Tema 6

Análisis Discriminante

6.1. Introducción

El análisis discriminante es una técnica estadística multivariante que se utiliza para realizar clasificaciones de individuos en grupos definidos previamente. Para aplicar esta técnica debemos disponer, por una parte, de información observable en una muestra y, por otra parte, de conocimiento del grupo de pertenencia de los individuos de la muestra. Este conocimiento del grupo de pertenencia debe ser ajeno a los datos observados, es decir, el conocimiento de pertenencia al grupo debe darse mediante otros mecanismos que no sean la mera observación de las variables consideradas. En estas condiciones, estableceremos la relación existente entre el grupo de pertenencia y los valores observados de las variables para que, cuando un nuevo individuo deba ser clasificado, solamente observando las variables de interés podamos clasificarlo en el grupo más probable y podamos calcular esta probabilidad de pertenencia. La información de partida es, por tanto, la de un grupo cuya clasificación es conocida y sobre la que se pueden observar variables que están relacionadas con su pertenencia al grupo pero que no permiten por sí solas realizar la clasificación. Una vez realizado el análisis discriminante obtendremos un criterio de clasificación para maximizar y cuantificar la probabilidad de clasificar nuevos individuos sobre los que observaremos las mismas variables de interés.

Un ejemplo de la utilidad que puede llegar a tener esta técnica se encuentra en la clasificación de pacientes en un grupo diagnóstico a partir de síntomas o pruebas fácilmente realizables, o la clasificación de clientes bancarios en grupos de riesgo alto o bajo a la hora de concesión de créditos. Clasificar individuos a partir de la observación de un perfil es una herramienta frecuente en muchos campos de aplicación. El análisis discriminante establece la forma de realizar esta clasificación optimizando la información disponible.

La primera aplicación de Análisis Discriminante aparece en un trabajo de 1936 [4] donde Fisher desarrolla y evalúa una función lineal para diferenciar las especies de Iris (Iris setosa, Iris versicolor e Iris virginica) en función de la morfología de sus flores. Hoy en día se

suele utilizar este conjunto de datos, llamado habitualmente «Iris de Fisher» como recurso para aprender el mecanismo de algunas técnicas de clasificación, aunque formalmente este conjunto no cumple con las hipótesis requeridas para realizar la clasificación que propuso el propio autor. Si bien en su artículo no se cumplían estrictamente todas las hipótesis necesarias para el correcto uso de la técnica del análisis discrimimnante lineal, Fisher formuló la idea de usar una variable que fuera combinación lineal de varias variables independientes para diferenciar entre grupos, que es la base del análisis discriminante lineal. Posteriormente, a partir de este tipo de análisis se han desarrollado otras muchas técnicas, que aportan diferentes mejoras a la original.

El análisis discriminante, dentro del grupo de técnicas multivariantes, se encuentra en el grupo de aquellas que tratan de modelizar el comportamiento de una o varias variables dependientes, al igual que MANOVA, correlación canónica o regresión múltiple. La diferencia en este caso es que la variable dependiente no es una variable métrica, puesto que contienen información acerca del grupo de pertenencia y, por ello, necesariamente es categórica. Además, el conjunto de datos inicial debe contener esta información no métrica observada de forma directa. Sin embargo, la búsqueda de combinaciones lineales de variables es uno de los pilares compartidos por diferentes técnicas. Por ejemplo, en el análisis de correlaciones canónicas y en el análisis factorial, se utilizan en busca de una mejor explicación de la relación entre variables y en regresión múltiple para obtener un modelo de explicación de una variable dependiente en función de un conjunto de independientes.

Los objetivos del análisis discriminante son dos: por una parte se pretende explicar, con la función discriminante, tanto la pertenencia de cada caso a su grupo como el peso de las variables en la discriminación; por otra parte, y como objetivo principal, se pretende predecir a qué grupo es más probable que pertenezca un individuo del que solo se conoce su perfil de variables. Para alcanzar estos objetivos existen dos enfoques diferentes: el primero de ellos se basa en la obtención de funciones discriminantes, de cálculo similar a las ecuaciones de regresión lineal múltiple. El otro enfoque emplea técnicas de correlación canónica y de componentes principales.

El fundamento para el análisis mediante funciones discriminantes será conseguir unas funciones lineales a partir de las variables explicativas, que tengan capacidad de clasificar nuevos individuos. Para ello, una vez observados los valores en las variables para el nuevo individuo, las funciones discriminantes evaluadas con estos valores indicarán cuál es el grupo más probable de pertenencia. El enfoque canónico intenta reducir la dimensión del problema utilizando el análisis de componentes principales para ello, encontrando un conjunto de variables sintéticas que permitan la representación de la información en una dimensión menor. Estas variables actúan también como variables discriminantes consiguiendo hacer máxima la relación entre la dispersión factorial y la dispersión residual, para separar los centros de gravedad tanto como se pueda, de forma que los valores obtenidos en los individuos de cada grupo estén lo más concentrados posible. Esta es también

la idea del análisis de la varianza para detectar diferencias entre medias.

En este tema nos centraremos en el primero de los enfoques, el de conseguir funciones discriminantes lineales para la clasificación. Este problema se resuelve de manera relativamente sencilla cuando el número de variables es muy reducido, pero su complejidad se hace mayor cuando crece el número de variables. Observaremos a lo largo del tema el caso de clasificación en dos grupos con una sola variable discriminante, con dos variables discriminantes y con p variables discriminantes ($p > 2$) así como la clasificación en más de dos grupos. El objetivo de estudiar el proceso en una dimensión reducida es que la generalización al caso de p variables y k grupos se comprenda mejor.

El punto de partida para realizar análisis discriminante es disponer de una matriz de datos de N individuos para los que se observan p variables cuantitativas independientes (explicativas o discriminantes), que definen el perfil de cada individuo, y una variable adicional, cualitativa, que es la variable dependiente. Esta última contiene la información del grupo de pertenencia para cada uno de los individuos. En estas circunstancias, para poder comenzar a plantear el análisis discriminante, partiremos de las siguientes hipótesis:

1. La variable dependiente divide a los individuos en al menos 2 grupos diferentes, con un número de individuos mayor a 1 en cada grupo.
2. El número de individuos N debe ser mayor que el número de variables independientes más 2 ($N > p + 2$).
3. Ninguna de las variables independientes es combinación lineal de otras variables independientes (esto es, no existe multicolinealidad).
4. No hay diferencias entre las matrices de varianza-covarianzas de cada grupo.
5. El vector de variables independientes tiene distribución normal multivariante.

6.2. Clasificación en dos grupos

Comenzamos estudiando el caso de discriminación en dos grupos G_I y G_{II}. Para realizar la clasificación, en primer lugar utilizaremos la información de una sola variable discriminante, a continuación abordamos el problema utilizando la información de dos variables discriminantes y terminaremos esta sección generalizando al caso en que queramos hacer una clasificación en dos grupuos conociendo la información de p variables discriminantes.

