UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

"ANÁLISIS MULTIVARIADO: TÉCNICAS Y APLICACIONES"

T E S I S

QUE PARA OBTENER EL TÍTULO DE

A C T U A R Í A

P R E S E N T A :

NORMA ARACELI ESTRADA MENDOZA

DIRECTOR DE TESIS

ACT. JAIME VÁZQUEZ ALAMILLA

2004

ACT. MAURICIO AGUILAR GONZÁLEZ
Jefe de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo escrito:

"Análisis multivariado : técnicas y aplicaciones"

realizado por Norma Araceli Estrada Mendoza.

con número de cuenta 9201465-3 **, quien cubrió los créditos de la carrera de:** Actuaría.

Dicho trabajo cuenta con nuestro voto aprobatorio.

A t e n t a m e n t e

Director de Tesis Propietario	Act. Jaime Vázquez Alamilla
Propietario	M. en A.P. María del Pilar Alonso Reyes.
Propietario	M. en C. José Antonio Flores Díaz.
Suplente	Act. Marisa Miranda Tirado.
Suplente	Dr. Luis Antonio Rincon Solis.

Consejo Departamental de Matemáticas

Act. Jaime Vázquez Alamilla.

FACULTAD DE CIENCIAS
CONSEJO DEPARTAMENTAL DE MATEMATICAS

AGRADECIMIENTOS

Antes que nada quiero agradecer a *Dios* por permitirme terminar este trabajo.

A mis padres *Enrique y María Guadalupe*, a mis hermanas *Leticia, Pita, Irene, Nancy e Isela*, por el apoyo a lo largo de toda mi carrera.

A mi "*abuelito*" desde donde este, gracias por todo su amor.

De manera muy especial a *Jaime Vázquez* ya que sin la paciencia, apoyo y confianza que me brindaste no hubiera podido estar aquí. Por aceptar ser mi director sin conocerme, y sobre todo por tus valiosos comentarios y acertadas observaciones para este trabajo.

A *Pilar Alonso* por sus valiosas observaciones y correcciones. Por mostrarme lo maravillosos de la *Estadística.*

A *José Antonio* y *Luis Rincón* por aceptar y darse el tiempo de revisar este trabajo.

Sin olvidar a *Margarita Chávez* y *Marisa Miranda* por sus palabras, su confianza. Por mostrarme lo fascinante de la Estadística y la Probabilidad, también por permitirme trabajar con ellas y desarrollarme en la docencia.

Al profesor *Miguel Gutierrez* por su tiempo, sus palabras y su apoyo.

A *Miguel Santa Rosa* que ha sido mi amigo y compañero desde el inicio de nuestras carreras.

A mis amigos: *Jean Paul, David, Dalia, Karina, Katia, Erika, Mónica, Yareli, Mauricio,* y a los que olvide mencionar.

A *Claudia y Rafael* por su gran apoyo, y por haber sido tan buenos alumnos.

A la *UNAM.*

GRACIAS

Índice General

Prefacio

La teoría del análisis multivariado tiene su inicio en 1930. El término análisis multivariado se refiere al estudio de muchas variables que efectivamente incluyen varios casos, es decir, se refiere al estudio de n vectores aleatorios ($\underline{X}_1$ a $\underline{X}_n$) cada uno de dimensión p.

Surgen varias preguntas, como: ¿Qué es el análisis multivariado?, ¿Para qué sirve? y ¿Qué beneficios se pueden obtener?.

Mucho de este tema se refiere al estudio y análisis de los datos, en relación con la inferencia estadística. Las técnicas multivariadas son aplicadas en la industria, la investigación antropológica y zoológica, entre otras áreas.

El análisis multivariado no es fácil de definir, precisamente debido al término *multivariado*, el cual no es usado frecuentemente, ya que se refiere a métodos estadísticos. Cualquier análisis de más de dos variables puede ser considerado vagamente como análisis múltiple; como tal, muchas técnicas multivariadas se generalizaron del análisis univariado (análisis de una sola variable) y del análisis bivariado (correlaciones, análisis de varianza y regresión simple). En varias ocasiones se tiene el supuesto de que todas las variables múltiples tienen la distribución normal.

Es imposible discutir las aplicaciones de las técnicas multivariadas sin discutir el desarrollo de la computación. La teoría de estas técnicas estuvieron bien desarrolladas antes del surgimiento de las computadoras, pero se hizo indispensable su uso por los complejos cálculos. A la mayoría de las personas la disponibilidad de diferentes programas para el análisis multivariado les facilita la compleja manipulación de la matriz de datos; esto es de mucha ayuda para el crecimiento de las técnicas multivariadas.

El propósito de este trabajo es proporcionar material didáctico que sirva de apoyo para cursos de Análisis Multivariado, desarrollando varias técnicas multivariadas y ejemplificándolas.

En el primer capítulo se mencionan las técnicas de graficación para datos multivariados, que proporcionan una idea de su aplicación; así como las distribuciones multivariadas y sus propiedades.

En el segundo capítulo se hace una revisión sobre la teoría de Componentes Principales y Análisis de Factores; así como algunos ejemplos.

El tercer capítulo trata sobre el tema de Análisis Discriminante, su aplicación y algunos ejemplos.

Y por último, en el capítulo cuatro se estudian las técnicas de Escalamiento Multidimensional y Análisis de Conglomerados.

Para la comprensión de este material se necesita un conocimiento de estadística matemática, así como de álgebra de matrices. Algunos resultados del álgebra de matrices se incluyen en el apéndice.

Capítulo 1

El análisis multivariado

1.1 Introducción

La teoría del análisis multivariado surge en 1930 y mucho de este trabajo fue restringido a la función de distribución normal multivariada. El término análisis multivariado se refiere al estudio de muchas variables que efectivamente incluyen varios casos, es decir, se refiere al estudio de n vectores aleatorios ($\underline{X}_1$ a $\underline{X}_n$) cada uno de dimensión p.

Igualmente importante es la generalización de las aplicaciones y/o técnicas estadísticas sobre el análisis multivariado. Las técnicas analíticas multivariadas son aplicadas en la industria y en la investigación.

Una razón para la dificultad de la definición del análisis multivariado es precisamente el término *multivariado*, el cual no es usado frecuentemente; algunos investigadores estudian la relación entre dos a más variables. Otra razón es suponer que todas las variables múltiples tienen la distribución normal.

1.2 Técnicas de graficación

Muchos procedimientos estadísticos tienen fundamento en algunos supuestos, y si éstos son falsos los cálculos numéricos conducen a conclusiones erróneas; en este sentido las gráficas pueden pensarse como un elemento intuitivo en la

verificación de los supuestos y un punto de apoyo para el desarrollo numérico. Existen diferentes gráficas para datos multivariados como son:

1.- Diagrama de tallos y hojas:

John Tukey en 1977 creó esta técnica, la cual consiste en sustituir las cantidades de cada intervalo de una tabla de distribución de frecuencias por símbolos. Puede haber diferentes símbolos, de tal manera que las barras de un histograma están formadas por todos los valores de cada variable. Por un lado, se tienen las variables y en el otro las frecuencias de cada una de ellas.

Cada renglón es un tallo, cada símbolo en un tallo es una hoja. Cuando los datos tienen más de una cifra, puede ser confuso emplearlos como símbolos; por lo que es preferible emplear como tallo la primera parte de la cifra y el restante como hoja.

Con respecto a los árboles que emplean más de un renglón por tallo, Tukey distingue el primero del segundo mediante un asterisco y un punto; el asterisco es asignado para los valores del 0 al 4 y el punto para los valores del 5 al 9.

La principal ventaja de esta técnica es que nos permite conocer el total de datos, así como el valor de la mediana y los cuartiles.

Ejemplo 1.1 *Considere que se tienen los siguiente datos:*

4 5 6 7 8 1 1 1 3 4 3 3 4 5 8 8 9 8 6

A continuación se ilustra el diagrama de tallos y hojas que corresponden a los datos anteriores.

Valores		Símbolo
1	|	111
2	|	
3	|	333
4	|	444
5	m	55
6	|	66
7	|	7
8	|	8888
9	|	9

Se puede observar que el valor de la moda es el 8 y el de la mediana (m) es el 5.

Ejemplo 1.2 *Los siguientes datos, muestran el número de "extirpaciones del útero" (Histerectomías) de una muestra de 15 cirujanos hombres. Utilizando la sugerencia de Tukey*

27 50 33 25 86 25 85 31 37 44 20 36 59 34 28

Entonces el diagrama correspondiente es:

Valores		Símbolo
2	|	0, 5, 5, 7, 8
3	m	1, 3, 4, 6, 7
4	|	4
5	|	0, 9
6	|	
7	|	
8	|	5, 6

Los valores máximo y mínimo son 86 y 20 respectivamente, el valor de la moda es 25, el valor de la mediana (m) es 34, esto es, que no existe simetría en los datos; y cada cirujano realiza en promedio 41.3 histerectomías (valor de la media).

2.- Diagrama de caja y brazos:

Para esta técnica es necesario conocer los valores de los cuartiles, la mediana y el valor máximo y mínimo. Esta información es facil de obtener de un diagrama de tallos y hojas.

Para la realización de este diagrama, es necesario graficar un rectángulo, de tal manera que el borde superior e inferior representen los cuartiles mayor y menor respectivamente; el segmento que divide al rectángulo es el valor de la mediana. Las líneas exteriores son las distancias que existen entre los cuartiles y los datos mínimo y máximo.

Si algún punto queda fuera de los límites mínimo y máximo se dice que es un punto aislado o un punto aberrante. Si la mediana divide al rectángulo en dos partes iguales se dice que hay una distribución simétrica en los datos; el ancho de la caja no tiene significado particular.

Ejemplo 1.3 *Los siguientes datos son las Histerectomías de una muestra de 10 cirujanos mujeres.*

5 7 10 14 18 19 25 29 31 33

Entonces para el diagrama de caja considere tambien los datos del *Ejemplo 1.2*

Figura 1.

Se puede observar que las cirujanas hacen en promedio menos operaciones que los cirujanos, el valor de la mediana puede dar una idea de si existe simetría en los datos o no. En los datos de los cirujanos como ya se había mencionado no existe simetría, caso contrario con los datos de las cirujanas.

3.- Diagramas de dispersión:

El hablar del comportamiento de dos variables sólo con fundamento numérico puede ser riesgoso si no se cuenta con el apoyo gráfico. Los diagramas de dispersión son una herramienta en la relación entre las variables.

3.1.- Diagramas de dispersión en dos dimensiones Una de las variables a graficar se conoce como la variable independiente, y se grafica en el eje de las abscisas, mientras que la otra variable es la dependiente y se grafica en el eje de las ordenadas. En caso de no existir dependencia entre las variables es indistinto como se grafiquen.

En un plano cartesiano, se grafican parejas ordenadas (x_i, y_i) con $i = 1, 2, \cdots n$ y n es el número de datos.

Ejemplo 1.4 *Los siguientes datos son el resultado de ozono del IMECA (Indice Metropolitano de Calidad del Aire), registrados en la zona noroeste y*

noreste durante febrero de 1990.

día	noroeste	noreste
1	108	63
2	172	84
3	98	53
4	155	68
5	94	88
6	155	102
7	138	102
8	144	136
9	212	108
10	225	108
11	114	40
12	130	79
13	184	96
14	117	83
15	38	70
16	141	74
17	117	102
18	123	78
19	112	103
20	58	68
21	45	80
22	106	51
23	57	47
24	91	65
25	11	48
26	141	61
27	136	80
28	110	90

Datos del ozono de 1990.

Utilizando el paquete Statistica, se obtiene la siguiente gráfica.

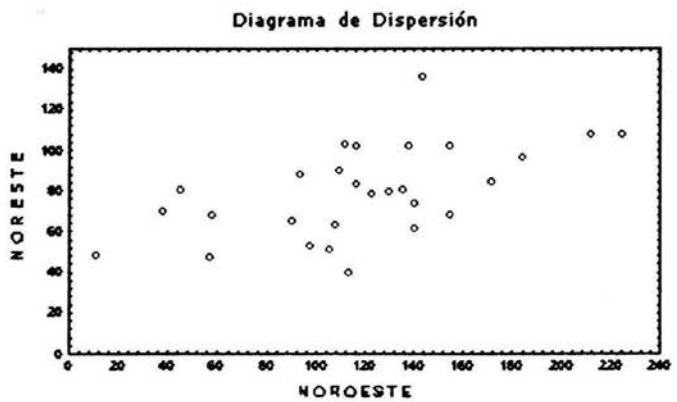

Figura 2.

Se graficó noroeste en el eje de las abscisas, y noreste en el eje de las ordenadas.

En la figura 2, se observa que en existe una relación lineal entre las variables, es decir, que si los resultados de ozono del IMECA aumentan en el noroeste, entonces tambien aumentan en el noreste.

3.2.- Diagramas de dispersión en tres dimensiones Por medio de estos diagramas, es posible comparar tres variables simultáneamente. De forma análoga que en los diagramas de dispersión de dos variables, en un plano cartesiano, se grafican parejas ordenadas (x_i, y_i); la tercer variable se representa con la altura, y de esta manera se está graficando en un espacio euclidiano y los puntos son de la forma (x_i, y_i, z_i).

Ejemplo 1.5 *Considere los datos del Ejemplo 1.4; pero agregando la variable centro. Los datos tambien son de febrero de 1990.*

dia	centro
1	166
2	146
3	129
4	147
5	108
6	115
7	145
8	167
9	171
10	144
11	145
12	126
13	134
14	93
15	64
16	126
17	126
18	99
19	166
20	47
21	71
22	65
23	67
24	124
25	164
26	128
27	130
28	98

Feb. de 1990.

Se obtiene la siguiente gráfica.

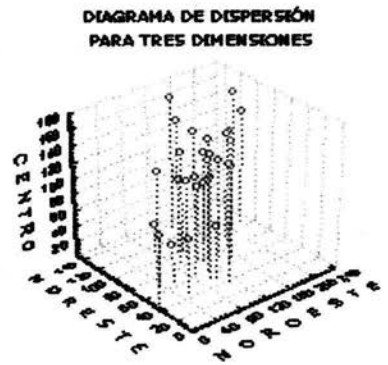

Figura 3.

Los datos se graficaron (*noroeste*, *noreste*, *centro*). En la zona noroeste los niveles de contaminación van de 11 a 225, de 40 a 136 para la zona noroeste y para la zona centro de 47 a 171.

Pero como existen puntos sobre puestos se pierden detalles de la gráfica y la interpretación puede ser confusa.

3.3.- Diagramas de dispersión múltiple Esta técnica se caracteriza por ser un arreglo de diagramas de dispersión donde cualquier par de gráficas adyacentes tienen un eje común.

La ventaja de esta técnica es la sencillez tanto en su elaboración como en su interpretación, la desventaja es el número de gráficas a elaborar cuando se tiene una elevada cantidad de variables.

Ejemplo 1.6 *Los siguientes datos son los resultados de ozono del IMECA registrados en la zona noroeste, noreste, centro, suroeste y sureste de febrero*

de 1990.

día	noroeste	noreste	centro	suroeste	sureste
1	108	63	166	228	109
2	172	84	146	142	110
3	98	53	129	145	95
4	155	68	147	178	137
5	94	88	108	91	95
6	155	102	115	157	69
7	138	102	145	179	94
8	144	136	167	197	87
9	212	108	171	193	172
10	225	108	144	172	97
11	114	40	145	130	102
12	130	79	126	192	110
13	184	96	134	121	54
14	117	83	93	100	39
15	38	70	64	82	29
16	141	74	126	193	120
17	117	102	126	126	55
18	123	78	99	118	73
19	112	103	166	124	79
20	58	66	47	56	34
21	45	80	71	81	97
22	106	51	65	100	66
23	57	47	67	96	97
24	91	65	124	183	124
25	11	48	164	185	109
26	141	61	128	195	110
27	136	80	130	151	115
28	110	90	98	204	116

Datos de febrero de 1990.

Se obtiene la siguiente gráfica.

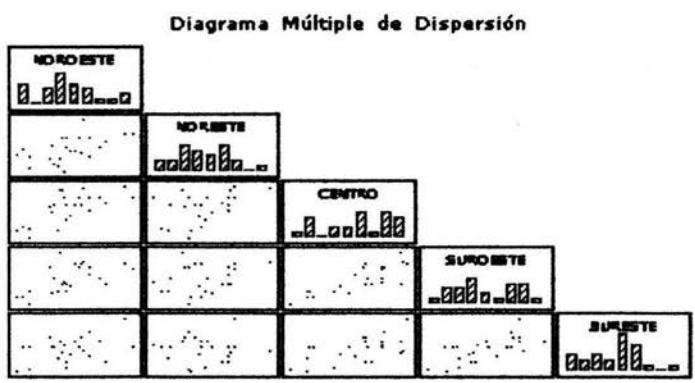

Figura 4.

De la gráfica 4 se observa que existe una relación lineal entre las variables *noroeste* y *centro*; así como del *sureste* con *suroeste*. Pero no hay relación de las variables *sureste* y *noroeste*, o *sureste* y *noreste*.

Esto es que si el ozono aumenta en el noroeste, tambien aumenta en el centro. Y si aumenta en el sureste, entonces aumenta en el suroeste, pero no hay cambios en el noroeste o noreste.

4.- Diagramas simbólicos:

Se utilizan los valores de los datos como parámetros de tal manera que sus variaciones causan la apariencia de un símbolo.

4.1.- Soles y estrellas: Cada diagrama representa un caso, y a la vez habrá tantos rayos como variables, manteniendo igual espacio entre uno y otro sobre un círculo o semicírculo de tal manera que se pueden representar como:

i. círculo $\theta_j = \dfrac{2\pi(j-1)}{p}$ ii. semicírculo $\theta_j = \dfrac{\pi(j-1)}{p-1}$

Para cada una de las variables, se deben considerar valores en el intervalo $(0,1)$; después se localiza el centro del símbolo en el origen de un plano cartesiano, y el punto en coordenadas polares $P_{ij} = (X_{ij}\cos\theta_j, X_{ij}\,sen\,\theta_j)$ que corresponde a la j-ésima variable sobre el i-ésimo elemento (X_{ij}).

La semejanza entre los símbolos, refleja una semejanza en el comportamiento de los datos.

Estos diagramas son de elaboración sencilla, sin embargo puede ser tedioso cuando el número de datos es demasiado grande, y resulta confuso cuando el número de variables es excesiva.

Ejemplo 1.7 *Se tienen 10 grupos de compañías aceiteras norteamericanas, las cuales consideran 15 variables para poder saber cual es la que tiene mayor ganacia; los datos son los siguientes:*

	C O M P A Ñ I A S									
	A	U	G	M	T	C	GU	AM	S	E
V1	.56	.53	.54	1.21	1.16	.84	1.01	.66	.97	1.44
V2	1.1	1.2	1.0	2.8	2.7	1.2	2.2	1.3	1.7	2.9
V3	.78	.49	.32	.50	.56	1.16	.67	.66	1.59	1.02
V4	306	203	197	211	176	378	219	258	336	250
V5	49	47	31	50	66	70	65	53	95	84
V6	10	4	11	8	8	13	11	8	13	8
V7	4.5	4.2	4.0	3.9	7.8	5.8	4.1	7.3	3.6	5.2
V8	.38	1.22	.67	1.04	.31	.70	1.53	.45	1.90	.99
V9	66	103	51	68	56	197	338	37	430	276
V10	62	99	57	78	50	141	235	44	378	199
V11	.11	.19	.11	.06	.04	.17	.23	.07	.39	.14
V12	174	527	160	339	277	355	481	213	656	609
V13	.84	.98	.38	.81	.91	.50	.83	.31	.38	.58
V14	2.8	8.5	2.8	4.0	2.5	1.6	2.9	2.7	2.7	4.3
V15	35	38	26	25	34	32	37	30	54	36

donde:

$A = ARCO$ $C = CHEVRON$
$U = UNION$ $GU = GULF$
$G = GETTY$ $AM = AMOCO$
$M = MOBIL$ $S = SHELL$
$T = TEXACO$ $E = EXXON$

y las variables son:

V1 Total neto en dolares pagados.

V2 Promedio total de dolares pagados sobre la segunda oferta más alta.

V3 Total de acres rentados (millones).

V4 Número de rentas ganadas.

V5 Porcentaje promedio de propiedades de rentas.

V6 Porcentaje de rentas encontradas últimamente ganado de la compañía.

V7 Promedio de años entre la venta y la primera producción.

V8 Producción neta de gas.

V9 Producción neta de líquido.

V10	Pago real neto al gobierno.
V11	Número real de años de producción.
V12	Cuadrado del coeficiente de correlación sobre la producción y el tiempo real.
V13	Producción anual real por acre.
V14	Porcentaje de rentas pagadas terminadas.

Utilizando el paquete Statistica se obtiene la siguiente gráfica.

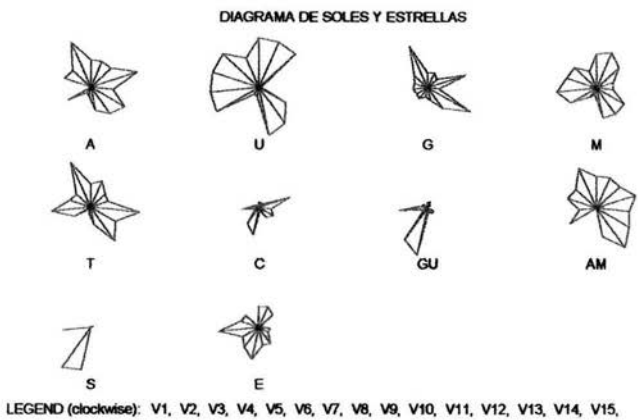

Figura 5.

Se intenta analizar el total de ganacias así como la producción de cada una de las 10 compañías. Cada símbolo es una de las compañías; en este caso GULF y SHELL pueden pertenecer al mismo conjunto, ya que su producción neta de gas y de líquido fue más alta en comparación a las otras compañías.

EXXON y CHEVRON a otro conjunto distinto ya que tienen el mayor número de años entre la primera producción y su venta.

Mientras que UNION, MOBIL y AMOCO pertenencen a otro conjunto por que es muy parecido el promedio del número total de acres rentados, las rentas ganadas y las propiedades de renta.

Y finalmente ARCO, TEXACO y GETTY están en otro conjunto debido a que el pago neto al gobierno y las rentas pagadas es muy parecido para las 3 compañías.

4.2.- Girasoles: Fueron desarrollados para saber cuando existen datos sobre puestos:

Un punto	representa	*una observación,*
Un punto con dos segmentos	representa	*dos observaciones,*
Un punto con tres lineas	representa	*tres observaciones, etc.*

Ejemplo 1.8 *Considereque se tienen los siguientes datos:*

4 5 6 7 8 1 1 1 3 4 3 3 4 5 8 8 9 8 6

Figura 6.

Con esta gráfica es fácil saber la densidad de las variables, esto es, que el valor que más se repite es el 8, el 7 y el 9 solo una vez, los valores 1, 3 y 4 tres veces y el 5 y el 6 dos veces.

5.- Diagrama de Caras de Chernoff:

Con el objeto de estudiar datos multidimensionales, dada la dificultad que implica la representación gráfica, Herman Chernoff (1971) ideó una técnica con la cual logró graficar 18 variables; la elaboración de las caras toma 5 aspectos básicos, la forma de la cara, la boca, la nariz, los ojos, las cejas. En 1978 Bruckner aumentó la característica de las orejas.

Según Chernoff las "caritas sonrientes" pueden relacionarse con el éxito, mientras que las otras reflejan dificultades. En 1988 Bernhard Flury duplicó

el número de variables, olvidando la simetría, y así en la parte izquierda se pueden graficar 18 variables y en la parte derecha otras 18 logrando así un total de 36 variables.

El desarrollo de las caras, parte de un punto central O se dibuja un rayo horizontal hacia un punto P y otro punto P' en el otro extremo, de tal manera que $O\ P$ y $O\ P'$ sean simétricos y así determinar la amplitud de la cara. En la parte superior e inferior de lo que va a ser la cara, se marcan dos puntos S e I, de tal manera que los puntos $O\ S$ y $O\ I$ sean verticales y de igual longitud.

La parte superior de la cara es una elipse determinada por cierta excentricidad para los puntos $P\ S\ P'$ y análogamente la parte inferior se determina por los puntos $P'I\ P$ y otra excentricidad.

La nariz se forma con un triángulo centrado en el punto O y su longitud está controlada por una variable. La boca es un arco de circunferencia centrado en el eje vertical; se debe tener en cuenta su posición, curvatura y ancho, que son determinadas por diferentes variables.

Los ojos son elipses y su posición, separación, inclinación, excentricidad y tamaño pueden variar. Las cejas son segmentos lineales localizados simétricamente sobre el eje vertical a través del punto O, se debe determinar posición, inclinación y tamaño. Las pupilas se localizan en la misma distancia horizontal desde el centro de los ojos. Las orejas están representadas por círculos tangentes a P y P' cuyo radio depende del valor de una variable.

A fin de evitar diagramas deformes o irregulares es conveniente hacer una normalización de los datos, es decir, deben tener la misma unidad de medida. La asignación de las variables para los rasgos de las caras es:

Número de variable	Característica de la Cara
1	Amplitud de la cara.
2	Nivel de las orejas.
3	Altura de la cara.
4	Excentricidad de la parte superior de la cara

5	Excentricidad de la parte inferior de la cara
6	Longitud de la nariz.
7	Nivel de la boca.
8	Curvatura de la boca.
9	Longitud de la boca.
10	Separación de los ojos.
11	Inclinación de los ojos.
12	Nivel de los ojos.
13	Posición de las pupilas.
14	Altura de las cejas.
15	Longitud de la ceja
16	Diámetro de las orejas.
17	Ancho de la nariz.

Ejemplo 1.9 *Este ejemplo se conoce como las Irises de Fisher, el cual consiste en 3 tipos de flor de Iris, en donde se considera el largo y ancho del sépalo, así como el largo y ancho del pétalo. Usando el paquete Statistica, se obtiene:*

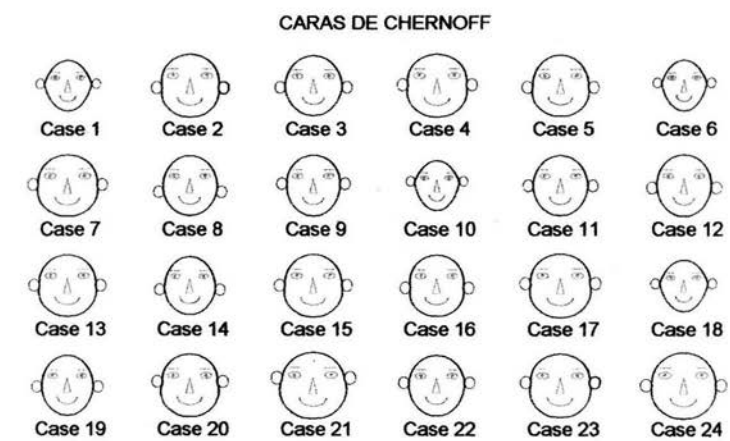

Figura 7.

en donde se tienen las siguientes variables y sus respectivas características asociadas a la cara.

