ANÁLISIS DISCRIMINANTE

Otra vía de hacer extracción de conocimiento es mediante la identificación de las variables que mejor identifican una similitud. Es decir discriminan, diferencian, apartan entre grupos. El análisis más sencillo es entre dos grupos (bivariante) y el más complejo discrimina entre muchos grupos (multivariado). Esto es Análisis Discriminante (AD). Ej.: Identificar los factores que diferencien una situación de déficit o no de viviendas en un centro poblado y a su vez, predice el nivel deficitario, o también, clasificar de acuerdo al riesgo financiero.

Esta selección se refiere a procesos donde el espacio original de los datos es transformado a un nuevo espacio de variables que expresan la combinación lineal de las variables originales que tiene una alta relación entre ellas.

- Es una técnica adicional de reducción de variables

- Así como la varianza total juega un papel importante en los ACP, en esta técnica la varianza entre los grupos y dentro del grupo definen la discriminación. *Mientras mayor es la razón entre la suma de cuadrados entre los grupos y la suma de cuadrados dentro del grupo,*

mayor es el nivel de discriminación.

- El resultante es una función discriminante que transforma las variables originales x_i en una nueva variable Z denominada variable discriminante.

- El análisis discriminante puede comprender desde una función (caso de dos variables) a varias funciones (discriminante múltiple)

- El análisis discriminante es clasificador.

GEOMETRÍA

Para comprender como se interpreta el AD, veamos el caso más simple: dos grupos. Se deben reunir los tres siguientes objetivos:

a) *Identificar el conjunto de variable que mejor discriminan entre dos grupos; a estas variables se le conocen como variables discriminatorias*
b) *Identificar los nuevos ejes,* **z**, *tal que la nueva variable z, dada por la proyección de las observaciones en los nuevos ejes, provea la máxima separación o discriminación entre los dos grupos.*
c) *Usar esta nueva variable para desarrollar una regla de clasificación*

entre los dos grupos para las nuevas observaciones.

ANÁLISIS DISCRIMINANTE PARA DOS GRUPOS

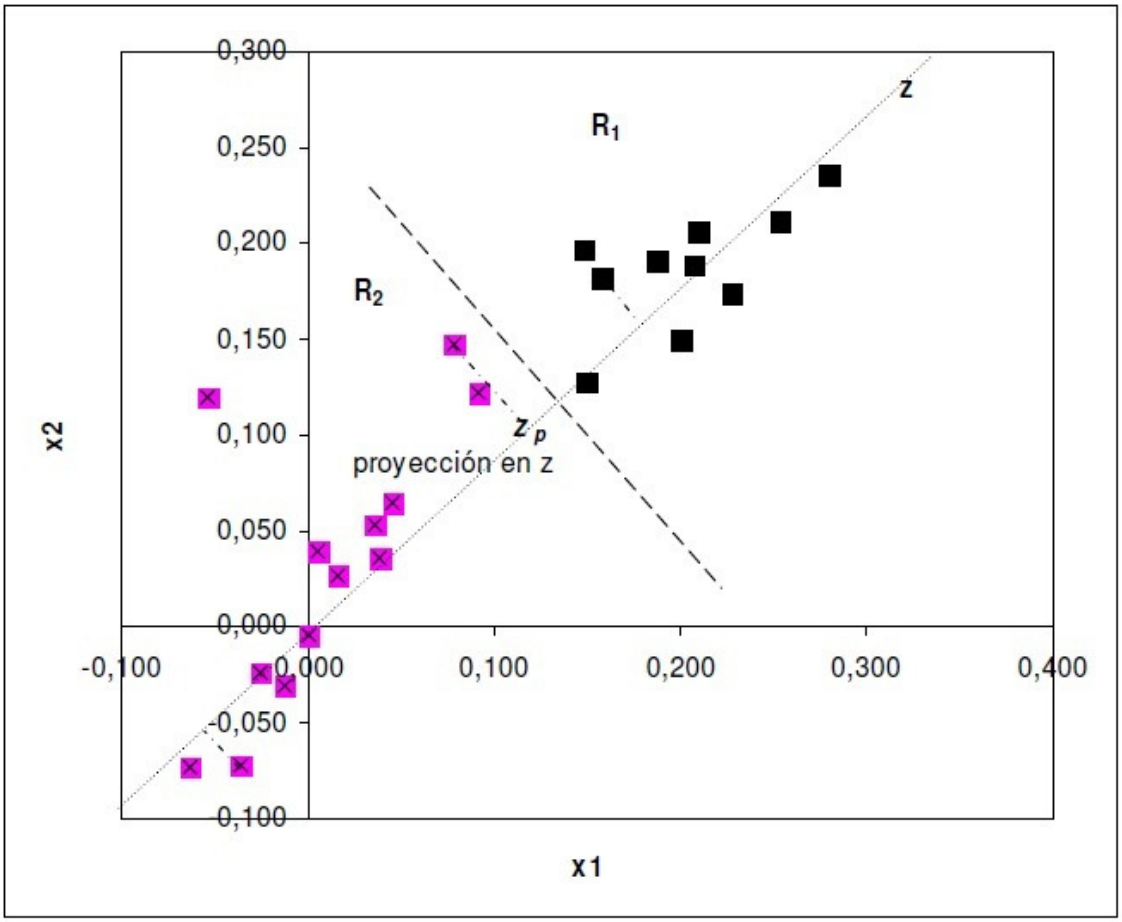
Supongamos que tenemos dos variables x_1 y x_2 que representan las variables originales. Estas variables han sido ordenadas de acuerdo a una característica de discriminación. Siguiendo el ejemplo de la tabla en la página 238 en Subash, Capítulo 8, se puede obtener una idea mediante un ejemplo básico con dos variables solamente y 12 observaciones para cada caso de discriminación.