6.2.1. Clasificación a partir de una variable discriminante

Supongamos que disponemos de una variable discriminante X para clasificar un nuevo individuo entre dos grupos posibles G_I y G_{II}. Nuestro objetivo es encontrar una función

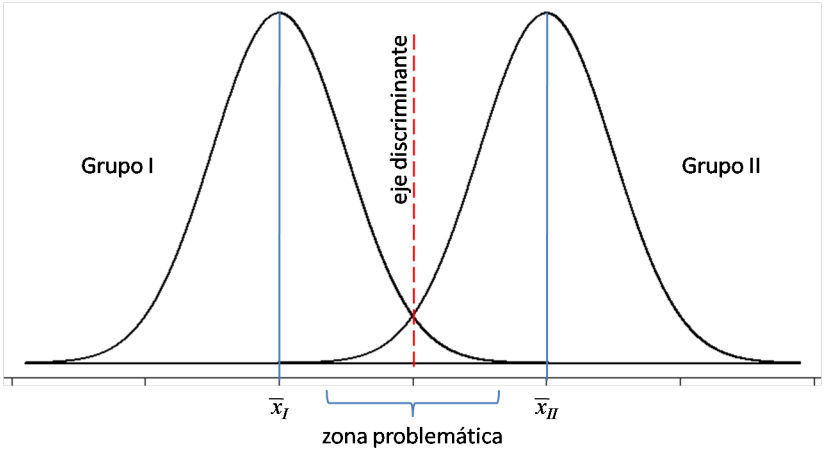

Figura 6.1: Análisis del problema en dos grupos con una sola variable discriminante.

lineal de la variable discriminante X que nos permita clasificar cada observación en uno de los grupos, minimizando el error de clasificación. Para ello tenemos en cuenta que, según las hipótesis de partida, la distribución de X en cada uno de los grupos solo se diferencia en su posición, ya que tiene la misma forma y dispersión para cada grupo.

Si nos fijamos en la Figura 6.1, podemos tomar como función o eje discriminante la recta $C = \dfrac{\overline{X}_I + \overline{X}_{II}}{2}$ siendo $\overline{X}_I$ la media muestral de la variable X en el grupo G_I y siendo $\overline{X}_{II}$ la media muestral de la variable X en el grupo G_{II}. De esta manera, la zona problemática que se señala en la Figura 6.1 queda dividida y el error que se comete con una clasificación para un individuo que pertenezca a esta zona se minimiza, es decir, usando esta función discriminante se cometerá un error mínimo para el problema de clasificación.

En este problema obtenemos una solución inmediata teniendo en cuenta las hipótesis de partida, aunque podríamos variar la solución en el caso en que los grupos de partida no tengan el mismo tamaño, cosa que hemos asumido pero no tiene en principio por qué ocurrir. Para minimizar la probabilidad de error, debemos desplazar el eje C para que quede más cerca del centro del grupo con menor número de individuos. Para eso habría que ponderar, obteniendo el punto de corte como $C = \dfrac{n_I \overline{X}_I + n_{II} \overline{X}_{II}}{N}$ con n_I y n_{II} tales que $n_I + n_{II} = N$ los tamaños de los grupos I y II correspondientes.

6.2.2. Clasificación a partir de dos variables discriminantes

Supongamos ahora que disponemos de dos variables discriminantes (X_1, X_2). Queremos encontrar una función lineal de las variables discriminantes X_1 y X_2 que permita

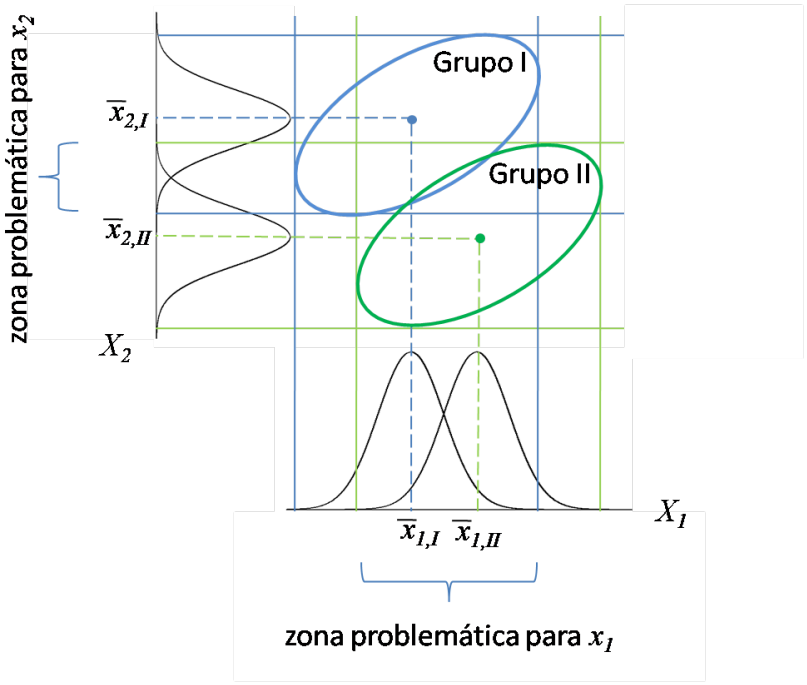

Figura 6.2: Análisis del problema en dos grupos con dos variables discriminantes.

clasificar cada observación en G_I o en G_{II}, minimizando el error de clasificación.

Para ello, consideramos en primer lugar la variable X_1 y proyectamos sobre el eje correspondiente los datos observados, de forma que podemos usar la misma solución que hemos obtenido en el caso de una variable discriminante. Al hacer esto, se crea una zona problemática debido a la superposición de las distribuciones normales de los dos grupos. A continuación, hacemos lo mismo con X_2, creando otra nueva zona problemática. Por tanto, al proyectar en cada uno de los ejes (variables) tenemos dos zonas problemáticas distintas como puede verse en la Figura 6.2.

Si consideramos el ejemplo de la Figura 6.2, se tiene que la zona problemática creada por la variable discriminante X_2 es menor que la generada por X_1. En consecuencia, X_2 discrimina mejor que X_1. Pero, dado que el objetivo es minimizar la región problemática, podremos reducir aún mas la región problemática si buscamos una función lineal de X_1 y X_2

$$D = \omega_1 X_1 + \omega_2 X_2$$

que minimice dicha región, como se muestra en la Figura 6.3.

