma valores en $[-\pi/2, \pi/2]$. Note que en el contexto de la estadística circular $\bar{\theta}$ no es una media $(\theta_1 + \ldots + \theta_n)/n$ (que no está bien definida, ya que depende de donde se corta el círculo).

Se deduce de (A.1) y (A.2) que:

$$\frac{1}{n}\sum_{j=1}^{n}\cos\left(\theta_j - \bar{\theta}\right) = \bar{R} \tag{A.5}$$

y (para $\bar{R} > 0$)

$$\sum_{j=1}^{n}\sin\left(\theta_j - \bar{\theta}\right) = 0 \tag{A.6}$$

La ecuación (A.6) es análoga a:

$$\sum_{j=1}^{n} \left(x_j - \bar{x}\right) = 0 \tag{A.7}$$

para observaciones $x_1, \ldots, x_n$ en la recta con media muestral $\bar{x}$. Las ecuaciones (A.6) y (A.7) establecen que la suma de desviaciones alrededor de la media son cero.

Ahora se considerará el efecto de rotaciones en la media muestral direccional. Suponga que una nueva dirección inicial es elegida, haciendo un ángulo α con la dirección inicial original. Entonces los puntos de los datos corresponden a los ángulos

$$\theta'_j = \theta_j - \alpha, \quad j = 1, \ldots, n \tag{A.8}$$

en este nuevo sistema de coordenadas.

Se escibe:

$$\bar{C}' = \frac{1}{n}\sum_{j=1}^{n} \cos \theta'_j \quad \bar{S}' = \frac{1}{n}\sum_{j=1}^{n} \sin \theta'_j \tag{A.9}$$

entonces

$$\bar{C}' = \bar{R}\cos\left(\bar{\theta} - \alpha\right), \quad \bar{S}' = \bar{R}\sin\left(\bar{\theta} - \alpha\right) \tag{A.10}$$

Si las coordenadas polares de $\left(\bar{C}', \bar{S}'\right)$ son $\left(\bar{R}', \bar{\theta}'\right)$, entonces

$$\bar{C}' = \bar{R}'\cos\bar{\theta}', \quad \bar{S}' = \bar{R}'\sin\bar{\theta}' \tag{A.11}$$

La comparación de (A.9) y (A.10) da

$$\bar{\theta}' = \bar{\theta} - \alpha, \quad \bar{R}' = \bar{R} \tag{A.12}$$

Por lo tanto, la dirección media de $\theta_1 - \alpha, \ldots, \theta_n - \alpha$ es $\bar{\theta} - \alpha$, es decir, la dirección media muestral es equivalente bajo una rotación.

A.4. Medidas de concentración y dispersión

A.4.1. La longitud media resultante y la varianza circular

La longitud media resultante $\bar{R}$ fue introducida en (A.3) como la longitud del vector de centro de masa $\bar{x}$, y esta dado por

$$\bar{R} = \left(\bar{C}^2 + \bar{S}^2\right)^{1/2}$$

como $x_1, \ldots, x_n$ son vectores unitarios

$$0 \leq \bar{R} \leq 1 \tag{A.13}$$

Si las direcciones $\theta_1, \ldots, \theta_n$ están fuertemente agrupadas entonces $\bar{R}$ podría ser a lo más uno. Por otro lado, si $\theta_1, \ldots, \theta_n$ están muy dispersos entonces $\bar{R}$ puede ser a lo más cero, por lo tanto, $\bar{R}$ es una medida de concentración de un conjunto de datos. Note que cualquier conjunto de datos de la forma $\theta_1, \ldots, \theta_n, \theta_1 + \pi, \ldots, \theta_n + \pi$ tiene $\bar{R} = 0$. Se sigue que $\bar{R} \simeq 0$ no implica que las direcciones se distribuyen casi por igual alrededor del círculo.

La longitud resultante R es la longitud del vector resultante $x_1 + \ldots + x_n$, así:

$$R = n\bar{R} \tag{A.14}$$

Para propósitos más descriptivos e inferenciales, la longitud media resultante $\bar{R}$ es más importante que cualquier medida de dispersión, sin embargo, para propósitos de comparación con datos sobre la recta, a veces esto es útil para considerar medidas de dispersión para datos circulares, la más simple de éstas es la varianza circular muestal definida como

$$V = 1 - \bar{R} \tag{A.15}$$

Se deduce de (A.13) que

$$0 \leq V \leq 1 \tag{A.16}$$

A.4.2. Descomposición de la dispersión

Una medida usada de la distancia entre dos ángulos θ y ξ es

$$1 - \cos(\theta - \xi) \tag{A.17}$$

Entonces un camino de medición de dispersión de ángulos $\theta_1, \ldots, \theta_n$ sobre un ángulo dado α es

$$D(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\{1 - \cos(\theta_i - \alpha)\} \tag{A.18}$$