Nombre de la variable	Característica de la Cara
SEPALLEN	Amplitud de la cara.
SEPALWID	Nivel de las orejas.
PETALLEN	Altura de la cara.
PETALWID	Excentricidad de la parte superior de la cara
IRISTYPE	Excentricidad de la parte inferior de la cara

Como las cara de chernoff todas estan "sonrientes", esto quiere decir que el largo y ancho del pétalo y del sépalo son muy parecidos, pero algunas *caras* están más alargada y otras más redondas; estas diferencias son por las variables asignadas a las características y quizás especifiquen los distintos grupos existentes.

6.- Curvas de Andrews:

Permite utilizar una gráfica para la identificación de cúmulos, puntos lejanos y otras características de la distribución de los datos. Supone que cada una de las n observaciones de X define una función, (X es de dimensión p):

$$f_X(t) = \frac{X_1}{2} + X_2\, sen\,(t) + X_3\, \cos(t) + X_4\, sen(2\,t) + X_5\, \cos(2\,t)\, \ldots$$

con $t \in [-\pi, \pi]$. De esta manera, las n observaciones aparecen como un conjunto de líneas dibujadas a través de la gráfica; por lo cual los puntos que permanecen juntos en el espacio p-dimensional quedan representados por curvas cercanas entre sí en todo el intervalo de t.

Con esta curva se pueden identificar los cúmulos o puntos lejanos de los datos.

Ejemplo 1.10 *considere los datos del Ejemplo 1.7. Se obtiene la siguiente gráfica.*

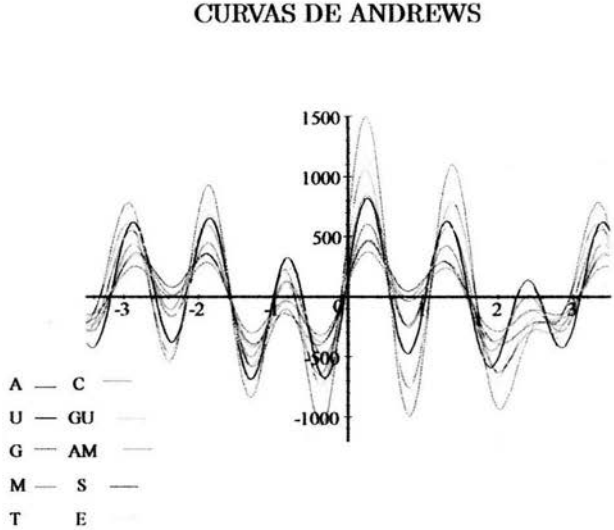

Figura 8.

En la gráfica se puede observar que es muy parecido el comportamiento de las 10 compañías, sin embargo la compañía SHELL tiene mayores beneficios y la que tiene menos es GETTY.

1.3 Distribuciones multivariadas

Varias de las distribuciones univariadas continuas tienen distribuciones análogas para el caso multivariado. La distribución Normal multivariada es una de las distribuciones más utilizadas en estadística. En el análisis multivariado muchos de los procedimientos se han hecho bajo el supuesto de normalidad

1.3.1 Normal multivariada

Es muy importante la densidad normal para el caso multivariado debido a la fácil generalización para la distribución conjunta de p variables, además de que las distribuciones marginales y condicionales de normales también tienen distribucion normal.

La distribución *normal multivariada* ha sido un modelo fundamental en el desarrollo de las distintas técnicas multivaridas como, por ejemplo, el análisis de varianza, el diseño de experimentos y el análisis discriminante.

Definición 1.1 *Sea* $\underline{X}$ *un vector aleatorio de dimensión* p, *se dice que* $\underline{X}$ *tiene distribución normal multivariada con vector de medias* $\underline{\mu}$ *y matriz de covarianzas* Σ; *si su función de densidad está dada por:*

$$f_{\underline{x}}(\underline{X}, \underline{\mu}, \Sigma) = \mid 2\pi\Sigma \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\underline{X} - \underline{\mu})^{t}\, \Sigma^{-1}(\underline{X} - \underline{\mu})\right\} \tag{1.1}$$

donde: $\underline{\mu} \in R^{P}$, $\Sigma_{p \times p}$ es una matriz definida positiva. Se denota por $\underline{X} \sim N_{p}(\underline{\mu}, \Sigma)$.

Note que por la Definición A.23, en el exponente de la normal multivariada, se tiene una forma cuadrática, dada por:

$$\phi = (\underline{X} - \underline{\mu})^{t}\, \Sigma^{-1}(\underline{X} - \underline{\mu}). \tag{1.2}$$

Definición 1.2 *Sea* $\underline{X} \in \Re^{p}$ *un vector aleatorio. La función generadora de momentos* (fgm) *de* $\underline{X}$ *está definida como:*

$$m_{\underline{X}}(\underline{t}) = E\left(\exp\left\{\underline{t}^{t}\, \underline{X}\right\}\right). \tag{1.3}$$

Teorema 1.3 *Sea* $\underline{X}$ *un vector aleatorio de dimensión* p, *donde cada uno de sus componentes* $\underline{X}$ *tiene la distribución* $N_{p}(\underline{\mu}, \Sigma)$, *entonces la función generadora de momentos de* $\underline{X}$ *está dada por:*

$$m_{\underline{X}}(\underline{t}) = \exp\left\{\underline{t}^{\,t}\,\underline{\mu} + \frac{1}{2}\,\underline{t}^{\,t}\,\Sigma\,\underline{t}\right\}. \tag{1.4}$$

Demostración.
Considere la siguiente transformación $\underline{Y} = \Sigma^{-1/2}(\underline{X} - \underline{\mu})$, entonces:

$$\underline{X} = \Sigma^{1/2}\underline{Y} + \underline{\mu} \quad \text{y} \quad |J| = \left|\Sigma^{1/2}\right|.$$

Por lo que la función generadora de momentos está dada por:

$$\begin{aligned}
m_{\underline{Y}_i}(\underline{t}) &= E\left(\exp\{\underline{t}_i\,\underline{Y}_i\}\right) \\
&= \int_{-\infty}^{\infty} \exp(\underline{t}_i\,y_i)\,(2\pi)^{-1/2}\exp\left\{-\frac{1}{2}(y_i^2)\right\}dy_i \\
&= (2\pi)^{-1/2}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2}\,\underline{y}_i^{\,2} + \underline{t}_i\,y_i\right\}dy_i \\
&= (2\pi)^{-1/2}\exp\left(\frac{1}{2}\,\underline{t}_i^{\,2}\right)\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2}(y_i - \underline{t}_i)^2\right\}dy_i \\
&= \exp\left\{\frac{1}{2}(\underline{t}_i^{\,2})\right\}.
\end{aligned}$$

Entonces

$$\begin{aligned}
m_{\underline{Y}}(\underline{t}) &= E\left(\exp\{\underline{t}^{\,t}\,\underline{Y}\}\right) \\
&= \prod_{i=1}^{p} m_{\underline{Y}_i}(\underline{t}) \\
&= \exp\left\{\frac{1}{2}(\underline{t}^{\,t}\,\underline{t})\right\}.
\end{aligned}$$

Por lo que:

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

$$= \exp\left\{\underline{t}^{\,t}\,\underline{\mu} + \frac{1}{2}\,\underline{t}^{\,t}\,\Sigma\,\underline{t}\right\} \qquad \blacksquare$$

Teorema 1.4 *Si $\underline{X}_1$, $\underline{X}_2$ son dos vectores aleatorios independientes de dimensión p_1 y p_2 respectivamente, entonces la función generadora de momentos se factoriza como el producto de las funciones generadoras de momentos marginales de $\underline{X}_1$, $\underline{X}_2$.*

Demostración.

Sea $\underline{X} = \begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \end{pmatrix}$, por la Definición 1.2 se tiene:

$$\begin{aligned} m_{\underline{X}}(\underline{t}) &= E\left(\exp\{\underline{t}^{\,t}\,\underline{X}\}\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \exp\{\underline{t}^{\,t}\,\underline{X}\}\, f_{\underline{X}}(x)\, dx_1 \ldots dx_p \\ &= \int_{-\infty}^{\infty} \exp\{\underline{t}_1^{\,t}\,\underline{X}_1 + \underline{t}_2^{\,t}\,\underline{X}_2\}\, f_{\underline{X}_1}(x_1)\, f_{\underline{X}_2}(x_2)\, dx_1\, dx_2 \\ &= m_{\underline{X}_1}(\underline{t}_1)\, m_{\underline{X}_2}(\underline{t}_2). \end{aligned} \qquad \blacksquare$$

Teorema 1.5 *Sea $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$ y $\underline{X}_1, \underline{X}_2$ tales que $\underline{X} = \begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \end{pmatrix}$, $\underline{\mu} = \begin{pmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{pmatrix}$, $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, donde $\underline{X}_i$, $\underline{\mu}_i$ tienen dimensión p_i y cada Σ_{ij} (i,j =1,2) tiene dimensión $p_i \times p_j$ $(p = p_1 + p_2)$; entonces los vectores $\underline{X}_1$, $\underline{X}_2$ son independientes si y sólo si $\Sigma_{12} = 0$ y $\Sigma_{21} = 0$.*

Demostración.

Si $\underline{X}_1$ y $\underline{X}_2$ son independientes, entonces por el Teorema 1.3 la función generadora de momentos de $\underline{X}$ está dada por:

$$m_{\underline{X}}(\underline{t}) = \exp\left\{\underline{t}^{\,t}\,\underline{\mu} + \tfrac{1}{2}\,\underline{t}^{\,t}\,\Sigma\,\underline{t}\right\}$$

y la función generadora de momentos de cada $\underline{X}_i$ está dada por:

$$m_{\underline{X}_1}(\underline{t}) = E\left(\exp\{\underline{t}_1^{\,t}\,\underline{X}_1\}\right) = \exp\left\{\underline{t}_1^{\,t}\,\underline{\mu}_1 + \frac{1}{2}\,\underline{t}_1^{\,t}\,\Sigma_{11}\,\underline{t}_1\right\}$$

$$m_{\underline{X}_2}(\underline{t}) = E\left(\exp\{\underline{t}_2^{\,t}\,\underline{X}_2\}\right) = \exp\left\{\underline{t}_2^{\,t}\,\underline{\mu}_2 + \frac{1}{2}\,\underline{t}_2^{\,t}\,\Sigma_{22}\,\underline{t}_2\right\}$$

por la independencia se tiene:

$m_{\underline{X}}(\underline{t}) = m_{\underline{X}_1}(\underline{t})\, m_{\underline{X}_2}(\underline{t})$ esto se cumple *si y sólo si*

$$\frac{1}{2}\,\underline{t}^{\,t}\,\Sigma\,\underline{t} = \frac{1}{2}\,\underline{t}_1^{\,t}\,\Sigma_{11}\,\underline{t}_1 + \frac{1}{2}\,\underline{t}_2^{\,t}\,\Sigma_{22}\,\underline{t}_2. \tag{1.5}$$

Definiendo $\underline{t} = \begin{pmatrix} \underline{t}_1 \\ \underline{t}_2 \end{pmatrix}$ y $\underline{t}^{\,t} = \begin{pmatrix} \underline{t}_1^{\,t} & \underline{t}_2^{\,t} \end{pmatrix}$ se tiene que:

$$\begin{aligned} \underline{t}^{\,t}\,\Sigma\,\underline{t} &= \begin{pmatrix} \underline{t}_1^{\,t} & \underline{t}_2^{\,t} \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} \underline{t}_1 \\ \underline{t}_2 \end{pmatrix} \\ &= \underline{t}_1^{\,t}\,\Sigma_{11}\,\underline{t}_1 + \underline{t}_2^{\,t}\,\Sigma_{12}\,\underline{t}_2 + \underline{t}_1^{\,t}\,\Sigma_{21}\,\underline{t}_1 + \underline{t}_2^{\,t}\,\Sigma_{22}\,\underline{t}_2. \end{aligned}$$

Entonces por (1.5), se debe cumplir que:

$$\begin{aligned} \Sigma_{12} &= 0 \\ \Sigma_{21} &= 0 \end{aligned}$$

Inversamente, $\Sigma_{12} = 0$ y $\Sigma_{21} = 0$, entonces la forma cuadrática de la densidad de $\underline{X}$ puede expresarse como:

$$\begin{aligned} Q &= (\underline{X} - \underline{\mu})^{\,t}\,\Sigma^{-1}(\underline{X} - \underline{\mu}) \\ &= \begin{pmatrix} \underline{X}_1 - \underline{\mu}_1 \\ \underline{X}_2 - \underline{\mu}_2 \end{pmatrix}^{t} \begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix} \begin{pmatrix} \underline{X}_1 - \underline{\mu}_1 \\ \underline{X}_2 - \underline{\mu}_2 \end{pmatrix} \\ &= (\underline{X}_1 - \underline{\mu}_1)^{\,t}\,\Sigma_{11}^{-1}(\underline{X}_1 - \underline{\mu}_1) + (\underline{X}_2 - \underline{\mu}_2)^{\,t}\,\Sigma_{22}^{-1}(\underline{X}_2 - \underline{\mu}_2) \\ &= Q_1 + Q_2. \end{aligned}$$

donde $Q_1 = (\underline{X}_1 - \underline{\mu}_1)^{\,t}\,\Sigma_{11}^{-1}\,(\underline{X}_1 - \underline{\mu}_1)$, y
$Q_2 = (\underline{X}_2 - \underline{\mu}_2)^{\,t}\,\Sigma_{22}^{-1}\,(\underline{X}_2 - \underline{\mu}_2)$.
por lo que la función de densidad de $N_p(\underline{\mu}, \Sigma)$ se reduce a :

$$f_{\underline{X}}(\underline{X}, \underline{\mu}, \Sigma) = |2\pi|^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(Q_1 + Q_2)\right\}$$
$$= |2\pi|^{-\frac{p_1}{2}} |\Sigma_{11}|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(Q_1)\right\} \times |2\pi|^{-\frac{p_2}{2}} |\Sigma_{22}|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(Q_2)\right\}$$
$$= f_{\underline{X}_1}(\underline{X}_1, \underline{\mu}_1, \Sigma_{11})\, f_{\underline{X}_2}(\underline{X}_2, \underline{\mu}_2, \Sigma_{22}).$$

Por lo tanto $\underline{X}_1$ y $\underline{X}_2$ son independientes. ∎

Definición 1.6 *Sea $\underline{X} \in \Re^p$ un vector aleatorio. La función característica de $\underline{X}$ está definida como:*

$$\phi_{\underline{X}}(\underline{t}) = \exp\left\{i\, \underline{t}^t\, \underline{X}\right\}. \tag{1.6}$$

Teorema 1.7 *Sea $\underline{X}$ un vector aleatorio de dimensión p, donde $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$, entonces la función característica de $\underline{X}$ está dada por:*

$$\phi_{\underline{X}}(\underline{t}) = \exp\left\{i\, \underline{t}^t\, \underline{\mu} - \frac{1}{2}\, \underline{t}^t\, \Sigma\, \underline{t}\right\}. \tag{1.7}$$

Demostración.

Siguiendo un razonamiento análogo al del Teorema 1.3, considere la transformación $\underline{Y} = \Sigma^{-1/2}(\underline{X} - \underline{\mu})$, entonces:

$$\underline{X} = \Sigma^{1/2}\, \underline{Y} + \underline{\mu} \quad \text{y} \quad |J| = \left|\Sigma^{1/2}\right|.$$

Por lo que la *función característica*, está dada por:

$$\begin{aligned}
\phi_{\underline{Y}_j}(\underline{t}) &= E\left(\exp\left\{(i\, \underline{t}_j\, \underline{Y}_j)\right\}\right) \\
&= \int_{-\infty}^{\infty} \exp\left\{(i\, \underline{t}_j\, y_j)\right\} (2\pi)^{-1/2} \exp\left\{-\frac{1}{2}(y_j^2)\right\} dy_j \\
&= (2\pi)^{-1/2} \exp\left\{-\frac{1}{2}(\underline{t}_j^{\,2})\right\} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}(y_j - i\, \underline{t}_j)^2\right\} dy_j \\
&= \exp\left\{-\frac{1}{2}(\underline{t}_j^{\,2})\right\}.
\end{aligned}$$

Entonces:

$$\phi_{\underline{Y}}(\underline{t}) = E\left(\exp\{\underline{t}^{t}\underline{Y}\}\right) = \prod_{j=1}^{p} \phi_{\underline{Y}_j}(\underline{t})$$
$$= \exp\left\{-\frac{1}{2}(\underline{t}^{t}\,\underline{t})\right\}.$$

Por lo que

$$\begin{aligned}
\phi_{\underline{X}}(\underline{t}) &= E\left(\exp\left\{i\,\underline{t}^{t}\,\underline{X}\right\}\right) \\
&= E\left(\exp\left\{i\,\underline{t}^{t}\,(\Sigma^{1/2}\underline{Y}+\underline{\mu})\right\}\right) \\
&= E\left(\exp\left\{i\,\underline{t}^{t}\,\underline{\mu}\right\}\right) E\left(\exp\left\{i\,\underline{t}^{t}\,\Sigma^{1/2}\,\underline{Y}\right\}\right) \\
&= \exp\left\{i\,\underline{t}^{t}\,\underline{\mu}\right\} \times \exp\left\{-\frac{1}{2}\left(\underline{t}^{t}\,\Sigma^{1/2}\right)\left(\underline{t}^{t}\,\Sigma^{1/2}\right)^{t}\right\} \\
&= \exp\left\{i\,\underline{t}^{t}\,\underline{\mu} - \frac{1}{2}\,\underline{t}^{t}\,\Sigma\,\underline{t}\right\}.
\end{aligned}$$

■

1.3.2 Estimación de parámetros

Con la finalidad de estimar los parámetros de la distribución $N_p(\underline{\mu}, \Sigma)$, suponga que se tiene un conjunto de observaciones independientes $\underline{X}_1, \underline{X}_2, \cdots, \underline{X}_n$ donde cada $\underline{X}_i \sim N_p(\underline{\mu}, \Sigma)$ para $i = 1, \ldots, n$. A continuación se realizará la estimación mediante el método de máxima verosimilitud.

Teorema 1.8 *Sea* $\underline{X}_i \sim N_p(\underline{\mu}, \Sigma)$ *y* $\underline{X}_1, \underline{X}_2, \cdots, \underline{X}_n$ *una muestra aleatoria de la distribución. Los estimadores máximo verosimiles de* $\underline{\mu}$ *y* Σ *son:*

$$\widehat{\underline{\mu}} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1.8}$$

$$\widehat{\Sigma} = S = \frac{1}{n}\sum_{i=1}^{n}(\underline{X}_i - \overline{X})\,(\underline{X}_i - \overline{X})^{t}. \tag{1.9}$$

Demostración.
La función de verosimilitud está dada por la ecuación:

$$L(\mu, \Sigma; \underline{X}_n) = |2\pi\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} (\underline{X}_i - \underline{\mu})^t \Sigma^{-1} (\underline{X}_i - \underline{\mu})\right\}$$

donde $\underline{X}_n = \{\underline{X}_1, \underline{X}_2, \cdots, \underline{X}_n\}$, entonces aplicando logaritmo natural a la ecuación anterior.

$$\ln L(\mu, \Sigma; \underline{X}_n) = -\frac{n}{2}\ln |2\pi\Sigma| - \frac{1}{2}\sum_{i=1}^{n} (\underline{X}_i - \underline{\mu})^t \Sigma^{-1} (\underline{X}_i - \underline{\mu}). \quad (1.10)$$

Obteniendo las derivadas parciales con respecto a cada parámetro, se tiene que:

$$\frac{\partial \ln L(\mu, \Sigma; \underline{X}_n)}{\partial \mu} = -\frac{1}{2}\sum_{i=1}^{n} \{-2\,\Sigma_i^{-1}\,\underline{X} + 2\,\Sigma^{-1}\mu\}$$

$$= n\,\Sigma^{-1}(\overline{X} - \underline{\mu}) \quad (1.11)$$

y sea $V = \Sigma^{-1}$ por lo que del Teorema A.25:

$$\frac{\partial \ln |V|}{\partial V} = 2\,V^{-1} - Diag(V^{-1})$$

$$= 2\,\Sigma - Diag(\Sigma). \quad (1.12)$$

Por la Definición A.23, se sabe que $\sum_{i=1}^{n} (\underline{X}_i - \underline{\mu})^t \Sigma^{-1} (\underline{X}_i - \underline{\mu})$ es una forma cuadrática; la derivada parcial con respecto a Σ está dada por:

$$\frac{\partial \sum_{i=1}^{n} (\underline{X}_i - \underline{\mu})^t \Sigma^{-1} (\underline{X}_i - \underline{\mu})}{\partial \Sigma} =$$

$$\sum_{i=1}^{n} \{2\,[(\underline{X}_i - \underline{\mu})(\underline{X}_i - \underline{\mu})^t] - Diag\,[(\underline{X}_i - \underline{\mu})(\underline{X}_i - \underline{\mu})^t]\} =$$

$$2\sum_{i=1}^{n}\left[(\underline{X}_i-\underline{\mu})\,(\underline{X}_i-\underline{\mu})^t\right]-Diag\sum_{i=1}^{n}\left[(\underline{X}_i-\underline{\mu})\,(\underline{X}_i-\underline{\mu})^t\right] \qquad (1.13)$$

Por lo que de (1.12) y (1.13), se tiene que:

$$\frac{\partial \ln L(\mu,\ \Sigma;\ \underline{X}_n)}{\partial\Sigma}=\frac{n}{2}\left[2\,\Sigma-Diag(\Sigma)\right]-$$
$$\frac{1}{2}\left\{2\sum_{i=1}^{n}\left[(\underline{X}_i-\underline{\mu})\,(\underline{X}_i-\underline{\mu})^t\right]-Diag\sum_{i=1}^{n}\left[(\underline{X}_i-\underline{\mu})\,(\underline{X}_i-\underline{\mu})^t\right]\right\}$$
$$=n\,\Sigma-\frac{n}{2}\,Diag(\Sigma)-n\;S+\frac{n}{2}\,Diag(S)$$

$$=n\,(\,\Sigma-S\,)-\frac{n}{2}\,(\,\Sigma-S\,). \qquad (1.14)$$

Donde

$$S=\frac{1}{n}\sum_{i=1}^{n}\left[(\underline{X}_i-\underline{\mu})\,(\underline{X}_i-\underline{\mu})^t\right].$$

Para obtener los estimadores, se debe igualar a cero (1.11). Esto es:

$$n\,\Sigma^{-1}(\overline{X}-\underline{\widehat{\mu}})=0$$

Y esta ecuación se anula cuando

$$\underline{\widehat{\mu}}=\overline{X}$$

lo que demuestra (1.8). Sustituyendo este resultado en (1.12), e igualando a cero se obtiene:

$$n\,(\widehat{\Sigma}-S)-\frac{n}{2}(\widehat{\Sigma}-S)=0$$

en donde la ecuación se anula cuando $\widehat{\Sigma}=S$.

Para probar que $\underline{\widehat{\mu}}$ maximiza la función de verosimilitud, es necesario minimizar la siguiente ecuación:

$$\sum_{i=1}^{n} (\underline{X}_i - \underline{\mu})^t \Sigma^{-1} (\underline{X}_i - \underline{\mu})$$

$$= \sum_{i=1}^{n} (\underline{X}_i - \overline{X})^t \Sigma^{-1} (\underline{X}_i - \overline{X}) + n(\overline{X} - \underline{\mu})^t \Sigma^{-1} (\overline{X} - \underline{\mu})$$

como Σ^{-1} es definida positiva, entonces $n\,(\,\overline{X} - \underline{\mu})^t \Sigma^{-1} (\,\overline{X} - \underline{\mu}) > 0$ y la ecuación es cero *si y sólo si*

$$\underline{\hat{\mu}} = \overline{X}$$

por lo que maximiza la función de verosimilitud.