La solución que aportó Fisher a este problema fue crear una función que maximiza la separación entre los grupos, maximizando la distancia entre sus medias, y minimiza la dispersión dentro de los grupos. La función lineal que se obtiene al aplicar estas condiciones

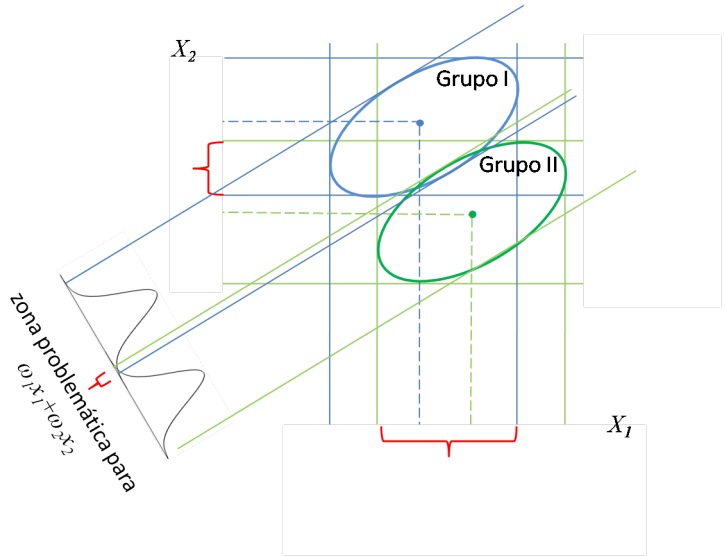

Figura 6.3: Función lineal que minimiza la zona problemática con dos variables discriminantes.

se denomina «función discriminante lineal de Fisher».

A continuación veremos cómo se obtiene dicha función para el caso general de p variables discriminantes.

6.2.3. Clasificación a partir de p variables discriminantes

El objetivo es encontrar una función discriminante capaz de minimizar la variabilidad dentro de los grupos y maximizar la variabilidad entre los grupos y que sea combinación lineal de las p variables de las que se dispone:

$$D = \omega_1 X_1 + \omega_2 X_2 + \ldots + \omega_p X_p$$

y para ello debemos calcular los coeficientes ω_j.

Consideremos el caso de tener N observaciones. Para cada observación $i = 1, \ldots, N$, la función discriminante tiene la siguiente forma:

$$D_i = \omega_1 X_{1i} + \omega_2 X_{2i} + \ldots + \omega_p X_{pi}.$$

Si expresamos las N funciones discriminantes anteriores de forma matricial obtenemos:

$$\begin{pmatrix} D_1 \\ D_2 \\ \vdots \\ D_N \end{pmatrix} = \begin{pmatrix} X_{11} & X_{21} & \ldots & X_{p1} \\ X_{12} & X_{22} & \ldots & X_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1N} & X_{2N} & \ldots & X_{pN} \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_p \end{pmatrix} \tag{6.1}$$

Si se reescribe la expresión (6.1), en función de las desviaciones de la media, se obtiene:

$$\begin{pmatrix} D_1 - \overline{D} \\ D_2 - \overline{D} \\ \vdots \\ D_N - \overline{D} \end{pmatrix} = \begin{pmatrix} X_{11} - \overline{X}_1 & X_{21} - \overline{X}_2 & \dots & X_{p1} - \overline{X}_p \\ X_{12} - \overline{X}_1 & X_{22} - \overline{X}_2 & \dots & X_{p2} - \overline{X}_p \\ \vdots & \vdots & \ddots & \vdots \\ X_{1N} - \overline{X}_1 & X_{2N} - \overline{X}_2 & \dots & X_{pN} - \overline{X}_p \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_p \end{pmatrix} \tag{6.2}$$

siendo

$$\overline{D} = \omega_1 \overline{X}_1 + \omega_2 \overline{X}_2 + \ldots + \omega_p \overline{X}_p.$$

La expresión (6.2) se resume en la función discriminante en diferencias:

$$d = X\omega.$$

A partir de d vamos a obtener la variabilidad de la función discriminante, es decir, la suma de cuadrados de las desviaciones de las variables discriminantes con respecto de su media:

$$d'd = \omega' X' X \omega,$$

siendo $X'X$ la matriz simétrica que expresa las desviaciones cuadráticas respecto de la media de las variables.

A continuación se enuncia un teorema que nos resultará útil para continuar con los cálculos.

Teorema. *La matriz de desviaciones cuadráticas $X'X$ se puede descomponer como la suma entre la matriz que incluye las variabilidades entre los grupos para las variables, E, y la matriz que incluye las variabilidades dentro de cada grupo o intragrupos I. Es decir*

$$X'X = E + I,$$

con

$$E = n_I(\overline{X}_I - \overline{X})(\overline{X}_I - \overline{X})' + n_{II}(\overline{X}_{II} - \overline{X})(\overline{X}_{II} - \overline{X})'$$

donde:

$$\overline{X} = \frac{n_I \overline{X}_I + n_{II} \overline{X}_{II}}{N} \quad con \quad N = n_I + n_{II}.$$

$$\overline{X}_I = (\overline{X}_1^I, \ldots, \overline{X}_p^I)' \quad con \quad \overline{X}_j^I = \frac{\sum_{i=1}^{n_I} X_{ij}}{n_I}$$

$$\overline{X}_{II} = (\overline{X}_1^{II}, \ldots, \overline{X}_p^{II})' \quad con \quad \overline{X}_j^{II} = \frac{\sum_{i=1}^{n_{II}} X_{ij}}{n_{II}}$$

$$I = \sum_{i=1}^{n_I}(X_{iI} - \overline{X}_I)(X_{iI} - \overline{X}_I)' + \sum_{i=1}^{n_{II}}(X_{iII} - \overline{X}_{II})(X_{iII} - \overline{X}_{II})',$$

denotando:

$$X_{iI} = (X_{i1}^I, \ldots, X_{ip}^I)'$$
$$X_{iII} = (X_{i1}^{II}, \ldots, X_{ip}^{II})'$$

siendo X_{ij}^I el individuo i observado en la variable j en el grupo I y siendo X_{ij}^{II} el individuo i observado en la variable j en el grupo II.

Nota. *Denotaremos como $\overline{X}_I$ y $\overline{X}_{II}$ los centroides de los grupos I y II, respectivamente.*

$$d'd = \omega' X' X \omega = \omega'(E + I)\omega = \omega' E \omega + \omega' I \omega.$$

Siguiendo la idea de Fisher, para encontrar las funciones D_i que consigan discriminar de la mejor manera posible, hay que maximizar la varianza entre grupos y minimizar la varianza dentro de los grupos. Esto se resume en hallar

$$\text{máx}\left(\frac{\omega' E \omega}{\omega' I \omega}\right).$$

Calcular este máximo es equivalente a calcular máx$(\omega' E \omega)$ con $\omega' I \omega = 1$, pues la función $\dfrac{\omega' E \omega}{\omega' I \omega}$ es invariante frente a cambios de escala.