Veamos la tabla financiera de veinticuatro empresas, de las cuales hay doce más aventajadas y las otra doce menos. Se muestran la variables x_1 que identifica la razón financiera *ganancias antes de los intereses a los activos totales* y x_2 es la razón financiera que representa *el retorno al capital total*.

Grupo 1: Más aventajadas				Grupo2: Menos aventajadas			
Empresa	x1	x2	Z	Empresa	x1	x2	Z
1	0,158	0,182	0,240	13	-0,012	-0,031	-0,030
2	0,210	0,206	0,294	14	0,036	0,053	0,063
3	0,207	0,188	0,279	15	0,038	0,036	0,052
4	0,280	0,236	0,365	16	-0,063	-0,074	-0,097
5	0,197	0,193	0,276	17	-0,054	-0,119	-0,122
6	0,227	0,173	0,283	18	0,000	-0,005	-0,004
7	0,148	0,196	0,243	19	0,005	0,039	0,031
8	0,254	0,212	0,329	20	0,091	0,122	0,151
9	0,079	0,147	0,160	21	-0,036	-0,072	-0,076
10	0,149	0,128	0,196	22	0,045	0,064	0,077
11	0,200	0,150	0,247	23	-0,026	-0,024	-0,035
12	0,187	0,191	0,267	24	0,016	0,026	0,030

Pensemos que el resultado esté representado, tal como lo muestra el grafico más abajo, por las variables x_1 y x_2 define los ejes en un plano bidimensional. Una primera aproximación sería la discriminación de ambos grupos de acuerdo a la razón propiamente dicha (R_1 y R_2). En este caso se aprecia que los puntos negros se refieren a R_1 y los puntos rojos a la razón financiera R_2. Al examinar las diferencias entre grupos con respecto a una variable se le conoce como *univariante*. Cada razón discrimina entre los dos grupos. El término univariante resalta el hecho de que las diferencias entre los dos grupos son evaluadas para cada variable, independiente de las variables restantes. Se ve claro que los dos grupos están bien separados en un espacio bidimensional, lo cual implica que ambas razones combinadas o conjuntamente proveen una buena separación

de los dos grupos de empresas. Cuando la diferencia es con respecto a dos o más variables simultáneamente, se refiere a análisis *multivariante*. El término multivariante es para resaltar que las diferencias en las medias de los dos grupos son evaluadas simultáneamente para todas las variables.

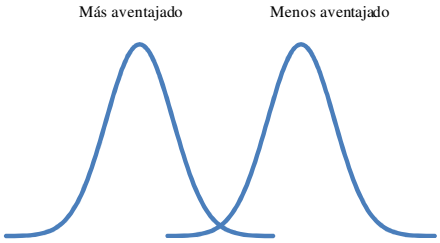


Los datos originales de las razones financieras $\boldsymbol{x}_1$ y $\boldsymbol{x}_2$ con su proyección sobre el nuevo eje $\mathbf{z}$ para definir la función discriminante z

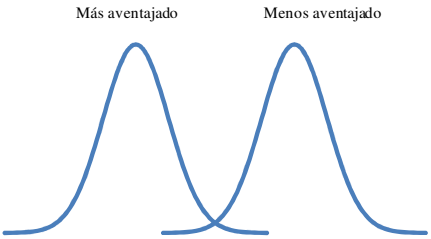
Consideremos, por ejemplo, que las observaciones para los dos grupos de

empresas además de tabular dos razones financieras, se agregan dos más. Esto es, fueron tabuladas x_1, x_2, x_3 y x_4. Para dar una idea de cómo se pueden identificar las variables que mejor discriminan, en la figura más abajo se muestra como sería la distribución de cada razón financiera. En las dos primeras se puede observar que hay una aparente diferencia entre las empresas más y menos aventajadas con respecto a las razones x_1 y x_2 que con x_3 y x_4. Esto es, las razones x_1 y x_2 son las que proveen mayor discriminación entre los dos grupos. *De ahí que el primer objetivo es identificar el conjunto de variable que mejor discriminan entre dos grupos; a estas variables se le conocen como variables discriminatorias.*