Falta probar que $\hat{\Sigma}$ maximiza la función de verosimilitud. Para ello, note que:

$$\sum_{i=1}^{n} (\underline{X}_i - \underline{\mu})(\underline{X}_i - \underline{\mu})^t = \sum_{i=1}^{n} \left[(\underline{X}_i - \overline{X}) + (\overline{X} - \underline{\mu})\right] \left[(\underline{X}_i - \overline{X}) + (\overline{X} - \underline{\mu})\right]^t$$

$$= \sum_{i=1}^{n} \left[(\underline{X}_i - \overline{X})(\underline{X}_i - \overline{X})^t + (\underline{X}_i - \overline{X})(\overline{X} - \underline{\mu})^t + (\overline{X} - \underline{\mu})(\underline{X}_i - \overline{X})^t \right.$$
$$\left. + (\overline{X} - \underline{\mu})(\overline{X} - \underline{\mu})^t\right]$$

$$= \sum_{i=1}^{n} (\underline{X}_i - \overline{X})(\underline{X}_i - \overline{X})^t + \left[\sum_{i=1}^{n} (\underline{X}_i - \overline{X})\right] (\overline{X} - \underline{\mu})^t$$
$$+ (\overline{X} - \underline{\mu}) \left[\sum_{i=1}^{n} (\underline{X}_i - \overline{X})^t\right] + n\,(\overline{X} - \underline{\mu})(\overline{X} - \underline{\mu})^t$$

$$= \sum_{i=1}^{n} (\underline{X}_i - \overline{X})(\underline{X}_i - \overline{X})^t + n\,(\overline{X} - \underline{\mu})(\overline{X} - \underline{\mu})^t$$
$$= A + n\,(\overline{X} - \underline{\mu})(\overline{X} - \underline{\mu})^t$$

donde

$$A = \sum_{i=1}^{n} (\underline{X}_i - \overline{X})(\underline{X}_i - \overline{X})^t.$$

Por lo que:

$$tr\left(\sum_{i=1}^{n}(\underline{X}_i-\underline{\mu})^t\,\Sigma^{-1}(\underline{X}_i-\underline{\mu})\right)=tr\left(\sum_{i=1}^{n}\Sigma^{-1}(\underline{X}_i-\underline{\mu})\,(\underline{X}_i-\underline{\mu})^t\right)$$
$$=tr\,(\Sigma^{-1}A)+tr\left(\Sigma^{-1}\,n\,(\overline{X}-\underline{\mu})\,(\overline{X}-\underline{\mu})^t\right)$$
$$=tr\,(\Sigma^{-1}A)+n\,(\overline{X}-\underline{\mu})\,\Sigma^{-1}\,(\overline{X}-\underline{\mu})^t$$

Y así (1.10) se puede reescribir como:

$$\ln L(\mu,\Sigma;\underline{X}_n) = -\frac{p}{2}\ln(2\pi)-\frac{n}{2}\ln|\Sigma|$$
$$-\frac{1}{2}tr\left(\Sigma^{-1}A\right)-\frac{1}{2}n\,(\overline{X}-\underline{\mu})\Sigma^{-1}(\overline{X}-\underline{\mu})^t$$

Entonces para maximizar el segundo y tercer término de la ecuación anterior,

$$-n\ln|\Sigma|-tr\left(\Sigma^{-1}A\right),$$

se define:

$$f(\Sigma)=-n\ln|\Sigma|-tr\left(\Sigma^{-1}A\right),$$

la cual por el Teorema A.24, se maximiza si

$$\Sigma=\frac{1}{n}A,\ \text{con}\ A=\sum_{i=1}^{n}(\underline{X}_i-\overline{X})\,(\underline{X}_i-\overline{X})^t;$$

y por lo tanto,

$$\hat{\Sigma} = S$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left[(\underline{X}_i-\overline{X})\,(\underline{X}_i-\overline{X})^t\right]$$

maximiza la función de verosimilitud ∎

Para ver si los estimadores máximo verosimil son insesgados, se calculará su valor esperado

$$
\begin{aligned}
E(\underline{\hat{\mu}}) &= E(\overline{X}) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(\underline{X}_i) \\
&= E(X) \\
&= \underline{\mu}
\end{aligned}
$$

Por lo que $\underline{\hat{\mu}}$ *es insesgado.*

En cuanto a $\hat{\Sigma}$,

$$
\begin{aligned}
E(\hat{\Sigma}) &= E(S) \\
&= E(\frac{1}{n}\sum_{i=1}^{n}\left[(\underline{X}_i - \overline{X})\,(\underline{X}_i - \overline{X})^t\right]) \\
&= E(\frac{1}{n}\sum_{i=1}^{n}\left[(\underline{X}_i - \mu + \mu - \overline{X})\,(\underline{X}_i - \mu + \mu - \overline{X})^t\right]) \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left\{(\underline{X}_i - \underline{\mu})\,(\underline{X}_i - \underline{\mu})^t\right\} - E\left\{(\overline{X} - \underline{\mu})\,(\overline{X} - \underline{\mu})^t\right\} \\
&= \Sigma - \frac{1}{n^2} E\left\{\sum_{i=1}^{n}(\underline{X}_i - \underline{\mu})(\underline{X}_i - \underline{\mu})^t - \sum_{i\neq j}^{n}(\underline{X}_i - \underline{\mu})(\underline{X}_j - \underline{\mu})^t\right\} \\
&= \Sigma - \frac{1}{n}\Sigma = \frac{n-1}{n}\Sigma.
\end{aligned}
$$

Por lo tanto $\hat{\Sigma}$ no es insesgado

Si se desea tener un estimador insesgado, se puede proponer a

$$
\begin{aligned}
\hat{\Sigma}_\mu &= S_\mu \\
&= \frac{n}{n-1}S
\end{aligned}
$$

el cual corrige el sesgo de $\hat{\Sigma}$.

1.3.3 Transformaciones Lineales

Cuando se tiene una combinación lineal de un vector aleatorio que se distribuye normal multivariado, es necesario saber que función de densidad tiene esa combinación lineal.

Teorema 1.9 *Si* $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$ *sea* $\underline{Y} = A\,\underline{X} + C$, *donde* A *es una matriz cuadrada de constantes y* $C \in \Re^p$, *entonces* $\underline{Y} \sim N_p(A\,\underline{\mu} + C,\ A\,\Sigma\,A^t)$.

Demostración.
Por la Definición 1.2, se tiene que:

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

Se tiene la función de densidad de una normal con media $A\underline{\mu} + C$ y matriz de covarianzas $A\,\Sigma\,A^t$, faltaría demostrar que $A\,\Sigma\,A^t > 0$

Sea $\underline{t} \in \Re^p$ con $\underline{t} \neq 0$ es necesario probar que $\underline{t}^t\,A\,\Sigma\,A^t\,\underline{t} > 0$, por lo que:

$$\begin{aligned} \underline{t}^t A\,\Sigma\,A^t\,\underline{t} &= \left(\underline{t}^t A\right)\ \Sigma\ \left(A^t\,\underline{t}\right) \\ &= \underline{s}^t\,\Sigma\,\underline{s} \end{aligned} \tag{1.15}$$

donde $\underline{s}$ es un vector no nulo, por lo cual se tiene que:

$$A\,\underline{s} = A\,A^t\,\underline{t}$$

como $A\,A^t$ es una matriz cuadrada, entonces existe $(A\,A^t)^{-1}$, entonces de la ecuación anterior se tiene una expresión para $\underline{t}$

$$\underline{t} = \left(A\, A^{t}\right)^{-1} A\, \underline{s}$$

Si $\underline{s} = 0$, entonces $\underline{t} = 0$ lo cual es una contradicción ya que $\underline{t} \neq 0$ por lo cual $\underline{s} \neq 0$ y de (1.15) se concluye que $A\, \Sigma\, A^{t} > 0$ ∎

1.3.4 La distribución Ji-cuadrada

Existen diferentes formas cuadráticas de variables que tienen distribución Normal Multivariada, como es el caso de la *ji-cuadrada* (χ^{2}).

Definición 1.10 *Se dice que una variable aleatoria X tiene distribución χ^{2} con n grados de libertad si su función de densidad está dada por:*

$$f_X\,(x,\, n) = \left\{2^{\frac{n}{2}}\,\Gamma\,\left(\frac{n}{2}\right)\right\}^{-1} x^{\frac{n}{2}-1} \exp\left(-\frac{x}{2}\right) \qquad (1.16)$$

para $x > 0$, $n > 0$, donde $\Gamma\,(t)$ es la función gamma definida por:

$$\Gamma(t) = \int_0^{\infty} u^{t-1} \exp(-u) du$$

con $t > 0$ y será denota por $X \sim \chi^{2}_{(n)}$.

Teorema 1.11 *Si X es una variable aleatoria, con distribucuón $\chi^{2}_{(n)}$, entonces se cumple:*

(a) $\phi_X\,(t) = (1 - 2\, i\, t)^{-n/2}$.

(b) $M_X\,(t) = (1 - 2\, t)^{-n/2}$ para $t < \frac{1}{2}$.

Demostración.

(a) $\phi_X(t) = E(\exp(i\,t\,X))$

$$= \int_0^\infty \frac{1}{2^{n/2}\,\Gamma\left(\frac{n}{2}\right)}\, x^{\frac{n}{2}-1}\left(\exp\left\{-\frac{x}{2}(1-2\,i\,t)\right\}\right) dx$$

Realizando un cambio de variable, $u = (1-2\,i\,t)\,x$

$$= \frac{1}{(1-2\,i\,t)^{n/2}} \int_0^\infty \frac{u^{\frac{n}{2}-1}\exp -\frac{u}{2}}{2^{n/2}\,\Gamma\left(\frac{n}{2}\right)}\, du$$

$$= \frac{1}{(1-2\,i\,t)^{n/2}}.$$

Esto por que, $U \sim \chi^2_{(n)}$.

(b) $M_X(t) = E(\exp(\,t\,X\,))$

$$= \int_0^\infty \frac{x^{\frac{n}{2}-1}}{2^{n/2}\,\Gamma\left(\frac{n}{2}\right)}\left(\exp\left\{-\frac{x}{2}(1-2\,t)\right\}\right) dx.$$

$$= \frac{1}{(1-2t)^{n/2}} \int_0^\infty \frac{(1-2t)^{n/2}}{2^{n/2}\,\Gamma\left(\frac{n}{2}\right)}\, x^{\frac{n}{2}-1}\left(\exp\left\{-\frac{x}{2}(1-2t)\right\}\right) dx.$$

Si se realiza el cambio de variable $v = x\,(1-2\,t)$, se tiene la función de densidad de una $\chi^2_{(n)}$ sobre todo su espacio.

$= (1-2\,t)^{-n/2}$ si $t < 1/2$. ∎

Por lo que si $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$, y $Z = (\underline{X}-\mu)^t\,\Sigma^{-1}\,(\underline{X}-\mu)$, entonces $Z \sim \chi^2_{(p)}$.

1.3.5 La distribución Wishart

Esta distribución corresponde a la generalización de la distribución χ^2, que está dada por la forma cuadrática $\underline{X}^t\,\underline{X}$, donde $\underline{X}$ es una matriz de datos de una distribución $N_p(0, \Sigma)$.

Definición 1.12 *Sea M una matriz de $p \times p$ de tal forma que $M = \underline{X}^t\,\underline{X}$, donde $\underline{X}$ es una matriz de dimensión $n \times p$ y $\underline{X} \sim N_p(0, \Sigma)$, entonces se dice que M tiene distribución Wishart si su función de densidad está dada por:*

$$f_M(m) = \frac{|m|^{1/2(n-p-1)} \exp\left\{-\frac{1}{2} tr(\Sigma^{-1} m)\right\}}{2^{\frac{np}{2}} \pi^{\frac{p(p-1)}{4}} |\Sigma|^{\frac{n}{2}} \prod_{i=1}^{p} \Gamma\left[\frac{1}{2}(n+1-i)\right]} \tag{1.17}$$

donde la matriz $\Sigma > 0$, $m > 0$ y n son los grados de libertad; y se denota por $M \sim W_p(\Sigma, n)$.

Teorema 1.13 *Si $M \sim W_p(\Sigma, n)$ y B es una matriz de dimensión $p \times q$, entonces $B^t M B \sim W_q(B^t \Sigma B, n)$.*

Demostración.
Sea $\underline{X} \sim N_p(0, \Sigma)$, y $M = \underline{X}^t \underline{X}$, entonces:

$$B^t M B = B^t \underline{X}^t \underline{X} B = Y^t Y = W$$

en donde $\underline{Y} = \underline{X} B$; por el Teorema 1.9, se tiene que:

$$\underline{X} B \sim N_p(0, B^t \Sigma B)$$

y por la Definición 1.12,

$$W \sim W_q(B^t \Sigma B, n),$$

o bien,

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

■

Corolario 1.14 $\Sigma^{-1/2} M \Sigma^{1/2} \sim W_p(I, n)$.

Demostración.
Se sabe que $M = \underline{X}^t \underline{X}$, sea $\underline{Y} = \underline{X} \Sigma^{1/2}$, entonces

$$\begin{aligned} \Sigma^{-1/2} M \Sigma^{1/2} &= \Sigma^{-1/2} \underline{X}^t \underline{X} \Sigma^{1/2} \\ &= \underline{Y}^t \underline{Y} \end{aligned}$$

y por el Teorema 1.9 se sabe que $\underline{Y} \sim N_p(0, I)$, de tal forma que:

$$\Sigma^{-1/2} M \Sigma^{1/2} \sim W_p(I, n).$$

■

Corolario 1.15 *Si* $M \sim W_p(I, n)$ *y* B *es una matriz de dimensión* $p \times q$, *entonces* $B^t M B \sim W_p(I, n)$.

Demostración.
Por el Teorema 1.13 se sabe que:

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

Se puede observar que:

$$B^t I_p B = B^t B = I_q ,$$

entonces

$$B^t M B \sim W_p(I, n).$$

■

Teorema 1.16 *Si* $M \sim W_p(\Sigma, n)$ *y* a *es un vector de dimensión* p, *entonces* $\frac{a^t M a}{a^t \Sigma a} \sim \chi^2_{(n)}$.

Demostración.
Sea $M = \sum_{i=1}^{n} X_i X_i^t$ donde cada $X_i \sim N_p(0, \Sigma)$, entonces

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

donde

$$Z_i = X_i^t a = a^t X_i$$

entonces $Z_i \sim N(0, a^t \Sigma a)$ para $i = 1, \cdots, n$. Estándarizando a Z_i se tiene que:

$$\frac{Z_i}{\sqrt{a^t \Sigma a}} \sim N(0,1)$$
$$\frac{Z_i^2}{a^t \Sigma a} \sim \chi^2_{(1)}$$

y como las X_i son independientes, entonces

$$\sum_{i=1}^{n} \frac{Z_i^2}{a^t \Sigma a} \sim \chi^2_{(n)}$$

$$\frac{a^t M a}{a^t \Sigma a} \sim \chi^2_{(n)}$$

■

1.3.6 La distribución T^2-Hotelling

Esta distribución es de la forma $\underline{X}^t M^{-1} \underline{X}$ donde $\underline{X}$ tiene distribución normal, M tiene distribución Wishart y $\underline{X}$ y M son independientes.

Definición 1.17 *Si w puede escribirse como $m\, \underline{X}^t M^{-1} \underline{X}$, donde $\underline{X}$ y M son independientes y $\underline{X} \sim N_p(0, I)$ y $M \sim W_p(I, m)$ con $m > p$, entonces w tiene distribución T^2 −Hotelling con parámetros p y m y se denota por $w \sim T^2(p, m)$.*

Teorema 1.18 *Si $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$ y $M \sim W_p(\Sigma, m)$ con $m > p$ son independientes, entonces:*

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

Demostración.

Sea $Y = \Sigma^{-\frac{1}{2}}(\underline{X}-\underline{\mu}) \sim N_p(0\,, I)$ y se define $Z = \Sigma^{-\frac{1}{2}}\, M\, \Sigma^{-\frac{1}{2}}$, entonces por el Teorema 1.13

$$Z \sim W_p(I, m).$$

y por la Definición 1.17 se cumple que:

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

en donde

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

Por lo tanto

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

■

1.3.7 La distribución Lambda de Wilks (Λ)

La distribución Lambda de Wilks representa el caso multivariado de la distribución Beta, se utiliza principalmente en pruebas de Hipótesis de medias de la distribución Normal Multivariada.

Definición 1.19 *Sean W_1 y W_2 matrices aleatorias independientes con distribución $W_p(I, m)$ y $W_p(I, n)$, respectivamente y $m \geq p$. Si la variable $\underline{X}$ puede escribirse como:*

$$\begin{aligned} \underline{X} &= \frac{|W_1|}{|W_1 + W_2|} \\ &= \left| I + W_1^{-1}\, W_2 \right|^{-1} \end{aligned}$$

entonces $\underline{X}$ tiene distribución Λ de Wilks con parámetros p, m y n; y se denota como $\underline{X} \sim \Lambda(\ p,\ m,\ n)$.

Además $\underline{X} = \prod_{i=1}^{p} (1 + \lambda_i)^{-1}$ donde $\lambda_1, \lambda_2, \cdots, \lambda_p$ son los eigenvalores no negativos de W_1^{-1} y W_2.

Teorema 1.20 *Las distribuciones* $\Lambda(\ p,\ m,\ n)$ *y* $\Lambda\ (\ n,\ m+n-p,\ p)$ *son iguales.*

Demostración.

Si $\underline{Y} \sim \Lambda(\ p,\ m,\ n)$, se define $k = \min(n,\ p)$ y por la Definición 1.19, se tiene que:

$$\begin{aligned} \underline{Y} &= \left| I + W_1^{-1} W_2 \right|^{-1} \\ &= \prod_{i=1}^{p} (1 + \lambda_i)^{-1} \\ &= \prod_{i=1}^{k} (1 + \lambda_i)^{-1}. \end{aligned}$$

Como W_1^{-1} y W_2 son independientes y W_2 se distribuye Wishart, entonces $W_2 = \underline{X}^t\, \underline{X}$ donde $\underline{X} \sim N_p(0, I)$ y es una matriz de $n \times p$, entonces existe $A = \left(\underline{X}\, \underline{X}^t\right)^{-1/2} \underline{X}$, por lo que:

$$\begin{aligned} W_1^{-1} W_2 &= W_1^{-1}\, \underline{X}^t\, \underline{X} \\ &= W_1^{-1} A^t\, A\, \underline{X}^t\, \underline{X}\, A^t\, A \end{aligned}$$

que tiene los mismos eigenvalores de:

$$B^{-1}\, C = \left(A\, W_1^{-1}\, A^t\right) \left(A\, \underline{X}^t\, \underline{X}\, A^t\right)$$

donde $B \sim W_n(\ I,\ m+n-p)$ y $C \sim W_n(\ I,\ p)$ por lo que:

$$B^{-1}\, C \sim \Lambda(n,\ m+n-p, p)$$

y si $n < p$,

$$\Lambda(p,\ m,\ n) = \Lambda(n,\ m+n-p,\ p).$$

Finalmente si $n > p$

$$\Lambda(n,\ m+n-p,\ p) = \Lambda(p,\ m,\ n).$$

■

Capítulo 2

Análisis de componentes principales

2.1 Introducción

El objetivo principal del análisis de componentes principales ($A\ C\ P$) es realizar combinaciones lineales estandarizadas de tal manera que se acumule una proporción significativa de la varianza, pero sin perder información.

La técnica de las combinaciones lineales fue desarrollada por Hotelling en 1983, pero la idea original fue de Karl Pearson (1901). Los Componentes Principales son un método de la estadística multivariada; actualmente se utiliza en relaciones internacionales, sociología, medicina, y otras disciplinas.

2.2 Definición y propiedades de los componentes principales

Se considera una muestra de n individuos denotados $\underline{X}_i$, donde cada uno de estos individuos tienen asociadas p características; por lo que tenemos una matriz $\underline{X}$ de $n \times p$.

$$\underline{X} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}$$

Definición 2.1 *Si $\underline{X}$ es un vector aleatorio con media $\underline{\mu}$ y matriz de covarianzas Σ, entonces la transformación de los componentes principales está dada por:*

$$\underline{X} \longrightarrow \underline{Y} = \Gamma^t \, (\, \underline{X} - \underline{\mu} \,) \tag{2.1}$$

donde Γ es ortogonal, $\Gamma^t \, \Sigma \, \Gamma = \Lambda$ es una matriz diagonal de los eigenvalores de Σ, tal que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

El j-ésimo componente principal de $\underline{X}$ está definido como el j-ésimo elemento del vector $\underline{Y}$.

$$\underline{Y}_j = \Gamma_{(j)}^t \, (\underline{X} - \underline{\mu}).$$

Donde $\Gamma_{(j)}^t$ es la j-ésima columna de Γ.

Teorema 2.2 *Si $\underline{X}$ es un vector aleatorio con media $\underline{\mu} = \underline{0}$ y matriz de covarianzas Σ, donde Σ puede descomponerse como $\Sigma = \Gamma \, \Lambda \, \Gamma^t$ tal que $\Gamma\Gamma^t = I$ y $\underline{Y} = \Gamma^t \, (\underline{X} - \underline{\mu})$, entonces se satisfacen las siguientes propiedades:*

a. $E(\underline{Y}_j) = 0$ para toda j

b. $Var(\underline{Y}_j) = \lambda_j$

c. $Var(\underline{Y}_1) \geq Var(\underline{Y}_2) \geq ... \geq Var(\underline{Y}_p) \geq 0$

d. $\sum_{j=1}^{p} Var \, (\underline{Y}_j) = tr \, (\Sigma)$

e. $\prod_{j=1}^{p} Var \, (\underline{Y}_j) = |\Sigma|$

Demostración.

a. $E(\underline{Y}_j) = E(\Gamma^t_{(j)} \ (\underline{X} - \underline{\mu}))$

$$\begin{aligned} &= E(\Gamma^t_{(j)} \ \underline{X}) - E(\Gamma^t_{(j)} \ \underline{\mu}) \\ &= \Gamma^t_{(j)} \ E(\underline{X}) - \Gamma^t_{(j)} \ \underline{\mu} \\ &= 0. \end{aligned}$$

b. $Var(\underline{Y}) = Var(\ \Gamma^t \ (\underline{X} - \underline{\mu}))$

$$\begin{aligned} &= Var(\Gamma^t \ \underline{X}) \\ &= \Gamma^t \ Var(\underline{X}) \ \Gamma \\ &= \Gamma^t \ \Sigma \ \Gamma \\ &= \Gamma^t \ \Gamma \ \Lambda \ \Gamma^t \ \Gamma \\ &= \Lambda \end{aligned}$$

donde Λ es una matriz diagonal de los eigenvalores de Σ, y $\underline{Y} = (\underline{Y}_1, \ \underline{Y}_2, \cdots, \ \underline{Y}_p)^t$; por lo que

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

c. Por la Definición 2.1 y (b) de este Teorema:

$$Var(\underline{Y}_1) \geq Var(\underline{Y}_2) \geq ... \geq Var(\underline{Y}_p) \geq 0$$

d. $tr \ (\Sigma) = tr(\Gamma \ \Lambda \ \Gamma^t)$

$$\begin{aligned} &= tr(\Gamma^t \ \Gamma \ \Lambda) \\ &= tr(\Lambda) \\ &= \sum_{j=1}^{p} \lambda_j \\ &= \sum_{j=1}^{p} Var \ (\underline{Y}_j) \end{aligned}$$

e. $|\Sigma| = |\Gamma \ \Lambda \ \Gamma^t|$

$$
\begin{aligned}
&= |\Gamma|\ |\Lambda|\ |\Gamma^{t}| \\
&= |\Gamma\ \Gamma^{t}|\ |\Lambda| \\
&= |\Lambda| \\
&= \prod_{j=1}^{p} \lambda_{j} \\
&= \prod_{j=1}^{p} Var(\underline{Y}_{j})
\end{aligned}
$$

■

Definición 2.3 *Una combinación lineal estandarizada de* $\underline{X}$ *se define como:*

$$\underline{Y} = \underline{a}^{t}\ \underline{X}$$

donde $\underline{a}^{t}\ \underline{a} = 1$

Teorema 2.4 *Sea* $\underline{Y} = \underline{a}^{t}\ \underline{X}$ *una combinación lineal estandarizada de* $\underline{X}$, *donde* $E(\underline{X}) = \underline{\mu}$ *y* $V(\underline{X}) = \Sigma$, *entonces* $Var(\underline{Y}) \leq \lambda_{j}$ *y la descomposición espectral de* Σ *se define* $\Sigma = \Gamma\ \Lambda\ \Gamma^{t}$, $\Lambda = Diag(\lambda_{1}, \lambda_{2}, \ldots, \lambda_{p})$ *y* $\Gamma\ \Gamma^{t} = I$.

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

$\underline{a} = \Gamma\ \underline{b}$, con $\underline{b}^{t} = (b_{1}, b_{2}, \ldots, b_{p})$

$$
\begin{aligned}
\underline{a} &= \left(\Gamma_{(1)}\ \Gamma_{(2)}\ \ldots\ \Gamma_{(p)}\right)\ \underline{b} \\
&= \sum_{j=1}^{p} \underline{b}_{j}\ \Gamma_{(j)}\ .
\end{aligned}
$$

por la Definición 2.3 se tiene:

$$
\begin{aligned}
1 &= \underline{a}^{t}\ \underline{a} \\
&= (\Gamma\ \underline{b})^{t}\ (\Gamma\ \underline{b}) \\
&= \underline{b}^{t}\ \underline{b}
\end{aligned}
$$

entonces la $Var(\underline{Y})$ queda

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

y la varianza se maximiza para

$\underline{b} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ y así $\underline{a} = \Gamma_{(1)}$. Por lo que

$$\begin{aligned} Var(\underline{Y}) &= \Gamma_{(1)}^{\,t}\,\Lambda\,\Gamma_{(1)} \\ &= \lambda_1 \end{aligned}$$

■

2.3 Propiedades de los componentes principales

1. La proporción de variabilidad explicada por los primeros k componentes principales se expresa como: $\dfrac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$

2. Cuando algunas de las variables originales son linealmente dependientes, se tienen algunos eigenvalores iguales a cero. Por lo que si el $ran(\Sigma) = q$ y $q < p$, entonces la variabilidad de $\underline{X}$ es explicada por los q primeros componentes.

3. Si existen eigenvalores de multiplicidad $n > 1$, de tal manera que solamente se tengan t eigenvalores distintos ($t < p$); entonces la matriz Γ no es UNICA, ya que podría multiplicarse por otra matriz de la forma $A = Diag(A_1, A_2, \ldots, A_t)$, donde cada A_j es una matriz ortogonal de orden m_j siendo m_j la multiplicidad de los eigenvalores, y la matriz de covarianzas de Y está dada por: $Var(Y) = \Gamma^t \Sigma \Gamma = \Lambda$.

4. Los componentes principales de un vector aleatorio *no* son invariantes respecto a la escala; es decir, las variables pueden tener diferente unidad de medida y la interpretación de los datos cambia al realizar cambios en la escala, como por ejemplo al tratar de que todas las variables tengan la misma unidad de medida.

2.4 Componentes principales a partir de la muestra

Generalmente la matriz de covarianzas Σ es desconocida, pero se puede estimar a partir de la muestra.

2.4.1 Componentes principales con la matriz de covarianzas

Como la matriz Σ es desconocida el análisis de componentes principales se basa en la matriz de covarianzas muestral S_X.

$$S_{jk} = \frac{1}{n} \sum_{i=1}^{n} (\underline{X}_{ij} - \overline{X}_j)(\underline{X}_{ik} - \overline{X}_k)^t$$

donde $\overline{X}_j = \frac{1}{n} \sum_{i=1}^{n} \underline{X}_{ij}$ es la media de los valores observados para la j-ésima variable y $S_{jk} = S_X$ es la covarianza entre las variables $\underline{X}_j$ y $\underline{X}_k$.