Para calcular el máximo aplicaremos multiplicadores de Lagrange:

$$L = \omega' E \omega - \lambda(\omega' I \omega - 1) \Rightarrow \frac{\partial \mathrm{L}}{\partial \omega} = 2E\omega - 2\lambda I\omega = 0$$
$$\Rightarrow E\omega = \lambda I\omega \Rightarrow (I^{-1}E)\omega = \lambda\omega.$$

Como $E\omega = \lambda I\omega$, obtenemos que $\omega' E \omega = \omega' \lambda I \omega = \lambda$.

Por tanto, si tomamos el vector propio asociado al máximo valor propio obtendremos la función que mejor discrimina.

6.2.4. Clasificación

Supongamos que queremos clasificar un individuo **p** cuya observación viene dada por $\mathbf{x_0} = (x_1, \ldots, x_p)'$. Entonces, calculamos la función discriminante d_0, sustituyendo los valores correspondientes de las p variables en ella.

Podemos calcular la frontera discriminante:

$$C = \frac{\overline{D}_I + \overline{D}_{II}}{2},$$

siendo:

$$\overline{D}_I = \omega_1 \overline{X}_{1I} + \ldots + \omega_p \overline{X}_{pI}$$
$$\overline{D}_{II} = \omega_1 \overline{X}_{1II} + \ldots + \omega_p \overline{X}_{pII}$$

con $\omega_1, \ldots, \omega_p$ los elementos del vector propio asociado al mayor valor propio de la matriz $I^{-1}E$.

Por lo tanto, clasificaremos la observación en G_I si

$$d_0 < C \quad \Leftrightarrow \quad d_0 - C < 0,$$

y clasificaremos la observación en G_{II} si:

$$d_0 > C \quad \Leftrightarrow \quad d_0 - C > 0.$$

Esta regla es válida para minimizar los errores de clasificación, pero podemos observar que no se ha tenido en cuenta la posibilidad de que los grupos no tengan igual tamaño. Si esto ocurre, al usar C como frontera de clasificación, la proporción de casos mal clasificados en el grupo de menor tamaño será mucho mayor que en el grupo de mayor tamaño.

Para que esto no ocurra, cuando los tamaños son desiguales se puede usar una regla de clasificación que desplaza el punto de corte acercándolo al centroide del grupo de menor tamaño para igualar los errores de clasificación, como por ejemplo la distancia ponderada:

$$C = \frac{n_I \overline{D}_I + n_{II} \overline{D}_{II}}{n_I + n_{II}}.$$

Otra opción es calcular las funciones discriminantes para el grupo G_I y para el grupo G_{II} y clasificar la observación en el grupo en el cual la función tenga mayor valor.

6.3. Clasificación en más de dos grupos

En esta sección vamos a generalizar la idea expuesta la sección anterior al caso en el que estemos trabajando con g grupos.

Consideremos $G_1, \ldots, G_g$ los grupos en los cuales tenemos definido un vector aleatorio continuo $\mathbf{X} = (X_1, \ldots, X_p)'$ de forma que se cumplen las hipótesis de normalidad, homoscedasticidad y ausencia de multicolinealidad que se propusieron al comienzo.

Nuestro problema es clasificar un nuevo individuo en uno de los g grupos conociendo su valor en las p variables. En este caso, no será posible encontrar una única función discriminante. Serán necesarias $g - 1$ funciones discriminantes. El número máximo de funciones que se pueden calcular viene dado por:

$$\text{mín}\,(p, g - 1).$$

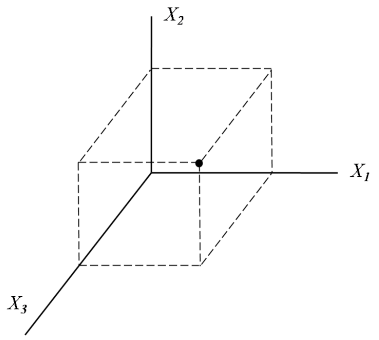

Figura 6.4: Posición de un punto en el espacio generado por las variables discriminantes X_1, X_2 y X_3.

6.3.1. Representación geométrica

Para comprender mejor el problema, vamos a visualizar geométricamente la situación. Para ello vamos a considerar una matriz de datos de N filas y p columnas, siendo N el número de individuos y p el número de variables discriminantes

$$\begin{pmatrix} x_{11} & x_{21} & \dots & x_{p1} \\ x_{12} & x_{22} & \dots & x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1N} & x_{2N} & \dots & x_{pN} \end{pmatrix}$$

Podemos considerar las variables discriminantes como ejes en el espacio p-dimensional que generan y cada individuo, es decir, cada fila de la matriz anterior, como un punto en dicho espacio. Por ejemplo, supongamos que tenemos únicamente tres variables discriminantes. Representando la idea anterior, obtenemos lo que se muestra en la Figura 6.4.

Por tanto, cuando los individuos pertenecen a un mismo grupo, tendrán una situación en el espacio similar, como se muestra en la Figura 6.5.

Entonces, si el problema que queremos resolver consiste en ver las diferencias existentes entre los grupos a partir de las variables discriminantes, bastará con analizar la posición de los centroides para ver si los grupos están bien diferenciados en el espacio generado por las p variables.

Si observamos la Figura 6.5, podemos construir un plano que pase por los tres centroides. Nuestro objetivo será determinar los ejes del plano que maximicen la distancia entre los centroides.

Intuitivamente, podemos construir un eje apuntando al punto en el que los centroides estén más separados, consiguiendo así la máxima dispersión posible. Para construir el se-

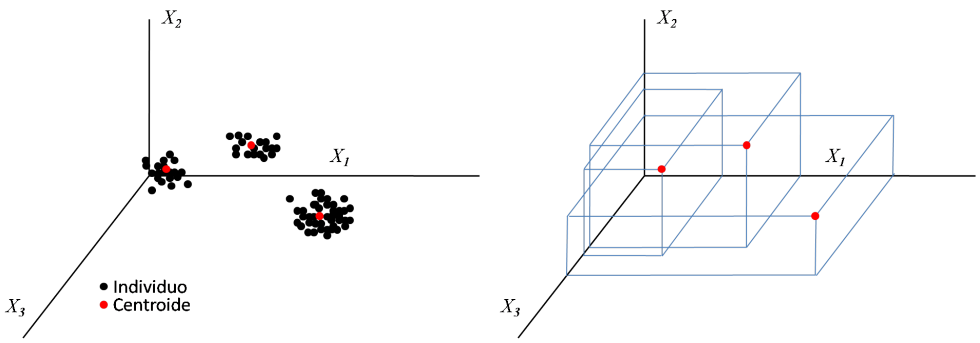

Figura 6.5: Posición de las observaciones y de los centroides.