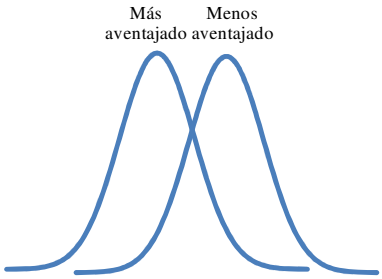
Razón financiera x_1



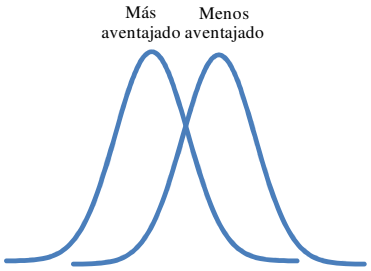
Razón financiera x_2



Razón financiera x_3



Razón financiera x_4



El segundo objetivo es identificar los nuevos ejes, **z***, tal que la nueva variable z, dada por la proyección de las observaciones en los nuevos ejes, provea la máxima separación o discriminación entre los dos grupos.*

Para el caso de dos grupos, esta recta debe ser la que mejor discrimine las variables originales. Para ello, como ya se mencionó, esa recta será la que la razón de la suma de cuadrados entre los dos grupos con la suma de cuadrados entre grupos sea mayor.

La función discriminante para dos grupos es de la forma $z_p = \cos\theta.x_1 + sen\theta.x_2$, donde θ, para el ejemplo, describe el ángulo entre el eje $\mathbf{x_1}$ y la recta de proyección **z**. La selección del valor de θ optimiza la relación entre las suma de cuadrados antes descrita.
Todos los valores de las dos variables x_1 y x_2 deben ser proyectados al nuevo eje. Estos vienen representados por la función $z = \cos\theta.x_1 + sen\theta.x_2$. Si por ejemplo θ es 45°, entonces la función resultaría en $z = x_1 \cos 45 + x_1 sen 45$ o lo que es lo mismo $Z = 0.707.x_1 + 0.707.x_2$. Los valores de z que se corresponden, aparecen en la tabla anterior en las columnas Z para cada grupo..

Los valores de la sumas de cuadrados (dentro del grupo (SS_w) y entre

grupos (SS_b)) se calculan de acuerdo a los valores suministrados por la función z y al valor de θ utilizado. Para el caso del ejemplo mostrado donde θ es 45°, los valores de las sumas de cuadrados de la variable z entre los grupos (SS_b) y dentro del grupo (SS_w), son como sigue.

N°	Tamaño	Z				Razón	Teta
grupo	grupo	Media	SSb	SSw	SSt	SSb/SSw	
1	12	0,265	0,411	0,1021	0,513	4,0195	45°
2	12	0,003					
Total	24	0,134					

Este cálculo es pertinente debido a que la selección de la mejor discriminación sería aquella correspondiente al máximo valor de la razón SS_b/SS_w del conjunto de valores de la variable z obtenidos para cada ángulo θ con respecto a la variable x_1.

En este ejercicio la mejor razón la determina θ=21° y la función discriminante es $z = x_1 \cos 21 + x_1 sen 21 = 0.934 x_1 + 0.358 x_2$.

Entonces la máxima separación es satisfecha cuando se cumple estas dos condiciones:

- La media de z para los dos grupos serían lo más alejadas cuanto sea posible. Esto equivale a tener un máximo valor de SS_b.
- Los valores de z para cada grupo serían tan homogéneos como sea posible. Esto equivale a tener un máximo valor de SS_w.

Lo similar de este análisis con ACP es que en ambos casos se identifica un nuevo eje y las nuevas variables (variables latentes) son una combinación lineal de las variables originales.

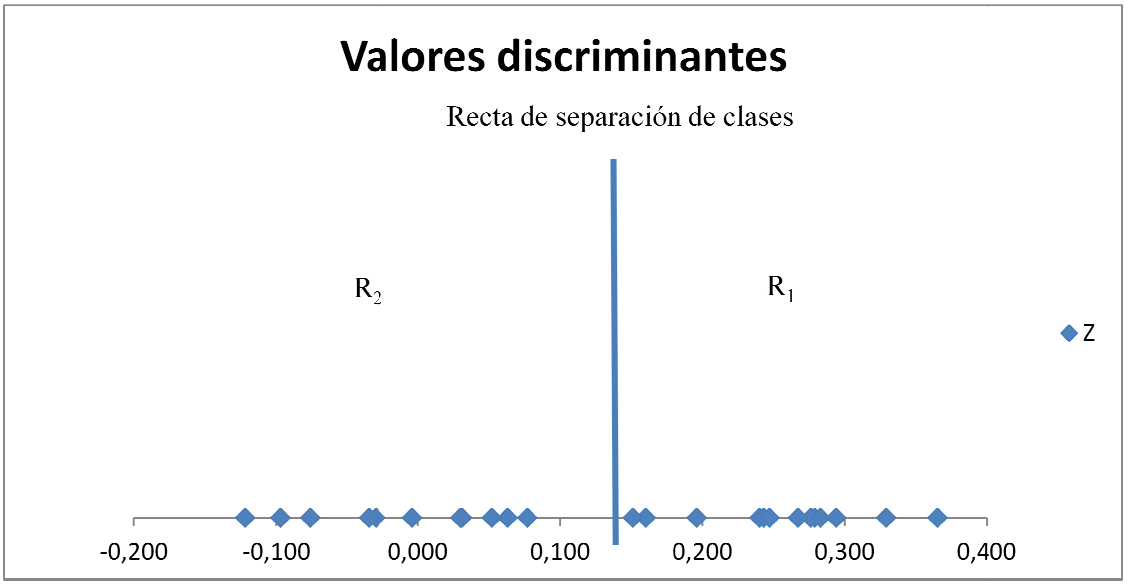
La diferencia entre AD y ACP es el criterio que se emplea para identificar los nuevos ejes.