Definición 2.5 *Sea U una matriz ortogonal cuyos elementos en la diagonal son positivos tales que $U^t S_X U = \widetilde{\Lambda}$, donde $\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_p$ son los valores propios ordenados que corresponden a la matrix S_X, y $U^t U = I$, entonces la transformación de los componentes principales está dada por:*

$$\underline{Y} = U^t (\underline{X} - \overline{X})$$

por lo que la i-ésima observación es: $\underline{Y}_i = U^t (\underline{X}_i - \overline{X})$

La media muestral de $\underline{Y}$ está dada:

$$\begin{aligned} \overline{Y} &= \frac{1}{n} \sum_{i=1}^{n} \underline{Y}_i \\ &= \frac{1}{n} \sum_{i=1}^{n} U^t (\underline{X}_i - \overline{X}) \\ &= \frac{1}{n} U^t \sum_{i=1}^{n} (\underline{X}_i - \overline{X}) \\ &= 0 \end{aligned}$$

y la matriz de covarianzas muestral S_Y puede definirse como:

$$\begin{aligned} S_Y &= \frac{1}{n} \sum_{i-1}^{n} (\underline{Y}_i - \overline{Y}) (\underline{Y}_i - \overline{Y})^t \\ &= \frac{1}{n} \sum_{i=1}^{n} U^t (\underline{X}_i - \overline{X}) (\underline{X}_i - \overline{X})^t U \\ &= U^t S_X U \\ &= L , \end{aligned}$$

en donde el j-ésimo componente principal está definido por:

$$\underline{Y}_{ij} = U_{(j)}^t (\underline{X}_i - \overline{X}).$$

2.4.2 Componentes principales a partir de variables estandarizadas

Las variables que se tienen están medidas en unidades diferentes por lo que es necesario hacer una estandarización y así poder evitar que las diferentes escalas influyan en la determinación de los componentes principales.

La estandarización de $\underline{X}$ queda definida como:

$$\underline{Z} = \left(D^{\frac{1}{2}}\right)^{-1} (\underline{X} - \overline{X})$$

para cada $\underline{X}_i$, donde la matriz $D = Diag(S_X)$ y $\underline{Z}$ tiene media *cero* y varianza

$$\begin{aligned} Var(\underline{Z}) &= Var\left(\left(D^{\frac{1}{2}}\right)^{-1} (\underline{X} - \overline{X})\right) \\ &= \left(D^{\frac{1}{2}}\right)^{-1} Var(\underline{X}) \left(D^{\frac{1}{2}}\right)^{-1} \\ &= R\,, \end{aligned}$$

donde R es la matriz de correlación muestral, cuya descomposición espectral es:

$$R = \widetilde{U}\ \widetilde{L}\ \widetilde{U}^{\,t}.$$

Definición 2.6 *La transformación de los componentes principales está dada por:*

$$\underline{Y} = \widetilde{U}^{\,t}\ \underline{Z}$$

y la varianza puede definirse como:

$$\begin{aligned} Var(\underline{Y}) &= Var(\widetilde{U}^{\,t}\ \underline{Z}) \\ &= \widetilde{U}^{\,t}\ Var(\underline{Z})\ \widetilde{U} \\ &= \widetilde{U}^{\,t}\ R\ \widetilde{U} \\ &= \widetilde{L}. \end{aligned}$$

2.4.3 Estructura de la correlación

Es importante examinar la *correlación* entre el punto $\underline{X}$ y el vector de componentes principales $\underline{Y}$. La *covarianza* entre $\underline{X}$ y $\underline{Y}$ se calcula:

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

La *covarianza* entre $\underline{X}_i$ y $\underline{Y}_j$ es: $C(\underline{X}_i, \underline{Y}_j) = \Gamma_{ij}\ \lambda_j$.

La $Var(\underline{X}) = \Sigma = \{\sigma_{ij}\}$, $i, j = 1, \ldots, p$ y Λ es la matriz diagonal de las varianzas de $\underline{Y}$, por lo que la *correlación* entre las variables $\underline{X}_i$ y $\underline{Y}_j$ se obtiene como:

$$\begin{aligned} \rho_{ij} &= \frac{\Gamma_{ij}\ \lambda_j}{\sqrt{\sigma_{ii}\ \lambda_j}} \\ &= \frac{\Gamma_{ij}\ \sqrt{\lambda_j}}{\sqrt{\sigma_{ii}}}. \end{aligned}$$

Entonces la proporción de variabilidad explicada de $\underline{X}_i$ por la componente $\underline{Y}_j$ es:

$$\rho^{2}_{ij} = \frac{\Gamma^{2}_{ij}\ \lambda_j}{\sigma_{ii}}.$$

2.5 Interpretación gráfica de los componentes principales

La forma cuadrática $(\underline{X} - \underline{\mu})^{t}\ \Sigma^{-1}(\underline{X} - \underline{\mu}) = c$; es un elipsoide de dimensión p, donde c es una constante y si se hace variar se genera una familia de estos elipsoides. Por el teorema A.21 $\Sigma^{-1} = U\ \Lambda^{-1}\ U^{t}$ se tiene:

$$\begin{aligned} c &= (\underline{X}-\underline{\mu})\,{}^{t}U\,\Lambda^{-1}\,U^{t}\,(\underline{X}-\underline{\mu}) \\ &= \underline{Y}^{\,t}\,\Lambda^{-1}\underline{Y} \end{aligned} \tag{2.2}$$

y los ejes principales son los vectores propios de Σ.

Ejemplo 2.1 *Encontrar la elipse de concentración del 95% para* $\underline{X}$ *que se distribuye normal bivariada con los siguientes parámetros.*

$$\underline{\mu} = \begin{pmatrix} -4 \\ 12 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 4 & 5 \\ 5 & 9 \end{pmatrix} \text{ donde } \Lambda^{-1} = \begin{pmatrix} 12.0902 & 0 \\ 0 & .90983 \end{pmatrix}$$

De (2.2) se sigue

$$\begin{aligned} c &= \begin{pmatrix} X_1+4 \\ X_2-12 \end{pmatrix}^{t} U\,\Lambda^{-1}\,U^{t} \begin{pmatrix} X_1+4 \\ X_2-12 \end{pmatrix} \\ &= \begin{pmatrix} X_1+4 \\ X_2-12 \end{pmatrix}^{t} \begin{pmatrix} .5257 & .8506 \\ .8506 & -.5257 \end{pmatrix} \Lambda^{-1} \\ &\times \begin{pmatrix} .5257 & .8506 \\ .8506 & -.5257 \end{pmatrix} \begin{pmatrix} X_1+4 \\ X_2-12 \end{pmatrix} \\ &= \begin{pmatrix} .5257\,(X_1+4)+.8506\,(X_2-12) \\ .8506\,(X_1+4)-.5257\,(X_2-12) \end{pmatrix} \begin{pmatrix} 12.0902 & 0 \\ 0 & .90983 \end{pmatrix}^{-1} \\ &\times \begin{pmatrix} .5257\,(X_1+4)+.8506\,(X_2-12) \\ .8506\,(X_1+4)-.5257\,(X_2-12) \end{pmatrix} \end{aligned}$$

Para obtener la elipse de concentración del 95% hay que calcular

$$P\left[\underline{Y}^{\,t}\,\Lambda^{-1}\underline{Y} \leq c\right] = .95$$

en donde $c = 5.99$ es el cuartil de una $\chi^2_{(2)}$; por lo los ejes principales se calculan:

$$\begin{aligned} \underline{Y}_1 &= 0, \text{ entonces } \underline{Y}_2 = \pm\sqrt{(.90983)\,(5.99)} \text{ y} \\ \underline{Y}_2 &= 0, \text{ entonces } \underline{Y}_1 = \pm\sqrt{(12.09021)\,(5.99)} \end{aligned}$$

finalmente, $\underline{Y}_1 = 8.51$ y $\underline{Y}_2 = 2.33$

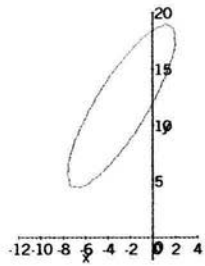

Figura 1. Elipse de concentración del 95%

2.6 Aplicaciones

Ejemplo 2.2 *Se analizará el ejemplo de las Irises de Fisher con una muestra de 50 observaciones por cada tipo de flor, el cual consiste en 3 tipos de flor de Iris (Iris Setosa, Iris Virgínica e iris Versicolor), y para cada uno de estos tipos de flor se considera el largo del sépalo(SEPALLEN), ancho del sépalo (SEPALWID), el largo del pétalo (PETALLEN) y ancho del pétalo (PETALWID).*

Utilizando el paquete Statistica, se observa que la variable PETELLEN (largo del pétalo) tiene mayor variación, es decir, que existe mayor variabilidad en el tamaño de las flores.

Means and Standard Deviations (irisdat.sta)
Casewise deletion of MD
N=150

	Means	Std.Devs
SEPALLEN	5.8433	0.8281
SEPALWID	3.0573	0.4359
PETALLEN	3.7580	1.7653
PETALWID	1.1993	0.7622

Tabla 1.

Con la matriz de correlaciones se sabe si existe relación entre las variables, dada en la siguiente tabla.

Correlations (irisdat.sta)
Casewise deletion of MD
N=150

	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.00	-0.12	0.87	0.82
SEPALWID	-0.12	1.00	-0.43	-0.37
PETALLEN	0.87	-0.43	1.00	0.96
PETALWID	0.82	-0.37	0.96	1.00

Tabla 2. Correlaciones.

por lo que las variables altamente correlacionadas son: largo y ancho del pétalo (PETALLEN y PETAIWID); también largo del pétalo y ancho del sépalo (PETALLEN y SEPALLEN); y finalmente ancho del pétalo y ancho del sépalo (PETAIWID y SEPALLEN) tienen correlación alta.

Es decir, las flores que tienen pétalos largos también tienen pétalos anchos, e inclusive sépalos largos.

Con la siguiente gráfica se puede apreciar la relación lineal de las variables.

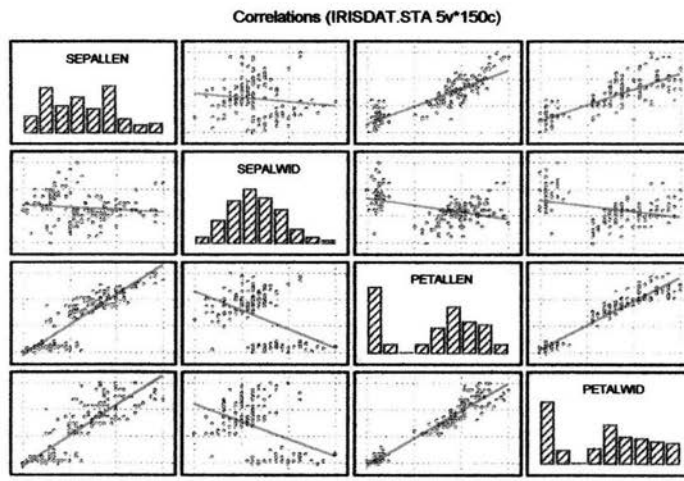

Figura 2.

Para poder saber el número de componentes principales necesarios, se utiliza la tabla de los eigenvalores.

Eigenvalues (irisdat.sta)
Extraction: Principal components

	Eigenval	% total Variance	Cumul. Eigenval	Cumul. %
1	2.9185	72.9624	2.9185	72.9624
2	0.9140	22.8508	3.8325	95.8132
3	0.1468	3.6689	3.9793	99.4821
4	0.0207	0.5179	4.0000	100.0000

Tabla 3. Eigenvalores.

y con 2 componentes se tiene más del 95% de la variación total de los datos.

Los correspondientes eigenvectores están dados en la siguiente tabla:

U_1	U_2	U_3	U_4	Λ
0.5210	-0.3792	-0.7199	0.2613	2.9185
-0.2693	-0.9225	0.2458	-0.1222	0.9140
0.5804	-0.0255	0.1458	-0.8014	0.1468
0.5648	-0.0672	0.6325	0.5257	0.0207

Tabla 4. Eigenvalores y eigenvectores.

Así que los componentes principales están dados como:

$$Y_1 = 0.5210\,X_1 - 0.2693\,X_2 + 0.5804\,X_3 + 0.5648\,X_4$$
$$Y_2 = -0.3792\,X_1 - 0.9225\,X_2 - 0.0255\,X_3 - 0.0672\,X_4$$

Ejemplo 2.3 *La siguiente tabla muestra los datos de 46 pacientes con diabetes; se realizó un análisis de componentes principales y se consideraron* 5 *variables que son:*

X_1	Intolerencia a la glucosa.
X_2	Respuesta de insulina a la glucosa oral.
X_3	Resistencia de la insulina.
Y_1	Peso relativo.

Y_2 Plasma y glucosa en ayunas.

Paciente	Y1	Y2	X1	X2	X3
1	0.81	80	356	124	55
2	0.95	97	289	117	76
3	0.94	105	319	143	105
4	1.04	90	356	199	108
5	1.00	90	323	240	143
6	0.76	86	381	157	165
7	0.91	100	350	221	119
8	1.10	85	301	186	105
9	0.99	97	379	142	98
10	0.78	97	296	131	94
11	0.90	91	353	221	53
12	0.73	87	306	178	66
13	0.96	78	290	136	142
14	0.84	90	371	200	93
15	0.74	86	312	208	68
16	0.98	80	393	202	102
17	1.10	90	364	152	76
18	0.85	99	359	185	37
19	0.83	85	296	116	60
20	0.93	90	345	123	50
21	0.95	90	378	136	47
22	0.74	88	304	134	50
23	0.95	95	347	184	91
24	0.97	90	327	192	124
25	0.72	92	386	279	74
26	1.11	74	365	228	235
27	1.20	98	365	145	158
28	1.13	100	352	172	140
29	1.00	86	325	179	145
30	0.78	98	321	222	99
31	1.00	70	360	134	90
32	1.00	99	336	143	105
33	0.71	75	352	169	32
34	0.76	90	353	263	165
35	0.89	85	373	174	78
36	0.88	99	376	134	80
37	1.17	100	367	182	54
38	0.85	78	335	241	175
39	0.97	106	396	128	80
40	1.00	98	277	222	186
41	1.00	102	378	165	117
42	0.89	90	360	282	160
43	0.98	94	291	94	71
44	0.78	90	269	121	29
45	0.74	93	318	73	42
46	0.91	86	328	106	56

Tabla 5.

Con la matriz de correlaciones es posible saber que tan correlacionadas son dichas variables. Utilizando el paquete Statistica se obtiene la siguiente tabla:

Correlations (e_cp1.sta)
Casewise deletion of MD
N=46

	Y1	Y2	X1	X2	X3
Y1	1.000	0.177	0.186	-0.034	0.372
Y2	0.177	1.000	0.037	-0.087	-0.095
X1	0.186	0.037	1.000	0.244	0.070
X2	-0.034	-0.087	0.244	1.000	0.507
X3	0.372	-0.095	0.070	0.507	1.000

Tabla 6. Correlaciones.

por lo que las variables que tiene relación son: resistencia de la insulina con respuesta de insulina a la glucosa oral y peso relativo, (X_2 con X_3 y X_3 con Y_1); pero no una correlación ni siquiera positiva con Y_1 y X_2. Tambien hay relación de la intolerancia a la glucosa y respuesta de insulina a la glucosa oral, (X_1 con X_2).

Para saber el número de componentes a utilizar, así como el porcentaje de varianza explicada, se utiliza la tabla de eigenvalores.

Eigenvalues (e_cp1.sta)
Extraction: Principal components

	Eigenval	% total Variance	Cumul. Eigenval	Cumul. %
1	1.715	34.294	1.715	34.294
2	1.223	24.460	2.938	58.754
3	0.953	19.059	3.891	77.813
4	0.810	16.199	4.701	94.012
5	0.299	5.988	5.000	100.000

Tabla 7. Eigenvalores.

Es necesario utilizar 3 componentes para obtener el 77.81% de la variación total de los datos. El porcentaje obtenido no es muy alto, pero se disminuye la dimensión de 5 variables a 3; por lo que el análisis se realiza con 3 componentes.

Los respectivos eigenvectores son:

U_1	U_2	U_3	U_4	U_5	Λ
0.6395	0.0902	0.4109	0.0892	0.6373	1.715
0.3530	-0.1744	-0.8169	-0.3429	0.2451	1.223
-0.0387	-0.6847	-0.1304	0.7057	0.1211	0.953
0.3885	-0.5951	0.3312	-0.4158	-0.4609	0.810
0.5604	0.3721	-0.1928	0.4512	-0.5538	0.299

Tabla 8. Eigenvectores y eigenvalores.

Y los componentes principales están dados como:

$$
\begin{aligned}
Y_1 &= 0.6395\,X_1 + 0.3530\,X_2 - 0.0387\,X_3 + 0.3885\,X_4 + 0.5604\,X_5 \\
Y_2 &= 0.0902\,X_1 - 0.1744\,X_2 - 0.6847\,X_3 - 0.5951\,X_4 + 0.3721\,X_5 \\
Y_3 &= 0.4109\,X_1 - 0.8169\,X_2 - 0.1304\,X_3 + 0.3312\,X_4 - 0.1928\,X_5
\end{aligned}
$$

2.7 Análisis de factores

En 1904, Spearman publicó un artículo el cual se pensó que era el origen del análisis de factores; y es utilizado en los "cambios de población" como natalidad, mortalidad, migración; además Kaiser desarrolló el método *Varimax* para rotaciones ortogonales. Sin embargo en 1977 Hills dice: "El Análisis de Factores ($A\ F$) no justifica el tiempo requerido para realizarlo y comprenderlo..."

En el análisis de factores se intenta representar a las variables $\underline{Y}_1, \underline{Y}_2, \cdots, \underline{Y}_p$ como combinaciones lineales de otras variables aleatorias $f_1, f_2, \cdots, f_m$ llamadas *factores* en donde $m < p$.

El objetivo del análisis de factores es caracterizar la repetición entre las variables por un número menor de factores; es decir, si las variables originales $\underline{Y}_1, \underline{Y}_2, \cdots, \underline{Y}_p$ están correlacionadas, entonces la dimensionalidad es menor que p.

Suponga que en la matriz de correlaciones las variables se pueden dividir en subconjuntos de tal forma que tengan correlaciones altas entre ellas, pero

pequeñas con todas las otras variables, entonces las variables de ese subconjunto estarán representadas por un *factor subyacente*. Si las otras variables pueden ser agrupadas similarmente en subconjuntos, tendremos que unos "pocos" factores pueden representar a las variables.

Ejemplo 2.4 *Se tiene la siguiente matriz de correlaciones*

$$\begin{pmatrix} 1 & .85 & .3 & .25 & .92 \\ .85 & 1 & .07 & -.16 & .88 \\ .3 & .07 & 1 & \mathbf{.9} & .29 \\ .25 & -.16 & \mathbf{.9} & 1 & .39 \\ .92 & .88 & .29 & .39 & 1 \end{pmatrix}$$

entonces las variables 1, 2 y 5 corresponden a un factor y las variables 3 y 4 corresponden a otro factor.

2.8 Modelo de factores ortogonales

En el modelo de factores se tiene una muestra aleatoria $\underline{Y}_1, \underline{Y}_2, \cdots, \underline{Y}_p$ de una población con vector de medias $\underline{\mu}$ y matriz de covarianzas Σ.

El modelo de análisis de factores expresa a cada variable como una combinación lineal de *factores comunes* o "*subyacentes*" $f_1, f_2, \cdots, f_m$ con un término de error.

Lo que buscamos es que $m < p$, en caso contrario no se ha encontrado una descripción de las variables como funciones de unos pocos *factores subyacentes*.

Definición 2.7 *Sea $\underline{Y}$ un vector aleatorio con media $\underline{\mu}$ y matriz de covarianzas Σ ($\Sigma > 0$); se tienen m factores subyacentes ($m < p$), $f_1, f_2, \cdots, f_m$ de tal forma que $\underline{Y}$ puede escribirse como:*

$$\underline{Y} - \underline{\mu} = \Lambda f + \varepsilon \tag{2.3}$$

donde Λ es una matriz de constantes de $p \times m$, f y ε son vectores aleatorios, los elementos de ε son los factores específicos; con los siguientes supuestos:

$$E(f) = 0 \qquad Var(f) = I \qquad E(\varepsilon) = 0$$
$$Var(\varepsilon) = \Psi \qquad cov(f\ , \varepsilon) = 0 \qquad cov(\varepsilon_{\ i}\ ,\ \varepsilon_{\ j}) = 0 \text{ para } i \neq j$$

Cada $\varepsilon_{\ i}$ indica la variación de la i-ésima variable, y $\Psi_{\ i}$ es su varianza; $\lambda_{\ i\ j}$ son los elementos de la matriz Λ llamados *cargas* e indican la importancia de las f con las $\underline{Y}_{\ i}$.

Es necesario poder expresar a Σ en términos de Λ y Ψ por lo que:

$$\begin{aligned} \Sigma &= Var(\underline{Y}) \\ &= Var(\Lambda\ f + \varepsilon) \\ &= Var(\Lambda\ f) + Var(\varepsilon) \\ &= \Lambda\ Var(f)\ \Lambda^{\ t} + \Psi \\ &= \Lambda\ \Lambda^{\ t} + \Psi. \end{aligned} \tag{2.4}$$

Con respecto a la varianza se tiene:

$$\begin{aligned} Var(\underline{Y}_{\ i}) &= \lambda^{\ 2}_{\ i\ 1} + \lambda^{\ 2}_{\ i\ 2} + \cdots + \lambda^{\ 2}_{\ i\ m} + \Psi_{\ i} \\ &= h^{\ 2}_{\ i} + \Psi_{\ i}. \end{aligned}$$

en donde $\Psi_{\ i}$ es la varianza especifica y $h^{\ 2}_{\ i} = \lambda^{\ 2}_{\ i\ 1} + \lambda^{\ 2}_{\ i\ 2} + \cdots + \lambda^{\ 2}_{\ i\ m}$ es la *comunalidad* o *varianza común*

En el análisis de factores se desea poder expresar las $\dfrac{p\ (p+1)}{2}$ varianzas y covarianzas de $\underline{Y}$ a partir de las $p \times m$ cargas $(\lambda_{\ i\ j})$ y las p varianzas específicas $\Psi_{\ i}$, si $m = p$ la matriz Σ puede ser expresada como en (2.4); sin embargo la mayoria de las matrices de covarianzas no pueden ser factorizadas de esa forma.

Si $m > 1$, entonces (2.3) puede ser multiplicado por una matriz ortogonal; sea $I = T\ T^{\ t}$ por lo que:

$$\begin{aligned} \underline{Y} - \mu &= \Lambda\ T\ T^{\ t}\ f + \varepsilon \\ &= \Lambda^{\ *}\ f^{\ *} + \varepsilon \end{aligned}$$

en donde $\Lambda^* = \Lambda T$ y $f^* = T^t f$ y estos nuevos factores satisfacen las condiciones iniciales.

De (2.4) se obtiene:

$$\begin{aligned}\Sigma &= \Lambda \Lambda^t + \Psi \\ &= \Lambda T T^t \Lambda^t + \Psi \\ &= \Lambda^* \Lambda^{*t} + \Psi.\end{aligned}$$

De tal forma que las nuevas cargas Λ^* reproducen a la matriz de covarianzas, exactamente como Λ.

2.9 Método del componente principal

El nombre *componente principal no* se refiere al análisis de componentes principales, los factores están relacionados con los m eigenvalores más grandes, entonces las *cargas* del j-ésimo factor son proporcionales a los coeficientes del j-ésimo componente principal y se podría suponer que se tiene la misma interpretación que en el análisis de componentes principales, sin embargo al realizar rotaciones en las *cargas* la interpretación es diferente.

2.9.1 Estimación de los parámetros

Teorema 2.8 *En una muestra aleatoria* $\underline{Y}_1, \underline{Y}_2, \cdots, \underline{Y}_p$ *de una población con vector de medias* $\underline{\mu}$ *y matriz de covarianzas* Σ, *es muy común tener la matriz de covarianzas muestral* S, *entonces es necesario encontrar un estimador de* $\widehat{\Lambda}$ *y de* $\widehat{\Psi}$ *tal que se tenga la siguiente relación:*

$$S \approxeq \widehat{\Lambda}\, \widehat{\Lambda}^t + \widehat{\Psi}.$$

Demostración.

Primero se encontrará el estimador para Λ, utilizando la descomposición espectral, Teorema A.21, se tiene que:

$$S = C\ D\ C^{t}$$

donde C es una matriz ortogonal formada por los eigenvectores estandarizados de S y $D = Diag\,(\theta_1, \theta_2, \cdots, \theta_p)$, θ_i son los eigenvalores de la matriz S; se utiliza θ_i en vez de λ_i para evitar confusión con las *cargas*. Si se factoriza $D = D^{1/2}\ D^{1/2}$, entonces se tiene que:

$$\begin{aligned} S &= C\ D\ C^{t} \\ &= C\ D^{1/2} D^{1/2}\ C^{t} \\ &= C\ D^{1/2} (C\ D^{1/2})^{t}. \end{aligned}$$

Sería fácil proponer a $\widehat{\Lambda} = C\ D^{1/2}$, pero esta matriz es de $p \times p$ y se necesita que $\widehat{\Lambda}$ sea una matriz de $p \times m$, con $m < p$.

Por lo que se define una nueva variable D_1 que contenga los m eigenvalores más grandes $\theta_1 > \theta_2 > \cdots > \theta_m$ y sea C_1 la matriz de los m eigenvectores correspondientes a D_1. Y así

$$\widehat{\Lambda} = C_1\ D_1^{1/2}$$

y esta matriz sí es de $p \times m$; en donde el i-ésimo elemento de $\widehat{\Lambda}\ \widehat{\Lambda}^{t}$ está dado por:

$$\lambda_i\ \lambda_i^{t} = \sum_{j=1}^{m} \widehat{\lambda}_{ij}^{2}.$$

En donde se define el estimador para Ψ_i como:

$$\widehat{\psi}_{ii} = s_{ii} - \sum_{j=1}^{m} \widehat{\lambda}_{ij}^{2}$$

o bién,

$$\begin{aligned} \widehat{\sigma}_{ii} &= s_{ii} \\ &= \sum_{j=1}^{m} \widehat{\lambda}_{ij}^{2} + \widehat{\psi}_{ii} \\ &= \widehat{h}_i^{2} + \widehat{\psi}_{ii}. \end{aligned} \tag{2.5}$$

Por lo que:

$$S \cong \widehat{\Lambda}\,\widehat{\Lambda}^{\,t} + \widehat{\Psi}.$$

■

2.9.2 Estimación de los parámetros con la matriz de correlaciones

En la práctica R (la matriz de correlaciones) es más utilizada que S (la matriz de covarianzas), y frecuentemente se obtienen mejores resultados ya que la mayoría de los paquetes estadísticos utilizan R.