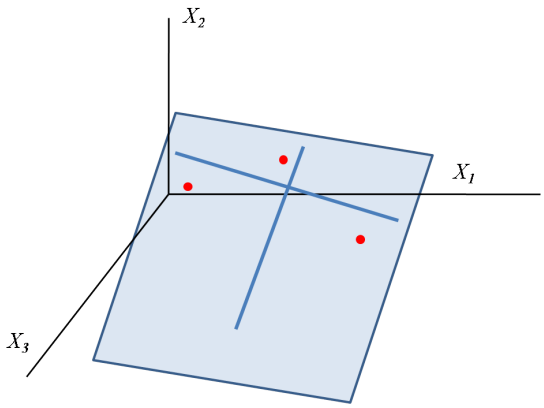

Figura 6.6: Ejes que consiguen la máxima dispersión para los tres centroides en el espacio generado por las variables discriminantes X_1, X_2 y X_3.

gundo eje, seguiremos el mismo criterio pero además este nuevo eje debe ser perpendicular al primero que hemos construido. De esta manera, conseguimos construir los dos ejes que proporcionan la mayor dispersión entre los tres grupos en el espacio, como se observa en la Figura 6.6.

Cada uno de estos ejes será una función discriminante.

6.3.2. Clasificación en g grupos con p variables discriminantes

Para el caso que estamos tratando, donde hay p variables discriminantes, podemos generalizar la idea expuesta en el apartado anterior. Para ello supongamos que tenemos g grupos donde se asignan una serie de individuos y p variables medidas sobre ellos, $(X_1, \ldots, X_p)'$.

El objetivo es encontar funciones discriminantes que nos permitan clasificar a cada individuo en su correspondiente grupo. Buscamos funciones discriminantes $(D_1, \ldots, D_m)'$

que sean funciones lineales de $(X_1, \ldots, X_p)'$, es decir:

$$\begin{array}{rcccccccc} D_1 & = & \omega_{11}X_1 & + & \omega_{12}X_2 & + & \ldots & + & \omega_{1p}X_p \\ D_2 & = & \omega_{21}X_1 & + & \omega_{22}X_2 & + & \ldots & + & \omega_{2p}X_p \\ \vdots & & \vdots & & \vdots & & \ddots & & \vdots \\ D_m & = & \omega_{m1}X_1 & + & \omega_{m2}X_2 & + & \ldots & + & \omega_{mp}X_p, \end{array}$$

siendo $m = \text{mín}(g-1, p)$.

Queremos que estas funciones discriminen lo máximo posible a los g grupos. Luego, basándonos en la idea de Fisher, las combinaciones lineales de las p variables tienen que maximizar la varianza entre grupos y minimizar la varianza dentro de los grupos para las N observaciones de las que se dispone.

Para obtener las funciones discriminantes, entonces, debemos obtener m funciones $(D_1, \ldots, D_m)'$ a partir de $(X_1, \ldots, X_p)'$ variables observadas en g grupos tal que:

$$D_j = \omega_{j1}X_1 + \omega_{j2}X_2 + \ldots + \omega_{jp}X_p, \quad j = 1, \ldots, m,$$

siendo $m = \text{mín}(g-1, p)$ y verificando $Corr(D_j, D_k) = 0 \quad \forall j \neq k$.

Entonces, calcularemos $(D_1, \ldots, D_m)'$ de forma que:

- D_1 será la combinación lineal de $(X_1, \ldots, X_p)'$ que proporcione la mayor discriminación entre los centroides de los grupos.
- D_2 será la combinación lineal de $(X_1, \ldots, X_p)'$ que proporcione la mayor discriminación entre los centroides de los grupos, después de D_1, y que verifique $Corr(D_1, D_2) = 0$.
- En general, D_j será la combinación lineal de $(X_1, \ldots, X_p)'$ que proporcione la mayor discriminación entre los centroides de los grupos, después de D_{j-1}, y que verifique $Corr(D_j, D_k) = 0$, con $\forall k = 1, \ldots, (j-1)$.

Ahora vamos a expresar matricialmente lo expuesto anteriormente. Buscaremos una función lineal de $(X_1, \ldots, X_p)'$ de manera que D_j proporcione la mayor discriminación entre los centroides de los grupos y las correlaciones entre los distintos grupos sean nulas.

Hacemos un cambio de notación, refiriéndos a las matrices I y E de la siguiente forma:

$$E = \sum_{j=1}^{g} n_j(\overline{X}_j - \overline{X})(\overline{X}_j - \overline{X})'$$

$$I = \sum_{j=1}^{g}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_j)(X_{ij} - \overline{X}_j)'.$$

Como la varianza total es igual a la suma de la varianza entre los grupos más la varianza dentro de los grupos, obtenemos que:

$$Var[D_j] = Var[\omega_j' X] = \omega_j' Var[X]\omega_j = \omega_j'(E+I)\omega_j = \omega_j' E\omega_j + \omega_j' I\omega_j.$$

De forma análoga al razonamiento seguido en la sección anteior, para buscar las funciones D_i que proporcionen la mayor discriminación calcularemos

$$\text{máx}\left(\frac{\omega_j' E\omega_j}{\omega_j' I\omega_j}\right).$$

La función $\dfrac{\omega_j' E\omega_j}{\omega_j' I\omega_j}$ es invariante frente a cambios de escala, luego calcular máx $\left(\dfrac{\omega_j' E\omega_j}{\omega_j' I\omega_j}\right)$ es equivalente a calcular máx$(\omega_i' E\omega_j)$, con $\omega_j' I\omega_j = 1$.

Aplicamos multiplicadores de Lagrange y obtenemos:

$$L = \omega_j' E\omega_j - \lambda(\omega_j' I\omega_j - 1) \Rightarrow \frac{\partial L}{\partial \omega_j} = 2E\omega_j - 2\lambda I\omega_j = 0$$
$$\Rightarrow E\omega_j = \lambda I\omega_j \Rightarrow (I^{-1}E)\omega_j = \lambda\omega_j.$$

Deducimos que el vector propio asociado a la primera función discriminante es vector propio de la matriz $I^{-1}E$, que no es simétrica en general.

Como $E\omega_j = \lambda I\omega_j$, entonces $\omega_j' E\omega_j = \lambda\omega_j' I\omega_j = \lambda$. Por tanto, si tomamos el vector propio asociado al mayor valor propio, obtenemos la función que mejor discrimina.

Para obtener más funciones discriminantes, obtenemos los vectores propios de la matriz $I^{-1}E$ asociados a los valores propios elegidos en orden decreciente. El procedimiento a seguir es:

(1) Calcular el mayor valor propio de $I^{-1}E$, λ_1, y su vector propio asociado, ω_1.

(2) Calcular el segundo mayor valor propio de $I^{-1}E$, λ_2, y su vector propio asociado, ω_2.

$\vdots$

$\vdots$

(m) Calcular el m-ésimo mayor valor propio de $I^{-1}E$, λ_m, y su vector propio asociado, ω_m, con $m = \text{mín}(g-1, p)$.

Los vectores $\omega_1, \ldots, \omega_m$ son linealmente independientes entre sí por estar asociados a valores propios distintos. Dan lugar a funciones discriminantes incorreladas entre sí, siempre que $I^{-1}E$ sea simétrica. Si esto último no ocurre, los ejes no tienen por qué ser perpendiculares entre sí.