- En ACP, un nuevo eje se identifica cuando la proyección de los puntos en ese eje contabiliza una máxima varianza en los datos (equivale a maximizar SS_t), mientras que,
- AD no contabiliza la máxima varianza en los datos (máximo SS_t), sino maximiza la razón entre SS_b y SS_w (por ejemplo SS_b/SS_w)n que resulta en la mejor discriminación entre los grupos.

El tercer objetivo del análisis es clasificar las nuevas observaciones entre los grupos.

La clasificación para futuras observaciones se hace mediante los scores. En la figura siguiente muestra los valores (scores) en un plano unidimensional que son producto de resolver la expresión de la función discriminante z para cada par observación en cada uno de los grupos. En nuestro caso si la función óptima fuera para θ=45°; es decir $z = 0.707x_1 + 0.707x_2$, entonces los

valores discriminantes obtenidos en la tabla anterior, serían los de más abajo. La clasificación de las observaciones se hace



Clasificación de la función discriminante en dos regiones

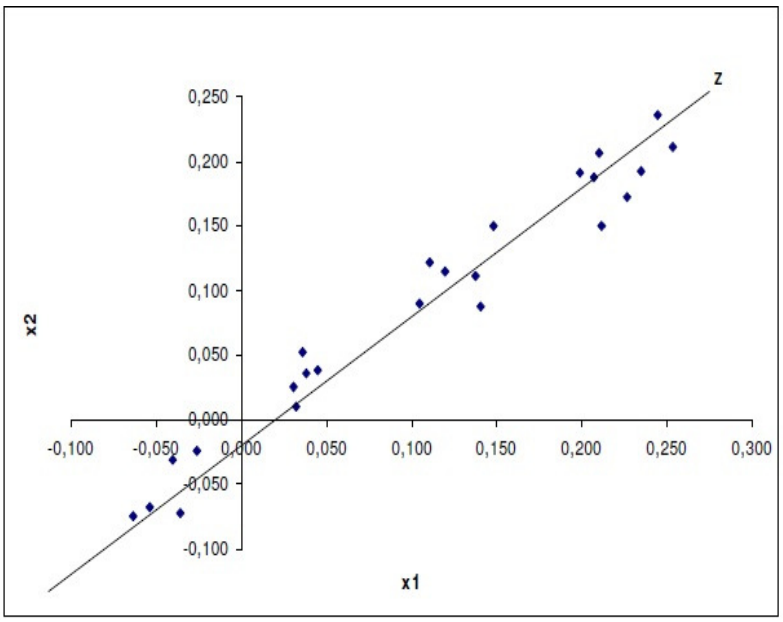
ANÁLISIS DISCRIMINANTE PARA MÁS DE DOS GRUPOS

Es similar al caso para dos grupos, pero están envueltas más de dos clases. Por ejemplo, sea el caso que desea discriminar la banca universal de acuerdo a su gestión de riesgo global en niveles de riesgo bajo, moderado y alto.

La principal diferencia en los objetivos que se persiguen con el caso de dos grupos es que la función discriminante no se puede construir como

una única función para todas las diferencias entre los grupos, sino hay que identificar funciones adicionales. Esto es, adicional a los tres objetivos anteriores, en este caso, se debe *identificar el mínimo número de funciones discriminantes que puedan proveer la mayor cantidad de discriminaciones entre grupos.*

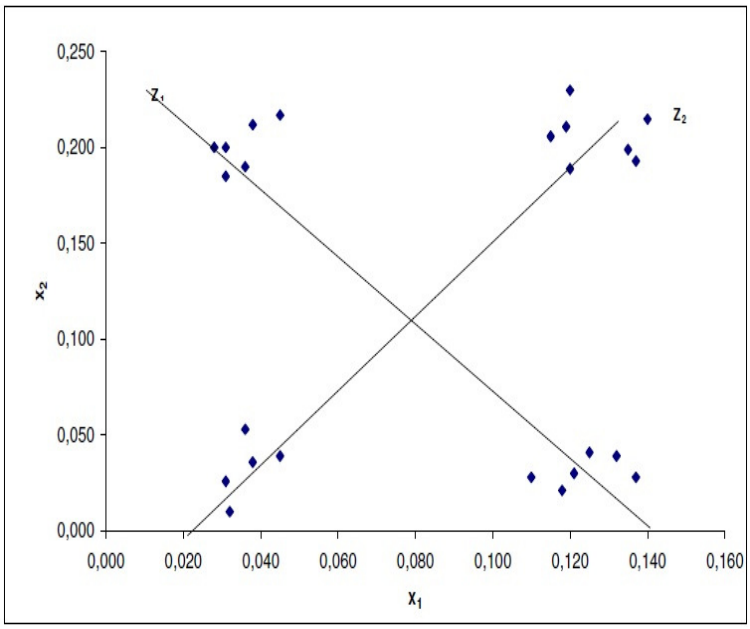
Supongamos que tenemos un gráfico como el que sigue. En él se pueden observar cuatro conjunto de observaciones representadas en dos es variables x_1 y x_2, que se pueden diferenciar de acuerdo a una función discriminante z, pudiéndose proyectar cada una de las observaciones sobre ese eje discriminante, tal como se vio en el caso de discriminación en dos grupos.