Cuando se utiliza la matriz de correlaciones R, los eigenvalores y eigenvectores son los que se utilizan para estimar a $\widehat{\Lambda}$

Y el desarrollo es exactamente el mismo por lo que:

$$R \cong \widehat{\Lambda}\,\widehat{\Lambda}^{\,t} + \widehat{\Psi}.$$

Ya que el objetivo del análisis de factores consiste en reproducir las covarianzas o correlaciones en lugar de las varianzas.

La varianza de la i-ésima variable está particionada, una parte corresponde a los factores y la otra unicamente a la variable. Esto se obtiene de (2.5).

Así el j-ésimo factor contribuye λ_{ij}^{2} a s_{ii}. La contribución del j-ésimo factor para la varianza muestral total está dada como:

$$tr(S) = s_{11} + s_{22} + \cdots + s_{pp}$$

que es la suma de cuadrados de las *cargas* de la j-ésima columna de Λ

$$\widehat{\lambda}_{1j}^{2} + \widehat{\lambda}_{2j}^{2} + \cdots + \widehat{\lambda}_{pj}^{2} = \theta_{j}$$

donde θ_{j} es el j-ésimo eigenvalor.

Entonces la proporción de la varianza muestral total del j-ésimo factor es:

$$\frac{\widehat{\lambda}_{1j}^{2} + \widehat{\lambda}_{2j}^{2} + \cdots + \widehat{\lambda}_{pj}^{2}}{tr(S)} = \frac{\theta_j}{tr(S)}.$$

Al utilizar R, la proporción correspondiente está dada por:

$$\frac{\widehat{\lambda}_{1j}^{2} + \widehat{\lambda}_{2j}^{2} + \cdots + \widehat{\lambda}_{pj}^{2}}{tr(R)} = \frac{\theta_j}{p}$$

donde p es el número de variables. Por lo que si la comunalidad es pequeña, la proporción de la varianza también será pequeña.

Ejemplo 2.5 *Utilizando el método del componente principal se tiene una muestra de 15 estudiantes a los que se les pidió un ensayo formal y otro informal, los datos son los siguientes:*

alumno	y_1	y_2	x_1	x_2
1	*148*	*20*	*137*	*15*
2	*159*	*24*	*164*	*25*
3	*144*	*19*	*224*	*27*
4	*103*	*18*	*208*	*33*
5	*121*	*17*	*178*	*24*
6	*89*	*11*	*128*	*20*
7	*119*	*17*	*154*	*18*
8	*123*	*13*	*158*	*16*
9	*76*	*16*	*102*	*21*
10	*217*	*29*	*214*	*25*
11	*148*	*22*	*209*	*24*
12	*151*	*21*	*151*	*16*
13	*83*	*7*	*123*	*13*
14	*135*	*20*	*161*	*22*
15	*178*	*15*	*175*	*23*

en donde las variables fueron:

y_1 = *número de palabras en el ensayo informal.*
y_2 = *número de verbos en el ensayo informal.*
x_1 = *número de palabras en el ensayo formal.*
x_2 = *número de verbos en el ensayo formal.*

Los siguientes resultados se obtuvieron del paquete Statistica.

Eigenvalues (factor.sta)
Extraction: Principal components

	Eigenval	% total Variance	Cumul. Eigenval	Cumul. %
1	2.6657	66.6436	2.6657	66.6436
2	0.8993	22.4834	3.5651	89.1270
3	0.3276	8.1910	3.8927	97.3180
4	0.1073	2.6820	4.0000	100.0000

Tabla 9. Eigenvalores.

Por lo que con dos factores tenemos más del 89% de la varianza total. A continuación se dan los resultados de las comunalidades así como las varianzas y las cargas de los factores.

Variables	Cargas $\widehat{\lambda}_{1j}$	Cargas $\widehat{\lambda}_{2j}$	Comunalidad $\widehat{h}_i^2$	Varianzas específicas $\widehat{\psi}_i$
y_1	0.8023	-0.5351	0.930	0.070
y_2	0.8557	-0.3264	0.839	0.161
x_1	0.8832	0.2701	0.853	0.147
x_2	0.7140	0.6584	0.943	0.057
Contribución de la varianza total	2.6657	0.8993	3.5650	
Proporción de la varainza total	0.6664	0.2248	0.8912	
Proporción acumulada	0.6664	0.8912	0.8912	

(2.6)

La interpretación apropiada de los factores se obtiene después de que se han rotado los ejes; pero se verá más adelante.

2.10 Método del factor principal

El método del factor principal o del eje principal utiliza un estimador $\widehat{\Psi}$ y el análisis se hace sobre $R - \widehat{\Psi}$ o $S - \widehat{\Psi}$; donde R es la matriz de correlaciones y S es la matriz de covarianzas.

2.10.1 Estimación de los parámetros

Como ya se mencionó en la práctica es más utilizada R que S. De manera análoga al *método del componente principal* es necesario tener un estimador inicial de $\widehat{\Psi}$ y el análisis se desarrolla sobre $S - \widehat{\Psi}$ o $R - \widehat{\Psi}$.

$$R - \widehat{\Psi} \approx \widehat{\Lambda}\,\widehat{\Lambda}^{\,t}.$$

$$S - \widehat{\Psi} \approx \widehat{\Lambda}\,\widehat{\Lambda}^{\,t}.$$

Los elementos en la diagonal de $S - \widehat{\Psi}$ ($h_i^{\,2} = 1 - \widehat{\psi}_{\,i}$) son las *comunalidades*; y cuando se utiliza S o R los valores de $\widehat{\psi}_{\,i}$ y $\widehat{h}_i^{\,2}$ son diferentes.

$$S - \widehat{\Psi} = \begin{pmatrix} \widehat{h}_1^{\,2} & s_{1\,2} & \cdots & s_{1\,p} \\ s_{2\,1} & \widehat{h}_2^{\,2} & \cdots & s_{2\,p} \\ \vdots & \vdots & \cdots & \vdots \\ s_{p\,1} & s_{p\,2} & \cdots & \widehat{h}_p^{\,2} \end{pmatrix}$$

$$R - \widehat{\Psi} = \begin{pmatrix} \widehat{h}_1^{\,2} & r_{1\,2} & \cdots & r_{1\,p} \\ r_{2\,1} & \widehat{h}_2^{\,2} & \cdots & r_{2\,p} \\ \vdots & \vdots & \cdots & \vdots \\ r_{p\,1} & r_{p\,2} & \cdots & \widehat{h}_p^{\,2} \end{pmatrix}$$

Por lo que un estimador para la comunalidad de $R - \widehat{\Psi}$ es:

$$\widehat{h}_i^{\,2} = r_{i\,i} - \frac{1}{r^{\,i\,i}}$$

donde $r_{i\,i}$ ($s_{i\,i}$) es el i-ésimo elemento de la diagonal de R (S), $r^{\,i\,i}$ ($s^{\,i\,i}$) es el i-ésimo elemento de la diagonal de $R^{\,-1}$($S^{\,-1}$). Para poder utilizar el estimador, es necesario que R sea una matriz singular, de no ser así el estimador de la comunalidad puede ser el valor absoluto, o el cuadrado de la correlación más grande del i-ésimo renglón.

2.10.2 Aplicación del modelo

La proporción de la varianza explicada por el j-ésimo factor es:

$$\frac{\theta_j}{tr(R-\widehat{\Psi})} = \frac{\theta_j}{\sum_{i=1}^{p}\theta_i}$$

$$\frac{\theta_j}{tr(S-\widehat{\Psi})} = \frac{\theta_j}{\sum_{i=1}^{p}\theta_i}$$

donde θ_j es el j-ésimo eigenvalor de $R-\widehat{\Psi}$ o $S-\widehat{\Psi}$, estas matrices frecuentemente tendrán eigenvalores negativos; en este caso la proporción de la varianza excedera de 1.

Ejemplo 2.6 *Para el método del factor principal se utilizan los datos del Ejemplo 2.5.*

Los siguientes resultados se obtuvieron del paquete Statistica.

Eigenvalues (factor.sta)
Extraction: Principal axis factoring

	Eigenval	% total Variance	Cumul. Eigenval	Cumul. %
1	2.5213	63.0320	2.5213	63.032
2	0.7563	18.9063	3.2775	81.938
3	0.1669	4.1731	3.4445	86.111

Tabla 10. Eigenvalores.

Por lo que con tres factores apenas si se tiene el 86%. Pero el análisis se realizará considerando solo dos factores. A continuación se dan los resultados de las comunalidades así como de las cargas de los factores y se comparan con las obtenidas en el método de componentes principales.

Variables	Cargas componente del principal f_1	f_2	Cargas factor del principal f_1	f_2	Comunalidad del factor principal
y_1	0.8023	-0.5351	0.8309	-0.5629	0.9927
y_2	0.8557	-0.3264	0.7681	-0.2037	0.6315
x_1	0.8832	0.2701	0.8190	0.1865	0.7056
x_2	0.7140	0.6584	0.7231	0.6520	0.9480
Contribución de la varianza	2.6657	0.8993	2.4741	0.8183	
Proporción de la varianza	0.6664	0.2248	0.8089	0.2426	
Proporción acumulada	0.6664	0.8912	0.8089	*1.052*	

(2.7)

Como uno de los eigenvalores es negativos, la proporción de la varianza excede del valor 1. Los dos conjuntos de las cargas son muy similares, debido a que el valor de las comunalidades es grande.

2.11 Rotación de factores

Al obtener los factores comunes, éstos están correlacionados en mayor o menor medida con cada una de las variables originales; con los factores rotados se trata de que cada una de las variables originales tenga una correlación lo más cercana al valor 1 con uno de los factores y correlaciones cercanas al valor 0 con el resto de los factores.

Existen 2 formas básicas para realizar la rotación de los factores, la *rotación ortogonal* y *la rotación oblicua.*

2.11.1 Rotación ortogonal

En la rotación ortogonal los ejes se rotan de forma que los nuevos ejes son perpendiculares. El método más utilizado es el método *Varimax,* (en donde la

*Var*ianza se *max*imiza); el cual se obtiene maximizando la suma de varianzas de las cargas factoriales al cuadrado dentro de cada factor.

Una de las propiedades del método *Varimax* es que no se altera la varianza total explicada por los factores ni la comunalidad de cada una de las variables.

Existen otros métodos de rotación ortogonal que se utilizan menos como el método *Equamax* y *Quartimax*.

2.11.2 Rotación Oblicua

Algunas veces se tienen factores no correlacionados entre sí, por lo que es necesario que las rotaciones no conserven el ángulo recto, es decir, los ejes *no* son perpendiculares. El método más utilizado del rotación oblicua es el método *Oblimin*.

Cuando se realizan rotaciones oblicuas, la matriz factorial original se convierte en dos matrices diferentes la matriz de ponderaciones y la matriz de correlaciones entre los factores y las variables.

Ejemplo 2.7 *Utilizando el método de componentes principales, se realizó una rotación ortogonal; por lo que considere los datos del Ejemplo 2.5.*

Los resultados se obtuvieron del paquete Statistica.

	Cargas		Método	Varimax	Comunalidad
Variables	f_1	f_2	f_1	f_2	$\widehat{h}_i^2$
y_1	0.8023	-0.5351	**0.9558**	0.1288	0.930
y_2	0.8557	-0.3264	**0.8579**	0.3206	0.839
x_1	0.8832	0.2701	0.4842	**0.7864**	0.853
x_2	0.7140	0.6584	0.1006	**0.9660**	0.943
Contribución de la varianza	2.6657	0.8993	1.8941	1.6710	3.5650
Proporción de la varianza	0.6664	0.2248	0.4735	0.4178	
Proporción acumulada	0.6664	**0.8912**	0.4735	**0.8913**	

(2.8)

Al comparar (2.6) con (2.8), se puede observar que efectivamente las comunalidades no cambian despues de haber realizado la rotación Varimax.

Ejemplo 2.8 *Se tienen* 11 *personas, y* 5 *variables que corresponden a los tiempos de respuesta para la i-ésima palabra clave en una oración* ($i = 1, \cdots, 5$)

número	Y_1	Y_2	Y_3	Y_4	Y_5
1	*51*	*36*	*50*	*35*	*42*
2	*27*	*20*	*26*	*17*	*27*
3	*37*	*22*	*41*	*37*	*30*
4	*42*	*36*	*32*	*34*	*27*
5	*27*	*18*	*33*	*14*	*29*
6	*43*	*32*	*43*	*35*	*40*
7	*41*	*22*	*36*	*25*	*38*
8	*38*	*21*	*31*	*20*	*16*
9	*36*	*23*	*27*	*25*	*28*
10	*26*	*31*	*31*	*32*	*36*
11	*29*	*20*	*25*	*26*	*25*

Los resultados se obtuvieron del paquete Statistica. Primeramente se observa que con las corrrelaciones se tienen 2 factores que corresponden a las variables Y_1, Y_3 y Y_5 para el primer factor y las variables Y_2 y Y_4 para el segundo factor. Aunque todas las variables están altamente correlacionadas.

Correlations
Casewise deletion of MD
N=11

	Y1	Y2	Y3	Y4	Y5
Y1	1.000	0.614	0.757	0.575	0.413
Y2	0.614	1.000	0.547	0.750	0.548
Y3	0.757	0.547	1.000	0.605	0.692
Y4	0.575	0.750	0.605	1.000	0.524
Y5	0.413	0.548	0.692	0.524	1.000

Tabla 11. Correlaciones.

Con la tabla 12 se sabe el porcentaje de varianza total.

Eigenvalues
Extraction: Principal components

	Eigenval	% total Variance	Cumul. Eigenval	Cumul. %
1	3.4165	68.3299	3.4165	68.3299
2	0.6144	12.2886	4.0309	80.6185
3	0.5723	11.4455	4.6032	92.0640
4	0.2712	5.4242	4.8744	97.4882
5	0.1256	2.5118	5	100

Tabla 12. Eigenvalores.

Por lo efectivamente el utilizar 2 factores es adecuado ya que se tiene el 80.61% de la varianza total. Las siguiente tabla contiene los resultados de las comunalidades, así como las varianzas y las cargas de los factores. Por el método de Componentes Principales.

	Cargas		Comunalidad	Varianzas específicas
Variables	$\widehat{\lambda}_{1j}$	$\widehat{\lambda}_{2j}$	$\widehat{h}_i^2$	$\widehat{\psi}_i$
Y_1	0.817	-0.157	0.692	0.308
Y_2	0.838	-0.336	0.815	0.185
Y_3	0.874	0.288	0.847	0.153
Y_4	0.838	-0.308	0.798	0.202
Y_5	0.762	0.547	0.879	0.121
Contribución de la varianza total	3.416	0.614	4.031	
Proporción de la varainza total	0.638	0.123	0.806	
Proporción acumulada	0.638	0.806	0.806	

¿Será conveniente una rotación?. Analizando la figura 3.

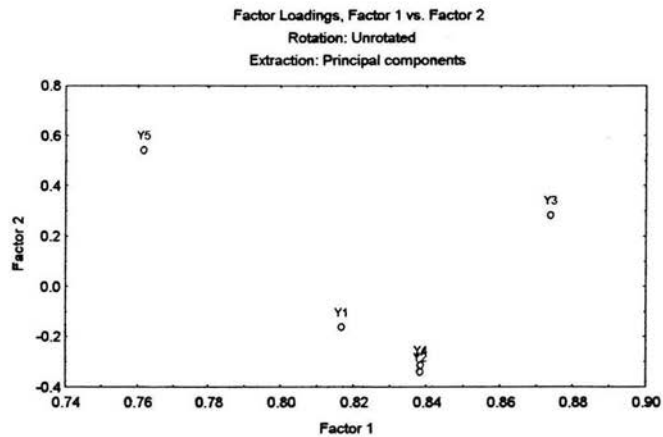

Figura 3. Factor 1 vs. Factor 2

En donde las cargas de Y_2 y Y_4 están duplicadas; como se mencionó con la matriz de correlaciones. La figura 4 corresponde a los factores, despues de realizar una rotación ortogonal, por el método varimax.

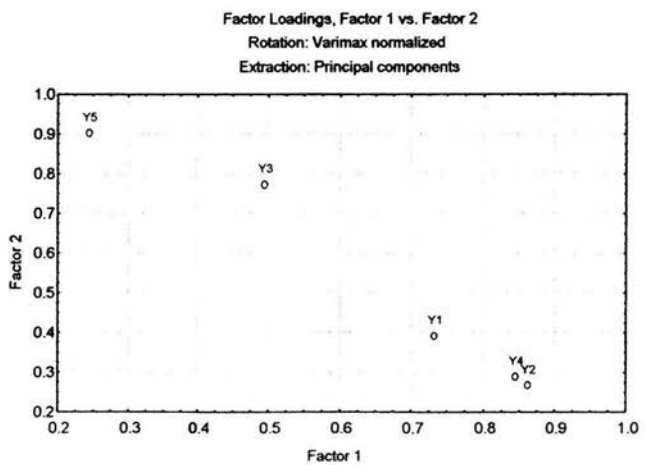

Figura 4.

Y los resultados de las comunalidades y las cargas de los factores después de la rotación son:

	Cargas		Método	Varimax	Comunalidad
Variables	f_1	f_2	f_1	f_2	$\widehat{h}_i^2$
Y_1	0.817	-0.157	0.732	0.395	0.692
Y_2	0.838	-0.336	**0.861**	0.271	0.815
Y_3	0.874	0.288	0.494	0.776	0.847
Y_4	0.838	-0.308	**0.844**	0.292	0.798
Y_5	0.762	0.547	0.244	0.905	0.879
Contribución de la varianza	3.416	0.614	2.294	1.736	4.031
Proporción de la varianza	0.638	0.123	0.459	0.347	
Proporción acumulada	0.638	0.806	0.459	0.806	

En donde las cargas de Y_2 y Y_4 son muy similares.

2.12 Diferencias entre análisis de componentes principales y análisis de factores

- El análisis de componentes principales es una transformación de los datos,no hace suposiciones acerca de la matriz de covarianzas. Y el análisis de factores hay que definir un nuevo modelo ($\underline{X} = \underline{\mu} + \Lambda\ f\ + \varepsilon$)
- En el análisis de componentes principales el énfasis es la transformación de las variables observadas en los componentes principales $\underline{Y} = \Gamma^{\ t}\ \underline{X}$, mientras que en el análisis de factores el énfasis es sobre una transformación de los factores a las variables observadas.
- Cuando las varianzas específicas se suponen son cero, el análisis de factores es equivalente al análisis de componentes principales. Por lo que si el modelo del análisis de factores se satisface y las varianzas específicas son pequeñas, es de esperarse que el análisis de componentes principales y el análisis de factores den resultados similares.

Capítulo 3

Análisis Discriminante

3.1 Introducción

El análisis discriminante es una técnica de clasificación y asignación de un individuo a un grupo, ya que se conocen sus características, por lo que se pueden realizar predicciones. Tiene muchas aplicaciones para los diagnósticos médicos, por ejemplo, en operaciones para enfermos de cáncer o en el desarrollo de la neumonía.

El análisis discriminante es un conjunto de técnicas cuyo propósito es describir y clasificar individuos en diferentes grupos, a partir de las observaciones hechas de sus distintas variables, es decir, dado un vector de medidas tomadas del individuo, el problema básico es encontrar alguna función de dichas medidas que ayuden a asignar al individuo dentro de uno de los grupos que se tienen. Dichas poblaciones tienen la característica de ser mutuamente excluyentes por lo que un individuo solamente puede ser asignado a uno y sólo un grupo. Se supondrá que los datos tienen distribución normal multivariada.

3.1.1 Clasificación de los individuos

Para poder clasificar a los individuos dentro del análisis discriminante se puede considerar:

- La regla de verosimilitud.- Se utiliza la función de densidad de la normal multivariada, suponiendo que se tiene diferente matriz de covarianzas.
- La función discriminante de Fisher.- Se utiliza la función de densidad de la normal multivariada, suponiendo que se tiene la misma matriz de covarianzas.

3.1.2 Regla de asignación

Sea $\underline{X}$ un vector de dimensión p, y suponga que se tienen g poblaciones o grupos, denotados por $\Pi_1, \cdots, \Pi_g$ cada uno con función de densidad de probabilidad $f_i(\underline{X})$, se desea asignar al vector $\underline{X}$ a *uno y sólo uno* de esos g grupos considerando sus p características.

Una regla discriminante corresponde a dividir el espacio $\Re^p$ en g regiones mutuamente excluyentes $R_1, \cdots, R_g$, es decir, $\bigcup_{i=1}^{g} R_i \equiv \Re^p$ y $R_i \cap R_j = \emptyset$ para $i \neq j$.

Entonces la regla discriminante d queda definida como:

$$d = asignar\ \underline{X}\ a\ \Pi_i\ si\ \underline{X} \in R_i\ para\ i = 1, 2, \cdots, g. \tag{3.1}$$

Difícilmente se conoce la función de densidad de probabilidad, por lo que se estiman algunos parámetros.

3.1.3 Errores de asignación.

Supóngase que se tienen dos grupos, Π_1 y Π_2, donde para $j = 1, 2$ la función de densidad de probabilidad será $f_j(\underline{X})$ si $\underline{X}$ proviene del grupo Π_j.

Considere la regla dada en (3.1) donde las regiones R_1 y R_2 son mutuamente excluyentes, es decir, cumplen con que $R_1 \cap R_2 = \emptyset$ y $R_1 \cup R_2 = R^p$.

Si se quisiera asignar un nuevo individuo $\underline{X}$ a uno de estos dos grupos, se podría cometer uno de dos errores:

$$d = \begin{cases} \text{Asignar } \underline{X} \text{ a } \Pi_2, \text{cuando } \underline{X} \text{ pertenece a } \Pi_1 \\ \text{Asignar } \underline{X} \text{ a } \Pi_1, \text{cuando } \underline{X} \text{ pertenece a } \Pi_2 \end{cases} \tag{3.2}$$

La probablidad de cometer el primer error se puede expresar de la siguiente manera:

$P(\text{Asignar } \underline{X} \text{ a } \Pi_2 \mid \underline{X} \text{ pertenece a } \Pi_1) = \int_{R_2} f_1(\underline{X})\, dx.$

La probabilidad de cometer el segundo error se puede expresar así:

$P(\text{Asignar } \underline{X} \text{ a } \Pi_1 \mid \underline{X} \text{ pertenece a } \Pi_2) = \int_{R_1} f_2(\underline{X})\, dx.$

3.2 Discriminación cuando las poblaciones son conocidas

Sea $\underline{X}$ un vector de dimensión p, y suponga que los parámetros de la función de densidad de probabilidad de los g grupos son conocidos.

3.2.1 Regla discriminante de máxima verosimilitud

Definición 3.1 *La Regla Discriminante de Máxima Verosimilitud es asignar $\underline{X}$ a la población que tenga la verosimilitud más grande, es decir, la Regla Discriminante de Máxima Verosimilitud dice que se asigne $\underline{X}$ a Π_j si:*

$$L_j(\underline{X}) = \max_i L_i(\underline{X}).$$

En otras palabras se debe encontrar la población i en donde la verosimilitud de $\underline{X}$ sea máxima, esto es:

$$R_j = \{\underline{X} \in \Re^p \mid L_j(\underline{X}) \geq L_i(\underline{X}) \quad i = 1, 2, \cdots, g\}.$$

3.2.2 Discriminación lineal

Se considera que $\underline{X}_i \sim N_p\left(\underline{\mu}_i,\ \Sigma\right)$, es decir, la matriz de covarianzas Σ, es común para las g poblaciones.

Teorema 3.2 *En el caso en el que Π_i tiene asociada una densidad $N_p\left(\underline{\mu}_i,\ \Sigma\right)$, la regla de asignación máximo verosimil asigna $\underline{X}$ a Π_i si:*

$$L_i\ (\underline{X}) = \underline{a}_i^{\,t}\left(\underline{X} - \frac{1}{2}\,\underline{\mu}_i\right) \quad \text{con } i = 1,\ 2, \cdots,\ g. \tag{3.3}$$

donde $L_j\ (\underline{X}) = \max\limits_i L_i\ (\underline{X})$.

Demostración.
Para maximizar la función de verosimilitud $L_j\ (\underline{X})$, se tiene que:

$$\begin{aligned} L_j\ (\underline{X}) &= \max_i\ L_i(\underline{X}) \\ &= \max_i \left\{ \frac{1}{|2\ \pi\ \Sigma|^{\,1/2}} \exp\left(-\frac{1}{2}(\underline{X} - \underline{\mu}_i)^{\,t}\ \Sigma^{-1}\ (\underline{X} - \underline{\mu}_i) \right) \right\} \\ &= \max_i \left\{ \exp\left(-\frac{1}{2}(\underline{X} - \underline{\mu}_i)^{\,t}\ \Sigma^{-1}\ (\underline{X} - \underline{\mu}_i) \right) \right\} \end{aligned}$$

y es equivalente a minimizar

$$\begin{aligned} &= \min_i \left(\frac{1}{2}(\underline{X} - \underline{\mu}_i)^{\,t}\ \Sigma^{-1}\ (\underline{X} - \underline{\mu}_i) \right) \qquad (3.4) \\ &= \min_i \left\{ \underline{X}^{\,t}\ \Sigma^{-1} \underline{X} - \underline{X}^{\,t}\ \Sigma^{-1}\ \underline{\mu}_i - \underline{\mu}_i^{\,t}\ \Sigma^{-1}\ \underline{X} + \underline{\mu}_i^{\,t}\ \Sigma^{-1} \underline{\mu}_i \right\} \\ &= \min_i \left\{ \underline{X}^{\,t}\ \Sigma^{-1}\ \underline{X} - 2\ \underline{\mu}_i^{\,t}\ \Sigma^{-1}\ \underline{X} + \underline{\mu}_i^{\,t}\ \Sigma^{-1}\ \underline{\mu}_i \right\} \\ &= \min_i \left\{ \underline{\mu}_i^{\,t}\ \Sigma^{-1}\ \underline{\mu}_i - 2\ \underline{\mu}_i^{\,t}\ \Sigma^{-1}\ \underline{X} \right\}. \end{aligned}$$

Entonces, se tiene que:

$$L_j(\underline{X}) = \min_i \left\{ \underline{\mu}_i^t \Sigma^{-1} \underline{\mu}_i - 2\, \underline{\mu}_i^t \Sigma^{-1} \underline{X} \right\}.$$

Multiplicando por $-\frac{1}{2}$ se tiene:

$$\begin{aligned} L_j(\underline{X}) &= \max_i \left\{ \underline{\mu}_i^t \Sigma^{-1} \underline{X} - \frac{1}{2} \underline{\mu}_i^t \Sigma^{-1} \underline{\mu}_i \right\} \\ &= \max_i \left\{ \underline{a}_i^t \left(\underline{X} - \frac{1}{2} \underline{\mu}_i \right) \right\} \end{aligned}$$

para $i = 1, \cdots, g.$, donde $\underline{a}_i^t = \underline{\mu}_i^t \Sigma^{-1}$ ∎

La probabilidad de que dos verosimilitudes tomen el mismo valor máximo es *cero*.