El valor propio asociado a la función discriminante indica la proporción de varianza total explicada por las m funciones discriminantes. Se tiene que $\sum_{i=1}^{m} \lambda_i$ es la proporción de varianza total explicada. Por tanto, el porcentaje explicado por D_i del total de varianza explicada por $(D_1, \ldots, D_m)'$ viene dado por:

$$\frac{\lambda_i}{\sum_{i=1}^{m} \lambda_i} 100\,\%.$$

Nota. *Si las variables discriminantes $(X_1, \ldots, X_p)'$ están tipificadas las funciones*

$$D_i = \omega_{i1} X_1 + \omega_{i2} X_2 + \ldots + \omega_{ip} X_p$$

se denominan discriminantes canónicas.

6.3.3. Clasificación

Una vez halladas las funciones discriminantes podemos clasificar los individuos usados para crear dichas funciones para ver el grado de eficacia en el ámbito de clasificar.

Otra opción más precisa, sería dividir la muestra en dos partes, la primera sería una muestra de entrenamiento con la que se construirán las funciones discriminantes y, la segunda, una muestra para probar el grado de eficacia de la clasificación.

Si los resultados son buenos, podemos usar las funciones discriminantes para clasificar nuevos individuos de los que se desconozca su procedencia conociendo sus valores en las variables discriminantes $(X_1, \ldots, X_p)'$ construyendo las funciones discriminantes para cada grupo y clasificando en el grupo en el que la puntuación discriminante sea mayor.

6.4. Aplicación con R

Para poder realizar un análisis de discriminante en R, tenemos que cargar el paquete *MASS* donde encontramos diferentes funciones de análisis.

6.4.1. Funciones de R para realizar el análisis discriminante lineal

Función «lda»

La primera función es `lda(x, ...)`

La sintaxis para utilizar esta función es la siguiente:

```
lda(x, grouping, prior = proportions, tol = 1.0e-4, method, CV = FALSE,...)
```

o bien

```
lda(formula, data, ... , subset, na-action)
```

donde:

- *formula:* es una fórmula del tipo `grupos~x1+x2+...` donde la variable dependiente es el factor de agrupación (`grupos`) y las independientes son las variables o independientes discriminantes (`x1, x2, ...`).
- *x:* matriz de datos.
- *grouping:* se indica aquí la variable que marca los grupos de la muestra. Es obligatorio introducirla si no se ha hecho en la formula.
- *data:* fichero de datos que se usa si x es una fórmula.
- *prior:* son las probabilidades de pertenencia a cada grupo. Si no se indica serán proporcionales al tamaño. Si se indican se harán en el mismo orden que los grupos.
- *tol:* valor de tolerancia para decidir si una variable discriminate forma parte del modelo.
- *subset:* vector de índices que marcan qué casos muestrales se usarán de prueba.
- *na.action:* indica qué hacer con los casos faltantes.
- *method:* indica el método para estimar la media y varianza de las variables. Puede ser por el de los momentos (moment) o máximo verosímil (mle).
- *cv:* es un valor lógico, tal que si es verdadero da la clase y probabilidad a posteriori para el procedimiento de validación cruzada al eliminar cada individuo.

Los objetos de esta clase que se devuelven mediante la expresión `objet$componente` son:

- *prior:* probabilidades a priori.
- *means:* medias de las variables por grupos.
- *scaling:* matriz de transformación de las funciones discriminates.
- *svd:* valores singulares, es decir, el poder discriminante de cada función.
- *N:* número de individuos considerados.

Función «predict»

Esta función, del paquete básico `stat` se usa para clasificar observaciones en combinación con la función anterior, proyectando los datos sobre las funciones discriminantes.

```
predict(object, newdata, prior = object$prior, dimen)
```

con

- *object:* es un objeto tipo lda.
- *newdata:* los datos a clasificar.
- *prior:* probabilidades a priori.
- *dimen:* dimensión que usaremos. Tiene que ser menor o igual que el mínimo entre número de variables y número de grupos menos 1.

los objetos resultantes son:

- *class:* indica el resultado de la clasificación.
- *posterior:* probabilidades a posteriori.
- *x:* puntuaciones discriminantes del individuo.

6.4.2. Opciones gráficas

Para representar de forma gráfica los resultados usaremos las funciones `plot`, `pairs` y `ldahist`.

Función «plot»

La sintaxis para utilizar esta función es:

```
plot(x, panel = panel.lda, ..., cex = 0.7, dimen, abbrev = FALSE,
    xlab = "LD1", ylab = "LD2")
```

Las opciones más interesantes son:

- *x:* objeto de la clase `lda`.
- *panel:* panel de funciones usado en el gráfico de datos.
- *dimen:* número de funciones discriminantes usadas en el gráfico. Hay que tener en cuenta que si la dimensión excede de 2, dibujará los pares de gráficos.

Función «pairs»

Con esta opción obtenemos una matriz de gráficos de las puntuaciones factoriales de los datos tomando las funciones discriminantes por parejas. La sintaxis es:

```
pairs(x, labels = colnames(x), panel = panel.lda, dimen,
      abbrev = FALSE, ..., cex=0.7, type = c("std", "trellis"))
```

Las ordenes más influyentes son:

- *x:* objeto de tipo lda.
- *labels:* etiquetas de las variables.
- *panel:* panel de los gráficos.
- *dimen:* número de funciones discriminantes utilizadas.

Función «ldahist»

Este paquete representa los histogramas y las densidades estimadas de los datos sobre una función discriminante de Fisher

```
ldahist(data, g, nbins = 25, h, x0 = - h/1000, breaks,xlim = range(breaks),
        ymax = 0, width, type = c("histogram", "density", "both"),
        sep = (type != "density"),col = 5,
        xlab = deparse(substitute(data)), bty = "n", ...)
```

donde:

- *data:* vector de datos.
- *g:* vector que especifique la pertenencia de un individuo a cada grupo.
- *type:* tipo de gráfico, histograma, de densidad o ambos.
- *sep:* si hay un gráfico para cada tipo o separado.