Los datos originales $\boldsymbol{x}_1$ y $\boldsymbol{x}_2$ y su proyección sobre el nuevo eje **z**

Sin embargo en el gráfico más abajo pareciera que la disposición de las observaciones correspondientes a las dos variables x_1 y x_2 no definiera dos grupos sino cuatro. Veamos, no como en el caso anterior, se pueden diferenciar dos grupos de acuerdo a la función discriminante $\mathbf{z_1}$. Los otros dos grupos no se pueden diferenciar y pareciera que se proyectaran sobre esta función $\mathbf{z_1}$, superponiéndose sobre ese eje discriminante. Si define uno nuevo $\mathbf{z_2}$, que no necesariamente debe ser ortogonal al $\mathbf{z_1}$, pero si no correlacionado, pudiera diferenciarse estos dos nuevos grupos.

De este modo en ambos casos, para lo que podría considerarse cuatros grupos que se pudieran discriminar, en un caso, con una sola función se pudo lograr dos grupos bien diferenciados. En el otro caso, con dos funciones se logra separar los cuatro grupos.



Las observaciones en $\mathbf{x_1}$ y $\mathbf{x}_2$ con sus propias proyecciones sobre los nuevos eje $\mathbf{z_1}$ y $\mathbf{z_2}$ para definir la funciones discriminantes Z_1 *y* Z_2

En el caso de múltiples grupos, el número de funciones discriminantes debe cumplir la regla de ***min(G-1,p).*** Donde *G* es el número de grupos y *p* el número de variables discriminantes.

Los nuevos ejes discriminantes no tienen porqué ser ortogonales. En este caso los valores de los z_i obtenidos de las funciones discriminantes $\boldsymbol{Z_i}$ deben ser *no correlacionados*.

Se restringe la capacidad de ver la clasificación eventualmente al caso de dos grupos. Es bastante complejo representar la clasificación de múltiples

grupos.

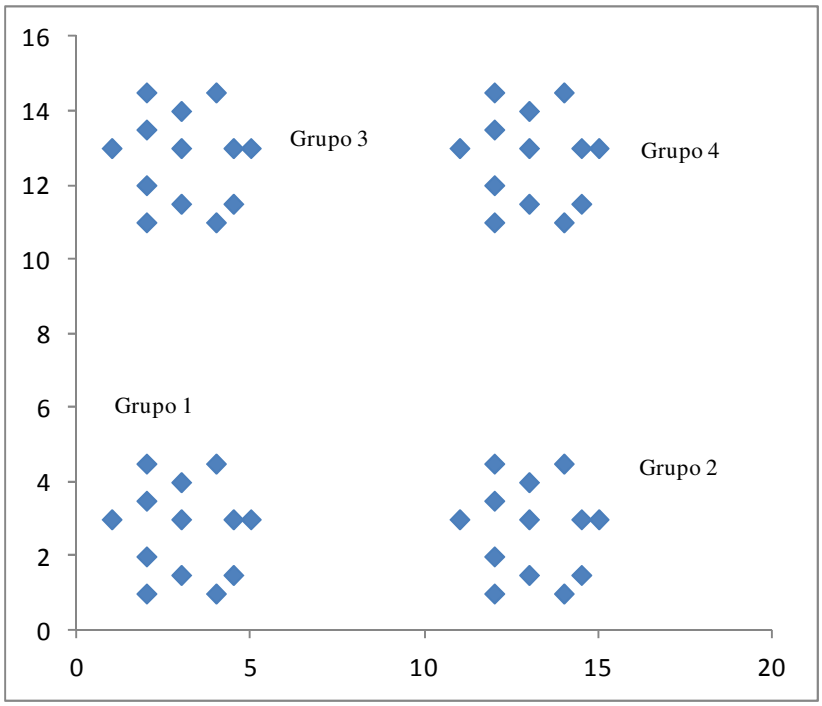
Veamos un ejemplo similar para visualizar gráficamente la discriminación de más de dos grupos.

Sea las siguientes observaciones teóricas.