Teorema 3.3

(a). Si $\underline{X}$ proviene de Π_i, donde $\Pi_i \sim N_p(\underline{\mu}_i, \Sigma)$ para $i = 1, \cdots, g$ y $\Sigma > 0$, entonces la *regla discriminante de máxima verosimilitud* asigna $\underline{X}$ a Π_j donde $j \in \{1, \cdots, g\}$, cuando se obtiene la *i-ésima* población que minimiza *la distancia de Mahalanobis.*

$$(\underline{X} - \underline{\mu}_i)^t \Sigma^{-1} (\underline{X} - \underline{\mu}_i). \tag{3.5}$$

(b). Cuando $g = 2$, la regla de asignación es

$$d = \begin{cases} \text{Asignar } \underline{X} \text{ a } \Pi_1 & \text{si } \alpha^t (\underline{X} - \underline{\mu}) > 0 \\ \text{Asignar } \underline{X} \text{ a } \Pi_2, & \text{en otro caso} \end{cases} \tag{3.6}$$

donde $\alpha = \Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$ y $\underline{\mu} = \frac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2)$.

Demostración.

(a). Vease (3.4) en la demostración del Teorema 3.2.

(b). Para asignar $\underline{X}$ a Π_1, entonces $L_1(\underline{X}) > L_2(\underline{X})$; esto se cumple *si y sólo si*

$$-\frac{1}{2}(\underline{X}-\underline{\mu}_1)^t\,\Sigma^{-1}\,(\underline{X}-\underline{\mu}_1) \;>\; -\frac{1}{2}(\underline{X}-\underline{\mu}_2)^t\,\Sigma^{-1}\,(\underline{X}-\underline{\mu}_2)$$
$$(\underline{X}-\underline{\mu}_1)^t\,\Sigma^{-1}\,(\underline{X}-\underline{\mu}_1) \;<\; (\underline{X}-\underline{\mu}_2)^t\,\Sigma^{-1}\,(\underline{X}-\underline{\mu}_2)$$

Desarrollando, se tiene

$$\underline{X}^t\Sigma^{-1}\underline{X}-2\underline{\mu}_1^t\Sigma^{-1}\underline{X}+\underline{\mu}_1^t\,\Sigma^{-1}\underline{\mu}_1 < \underline{X}^t\Sigma^{-1}\underline{X}-2\underline{\mu}_2^t\Sigma^{-1}\underline{X}+\underline{\mu}_2^t\,\Sigma^{-1}\underline{\mu}_2$$
$$-2\,\underline{\mu}_1^t\,\Sigma^{-1}\underline{X}+\underline{\mu}_1^t\,\Sigma^{-1}\underline{\mu}_1 < -2\,\underline{\mu}_2^t\,\Sigma^{-1}\underline{X}+\underline{\mu}_2^t\,\Sigma^{-1}\underline{\mu}_2$$
$$-\underline{\mu}_1^t\Sigma^{-1}\underline{X}+\frac{1}{2}\underline{\mu}_1^t\Sigma^{-1}\underline{\mu}_1 < -\underline{\mu}_2^t\Sigma^{-1}\underline{X}+\frac{1}{2}\underline{\mu}_2^t\Sigma^{-1}\underline{\mu}_2$$

si y sólo si

$$0 \;<\; \underline{\mu}_1^t\Sigma^{-1}\underline{X}-\frac{1}{2}\underline{\mu}_1^t\Sigma^{-1}\underline{\mu}_1-\underline{\mu}_2^t\Sigma^{-1}\underline{X}+\frac{1}{2}\underline{\mu}_2^t\Sigma^{-1}\underline{\mu}_2$$
$$0 \;<\; (\underline{\mu}_1^t-\underline{\mu}_2^t)\,\Sigma^{-1}\underline{X}-\frac{1}{2}(\underline{\mu}_1^t\,\Sigma^{-1}\underline{\mu}_1-\underline{\mu}_2^t\,\Sigma^{-1}\underline{\mu}_2)$$
$$0 \;<\; (\underline{\mu}_1^t-\underline{\mu}_2^t)\,\Sigma^{-1}\underline{X}-\frac{1}{2}\left((\underline{\mu}_1^t-\underline{\mu}_2^t)\,\Sigma^{-1}(\underline{\mu}_1+\underline{\mu}_2)\right)$$
$$0 \;<\; \left\{(\underline{\mu}_1^t-\underline{\mu}_2^t)\,\Sigma^{-1}\right\}\left\{\underline{X}-\frac{1}{2}\,(\underline{\mu}_1+\underline{\mu}_2)\right\}$$
$$0 \;<\; \left(\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)\right)^t\left(\underline{X}-\frac{1}{2}\,(\underline{\mu}_1+\underline{\mu}_2)\right).$$

si $\alpha=\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)$ y $\underline{\mu}=\frac{1}{2}(\underline{\mu}_1+\underline{\mu}_2)$. ∎

3.2.3 Discriminación cuadrática

Se considera el caso en que cada una de las g poblaciones tiene su matriz de covarianzas Σ_i, por lo que $\Pi_i \sim N_p\left(\underline{\mu}_i,\ \Sigma_i\right)$.

Teorema 3.4 *La Regla Discriminante de Máxima Verosimilitud es asignar la observación* $\underline{X}$ *a* Π_i *si:*

$$L_i(\underline{X}) > L_j(\underline{X}) \quad \text{para } i \neq j \tag{3.7}$$

esto pasa si y sólo si

$$R_i = \left\{\underline{X} \in \Re^p \mid (\underline{X}-\underline{\mu}_i)^t\, \Sigma_i^{-1}\, (\underline{X}-\underline{\mu}_i) \leq (\underline{X}-\underline{\mu}_j)^t\, \Sigma_j^{-1}\, (\underline{X}-\underline{\mu}_j)\right\}$$

para $i, j = 1, \cdots, g$.

Demostración.
La desigualdad $L_i(\underline{X}) > L_j(\underline{X})$ puede expresarse como:

$$|2\,\pi\,\Sigma_i|^{1/2} \exp\left(-\frac{1}{2}(\underline{X}-\underline{\mu}_i)^t\, \Sigma_i^{-1}\, (\underline{X}-\underline{\mu}_i)\right) >$$
$$|2\,\pi\,\Sigma_j|^{1/2} \exp\left(-\frac{1}{2}(\underline{X}-\underline{\mu}_j)^t\, \Sigma_j^{-1}\, (\underline{X}-\underline{\mu}_j)\right)$$

Esto ocurre *si y sólo si*

$$\exp\left(-\frac{1}{2}(\underline{X}-\underline{\mu}_i)^t \Sigma_i^{-1}\, (\underline{X}-\underline{\mu}_i)\right) > \exp\left(-\frac{1}{2}(\underline{X}-\underline{\mu}_j)^t\, \Sigma_j^{-1}\, (\underline{X}-\underline{\mu}_j)\right)$$

si y sólo si

$$(\underline{X}-\underline{\mu}_i)^t\, \Sigma_i^{-1}\, (\underline{X}-\underline{\mu}_i) < (\underline{X}-\underline{\mu}_j)^t\, \Sigma_j^{-1}\, (\underline{X}-\underline{\mu}_j). \qquad \blacksquare$$

3.3 Discriminación cuando los parámetros no son conocidos

En esta sección se deben estimar la media y la matriz de covarianzas.

3.3.1 Regla discriminante de máxima verosimilitud

Se considera la función de densidad $f_i(\underline{X} \mid \theta_i)$, donde θ_i es un vector de parámetro desconocido en cada uno de los g grupos.

Definición 3.5 *La Regla Discriminante muestral de Máxima Verosimilitud es asignar $\underline{X}$ a la población que tenga la verosimilitud más grande, es decir, la Regla Discriminante muestral de Máxima Verosimilitud dice que se asigne $\underline{X}$ a Π_j si:*

$$\widehat{L}_j\,(\underline{X}) = \max_i\ \widehat{L}_i\,(\underline{X}). \tag{3.8}$$

Por lo que la región de clasificación está dada por:

$$R_j = \left\{\underline{X} \in \Re^p \mid \widehat{L}_j\,(\underline{X}) \geq\ \widehat{L}_i\,(\underline{X}) \quad i = 1,\ 2,\ \cdots,\ g\right\}$$

3.3.2 Discriminación lineal

Se considera que $\underline{X}_i \sim N_p\left(\underline{\mu}_i,\ \Sigma\right)$, en donde $\underline{\mu}_i$ y Σ son desconocidas; la estimación de los parámetros queda:

$$\begin{aligned} \widehat{\underline{\mu}}_i &= \frac{1}{n}\sum_{j=1}^{n_i} \underline{X}_{ij} \\ &= \overline{X}_i \end{aligned} \tag{3.9}$$

para $i = 1, \cdots, g$.

$$\begin{aligned} W &= \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i} \left(\underline{X}_{ij} - \overline{X}_i\right)\left(\underline{X}_{ij} - \overline{X}_i\right)^t \\ &= \frac{1}{n}\sum_{i=1}^{g} n_i\, S_i \end{aligned} \tag{3.10}$$

donde S_i es el estimador de la matriz de covarianzas *entre grupos.*

$$S_i = \frac{1}{n_i} \sum_{j=1}^{n_i} (\underline{X}_{ij} - \overline{X}_i)(\underline{X}_{ij} - \overline{X}_i)^t. \tag{3.11}$$

Teorema 3.6 *En el caso en el que* Π_i *tiene asociada una densidad* $N_p\left(\underline{\mu}_i, \Sigma\right)$, *de parámetros desconocidos, la regla muestral por máxima verosimilitud asigna* $\underline{X}$ *a* Π_i *si*

$$L_j(\underline{X}) = \max_i L_i(\underline{X})$$

donde $L_i(\underline{X}) = \left\{\underline{a}_i^t \left(\underline{X} - \frac{1}{2}\overline{X}_i\right)\right\}$ *con* $i = 1, 2, \cdots, g$ *y* $\underline{a}_i = W^{-1}\overline{X}_i$.

Demostración.

Al maximizar la función de verosimilitud, se sustituyen los párametros estimados y la demostración es similar a la del Teorema 3.2 ∎

Teorema 3.7

1. Si $\underline{X}$ proviene de Π_i,donde $\Pi_i \sim N_p(\underline{\mu}_i, \Sigma)$ para $i = 1, 2, \cdots, g$ y $\Sigma > 0$, entonces la *regla discriminante muestral de máxima verosimilitud* asigna $\underline{X}$ a Π_j donde $j \in \{1, \cdots, g\}$, es el valor que de la *i-ésima* población que minimiza *la distancia de Mahalanobis* estimada.

$$(\underline{X} - \overline{X}_i)^t W^{-1} (\underline{X} - \overline{X}_i). \tag{3.12}$$

2. Cuando $g = 2$, la regla de asignación es:

$$\text{Asignar } \underline{X} \text{ a} = \begin{cases} \Pi_1 & \text{si y sólo si } \alpha^t\left(\underline{X} - \overline{X}\right) > 0 \\ \Pi_2 & \text{en otro caso} \end{cases} \tag{3.13}$$

donde $\alpha = W^{-1}(\overline{X}_1 - \overline{X}_2)$ y $\overline{X} = \frac{1}{2}(\overline{X}_1 + \overline{X}_2)$.

Demostración.

Sustituyendo los parámetros por los estimadores máximo verosímiles.

1. Vease la demostración del Teorema 3.2.

2. Se sigue del Teorema 3.3. ∎

3.3.3 Discriminación cuadrática

Se considera el caso en que cada una de las g poblaciones tiene su media y su matriz de covarianzas Σ_i, por lo que los estimadores máximo verosímil están dados como en (3.9) y (3.10).

Teorema 3.8 *En el caso en que* Π_i *tiene su media* $\underline{\mu}_i$ *y su matriz de covarianzas* Σ_i, *de parámetros desconocidos, la Regla Discriminante muestral de Máxima Verosimilitud asigna la observación* $\underline{X}$ *a* Π_i *si:*

$$L_i(\underline{X}) > L_j(\underline{X}) \quad \text{para } i \neq j$$

esto pasa si y sólo si

$$R_i = \left\{\underline{X} \in \Re^p \mid (\underline{X} - \overline{X}_i)\, S_i^{-1}\, (\underline{X} - \overline{X}_i) \leq (\underline{X} - \overline{X}_j)^t\, S_j^{-1}\, (\underline{X} - \overline{X}_j)\right\}.$$

para $i, j = 1, \cdots, g$.

Demostración.

Se sigue del Teorema 3.4, después de haber sustituido los estimadores máximo verosímiles. ∎

3.4 Regla discriminante de la razón de verosimilitudes

Una alternativa de la regla de asignación *muestral máximo verosimil*, es utilizar el criterio de *la razón de verosimilitud*, dada por Anderson (1958),

si $\underline{X}$ pertenece a Π_r y $\Pi_r \sim N_p\left(\underline{\mu}_r, \Sigma\right)$. La regla es calcular las i-verosimilitudes de las siguientes hipótesis:

$$H_r = \begin{cases} \underline{X} \text{ y los renglones de } \underline{X}_r \text{ provienen de } \Pi_r \\ \text{y las filas de } \underline{X}_j \text{ provienen de } \Pi_j \quad \text{para } r \neq j \end{cases}$$

De esta manera $\underline{X}$ es asignado a la población cuya hipótesis H_r tiene la verosimilitud más grande y está dada por:

$$\begin{aligned} & L_r(\underline{\mu}_1, \underline{\mu}_2, \cdots, \underline{\mu}_g, \Sigma) \qquad (3.14) \\ = & \prod_{i=1}^{g} \prod_{j=1}^{n_j} |2\pi\Sigma|^{1/2} \exp\left\{-\frac{1}{2}(\underline{X}_{ij} - \underline{\mu}_i)^t \Sigma^{-1}(\underline{X}_{ij} - \underline{\mu}_i)\right\} \\ & \times |2\pi\Sigma|^{1/2} \exp\left\{-\frac{1}{2}(\underline{X} - \underline{\mu}_j)^t \Sigma^{-1}(\underline{X} - \underline{\mu}_j)\right\}. \end{aligned}$$

Si los parámetros son conocidos, la región de clasificación en Π_r es:

$$R_r = \{\underline{X} \in \Re^p \mid L_r \geq L_j \text{ para } r = 1, \cdots, g\}$$

donde L_r es la función de verosimilitud de la *r-ésima* población dada en (3.14).

Si los parámetros son desconocidos, entonces los estimadores verosimil asociados al *j-ésimo* grupo según la hipótesis H_r, están dados por:

$$\underline{\hat{\mu}}_j = \begin{cases} \overline{X}_j & \text{si } j \neq r \\ \overline{X}_j = \dfrac{n_r \overline{X}_j + X}{n_r + 1} & \text{si } j = r \end{cases} \qquad (3.15)$$

$$\begin{aligned} \hat{\Sigma}_{H_r} &= W \qquad (3.16) \\ &= \frac{1}{n+1}\left[\sum_{j=1}^{g}(n_j + I_{(r)}(j))\, S_j\right]. \end{aligned}$$

donde

$$S_j = \begin{cases} S_j = \frac{1}{n_j} \sum_{l=1}^{n_l} (\underline{X}_{jl} - \overline{X}_j)\,(\underline{X}_{jl} - \overline{X}_j)^t & \text{para} \quad r \neq j \\ S_r = \frac{1}{n_r+1} \times \left\{ \sum_{l=1}^{n_r} (\underline{X}_{rl} - \overline{X}_r)(\underline{X}_{rl} - \overline{X}_r)^t + \right. & \\ \qquad \left. (\underline{X} - \overline{X}_r)(\underline{X} - \overline{X}_r)^t \right\} & \text{para} \quad r = j \end{cases}$$

Por lo que:

$$\begin{aligned} & \widehat{L}_r(\underline{X}) \qquad (3.17) \\ = & \prod_{i=1}^{g} \prod_{j=1}^{n_j} |2\,\pi\,\Sigma|^{1/2} \exp\left\{ -\frac{1}{2} (\underline{X}_{ij} - \overline{X})^t\, W^{-1} (\underline{X}_{ij} - \overline{X}) \right\} \\ & \times |2\,\pi\,\Sigma|^{1/2} \exp\left\{ -\frac{1}{2} (\underline{X} - \overline{X})^t\, W^{-1} (\underline{X} - \overline{X}) \right\}. \end{aligned}$$

Ejemplo 3.1 *Considere que se desea discriminar un nuevo individuo* $\underline{X}$ *a una de las dos poblaciones que se distribuyen normales multivariadas, denotadas por* Π_1 *y* Π_2 *donde* $\Pi_1 \sim N_p\left(\underline{\mu}_1, \Sigma\right)$ *y* $\Pi_2 \sim N_p\left(\underline{\mu}_2, \Sigma\right)$. *La hipótesis son:*

$$\begin{aligned} H_1 \;:\; & \underline{X} \text{ y los renglones de } \underline{X}_1 \text{ provienen de } \Pi_1 \\ & \text{y las filas de } \underline{X}_2 \text{ provienen de } \Pi_2 \\ & \qquad vs \\ H_2 \;:\; & \underline{X} \text{ y los renglones de } \underline{X}_2 \text{ provienen de } \Pi_2 \\ & \text{y las filas de } \underline{X}_1 \text{ provienen de } \Pi_1 \end{aligned}$$

Los estimadores máximo verosímiles para $\underline{\mu}_1$, $\underline{\mu}_2$ y Σ, bajo H_1 son:

$$\begin{aligned}\hat{\underline{\mu}}_1 &= \frac{n_1 \underline{X}_1 + \underline{X}}{n_1 + 1} \\ \hat{\underline{\mu}}_2 &= \underline{X}_2\end{aligned}$$

$$\hat{\Sigma}_{H_1} = \frac{1}{n_1 + n_2 + 1}\left\{W + \frac{n_1}{1 + n_1}(\underline{X} - \overline{X}_1)\,(\underline{X} - \overline{X}_1)^t\right\}$$

bajo H_2 son:

$$\begin{aligned}\hat{\underline{\mu}}_1 &= \underline{X}_1 \\ \hat{\underline{\mu}}_2 &= \frac{n_2 \underline{X}_2 + \underline{X}}{n_2 + 1}\end{aligned}$$

$$\hat{\Sigma}_{H_2} = \frac{1}{n_1 + n_2 + 1}\left\{W + \frac{n_2}{1 + n_2}(\underline{X} - \overline{X}_2)\,(\underline{X} - \overline{X}_2)^t\right\}$$

Por lo que el cociente de verosimilitudes es:

$$\left|\frac{\hat{\Sigma}_{H_2}}{\hat{\Sigma}_{H_1}}\right| = \left|\frac{\frac{1}{n_1 + n_2 + 1}\left\{W + \frac{n_2}{1 + n_2}(\underline{X} - \overline{X}_2)\,(\underline{X} - \overline{X}_2)^t\right\}}{\frac{1}{n_1 + n_2 + 1}\left\{W + \frac{n_1}{1 + n_1}(\underline{X} - \overline{X}_1)\,(\underline{X} - \overline{X}_1)^t\right\}}\right|$$

desarrollando y reacomodando términos se tiene:

$$\begin{aligned}\frac{\left|\hat{\Sigma}_{H_2}\right|}{\left|\hat{\Sigma}_{H_1}\right|} &= \frac{\left|1 + \frac{n_2}{1 + n_2}(\underline{X} - \overline{X}_2)^t\, W^{-1}\,(\underline{X} - \overline{X}_2)\right|}{\left|1 + \frac{n_1}{1 + n_1}(\underline{X} - \overline{X}_1)^t\, W^{-1}\,(\underline{X} - \overline{X}_1)\right|} \\ \frac{\left|\hat{\Sigma}_{H_2}\right|}{\left|\hat{\Sigma}_{H_1}\right|} &= \frac{1 + \frac{n_2}{1 + n_2}(\underline{X} - \overline{X}_2)^t\, W^{-1}\,(\underline{X} - \overline{X}_2)}{1 + \frac{n_1}{1 + n_1}(\underline{X} - \overline{X}_1)^t\, W^{-1}\,(\underline{X} - \overline{X}_1)}\end{aligned}$$

Se acepta H_1 y se asigna $\underline{X}$ a Π_1 *si y sólo si:*

$$\frac{n_2}{1+n_2}(\underline{X}-\overline{X}_2)^t W^{-1}(\underline{X}-\overline{X}_2) > \frac{n_1}{1+n_1}(\underline{X}-\overline{X}_1)^t W^{-1}(\underline{X}-\overline{X}_1)$$

En el caso especial en que $n_1 = n_2$ este criterio es equivalente a la *regla discriminante muestral máximo verosimil.* Si n_1 y n_2 son números grandes el procedimiento es *asintóticamente* equivalente. Pero si los tamaños muestrales son diferentes, entonces se clasifica a $\underline{X}$ a la población que tiene el tamaño de muestra más grande.

¿Tiene sentido el análisis discriminante?

Considere que se tienen g poblaciones o grupos, con función de densidad normal multivariada con la misma matriz de covarianzas, pero de parámetros desconocidos, los cuales se estiman a partir de (3.15) y (3.16). Si todas las medias son iguales $\underline{\mu}_1 = \underline{\mu}_2 = \cdots = \underline{\mu}_g$ no tendría sentido tratar de discriminar entre los grupos.

Sin embargo si las verdaderas medias son iguales, las medias muestrales $\underline{X}_1, \underline{X}_2, \cdots, \underline{X}_g$ serán diferentes, por lo que puede ser plausible llevar a cabo el análisis discriminante. Para saber si vale la pena realizar el análisis se necesita probar la hipótesis:

$$\underline{\mu}_1 = \underline{\mu}_2 = \cdots = \underline{\mu}_g \text{ dado que } \Sigma_1 = \Sigma_2 = \cdots = \Sigma_g$$

Esta hipótesis se conoce como el *análisis de varianza multivariado.*

Dos posibles pruebas de esta hipótesis son obtenidas particionando la matriz de "*suma de cuadrados y productos total*" (SSP) $T = \underline{X}^t H \underline{X}$ como:

$$T = W + B$$

donde W es la *matriz dentro de grupos* y B es la *matriz entre grupos.*

Por lo que la prueba de la Λ de Wilks y la raíz más grande están dadas por funciones de los eigenvalores de $W^{-1} B$. En particular si $g = 2$, $W^{-1} B$ tiene solamente un eigenvalor diferente de cero y las dos pruebas son la

misma y coinciden con la prueba T^2 de Hotelling para dos muestras. Bajo la hipótesis nula:

$$\left\{\frac{n_1 n_2(n-2)}{n}\right\} d^t W^{-1} d \sim T^2_{(p,\, n-p-1)}$$

y la hipótesis nula se rechaza para valores nulos de esta estadística.

3.5 Regla discriminante de Bayes

En algunas ocasiones es conveniente suponer que varias poblaciones tienen asignada una probabilidad *a priori*, por ejemplo en los diagnósticos médicos se puede pensar que un paciente que padece de presión arterial alta es más propenso a enfermedades cardiacas que otro paciente cuya presión arterial es normal.

La regla discriminante de Bayes utiliza las probabilidades a priori para la asignación de un individuo $\underline{X}$ a la población con mayor probabilidad *posterior*.

Definición 3.9 *Si las poblaciones* $\Pi_1, \cdots, \Pi_g$ *tienen probabilidades a priori* $\pi^t = (\pi_1, \cdots, \pi_g)$, *entonces la regla discriminante de Bayes (con respecto a* π*) asigna una observación* $\underline{X}$ *a la población que maximice:*

$$\pi_i L_i(\underline{X}).$$

Si todas las probabilidades iniciales fueran iguales, entonces la regla de máxima verosimilitud es un caso especial de la regla de Bayes. Cuando $g = 2$ poblaciones, las probabilidades consisten simplemente en aumentar el valor crítico de la función de discriminación en $\ln\left(\frac{\pi_2}{\pi_1}\right)$. Y la regla queda:

$$\text{Asignar } \underline{X} \text{ a} = \begin{cases} \Pi_1 & \text{si } h(\underline{X}) > \ln\left(\frac{\pi_2}{\pi_1}\right) \\ \Pi_2 & \text{en otro caso} \end{cases}$$

3.5.1 Propiedades óptimas

La regla discriminante de Bayes cumple con ciertas propiedades que son óptimas. Una regla discriminante aleatoria d, asigna una observación $\underline{X}$ a una población i con probabilidad $\phi_i(\underline{X})$, en donde $\phi_1, \cdots, \phi_g$ son funciones no negativas definidas en $\Re^p$, las cuales satisfacen:

$$\sum_{i=1}^{g} \phi_i(\underline{X}) = 1 \quad \text{para } \underline{X} \in \Re^p.$$

La regla de asignación determinísta es un caso particular de una regla de asignación aleatoria, tomando $\phi_i(\underline{X}) = 1$ para $\underline{X} \in \Re^p$ y $\phi_i(\underline{X}) = 0$ en cualquier otro caso.

Por ejemplo la regla de Bayes utilizando probabilidad a priori $\pi_1, \cdots, \pi_g$ está definida por:

$$\phi_i(\underline{X}) = \begin{cases} 1 & \text{si } \pi_i L_i(\underline{X}) = \max_k \{\pi_k L_k(\underline{X})\} \\ 0 & \text{en otro caso} \end{cases}$$

Definición 3.10 *La probabilidad de asignar un individuo a la población* Π_i, *cuando proviene de la población* Π_k, *está dada por:*

$$p_{ik} = \int \phi_i L_k(\underline{X})\, dx. \tag{3.18}$$

En particular la probabilidad de asignación correcta está dada por la expresión:

$$p_{ii} = \int \phi_i L_i(\underline{X})\, dx.$$

Definición 3.11 *Una regla discriminante* d *con probabilidad de asignación correcta* $\{p_{ii}\}$ *es tan buena como cualquier otra regla* d^* *con probabilidad* $\{p^*_{ii}\}$ *si:*

$$p_{ii} \geq p^*_{ii} \quad \text{para } i = 1, \cdots, g.$$

d *es menor que* d^* *si al menos una de las desigualdades es estricta.*

Definición 3.12 *si d es una regla para la cual no existe otra regla mejor, se dice entonces que d es una regla admisible.*

Teorema 3.13 *Todas las reglas discriminantes de Bayes son admisibles.*

Demostración.