6.4.3. Ejemplo de aplicación

En primer lugar tenemos que cargar la librería *MASS* y abrir el fichero de datos. En este caso, el fichero de datos (**discriminante.txt**) contiene para 37 individuos los datos correspondientes a 13 variables denominadas $X1, X2, \ldots, X13$ y una última variable que clasifica a las observaciones en tres grupos.

```
datos<-read.table("discriminante.txt", header=TRUE)
attach(datos)
datos
    X1   X2   X3   X4    X5   X6   X7    X8   X9  X10  X11 X12 X13 discriminate
1  5.7 4.67 17.6 1.50 0.104 1.50 1.88  5.15 8.40  7.5 0.14 205  24             1
2  5.5 4.67 13.4 1.65 0.245 1.32 2.24  5.75 4.50  7.1 0.11 160  32             1
3  6.6 2.70 20.3 0.90 0.097 0.89 1.28  4.35 1.20  2.3 0.10 480  17             1
4  5.7 3.49 22.3 1.75 0.174 1.50 2.24  7.55 2.75  4.0 0.12 230  30             1
5  5.6 3.49 20.5 1.40 0.210 1.19 2.00  8.50 3.30  2.0 0.12 235  30             1
6  6.0 3.49 18.5 1.20 0.275 1.03 1.84 10.25 2.00  2.0 0.12 215  27             1
7  5.3 4.84 12.1 1.90 0.170 1.87 2.40  5.95 2.60 16.8 0.14 215  25             1
8  5.4 4.84 12.0 1.65 0.164 1.68 3.00  6.30 2.72 14.5 0.14 190  30             1
9  5.4 4.84 10.1 2.30 0.275 2.08 2.68  5.45 2.40  0.9 0.20 190  28             1
10 5.6 4.48 14.7 2.35 0.210 2.55 3.00  3.75 7.00  2.0 0.21 175  24             1
11 5.6 4.48 14.8 2.35 0.050 1.32 2.84  5.10 4.00  0.4 0.12 145  26             1
12 5.6 4.48 14.4 2.50 0.143 2.38 2.84  4.05 8.00  3.8 0.18 155  27             1
13 5.2 3.48 18.1 1.50 0.153 1.20 2.60  9.00 2.35 14.5 0.13 220  31             2
14 5.2 3.48 19.7 1.65 0.203 1.73 1.88  5.30 2.52 12.5 0.20 300  23             2
...
```

Sunpondremos que se cumplen las hipótesis de normalidad, igualdad de varianzas-covarianzas y no multicolinealidad. La forma de comprobar estas hipótesis es similar a la que se utiliza para el resto de técnicas (contraste de normalidad, representación gráfica, etc).

Para realizar el análisis discriminante, utilizaremos la siguiente orden:

```
discrimi<-lda(discriminante~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10+X11+X12+X13,
       prior=c(0.33,0.33,0.34), method="moment", tol=0.001)
```

En este caso consideramos que las probabilidades de pertenencia a cada grupo son iguales (por redondeo, la última probabilidad será 0.34). Si quisiéramos que las probabilidades fueran proporcionales al tamaño, bastaría con no indicar nada en el argumento `prior`. Usaremos, además, el método de los momentos para la estimación.

Las salidas que nos ofrece R son las siguientes:

```
Prior probabilities of groups:
   1    2    3
0.33 0.33 0.34

Group means:
        X1       X2       X3       X4        X5       X6
1 5.666667 4.205833 15.89167 1.787500 0.1764167 1.609167
2 5.450000 3.097143 17.82143 1.564286 0.1464286 1.620000
```

```
3 5.500000 2.225455 16.34545 1.881818 0.1440909 1.690000

Coefficients of linear discriminants:
             LD1          LD2
X1  -0.438504888  2.794366461
X2  -2.319914185 -0.259285166
X3  -0.059255729 -0.201559933
X4   0.926748501 -0.906593013
X5   3.834968611 -2.106711106
X6  -1.408127592 -0.911935861
X7  -0.075797389  0.855037049
X8  -0.447283160 -0.427431158
X9  -0.085996563  0.267769265
X10  0.158941483 -0.051526319
X11  3.311102077 -0.505631483
X12 -0.001020368  0.004018043
X13  0.156474826  0.291613389

Proportion of trace:
   LD1    LD2
0.8982 0.1018
```

En primer lugar, se obtiene la tabla con las probabilidades a priori de pertenencia a cada grupo. Estas probabilidades las habíamos indicado al comienzo del análisis, forzando que fueran iguales (`prior=C(0.33,0.33,0.34)`). En segundo lugar, se tienen las puntuaciones medias en cada una de las variables para los grupos diferentes. Estos son los centroides de los grupos. A continuación aparecen los coeficientes de las funciones discriminantes. En nuestro caso son necesarias dos funciones discriminantes y cada una de ellas es función lineal de las 13 variables consideradas. Así, tenemos la expresión de las funciones siguiente (aproximando a la cuarta cifra decimal):

$$D_1 = -0{,}4385X_1 - 2{,}3199X_2 - 0{,}0593X_3 + 0{,}9267X_4 + 3{,}8350X_5 - 1{,}4081X_6$$
$$-0{,}0758X_7 - 0{,}4473X_8 - 0{,}0860X_9 + 0{,}1589X_{10} + 3{,}3111X_{11} - 0{,}0010X_{12} + 0{,}1565X_{13}$$

$$D_2 = 2{,}7944X_1 - 0{,}2593X_2 - 0{,}2016X_3 - 0{,}9066X_4 - 2{,}1067X_5 - 0{,}9119X_6$$
$$+0{,}8550X_7 - 0{,}4274X_8 + 0{,}2678X_9 - 0{,}0515X_{10} - 0{,}5056X_{11} + 0{,}0040X_{12} + 0{,}2916X_{13}$$

Finalmente, tenemos la proporción de varianza explicada por cada eje (el primer eje será mucho más discriminante que el segundo). Con la orden `svd` obtendremos la descomposición en valores singulares, que nos da una medida del poder discriminante de cada función. En nuestro caso se puede comprobar que la primera función es mucho más discriminante que la segunda.

```
discrimi$svd [1] 9.366061 3.153907
```

A continuación, trabajaremos con la función «predict». El objetivo será realizar una predicción sobre un par de individuos. Este par de individuos los registramos, dando sus valores en las 13 variables discriminantes:

```
Datos2<-rbind(c(5.6, 4.2, 15.8, 1.7, 0.17, 1.6, 2.3, 6.0, 4.0, 5.27,
    0.14, 216.25, 26.66), c(5.45,3.09, 17.82, 1.56, 0.14, 1.62, 1.97,
    5.21, 2.68, 7.45, 0.15, 285.71, 23.42))
Datos21<-data.frame(Datos2)
discrimi2<-predict(discrimi,newdata=Datos21,prior=discrimi$prior,2)
discrimi2
```

Obteniendo los siguientes resultados:

```
$class [1] 1 2 Levels: 1 2 3
$posterior
           1          2            3
1 0.98777190 0.01222749 6.094877e-07
2 0.01234011 0.98149260 6.167291e-03

$x
         LD1        LD2
1 -2.6839848  0.4190141
2 -0.1421247 -1.0286646
```

Podemos observar que la probabilidad para el primer individuo de pertenecer al grupo 1, es 0.9877 y la probabilidad del segundo individuo de pertenecer al grupo 2 es 0.9814, por lo que la clasificación de las funciones calculadas es que el primer individuo pertenece al grupo 1 y el segundo individuo pertenece al grupo 2.