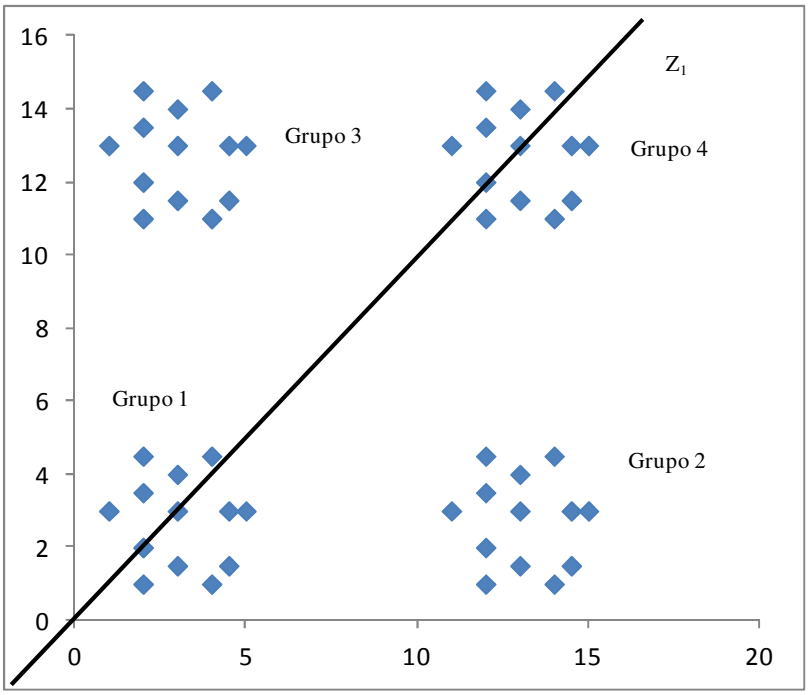
DATOS PARA CUATRO GRUPOS								
	Grupo 1		Grupo 2		Grupo 3		Grupo 4	
Obs.	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
1	1	3	11	3	1	13	11	13
2	2	1	12	1	2	11	12	11
3	4	1	14	1	4	11	14	11
4	5	3	15	3	5	13	15	13
5	4	4,5	14	4,5	4	14,5	14	14,5
6	2	4,5	12	4,5	2	14,5	12	14,5
7	3	1,5	13	1,5	3	11,5	13	11,5
8	4,5	1,5	14,5	1,5	4,5	11,5	14,5	11,5
9	4,5	3	14,5	3	4,5	13	14,5	13
10	3	4	13	4	3	14	13	14
11	2	3,5	12	3,5	2	13,5	12	13,5
12	2	2	12	2	2	12	12	12
13	3	3	13	3	3	13	13	13
Media	3,0769	2,7308	13,0769	2,7308	3,0769	12,7308	13,0769	12,7308
Desv. S	1,2391	1,2352	1,2391	1,2352	1,2391	1,2352	1,2391	1,2352

Dos variables, trece observaciones por grupo y cuatro grupos.

Si los graficamos podemos distinguir aparentemente cuatro grupos con cuatro medias para cada grupo aparentemente diferentes



Los grupos 1 y 4 claramente se pueden discriminar con un eje Z. En ese mismo eje no sucede así con los grupos 2 y 3 (se solaparían). Veamos.



Para ello requerimos definir un nuevo eje discriminante Z_2 que defina la separación de los grupos 3 y 2. Analíticamente, sería, tal como se mostró en el caso simple, hallar el valor del ángulo θ que maximice la relación entre las sumas de cuadrados SS_b / SS_w para hallar el primer eje Z_1. El opuesto, es aquel que minimice esa misma relación. Este sería el segundo eje.

Para el máximo lambda es

Grupo	Media	θ	SS_b	SS_w	λ
1	4,1011	46,1145	1300,0000	67,5340	19,2496
2	11,0335				
3	11,3082				
4	18,2406				
Total	11,1709				

y para el mínimo lambda es

Grupo	Media	θ	SS_b	SS_w	λ
1	4,1011	**46,115**	1300,0000	67,5340	19,2496
2	11,0334				
3	11,3083				
4	18,2406				
Total	11,1709				

El detalle para el cálculo de los valores de SS_b y SS_w para cada una de las variables *j*, es de acuerdo a las siguientes expresiones

$SSb_j = \sum_{g=1}^{G} n_g (\bar{x}_{jg} - \bar{x}_{j.})^2$ j=1,….p, donde G es el número de grupos, $\bar{x}_{jg}$ es la media de la variable *j* para el grupo *g*, $\bar{x}_{j.}$ es la media de la variable *j* para todas las observaciones.

$SS_w = \sum_{g=1}^{G} \sum_{j=1}^{ng} (Z_j - \bar{Z}_g)^2$, donde Z_j es el score *j* del grupo *g* y $\bar{Z}_g$ es la media del grupo correspondiente G.

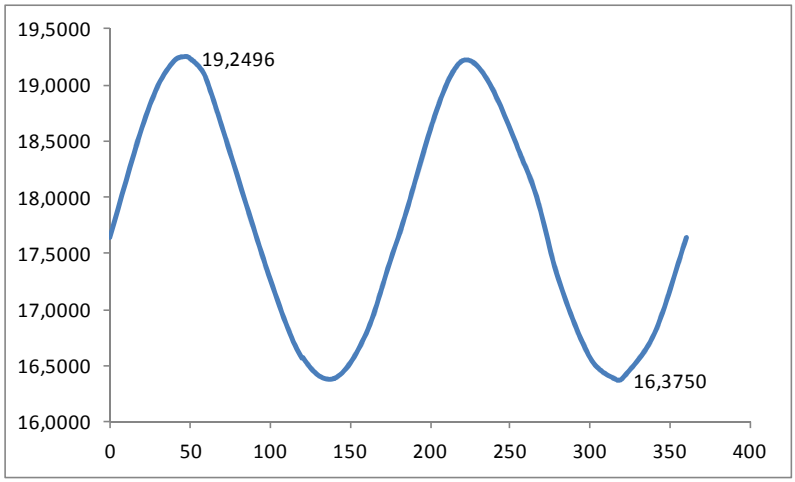
El detalle para el lambda máximo es como sigue.