Sea d^* una regla de bayes con probabilidad a priori $\pi_1, \cdots, \pi_g$. Suponga que existe otra regla d que es mejor que la regla anterior. Sean $\{p^*_{ii}\}$ y $\{p_{ii}\}$ las probabilidades de clasificación correcta para la regla d^* y d respectivamente.

Como d es mejor que d^* y $\pi_i > 0$ para toda i, se tiene que:

$$\sum_{i=1}^{g} \pi_i p_{ii} > \sum_{i=1}^{g} \pi_i p^*_{ii} .$$

Y por las Definiciones 3.11 y 3.12. ∎

3.6 La función lineal discriminante de fisher

Otra aproximación del análisis discriminante puede hacerse sin suponer alguna distribución paramétrica en las poblaciones $\Pi_1, \cdots, \Pi_g$ fué creada por Fisher (1936), que sugirió una función lineal $a^t \underline{X}$ la cual maximiza el cociente entre *la suma de cuadrados entre grupos* (B) y *la suma de cuadrados dentro de los grupos* (W), la combinación lineal se difine:

$$\underline{Y} = \underline{X}\, a$$

esto es:

$$\begin{pmatrix} \underline{X}_1\, a \\ \vdots \\ \underline{X}_g\, a \end{pmatrix} = \begin{pmatrix} \underline{Y}_1 \\ \vdots \\ \underline{Y}_g \end{pmatrix}$$

La *matriz de covarianzas total* para $\underline{Y}$, denotada por $T_{\underline{Y}}$, está definida como en (3.10).

$$
\begin{aligned}
T_{\underline{Y}} &= \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(\underline{Y}_{ij}-\overline{Y}_i\right)\left(\underline{Y}_{ij}-\overline{Y}_i\right)^t \\
&= a^t\frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(\underline{X}_{ij}-\overline{X}_i\right)\left(\underline{X}_{ij}-\overline{X}_i\right)^t a \\
&= a^t\, T_{\underline{X}}\, a.
\end{aligned}
$$

De esta forma *la matriz de covarianzas dentro de grupos* (W) para $\underline{Y}$ está dada por:

$$
\begin{aligned}
W_{\underline{Y}} &= \frac{1}{n}\sum_{i=1}^{g} n_i\,\left(\underline{Y}_{ij}-\overline{Y}_i\right)\left(\underline{Y}_{ij}-\overline{Y}_i\right)^t \\
&= a^t\frac{1}{n}\sum_{i=1}^{g} n_i\,\left(\underline{X}_{ij}-\overline{X}_i\right)\left(\underline{X}_{ij}-\overline{X}_i\right)^t a \\
&= a^t\, W_{\underline{X}}\, a.
\end{aligned}
$$

Así, análogamente *la matriz de covarianzas entre grupos* (B) para $\underline{Y}$ está dada por:

$$
\begin{aligned}
B_{\underline{Y}} &= \frac{1}{n}\sum_{i=1}^{g} n_i\,\left(\underline{Y}_{ij}-\overline{Y}_i\right)\left(\underline{Y}_{ij}-\overline{Y}_i\right)^t \\
&= a^t\frac{1}{n}\sum_{i=1}^{g} n_i\,\left(\underline{X}_{ij}-\overline{X}_i\right)\left(\underline{X}_{ij}-\overline{X}_i\right)^t a \\
&= a^t\, B_{\underline{X}}\, a.
\end{aligned}
$$

La razón de Fisher está dada por:

$$
\frac{a^t\, B_{\underline{X}}\, a}{a^t\, W_{\underline{X}}\, a}.
$$

Si a es el vector que maximiza la ecuación anterior, por lo que $a^t\,\underline{X}$ es llamada la función lineal, la función lineal discriminante de Fisher o la primera variable conónica.

Teorema 3.14 *El vector a en la función lineal discriminante de Fisher es el eigenvector de* $W^{-1}\,B$ *que corresponde al eigenvalor más grande.*

Demostración.
Vease Mardia (1995), pag. 319. ∎

Una vez que se ha calculado la función lineal discriminante, una observación $\underline{X}$ puede ser asignada a una de las g poblaciones con base en su puntaje discriminante o *score* $a^{t}\,\underline{X}$. La media muestral $\overline{X}_{i}$ tiene puntajes $a^{t}\,\overline{X}_{i} = \overline{Y}_{i}$. Entonces $\underline{X}$ es asignado a la población Π_{i} si:

$$\left|a^{t}\underline{X} - a^{t}\overline{X}_{j}\right| < \left|a^{t}\underline{X} - a^{t}\overline{X}_{i}\right| \text{ para } i \neq j.$$

La función discriminante de Fisher es más importante en el caso especial en el que $g = 2$ poblaciones. Entonces B tiene rango *uno* y puede escribirse como:

$$B = \left(\frac{n_1\, n_2}{n}\right) d\, d^{t},$$

donde $d = \left(\overline{X}_1 - \overline{X}_2\right)$, y $W^{-1}\,B$ tiene solamente un eigenvalor diferente de cero, que puede ser encontrado mediante la equación:

$$tr\left(W^{-1}\,B\right) = \left(\frac{n_1\, n_2}{n}\right) d^{t}\,W^{-1}\,d$$

y el correspondiente eigenvalor es:

$$a = W^{-1}\,d$$

así la regla discriminate es:

$$\text{Asignar } \underline{X} \text{ a} = \begin{cases} \Pi_1 & \text{si } d^{t}\,W^{-1}\left\{\underline{X} - \frac{1}{2}\left(\overline{X}_1 + \overline{X}_2\right)\right\} > 0 \\ \Pi_2 & \text{en otro caso} \end{cases} \qquad (3.19)$$

Observe que la regla anterior es exactamente la misma que la regla maximo verosímil muestral para dos grupos de la distribución normal multivariada

con la misma matriz de covarianzas; sin embargo en (3.13) existe el supuesto de normalidad, y en (3.19) se tiene solamente una regla basada en la función lineal de $\underline{X}$. Se espera que ésta regla sea apropiada en las que no se satisface exactamente la hipótesis de normalidad.

En general, $W^{-1}\,B$ tiene un $\min\{p,\ g-1\}$ de eigenvalores deferentes de cero; sus correspondientes eigenvectores definen la segunda, tercera ,etc. variable canónica. Las primeras k variables canónicas con $k \leq \min\{p,\ g-1\}$ son utilizadas cuando se espera que la diferencia entre grupos esté en k dimensiones.

3.7 Aplicaciones

Considere el caso particular en que $g = 2$ grupos.

Ejemplo 3.2 *La siguiente tabla contiene los datos de un banco que realizó prestamos a 16 personas y despues de un tiempo desea saber si le otorga el préstamo a dos nuevos clientes en base a la información que se tiene. Considere dos variables: los ingresos de los clientes, así como el total de la deuda. Los ingresos y la deuda son en millones; los ingresos son mensuales. El grupo uno considera a los clientes incumplidores y el grupo dos a los cumplidores.*

cliente	ingresos	deuda	grupo
1	1.3	4.1	1
2	3.7	6.9	1
3	5.0	3.0	1
4	5.9	6.5	1
5	7.1	5.4	1
6	4.0	2.7	1
7	7.9	7.6	1
8	5.1	3.8	1
9	5.2	1.0	2
10	9.8	4.2	2
11	9.0	4.8	2
12	12.0	2.0	2
13	6.3	5.2	2
14	8.7	1.1	2
15	11.1	4.1	2
16	9.9	1.6	2

Utilizando el paquete Statistica se obtiene los siguientes resultados.

Classification Matrix (sem.sta)
Rows: Observed classifications
Columns: Predicted classifications

	Percent Correct	G_1:1 p=.50000	G_2:2 p=.50000
G_1:1	100	8	0
G_2:2	87.5	1	7
Total	93.75	9	7

Tabla 1.

Se esperaría que un cliente del grupo 2 en realidad corresponda al grupo 1 ya que se tiene el 93.75% de los datos clasificados correctamente. ¿Pero cuál cliente será?

Classification Functions
grouping: GRUPO (sem.sta)

	G_1:1	G_2:2
INGR	0.7773204	1.8126000
DP	1.2956979	0.3639955
Constant	-5.8756928	-9.3958397

Tabla 2. Función de Fisher.

Con la Función discriminante descrita en la sección anterior; tenemos la regla de asignación al grupo 1 si los clientes obtiene valores negativos, y los que tengan valores positivos al grupo 2.

cliente	ingresos	deuda	grupo	discriminante	asignación
1	1.3	4.1	1	-5.994	1
2	3.7	6.9	1	-6.118	1
3	5.0	3.0	1	-1.139	1
4	5.9	6.5	1	-3.468	1
5	7.1	5.4	1	-1.201	1
6	4.0	2.7	1	-1.894	1
7	7.9	7.6	1	-2.422	1
8	5.1	3.8	1	-1.781	1
9	5.2	1.0	2	0.932	2
10	9.8	4.2	2	2.713	2
11	9.0	4.8	2	1.325	2
12	12.0	2.0	2	7.040	2
13	6.3	5.2	2	-1.843	1
14	8.7	1.1	2	4.462	2
15	11.1	4.1	2	4.152	2
16	9.9	1.6	2	5.239	2

Con esta nueva asignación ya se tiene el 100% de los datos clasificados correctamente.

Classification Matrix (sem.sta)
Rows: Observed classifications
Columns: Predicted classifications

	Percent Correct	G_1:1 p=.56250	G_2:2 p=.43750
G_1:1	100	9	0
G_2:2	100	0	7
Total	100	9	7

Tabla 3.

Pero lo que nos interesa saber es si se otorgan los dos nuevos prestamos.

cliente	ingresos	deuda	discriminante	asignación
1	10.1	6.8	0.601	2
2	9.7	2.2	4.472	2

Se utiliza la misma regla discriminante. Así que se otorgan los dos prestamos pero hay que observar que el segundo cliente tiene el valor discriminante más alto.

Ejemplo 3.3 *El gerente de un banco está molesto por el aumento de clientes morosos y desea reducir el número de clientes en esta situación, en base a 5 variables y tres grupos sabrá si le otorga un préstamo a una persona. La siguiente tabla contiene los datos.*

cliente	*grupo*	*ingresos*	*p_neto*	*c_propia*	*e_c*	*trab*
1	1	5450	56	1	1	0
2	1	3100	34	1	0	1
3	1	2100	8	0	1	1
4	1	6200	45	1	0	1
5	1	975	10	0	1	1
6	1	1250	22	1	1	1
7	1	4900	15	0	1	1
8	1	8900	38	1	1	1
9	1	3350	54	0	1	1
10	1	5200	80	1	1	0
11	1	2850	11	1	1	1
12	1	6500	22	1	1	1
13	1	7600	36	1	1	1
14	2	4350	39	1	1	0
15	2	1350	8	0	1	1
16	2	2100	19	1	0	1
17	2	1450	25	1	0	0
18	2	6400	4	0	1	1
19	2	3800	19	1	1	0
20	2	2450	24	0	1	0
21	3	3300	7	0	1	1
22	3	850	12	0	1	0
23	3	1200	5	1	1	1
24	3	1850	6	1	0	0
25	3	2650	25	0	0	0

donde:

grupo: Es el grado de cumplimiento del cliente.

1. Cliente cumplidor.
2. Cliente moroso.
3. Cliente incumplidor

ingresos: Son los ingresos anuales en miles.

p_neto: Patrimonio neto en millones.

c_porpia: Toma el valor

0. Si la casa es propia.
1. Si no lo es.

e_c: Toma el valor

0. Si es casado
1. Otro caso

trab:esta variable toma el valor

0. Si se tiene contrato
1. Si no tiene contrato, es eventual

Utilizando el paquete Statistica se obtienen los siguientes resultados:

Classification Matrix (grupos.sta)
Rows: Observed classifications
Columns: Predicted classifications

	Percent Correct	G_1:1 p=.52000	G_2:2 p=.28000	G_3:3 p=.20000
G_1:1	100	13	0	0
G_2:2	57.14	2	4	1
G_3:3	60.00	1	1	3
Total	80.00	16	5	4

Tabla 4.

Soló se tiene clasificado correctamente el 80% de los datos ¿Pero cuáles clientes habrá que cambiar?. Con la función discriminante se tiene la regla de asignación:

$$\text{Asignar } \underline{X} \text{ a} = \begin{cases} \Pi_1 & \text{si} \quad \underline{X} \leq -8 \\ \Pi_2 & \text{si} \quad -7 \leq \underline{X} < 0 \\ \Pi_3 & \text{si} \quad \underline{X} \geq 0 \end{cases}$$

la siguiente tabla muestra que efectivamente el 100% de los datos son asignados correctamente.

Classification Matrix (grupos.sta)
Rows: Observed classifications
Columns: Predicted classifications

	Percent Correct	G_1:1 p=.44000	G_2:2 p=.36000	G_3:3 p=.20000
G_1:1	100	11	0	0
G_2:2	100	0	9	0
G_3:3	100	0	0	5
Total	100	11	9	5

Tabla 5.

Considere 3 nuevos clientes, la siguiente tabla muestra los datos.

cliente	ingresos	p_neto	c_propia	e_c	trab
1	3000	40	0	1	1
2	1500	8	0	0	1
3	4850	38	0	1	0

¿El gerente autorizará los prestamos?. Utilizando la regla discriminante, los resultados son:

cliente	discriminante	asignación
1	-13	1
2	1	3
3	-2	2

El primer cliente como no paga renta eso le ayuda a poder solventar su deuda a pesar de que tiene trabajo eventual; el segundo cliente tiene ingresos menores y no tiene contrato por lo que posiblemente pueda no tener trabajo y sería un cliente *incumplidor*; el tercer cliente a pesar de tener buenos ingresos, no ser casado y tener casa propia puede ser considerado como un cliente *moroso*. Así que a los clientes 2 y 3 no se les otorga préstamo.

Ejemplo 3.4 *Se tienen 2 especies de escarabajos pulga, clasificados de acuerdo a 4 variables, la tabla 6 contiene los datos.*

	Haltica oleracea					Haltica carduorum			
número	Y1	Y2	Y3	Y4	número	Y1	Y2	Y3	Y4
1	189	245	137	163	1	181	305	184	209
2	192	260	132	217	2	158	237	133	188
3	217	276	141	192	3	184	300	166	231
4	221	299	142	213	4	171	273	162	213
5	171	239	128	158	5	181	297	163	224
6	192	262	147	173	6	181	308	160	223
7	213	278	136	201	7	177	301	166	221
8	192	255	128	185	8	198	308	141	197
9	170	244	128	192	9	180	286	146	214
10	201	276	146	186	10	177	299	171	192
11	195	242	128	192	11	176	317	166	213
12	205	263	147	192	12	192	312	166	209
13	180	252	121	167	13	176	285	141	200
14	192	283	138	183	14	169	287	162	214
15	200	294	138	188	15	164	265	147	192
16	192	277	150	177	16	181	308	157	204
17	200	287	136	173	17	192	276	154	209
18	181	255	146	183	18	181	278	149	235
19	192	287	141	198	19	175	271	140	192
					20	197	303	170	205

Tabla 6.

y las variables son:

Y_1 = *distancia de la ranura transversal desde el borde posterior del protorax (um).*

Y_2 = *longitud de alas en 0.01 mm..*

Y_3 = *longitud de la segunda conexión antenal (um).*

Y_4 = *longitud de la tercera conexión antenal (um).*

Los resultados se obtienen del paquete Statistica.

Classification Matrix (a_d2.sta)
Rows: Observed classifications
Columns: Predicted classifications

	Percent Correct	G_1:1 p=.48718	G_2:2 p=.51282
G_1:1	100	19	0
G_2:2	95	1	19
Total	97.436	20	19

Tabla 7.

Por lo que un escarabajo está mal clasificado; así que nuevamente se utiliza la función discriminante.

Classification Functions

	G_1:1	G_2:2
	p=.48718	p=.51282
Y1	0.9557	0.6105
Y2	-0.0209	0.1095
Y3	0.6843	0.7907
Y4	0.4353	0.5787
Constant	-178.335	-194.0887

Tabla 8. Función de Fisher.

La regla de asignación está dada como:

$$\text{Asignar } \underline{X} \text{ a} = \begin{cases} \Pi_1 & \text{si} \quad \underline{X} \geq 0 \\ \Pi_2 & \text{si} \quad \underline{X} < 0 \end{cases}$$

La siguiente tabla corresponde al 100% de los datos asignados correctamente. Originalmente había 19 escarabajos para el grupo 1 y 20 para el grupo 2. Despues de la clasificación con la función discriminante, hay 19 escarabajos para el grupo 2 y 20 para el grupo 1.

Classification Matrix (a_d2.sta)
Rows: Observed classifications
Columns: Predicted classifications

	Percent Correct	G_1:1 p=.51282	G_2:2 p=.48718
G_1:1	100	20	0
G_2:2	100	0	19
Total	100	20	19

Tabla 9.

Capítulo 4

Otras técnicas

En este capítulo se mencionaran otras técnicas multivariadas tales como:

- Análisis de conglomerados.
- Escalamiento multidimensional.

4.1 Análisis de conglomerados

El análisis de conglomerados ($A\ C$), tiene por objetivo identificar y agrupar objetos similares, de tal forma que se obtenga una serie de conjuntos homogéneos entre si, es decir, los objetos en un mismo conjunto son demasiado parecidos pero los objetos de 2 conjuntos distintos son diferentes. A estos conjuntos se les conoce como conglomerados.

Una vez que se tienen los conglomerados pueden considerarse cada uno de ellos como un solo individuo y así se reduce la dimensión de los datos.

4.1.1 Introducción

El análisis de conglomerados puede ser útil para reducir los datos definiendo grupos homogéneos de objetos, se utiliza en:

1. Biología.- Con animales o plantas cada especie pertenece a una serie de conglomerados de tamaño cada vez mayor con un número decreciente de características comunes, por ejemplo, el hombre pertenece a los primates, mamíferos, vertebrados y animales.

2. Medicina.- Un tipo particular de clasificación dentro de una enfermedad es la identificación de las etapas de severidad. Los grupos sanguíneos son un tipo de clasificación de la sangre.

3. Arqueología y antropología.- Al clasificar cráneos y objetos.

4.1.2 Funciones distancia

Para el desarrollo de las técnicas del análisis de conglomerados las funciones de distancia nos proporcionan un criterio para medir la similitud entre los objetos que se están estudiando.

Definición 4.1 *Sean s y r dos puntos y sea $d(s,r)$ una función, entonces $d(s,r)$ se dice que es una función distancia si cumple:*

1. $d(s,r) = d(s,r)$.
2. $d(s,r) \geq 0$.
3. $d(s,s) = 0$.
4. $d(s,r) = 0$ *si sólo si* $r = s$.
5. $d(s,r) \leq d(s,p) + d(p,r)$ (desigualdad del triángulo).
6. $d(s,r) \leq \max\{d(s,p), d(p,r)\}$ (desigualdad ultramétrica).

El concepto de similitud es fundamental para el análisis de conglomerados. La similitud interobjeto es una medida de la semejanza entre los objetos que serán conglomerados; existen tres métodos:

1. Medidas de correlación.- La medida de similitud interobjeto es el coeficiente de correlación entre un par de objetos sobre distintas variables; las columnas son los objetos y los renglones son las variables.

   Una correlación alta indica similitud y una baja correlación indica ausencia de la misma. Las medidas de correlación son usadas en otras técnicas multivariadas pero no son las más utilizadas en el análisis de conglomerados.

2. Medidas de asociación.- Son usadas para comparar objetos cuyas características tienen medidas ordinales o nominales (medidas no-métricas).

3. Medidas de distancia.- Vease la Definición (4.1). Las medidas de distancia más usadas son:

   (a) Distancia euclideana. Está definida como:

$$d_2(i,j) = \left[\sum_{k=1}^{n} (x_{ik} - x_{jk})^2 \right]^{1/2}$$

   (b) Distancia absoluta o distancia city-block. Está definida como:

$$d(i,j) = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$$

Definición 4.2 *Sea x_{ij} el valor de la i-ésima observación que corresponde a la j-ésima variable para $i = 1, \cdots, p$ y $j = 1, \cdots, N$, entonces la estandarización está dada como:*

$$z_{ij} = \frac{(x_{ij} - \overline{x}_j)}{s_j}$$

donde $\overline{x}_j$ y s_j son la media y la desviación estandar respectivamente de la j-ésima variable.

Definición 4.3 *La distancia euclideana de las observaciones estandarizadas se define como:*

$$d_{ih} = \left[\sum_{j=1}^{n} (z_{ij} - z_{hj})^2 \right]^{1/2}$$

4.1.3 La elección de variables

Es importante la elección inicial de las variables, la cual es una categorización de los datos y solamente hay guías matemáticas y estadísticas muy limitadas.

La primera duda es ¿Cuántas variables deben medirse en cada individuo?. En general no hay una base teórica para determinar el número de variables a usar y el problema debe resolverse empíricamente. El análisis de componentes principales es uno de los métodos más utilizados, el cual ha sido desarrollado en el *capítulo 2*. Los componentes principales indican la dimensión de los datos.

4.2 Técnicas jerárquicas

La clasificación jerárquica consiste en una serie de particiones que puede ir desde un solo conglomerado que contenga a todos los individuos, hasta n conglomerados y cada uno de ellos conteniendo a un solo individuo. Las clasificaciones jerárquicas pueden ser representadas por un diagrama bidimensional conocido como dendograma, el cual ilustra las divisiones que se realizaron.

4.2.1 Método liga simple

El método liga simple se basa en distancias mínimas, es decir, se calculan las distancias de los individuos u objetos y los que tengan la distancia más pequeña, son colocados en el primer conglomerado. La siguiente distancia más corta que se encuentra puede formar un nuevo conglomerado, o bien se integra al primer conglomerado; y así se continua hasta que todos los individuos formen parte del mismo conglomerado.

Definición 4.4 *Si C_1 y C_2 son dos conglomerados, la distancia que se define es la más pequeña disimilitud entre un elemento de C_1 y un elemento de C_2 y se calcula:*

$$d_{C_1 C_2} = \min \{ d_{r s} \mid r \in C_1,\ s \in C_2 \}$$

Ejemplo 4.1 *Considere la siguiente matriz de distancias*

$$D = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} \begin{bmatrix} 0 & 7 & 1 & 9 \\ 7 & 0 & 6 & 3 \\ 1 & 6 & 0 & 8 \\ 9 & 3 & 8 & 0 \end{bmatrix}$$

la mínima de ellas es $d_{13} = 1$, *por lo que los objetos* 1 *y* 3 *forman un conglomerado y*

$$\begin{aligned} d_{2(1,3)} &= \min\{d_{21}\, d_{23}\} = \min\{7,6\} = d_{23} = 6 \\ d_{4(1,3)} &= \min\{d_{41}\, d_{43}\} = \min\{9,8\} = d_{43} = 8 \end{aligned}$$

y la matriz distancia para los conglomerados es:

$$D = \begin{matrix} (1,3) \\ 2 \\ 4 \end{matrix} \begin{bmatrix} 0 & 6 & 8 \\ 6 & 0 & 3 \\ 8 & 3 & 0 \end{bmatrix}$$

como la distancia más pequeña es $d_{24} = 3$, *se tienen 2 conglomerados* $(1,3)$ *y* $(2,4)$. *Finalmente estos dos conglomerados se juntan para formar un sólo conglomerado.*

La siguiente figura ilustra este método. Es la forma vertical, aunque tambien está la forma horizontal.

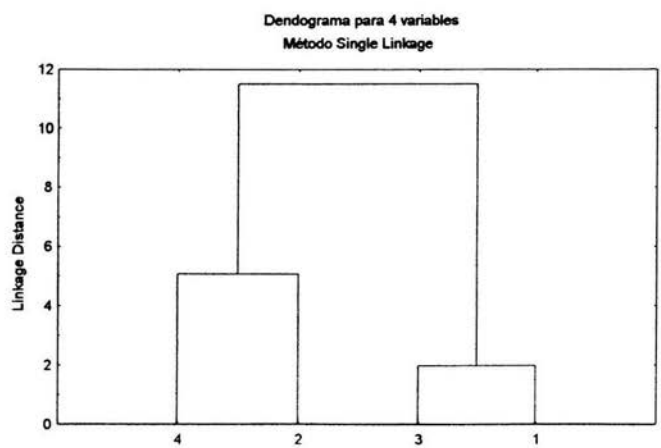

Figura 1.

4.2.2 Método Promedio entre grupos

Este método también se conoce como el de promedios, ya que el criterio para los conglomerados es la distancia promedio de todos los individuos de un conglomerado a todos los individuos de otro conglomerado.

Ejemplo 4.2 *Considere los datos del Ejemplo 4.1. El promedio mínimo es el de los individuos* 1 *y* 3, *y es* $d_{13} = 1$; *y las distancias promedio son:*

$$d_{(1,3)2} = \frac{1}{2}(d_{12} + d_{32}) = \frac{1}{2}(7+6) = 6.5$$

$$d_{(1,3)4} = \frac{1}{2}(d_{14} + d_{34}) = \frac{1}{2}(9+8) = 8.5$$

y la nueva matriz es:

$$D = \begin{matrix} \\ (1,3) \\ 2 \\ 4 \end{matrix} \begin{bmatrix} 0 & 6.5 & 8 \\ 6.5 & 0 & 3 \\ 8.5 & 3 & 0 \end{bmatrix}$$

por lo que la distancia más pequeña es $d_{24} = 1.5$, *y así, se tienen 2 conglomerados* $(1,3)$ *y* $(2,4)$. *Finalmente estos dos conglomerados se juntan para formar un sólo conglomerado.*

La siguiente figura ilustra este método. En el eje horizontal están representados los conglomerados y en el vertical las distancias entre ellos.

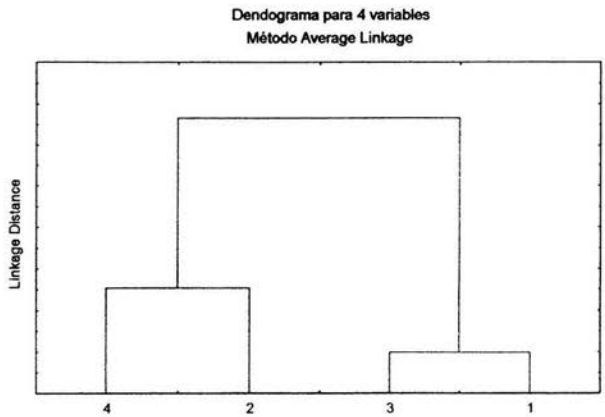

Figura 2.