Se deduce de (A.5) que

$$D\left(\bar{\theta}\right) = V \tag{A.19}$$

Desde la descomposición

$$n - \sum_{i=1}^{n}\cos(\theta_i - \alpha) = \left(n - n\bar{R}\right) + \left(n\bar{R} - \sum_{i=1}^{n}\cos(\theta_i - \alpha)\right) \tag{A.20}$$

junto con (A.19) y (A.5), se tiene

$$D(\alpha) = V + 2\bar{R}\left\{\sin\left(\frac{\bar{\theta} - \alpha}{2}\right)\right\}^2 \tag{A.21}$$

esto es análogo para la identidad

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - u)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + (\bar{x} - u)^2 \tag{A.22}$$

para observaciones $x_1, \ldots, x_n$ sobre la recta con media muestral $\bar{x}$.

A.4.3. La desviación estándar circular

A veces es útil disponer de un análogo para los datos circulares de la desviación estándar de los datos en la recta. Una manera de obtener una estadística es por la transformación de

la varianza muestral V. Una transformación estadística apropiada es la desviación estándar muestral circular dada por

$$v = \{-\log(1 - V)\}^{1/2} = \left\{-2\log \bar{R}\right\}^{1/2} \tag{A.23}$$

Note que v toma valores en $[0, \infty]$, mientras que V toma valores en $[0, 1]$.

Para valores pequeños de V, la ecuación (A.23) se reduce a

$$u \simeq (2V)^{1/2} = \left\{2\left(1 - \bar{R}\right)\right\}^{2} \tag{A.24}$$

APÉNDICE B

CÓDIGOS EN R

B.1. Gráficas

Figura 1.4:

```
x=seq(-1,12,length=1000)
y1=dnorm(x,mean=2,sd=1)
plot(x,y1,type="l",lwd=2,col="red")
y2=dnorm(x,mean=4,sd=1)
y=lines(x,y2,type="l",lwd=2,col="blue")
polygon(x,pmin(y1,y2),col="gray")
```

Figura 2.1 y 2.4:

```
library(circular)
windc <- circular(wind, type="angles", units="radians",
                  template='geographics')
par(mai=c(0.85, 0.85, 0.05, 0.05), cex.axis=1.1, cex.lab=1.3)
##diagrama de dispersión
```

```
plot(wind, pch=16, xlab="Observation number",
     ylab="Wind direction (in radians)")
```

```
##diagrama circular
plot(windc, cex=1.5, bin=720, stack=TRUE, sep=0.035, shrink=1.3)
##diagrama de rosa
axis.circular(at=circular(seq(0, 7*pi/4,pi/4)),
              labels=c("N","NE","E","SE","S","SW","W","NW"),
              zero=pi/2, rotation='clock', cex=1.5)
ticks.circular(circular(seq(0,2*pi,pi/8)), zero=pi/2,
               rotation='clock', tcl=0.075)
rose.diag(windc, bins=16, col="darkgrey", cex=1.5, prop=1.3, add=TRUE)
```

Para las densidades von-Mises, Figura 2.10:

```
library(circular)
#1er ejemplo
ff <- function(x) dvonmises(x, mu=circular(pi), kappa=2)
curve.circular(ff, join=TRUE, xlim=c(-1.38, 0.9), cex=1.3, lwd=2)
```

```
library(circular)
##2do ejemplo
ff <- function(x) dvonmises(x, mu=circular(pi), kappa=2)
curve.circular(ff, join=TRUE, xlim=c(-2, 2), ylim=c(-2, 1), cex=1.3, lwd=2)
```

Figura 2.11:

```
library(circular)
ff <- function(x) (dvonmises(x, mu=circular(pi/2),
                             kappa=12.5)+
                   dvonmises(x, mu=circular(pi), kappa=12.5))/2
curve.circular(ff, join=TRUE, cex=.000000001, ylim=c(-1.6,1.6),lwd=1)
```