También podemos calcular la matriz de confusión mediante la orden:

```
table(predict(discrimi)$class, discriminante)
discriminante
     1  2  3
  1 12  3  0
  2  0  9  0
  3  0  2 11
```

donde observamos que hay 5 individuos mal clasificados, todos pertenecientes al grupo 2; 3 se clasifican en el grupo 1 y 2 se clasifican en el grupo 3.

Finalmente, obtendremos las representaciones gráficas. Con la orden `plot(discrimi)` obtenemos el gráficos de las puntuaciones discriminantes que se puede ver en la Figura 6.7

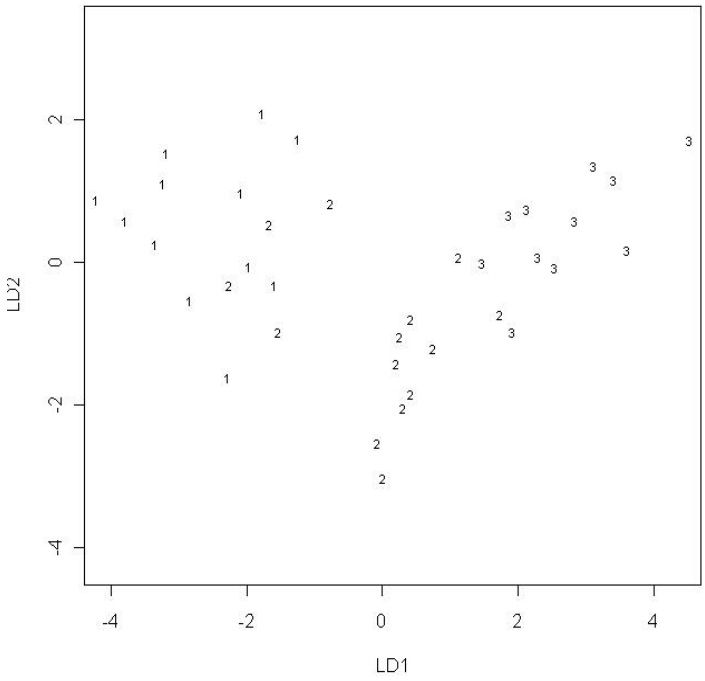

Figura 6.7: Gráfico de las puntuaciones discriminantes

La orden `pairs(discrimi)` produce los pares de gráficos de las puntuaciones discriminantes de los datos, como se aprecia en la Figura 6.8, que en este caso es similar a la ya obtenida con la función anterior pero que, en caso de estar trabajando con más de dos funciones discriminantes, será de utilidad.

El último conjunto de gráficos serán los histogramas de las variables dependientes frente a la variable de agrupación. Siempre es de utilidad representar todas las variables $X1, X2, \ldots, X13$ frente a *discriminante*. Para este ejemplo, realizamos solamente el gráfico de $X2$ frente a *discriminante*. Podemos observar en la Figura 6.9 que existe un comportamiento diferente de la variable dependiendo del grupo.

Una vez obtenidos los resultados, será preciso validar el modelo mediante la comprobación de las hipótesis iniciales de normalidad, homoscedasticidad y no multicolinealidad. Este último paso es común a las técnicas multivariantes habituales, por lo que en este tema no se van a revisar de nuevo.

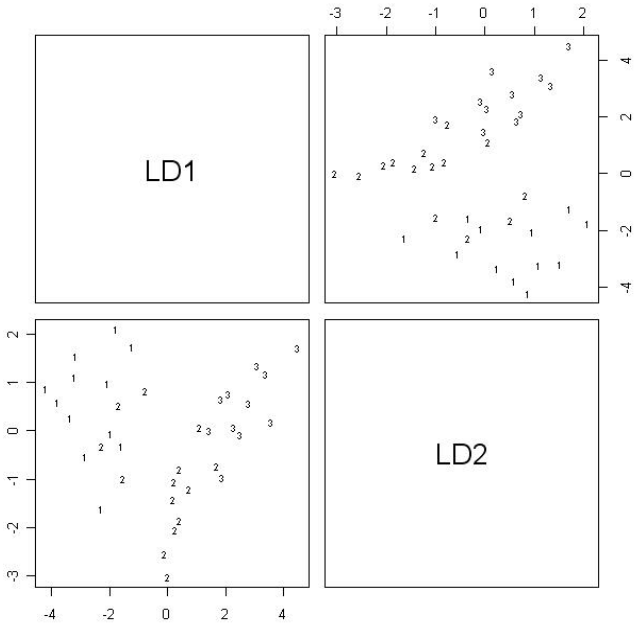

Figura 6.8: Gráfico de puntuaciones discriminantes de los datos

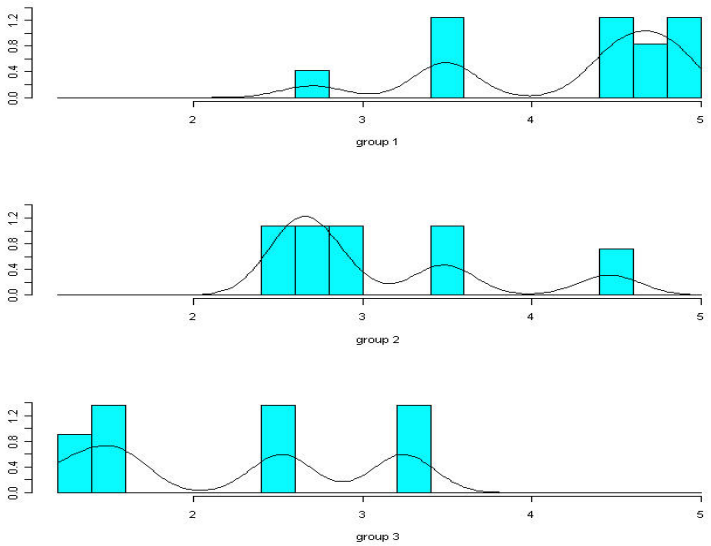

Figura 6.9: Histograma de la variable X_2 en los diferentes grupos

Bibliografía

[1] Carrasco, J.L.; Hernán, M.A. *Estadística multivariante en las ciencias de la vida. Fundamentos, métodos y aplicación*; Ciencia 3, D.L., 1993

[2] Cea, M.A. *Análisis Discriminante*; Colección Cuadernos Metodológicos nº54, Centro de Investigaciones Sociológicas, 2016.

[3] Cuadras, C.M. *Nuevos Métodos del Análisis Multivariante*; CMC, 2018.

[4] Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2), 179–188.

[5] Gnanadeskian, R. Methods for statistical data analysis of multivariate observations; (Vol 321) John Wiley Sons, 2011.

[6] Rencher, A.C. *Methods of Multivariate Analysis*; Wiley, N. York, 1995.

[7] Classification Methods Essentials. Discriminant Analysis Essentials in R `http://www.sthda.com/english/articles/36-classification-methods-essentials/146-discriminant-analysis-essentials-in-r/`