					$\cos\theta$	$sen\,\theta$		SS_w				
Grupo	x_1	x_2	°C	θ	w_1	w_2	Z_θ	$Z_\theta - \overline{Z}_{G1}$	$Z_\theta - \overline{Z}_{G2}$	$Z_\theta - \overline{Z}_{G3}$	$Z_\theta - \overline{Z}_{G4}$	$(Z_\theta - \overline{Z}_{G_i})$^2
1	1	3	46,1	0,8048	0,6932	0,7207	2,8554	-1,2458				1,5519
1	2	1	46,1	0,8048	0,6932	0,7207	2,1072	-1,9939				3,9758
1	4	1	46,1	0,8048	0,6932	0,7207	3,4937	-0,6075				0,3690
1	5	3	46,1	0,8048	0,6932	0,7207	5,6283	1,5272				2,3323
1	4	4,5	46,1	0,8048	0,6932	0,7207	6,0161	1,9150				3,6673
1	2	4,5	46,1	0,8048	0,6932	0,7207	4,6297	0,5285				0,2794
1	3	1,5	46,1	0,8048	0,6932	0,7207	3,1608	-0,9404				0,8843
1	4,5	1,5	46,1	0,8048	0,6932	0,7207	4,2006	0,0995				0,0099
1	4,5	3	46,1	0,8048	0,6932	0,7207	5,2817	1,1806				1,3937
1	3	4	46,1	0,8048	0,6932	0,7207	4,9626	0,8614				0,7420
1	2	3,5	46,1	0,8048	0,6932	0,7207	3,9090	-0,1922				0,0369
1	2	2	46,1	0,8048	0,6932	0,7207	2,8279	-1,2732				1,6211
1	3	3	46,1	0,8048	0,6932	0,7207	4,2418	0,1407				0,0198
2	11	3	46,1	0,8048	0,6932	0,7207	9,7877		-1,2458			1,5519
2	12	1	46,1	0,8048	0,6932	0,7207	9,0395		-1,9939			3,9758
2	14	1	46,1	0,8048	0,6932	0,7207	10,4260		-0,6075			0,3690
2	15	3	46,1	0,8048	0,6932	0,7207	12,5607		1,5272			2,3323
2	14	4,5	46,1	0,8048	0,6932	0,7207	12,9485		1,9150			3,6673
2	12	4,5	46,1	0,8048	0,6932	0,7207	11,5620		0,5285			0,2794
2	13	1,5	46,1	0,8048	0,6932	0,7207	10,0931		-0,9404			0,8843
2	14,5	1,5	46,1	0,8048	0,6932	0,7207	11,1330		0,0995			0,0099
2	14,5	3	46,1	0,8048	0,6932	0,7207	12,2141		1,1806			1,3937
2	13	4	46,1	0,8048	0,6932	0,7207	11,8949		0,8614			0,7420
2	12	3,5	46,1	0,8048	0,6932	0,7207	10,8413		-0,1922			0,0369
2	12	2	46,1	0,8048	0,6932	0,7207	9,7603		-1,2732			1,6211
2	13	3	46,1	0,8048	0,6932	0,7207	11,1742		0,1407			0,0198
3	1	13	46,1	0,8048	0,6932	0,7207	10,0625			-1,2458		1,5519
3	2	11	46,1	0,8048	0,6932	0,7207	9,3143			-1,9939		3,9758
3	4	11	46,1	0,8048	0,6932	0,7207	10,7008			-0,6075		0,3690
3	5	13	46,1	0,8048	0,6932	0,7207	12,8354			1,5272		2,3323
3	4	14,5	46,1	0,8048	0,6932	0,7207	13,2232			1,9150		3,6673
3	2	14,5	46,1	0,8048	0,6932	0,7207	11,8368			0,5285		0,2794
3	3	11,5	46,1	0,8048	0,6932	0,7207	10,3679			-0,9404		0,8843
3	4,5	11,5	46,1	0,8048	0,6932	0,7207	11,4077			0,0995		0,0099
3	4,5	13	46,1	0,8048	0,6932	0,7207	12,4888			1,1806		1,3937
3	3	14	46,1	0,8048	0,6932	0,7207	12,1697			0,8614		0,7420
3	2	13,5	46,1	0,8048	0,6932	0,7207	11,1161			-0,1922		0,0369
3	2	12	46,1	0,8048	0,6932	0,7207	10,0350			-1,2732		1,6211
3	3	13	46,1	0,8048	0,6932	0,7207	11,4489			0,1407		0,0198
4	11	13	46,1	0,8048	0,6932	0,7207	16,9948				-1,2458	1,5519
4	12	11	46,1	0,8048	0,6932	0,7207	16,2467				-1,9939	3,9758
4	14	11	46,1	0,8048	0,6932	0,7207	17,6331				-0,6075	0,3690
4	15	13	46,1	0,8048	0,6932	0,7207	19,7678				1,5272	2,3323
4	14	14,5	46,1	0,8048	0,6932	0,7207	20,1556				1,9150	3,6673
4	12	14,5	46,1	0,8048	0,6932	0,7207	18,7691				0,5285	0,2794
4	13	11,5	46,1	0,8048	0,6932	0,7207	17,3002				-0,9404	0,8843
4	14,5	11,5	46,1	0,8048	0,6932	0,7207	18,3401				0,0995	0,0099
4	14,5	13	46,1	0,8048	0,6932	0,7207	19,4212				1,1806	1,3937
4	13	14	46,1	0,8048	0,6932	0,7207	19,1020				0,8614	0,7420
4	12	13,5	46,1	0,8048	0,6932	0,7207	18,0484				-0,1922	0,0369
4	12	12	46,1	0,8048	0,6932	0,7207	16,9674				-1,2732	1,6211
4	13	13	46,1	0,8048	0,6932	0,7207	18,3813				0,1407	0,0198