B.2. Para las aplicaciones

Función vmf.da

```
vmf.da=function(x,ina,fraction=0.2,R=1000,seed=FALSE) {
  ## x es el conjunto de datos
  ##ina es el vector que indica el grupo de pertenencia
  ##del conjunto de datos x
  ## fraction denota el porcentaje de la muestra
  ##que se utilizará en el test
  ## R es el número de validaciones cruzadas
  x=as.matrix(x) ; p=ncol(x)
  ## p es la dimensionalidad de los datos
  per=numeric(R) ; n=nrow(x)
  ina=as.numeric(ina)
  frac=round(fraction*n)
  g=max(ina)
  mesi=matrix(nrow=g,ncol=p)
  k=numeric(g)
  ## si seed==TRUE el resultado siempre será el mismo
  if (seed==TRUE) set.seed(1234567)
  for (i in 1:R) {
    mat=matrix(nrow=frac,ncol=g)
```

```
  est=numeric(frac)
  nu=sample(1:n,frac) ; test=x[nu,]
  id=ina[-nu] ; train=x[-nu,]
  for (j in 1:g) {
    da=vmf(train[id==j,])
    ## estima los parámentros de von Mises-Fisher
    mesi[j,]=da$mu ## dirección media
    k[j]=da$k } ## concentración
  for (j in 1:g) {
    mat[,j]=(p/2-1)*log(k[j])
    +k[j]*test%*%mesi[j,]-0.5*p*log(2*pi)-
      log(besselI(k[j],p/2-1,expon.scaled=T))-k[j] }
  for (l in 1:frac) est[l]=which.max(mat[l,])
  per[i]=sum(est==ina[nu])/frac }
percent=mean(per)
s1=sd(per) ; s2=sqrt(percent*(1-percent)/R)
conf1=c(percent-1.96*s1,percent+1.96*s1) ## 1er I.C.
conf2=c(percent-1.96*s2,percent+1.96*s2) ## 2do I.C.
## A continuación se verifica si los límites de
##confianza exceden los límites permitidos
if (conf1[2]>1) conf1[2]=1
if (conf1[1]<0) conf1[1]=0
if (conf2[2]>1) conf2[2]=1
if (conf2[1]<0) conf2[1]=0
conf3=quantile(per,probs=c(0.025,0.975)) ## 3er I.C.
list(percentage=percent,sd1=s1,sd2=s2,
     conf.int1=conf1,conf.int2=conf2,conf.int3=conf3) }
```

Función vmfda.pred

```
vmfda.pred=function(xnew,x,ina) {
## xnew es la matriz de nuevas observaciones
## x es el conjunto de datos original
## ina es el vector que indica el grupo de pertenencia
x=as.matrix(x)
xnew=as.matrix(xnew)
if (ncol(xnew)==1) xnew=t(xnew)
p=ncol(x) ## p es la dimensionalidad de los datos
ina=as.numeric(ina)
g=max(ina)
mesi=matrix(nrow=g,ncol=p)
k=numeric(g)
nu=nrow(xnew)
mat=matrix(nrow=nu,ncol=g)
est=numeric(nu)
for (j in 1:g) {
da=vmf(x[id==j,]) ## se estiman los parámetros de von-Mises
mesi[j,]=da$mu ## dirección media
k[j]=da$k } ## concentración
for (j in 1:g) {
mat[,j]=(p/2-1)*log(k[j])+
  k[j]*xnew%*%mesi[j,]-0.5*p*log(2*pi)-
log(besselI(k[j],p/2-1,expon.scaled=T))-k[j] }
for (l in 1:nu) est[l]=which.max(mat[l,])
list(est.group=est) }
```

BIBLIOGRAFÍA

Balakrishnan, N., Kannan, N., and Nagaraja, H. N. (2007). *Advances in ranking and selection, multiple comparisons, and reliability: methodology and applications*. Springer Science & Business Media.

Figueiredo, A. (2009). Discriminant analysis for the von mises-fisher distribution. *Communications in Statistics-Simulation and Computation*, 38(9):1991–2003.

Fisher, N. I., Lewis, T., and Embleton, B. J. (1987). *Statistical analysis of spherical data*. Cambridge university press.

Flury, B. (2013). *A first course in multivariate statistics*. Springer Science & Business Media.

Jammalamadaka, S. R. and Sengupta, A. (2001). *Topics in circular statistics*, volume 5. World Scientific.

Mardia, K. V. and Jupp, P. E. (2009). *Directional statistics*, volume 494. John Wiley & Sons.

Martin, N. and Maes, H. (1979). *Multivariate analysis*. Academic press.

Moffitt, S. D. (1976). Discriminant analysis for directional and orientational data.

Morris, J. E. and Laycock, P. (1974). Discriminant analysis of directional data. *Biometrika*, 61(2):335–341.

Pewsey, A., Neuhäuser, M., and Ruxton, G. D. (2013). *Circular statistics in R*. Oxford University Press.

Romanazzi, M. (2014). Discriminant analysis with high dimensional von mises-fisher distributions. In *8th Annual International Conference on Statistics*, pages 1–16. Athens Institute for Education and Research.

Streit, F. (1997). Discriminant analysis for directional data exemplified in a concrete case. In *Classification and Knowledge Organization*, pages 208–214. Springer.

Webb, A. R. (2003). *Statistical pattern recognition*. John Wiley & Sons.