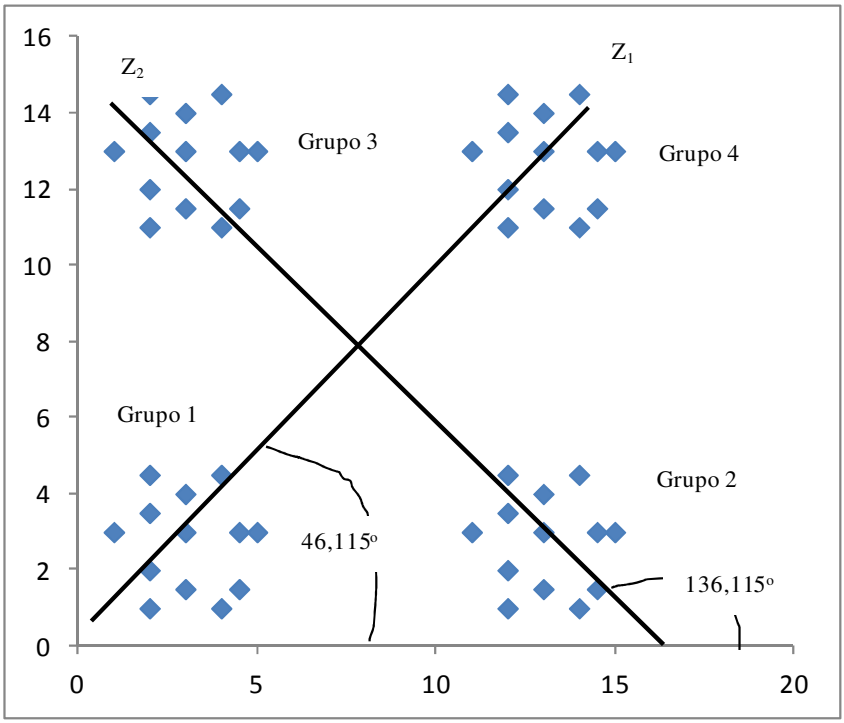
La siguiente tabla muestra los valores óptimos de lambda,

θ	SS_b	SS_w	λ
0	1300,0000	73,6923	17,6409
10	1300,0000	71,6526	18,1431
20	1300,0000	69,8311	18,6163
30	1300,0000	68,4475	18,9927
40	1300,0000	67,6686	19,2113
46,115	**1300,0000**	**67,5340**	**19,2496**
50	1300,0000	67,5883	19,2341
60	1300,0000	68,2164	19,0570
100	1300,0000	75,2699	17,2712
120	1300,0000	78,4753	16,5657
140	1300,0000	79,3348	16,3862
160	1300,0000	77,4463	16,7858
180	1300,0000	73,6934	17,6407
220	1300,0000	67,6688	19,2112
260	1300,0000	71,2174	18,2540
280	1300,0000	75,2689	17,2714
300	1300,0000	78,4747	16,5658
316,2	**1300,0000**	**79,3891**	**16,3750**
320	1300,0000	79,3350	16,3862
340	1300,0000	77,4471	16,7856
360	1300,0000	73,6945	17,6404

Gráficamente se observan los valores extremos



Los ejes que separarían los cuatro grupos, Z_1 y Z_2, gráficamente son



Este mismo ejercicio se puede practicar mediante MATLAB. Para ellos se desarrolló un laboratorio sencillo **lab_ad_2** que aparece con un vínculo en el sitio web.

Métodos alternativos, no paramétricos para realizar este tipo de clasificaciones es la red neuronal basada en funciones de base radial y máquinas de vectores de soporte.

Otro método de discriminación comúnmente utilizado es la Regresión Logística. En estos casos las observaciones que componen las variables independientes pueden ser cuantitativas o categóricas.