4.2.3 Método Liga Completa

Este método es similar al método liga simple, pero el criterio para los conglomerados se basa en la distancia máxima. Este método es el opuesto al método liga simple.

Definición 4.5 *Si C_1 y C_2 son dos conglomerados, la distancia que se define entre ellos es la mayor disimilitud entre un elemento de C_1 y un elemento de C_2 y se calcula:*

$$d_{C_1 C_2} = \max \{d_{r s} \mid r \in C_1,\ s \in C_2\}$$

Ejemplo 4.3 *Considere los datos del Ejemplo 4.1. En donde la máxima distancia es $d_{1 4} = 9$ por lo que los objetos 1 y 4 forman un conglomerado y*

$$d_{2\,(1,4)} = \max \{d_{2 1}\ d_{2 4}\} = \max \{7, 3\} = d_{2 1} = 7$$
$$d_{3\,(1,4)} = \max \{d_{3 1}\ d_{3 4}\} = \max \{1, 8\} = d_{3 4} = 8$$

y la matriz distancia para los conglomerados es:

$$D = \begin{matrix} \\ (1,4) \\ 2 \\ 3 \end{matrix} \begin{bmatrix} 0 & 7 & 8 \\ 7 & 0 & 6 \\ 8 & 6 & 0 \end{bmatrix}$$

como la distancia más grande es $d_{(1,4)\,3} = 8$, los objetos 1, 3 y 4 son un conglomerado y el objeto 2 es otro.

$$D = \begin{matrix} (1,4,3) \\ 2 \end{matrix} \begin{bmatrix} 0 & 7 \\ 7 & 0 \end{bmatrix}$$

Finalmente estos dos conglomerados se juntan y forman un sólo conglomerado.

Este método queda ilustrado en la siguiente figura.

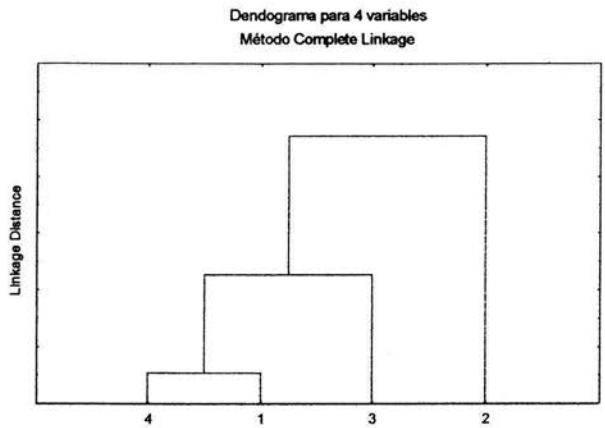

Figura 3.

4.3 Técnicas no jerárquicas

La clasificación no jerárquica tiene por objetivo realizar una sola partición de los individuos en k grupos, es decir, se debe especificar cuantos grupos se deben formar. Éste método utiliza los datos originales y no la matriz de distancias. En el método no jerárquico se tienen los siguientes métodos.

4.3.1 Método de Reasignación

Si un individuo ha sido asignado a un conglomerado, puede ser asignado a otro conglomerado en un paso posterior; esto mediante el método de *k-means* (k-medias); conocido también como método centroide o centro de gravedad.

Método k-means

Se divide un conjunto de individuos en k conglomerados de tal forma que cada individuo pertenece al conglomerado cuyo centro está más cercano a él. Este método se ilustrará en la siguiente sección.

4.4 Aplicaciones

Ejemplo 4.4 *Se realizó una prueba de aptitudes e inteligencia a* 12 *adultos, considerando 4 variables que son inteligencia, similitud, aritmética y rapidez; la escala va desde 0 "muy mal" hasta 14 "muy bién". Los datos son los siguientes.*

persona	*inteligencia*	*similitud*	*aritmética*	*rapidez*
1	*9*	*5*	*10*	*8*
2	*10*	*0*	*6*	*2*
3	*8*	*9*	*11*	*1*
4	*13*	*7*	*14*	*9*
5	*4*	*0*	*4*	*0*
6	*4*	*0*	*6*	*0*
7	*11*	*9*	*9*	*8*
8	*5*	*3*	*3*	*6*
9	*9*	*7*	*8*	*6*
10	*7*	*2*	*6*	*4*
11	*12*	*10*	*14*	*3*
12	*13*	*12*	*11*	*10*

El análisis se realizó en el paquete Statistica.

Con el Dendograma del método Liga Simple, se tiene:

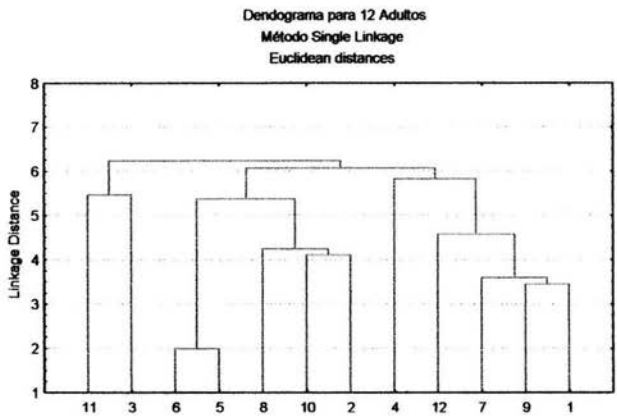

Figura 4.

Entonces se tienen cuatro grupos que están formados por: $(5,6)$, $(1,4,7,9,12)$, $(2,8,10)$ y $(3,11)$; esto significa que los resultados de la persona 5 son muy parecidos a los de la persona 6; pero diferentes a las demás personas.

Y de tener un grupo de 12 personas se obtienen 4 grupos; a pesar de que los grupos no sean homogéneos, es decir en un grupo hay 2 personas y en otro hay 5.

La siguiente figura muestra el método k-medias. Primeramente se considera $k = 4$.

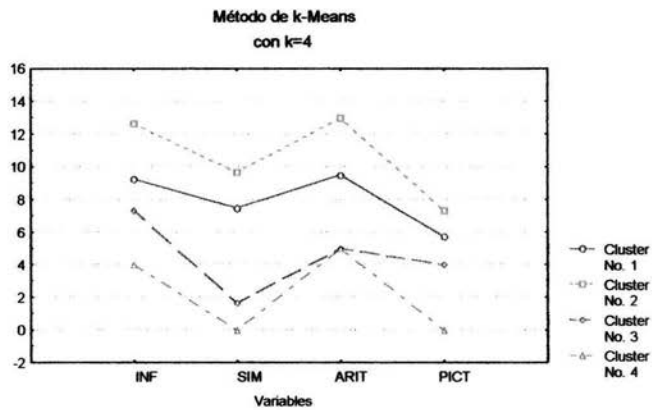

Figura 5.

los conglomerados 1 y 2 son muy parecidos, esto es por que se tienen 2 conglomerados con solo 2 individuos.

Si se considera $k = 5$, la gráfica del método k-medias es:

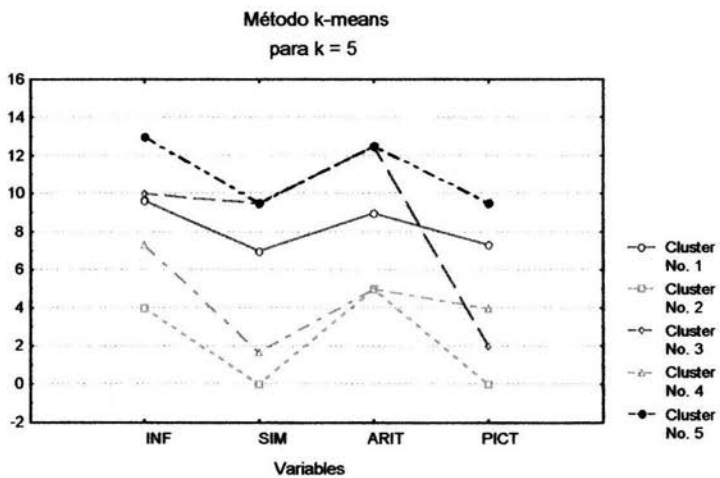

Figura 6.

el conglomerado 3 es muy diferente con respecto a los otros 4; y nuevamente los conglomerados 1 y 5 se parecen, así como los conglomerados 2 y 4.

Por lo que con el método Liga simple y el k-medias se consideran 4 conglomerados.

Ejemplo 4.5 *En una empresa de investigación de mercados desea saber en base a la opinión de* 50 *personas las prefencias de* 8 *tipos de comida que son:*

1 Japonesa	*5 Americana*
2 Cantonesa	*6 Española*
3 Francesa	*7 Italiana*
4 Mexicana	*8 Griega*

las variables utilizadas son, X_1 *que va del valor 1 "sencilla" hasta 7 "condimentada";* X_2 *su valor va de 1 "ligera" hasta 7 "pesada";* X_3 *nutritiva que va desde 1 "baja en calorias" hasta 7 "alta en calorias". La siguiente tabla contiene los promedios para las 3 variables:*

Comida	X_1	X_2	X_3
Japonesa	*2.8*	*3.2*	*3.4*
Cantonesa	*2.6*	*5.3*	*5.4*
Francesa	*3.5*	*4.5*	*5.1*
Mexicana	*6.4*	*4.3*	*4.3*
Americana	*2.3*	*5.8*	*5.7*
Española	*4.7*	*5.4*	*4.9*
Italiana	*4.6*	*6.0*	*6.2*
Griega	*5.3*	*4.7*	*6.0*

Nuevamente utilizando el paquete Statistica. Se obtuvo el siguiente Dendograma por el método Liga Simple:

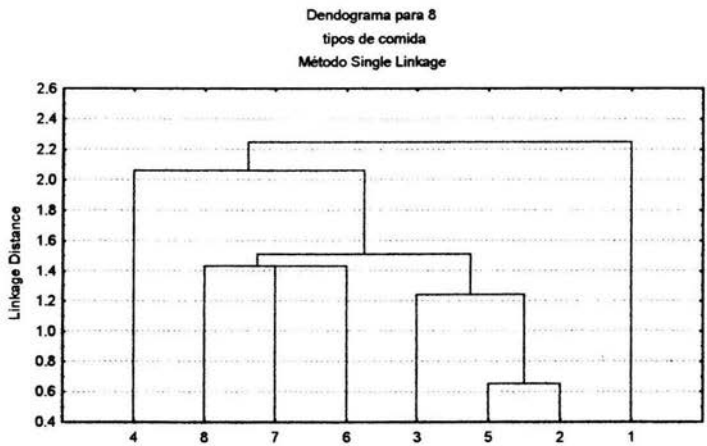

Figura 7.

Primeramente se tienen cuatro conglomerados formados por las comidas: (2, 3, 5), (7, 6, 8), (1) y (4); esto significa que las comidas Cantonesa, Americana y Francesa son muy similires en cuanto al sabor dulce que existe entre las primeras dos, pero el tipo de carne que utiliza la comida francesa aporta la cantidad de calorias que es muy similar a los otros 2 tipos de comida.

En los tipos de comida Española, Italiana y Griega es extraño que formen un conglomerado por la diferencia de sabores, y se unen sólo por la distancia de los valores.

Las comidas Japonesa y Mexicana forman conglomerados separados.

Comentarios

Es generalmente imposible saber que combinación de variables, medida de similaridad y técnica de conglomeración es apropiada para los datos. Más aún si la dimensión es grande no es fácil ver si se tiene una estructura de conglomerados.

4.5 Escalamiento multidimensional

El análisis de escalamiento multidimensional es un conjunto de procedimientos desarrollados para investigar, mediante algunas medidas las disimilaridades entre objetos; es decir, las medidas de disimilaridad son las distancias entre los objetos.

El análisis de escalamiento multidimensional busca la interpretación de coordenadas de matrices de distancias o disimilaridades; por ejemplo en mercadotecnia se logra determinar la competitividad entre productos y el criterio que la gente utiliza para comprarlos.

4.5.1 Introducción

El escalamiento multidimensional estudia la estructura de los individuos. El objetivo principal es tener un espacio métrico con el menor número de dimensiones posibles pero que permitan representar las proximidades entre objetos.

El escalamiento multidimensional puede utilizarse en antropología, psicología, sociología. Torgerson (1952-1958) fué el primero en publicar resultados sobre el escalamiento multidimensional.

4.6 Medidas de proximidad

El análisis de escalamiento multidimensional se basa en la comparación de objetos por ejemplo especies, zonas geográficas, productos. Las medidas de proximidad son un conjunto de números que indican el grado de semejanza o diferencia entre cada par de objetos en relación a ciertas características.

Definición 4.6 *Sea A el conjunto de n objetos, la similaridad es la medida de proximidad que indica el grado de semejanza entre el objeto i y el objeto j; y se denotará $s(i, j)$. Entre más grande es el valor de $s(i, j)$, mayor es la semejanza entre los objetos.*

Definición 4.7 *Sea A el conjunto de n objetos, la disimilaridad es la medida de proximidad que indica el grado de diferencia entre el objeto i y el objeto j; y se denotará* $\delta(i, j)$.

4.6.1 Medidas de proximidad derivadas

Las medidas de proximidad son índices derivados de otra información sobre los objetos, dando lugar a las medidas de proximidad derivadas

Coeficiente de correlación

Si se supone que se tienen variables cualitativas, el coeficiente de correlación de Pearson es una medida de proximidad derivada. Este coeficiente se calcula:

$$s(i,j) = \frac{\sum_{k=i}^{n} (x_{ik} - \overline{x}_i)(x_{jk} - \overline{x}_j)}{\left[\sum_{k=i}^{n} (x_{ik} - \overline{x}_i)^2 \sum_{k=i}^{n} (x_{jk} - \overline{x}_j)^2\right]^{1/2}}$$

donde x_i y x_j están relacionados linealmente. Y $s(i,j)$ es una medida de similaridad.

Distancias

Otro tipo de medidas de proximidad derivadas son las distancias como en la Definición (4.1), es decir, se define una medida de disimilaridad como $\delta(i,j) = d(i,j)$. Las más utilizadas son:

$$\delta_1(i,j) = d_1(i,j) = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$$

$$\delta_2(i,j) = d_2(i,j) = \left[\sum_{k=1}^{n} (x_{ik} - x_{jk})^2\right]^{1/2} \quad \text{distancia euclideana}$$

$$\delta_\infty(i,j) = d_\infty(i,j) = \sup_{1 \leq k \leq n} |x_{i\,k} - x_{j\,k}|$$

No existe una regla general para poder elegir una medida de proximidad que pueda servir para todos los casos del análisis de escalamiento multidimensional.

4.7 Solución clásica

Cuando se tiene una medida de disimilaridad, se obtiene una matriz de distancias D cuyos elementos son las medidas de disimilaridad entre los objetos.

Definición 4.8 *Una matiz de distancias D es euclideana si la medida de distancia de los elementos lo es.*

4.7.1 Similaridades

Algunas veces no se tiene la matriz de distancias, si no la de similaridad; para poder realizar el análisis de escalamiento multidimensional clásico se deben transformar las similaridades en distancias. La transformación es:

$$d_{i\,j} = (c_{i\,i} - 2\,c_{i\,j} + c_{j\,j})^{1/2} \tag{4.1}$$

En algunas situaciones la matriz de distancias D es no euclideana, por lo que algunos eigenvalores son negativos

4.7.2 Modelo métrico

Existen diferentes tipos del análisis de escalamiento multidimensional, esto se debe a los datos y a las medidas de proximidad; por lo que se tiene el modelo *métrico* y el *no métrico*. El primero supone que los datos están medidos en una escala intervalar, de razón o variables cuantitativas, mientras que el segundo supone que los datos son cualitativos principalmente ordinales.

El primer método del análisis de escalamiento multidimensional métrico se debe a Torgerson, el cual supuso que las medidas de disimilaridad son las distancias euclideanas:

$$\delta(i,j) = d_{ij} = d_2(i,j).$$

4.7.3 Modelo no métrico

El método del análisis de escalamiento multidimensional no métrico se debe a Kruskal. La hipótesis fundamental de este modelo es que las medidas de proximidad están relacionadas con las distancias entre los puntos mediante una función monótona. La cual se calculará:

$$\begin{aligned} \delta(i,j) &= f(d_{ij}) \\ &= f\left(\left[\sum_{k=1}^{n} (x_{ik} - x_{jk})^2\right]^{1/2}\right) \end{aligned}$$

en donde si $d_{ij} < d_{rs}$ entonces $f(d_{ij}) \leq f(d_{rs})$. Algunos ejemplos son las funciones lineales, la función exponencial y la función logaritmo.

Medidas de bondad de ajuste

En el análisis de escalamiento multidimensional no métrico, la medida de bondad de ajuste indica el grado de disimilaridad y existen tres conjuntos de parámetros.

El primero y más importante contiene las coordenadas x_{ik}.
El segundo contiene las distancias $d_{ij} = [\sum (x_{ik} - x_{jk})^p]^{1/p}$.
Y el tercero contiene las variables "dummy" o disparidades $\hat{d}_{ij}$, que son los valores más próximos a las distancias d_{ij} es decir si: $\delta_{ij} < \delta_{rs}$ entonces $\hat{d}_{ij} \leq \hat{d}_{rs}$.

Stress

Esta medida de bondad de ajuste fue propuesta por Kruskal. El Stress mide cuanto se desvían las disiparidades de las distancias; las medidas se calculan como:

$$S_1 = \left[\frac{\sum \left(\hat{d}_{ij} - d_{ij} \right)^2}{\sum d_{ij}^2} \right]^{1/2}$$

y

$$S_2 = \left[\frac{\sum \left(\hat{d}_{ij} - d_{ij} \right)^2}{\sum \left(\hat{d}_{ij} - d_{..} \right)} \right]^{1/2}$$

donde $d_{..} = \frac{1}{n^2} \sum_{ij} d_{ij}$.

Valores grandes de estas medidas indica un mal ajuste, un valor pequeño indica un buen ajuste.

4.7.4 Dimensionalidad y rotación

En el análisis de escalamiento multidimensional se necesita saber el número de dimensiones a conservar, la rotación de la solución obtenida y la interpretación del espacio obtenido. Para determinar la dimensionalidad se obtienen varias soluciones con diferentes dimensiones y en base a los criterios de *ajuste de los datos, interpretabilidad y reproductividad* se elige la dimensionalidad.

Ajuste de los datos Para determinar la dimensionalidad, se usa la gráfica *Scatter diagram*, la cual grafica en el eje de las abscisas los valores de las dimensiones y en el eje de las ordenadas los eigenvalores; se verá en la gráfica un codo una unidad adelante de la dimensionalidad.

Interpretabilidad En la interpretabilidad se debe conservar el espacio en el cual aparecen todas las características importantes de los objetos.

Reproductividad Se obtiene una solución para cada muestra y si hay k dimensiones que aparecen en todas ellas, entonces se tiene k dimensiones.

4.7.5 Rotación

Si en el análisis se utiliza la métrica euclideana, como en la solución clásica, entonces existe un problema de rotación. Si la solución no rotada no es facilmente interpretable se puede realizar una rotación objetiva.

Estas rotaciones son las utilizadas en el análisis de factores como son el método Varimax o Equimax.

4.8 Relación con otras técnicas

El análisis de escalamiento multidimensional está relacionado con otras técnicas multivariadas principalmente con las que se utilizan para reducir la dimensionalidad de los datos.

* Análisis de componentes principales.
* Análisis de factores.
* Análisis de conglomerados.

* Análisis de componentes principales

Dado un conjunto de variables $\underline{X}_1, \underline{X}_2, \cdots, \underline{X}_n$ mediante combinaciones lineales, se puede transformar en un conjunto más pequeño $\underline{Y}_1, \underline{Y}_2, \cdots, \underline{Y}_p$, con $p < n$ que describe la mayor parte de la varianza del conjunto original.

Si se utiliza el modelo métrico en el análisis de escalamiento multidimensional, entonces coincidirá con el de componentes principales; pero si se utiliza el modelo no métrico existirán diferencias en las técnicas.

* Análisis de factores

Los datos básicos en muchas de las aplicaciones del análisis de factores son medidas de proximidad entre pares de objetos. Cuando se utiliza el análisis de factores y el de escalamiento multidimensional para estudiar el mismo tema puede haber diferencias en las conclusiones debido a que el primero prefiere el coeficiente de correlación y el segundo las medidas de proximidad.

* Análisis de conglomerados

Dentro del análisis de conglomerados existe el método *jerárquico* que es el que está más relacionado con el escalamiento multidimensional, ya que los dos métodos son usados para investigar la estructura de los datos.

Las similitudes de los dos métodos son:

i. Con ambos métodos se pueden analizar matrices de proximidad.

ii. Ambos métodos se construyen sobre modelos de distancia.

Las diferencias de los éstos métodos son:

a. En el análisis de conglomerados la relación entre las proximidades $\delta(i, j)$ y las distancias $d(i, j)$ *no* es expresada como una combinación lineal o una función monótona como en el análisis de escalamiento multidimensional.

b. Las dimensiones de coordenadas en el análisis de escalamiento multidimensional son variables continuas, mientras que en el análisis de conglomerados las variables son discretas.

El análisis de escalamiento multidimensional y el análisis de conglomerados son considerados métodos competitivos, ya que Holman dice: "Si un método ajusta bien el otro ajusta pobremente"; pero Kruskal argumentó que los dos métodos son complementarios, es decir, si un método ajusta bien, el otro también.

4.9 Aplicaciones

Ejemplo 4.6 *La siguiente matriz de datos contiene la similaridad media entre 14 naciones tomadas de la respuesta de 10 personas en 1977. La escala va desde 1 para "muy diferente" hasta 9 para "muy similar".*

Euclidean distances (otro.sta)

	BRA	G_B	CHI	CUB	EGIPTO	INDIA	IND	ISR	JOR	P_B	POL	RUS	R_U	EU
BRA	0.000													
G_B	4.026	0.000												
CHI	5.335	4.725	0.000											
CUB	2.501	2.678	5.782	0.000										
EGIPTO	5.956	4.640	5.671	5.402	0.000									
INDIA	5.143	6.112	4.424	6.180	5.145	0.000								
IND	4.263	3.518	2.160	4.588	5.410	4.620	0.000							
ISR	5.663	4.794	6.359	5.080	1.633	5.334	5.891	0.000						
JOR	4.932	2.581	5.833	3.403	4.203	6.630	5.217	4.231	0.000					
P_B	4.708	5.312	5.218	5.223	3.837	3.154	4.763	3.735	5.754	0.000				
POL	5.475	4.285	1.053	5.624	5.327	4.834	2.107	6.094	5.419	5.277	0.000			
RUS	6.425	5.583	4.165	6.331	5.430	5.857	5.182	6.167	5.251	6.081	3.994	0.000		
R_U	3.710	4.549	5.033	3.923	5.636	5.548	4.473	5.668	5.060	4.475	5.036	5.151	0.000	
EU	6.607	7.741	7.520	6.948	6.839	6.675	7.794	6.747	7.060	5.994	7.672	5.876	4.694	0.000

Tabla 1.

Los resultados se obtuvieron del paquete Statistica, se utilizó el escalamiento multidimensional no métrico.

Se realizarón 37 iteraciones considerando solo 2 dimensiones; el valor del Stress es de .2434 que en realidad no es muy pequeño; posiblemente el análisis no es el más adecuado Los datos están en la siguienta tabla.

Final Configuration (otros.sta)
D-star: Raw stress = 19.42726; Alienation = .3109060
D-hat: Raw stress = 11.61886; Stress = .2434747

	DIM. 1	DIM. 2
BRA	-0.45049	1.00561
G_B	-0.29746	-0.48260
CHI	0.91490	0.56409
CUB	-0.01229	-0.66725
EGIPTO	-1.17857	-0.01489
INDIA	0.72500	-0.66511
IND	0.56467	0.69976
ISR	-0.74161	0.12721
JOR	-0.58182	-0.97885
P_B	0.28012	-1.23325
POL	0.92641	0.56098
RUS	1.12027	-0.01361
R_U	-1.07146	0.64368
EU	-0.19767	0.45423

Tabla 2.

Por lo que al realizar 34 iteraciones y tomando 3 dimensiones; el valor del Stress es de .1339 que en realidad es pequeño por lo que posiblemente el análisis es adecuado.

Final Configuration (otros.sta)
D-star: Raw stress = 5.258761; Alienation = .1632496
D-hat: Raw stress = 3.518101; Stress = .1339757

	DIM. 1	DIM. 2	DIM. 3
BRA	-0.95834	0.03185	0.47250
G_B	-0.46647	-0.79267	0.19099
CHI	0.91121	-0.07107	0.42043
CUB	-0.83730	-0.48963	0.05266
EGIPTO	-0.36955	0.95233	-0.39630
INDIA	0.28855	-0.67667	-0.61276
IND	0.42354	-0.31219	0.74915
ISR	0.06103	0.88268	-0.49161
JOR	-0.52365	-0.63372	-0.54003
P_B	0.23352	-0.11651	-1.09204
POL	0.90939	-0.06019	0.44190
RUS	0.95252	0.17369	-0.24214
R_U	-0.48328	0.92733	0.30744
EU	-0.14116	0.18477	0.73980

Tabla 3.

La siguiente gráfica corresponde a las tres dimensiones. Los números representan a los 14 países como:

1 Brasil	8 Israel
2 Gran Bretaña	9 Jordania
3 China	10 Países Bajos
4 Cuba	11 Polonia
5 Egipto	12 Rusia
6 India	13 Reino Unido
7 Indonesia	14 Estados Unidos

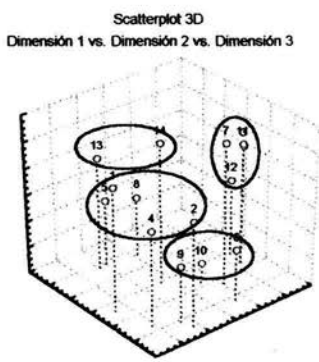

Figura 8.

Por lo que los países cercanos son parecidos en algún sentido. Es decir Reino Unido y Estados Unidos que son considerados paises desarrallodos; Brasil, Gran Bretaña, Israel, Cuba y Egipto son considerados países subdesarrollados;India, Jordania y Países Bajos son países subdesarrollados y no socialistas; y finalmente China, Indonesia, Polonia y Rusia ya que su régimen es socialista.

Conclusiones

En este trabajo se han revisado algunas de las técnicas del análisis estadístico multivariado con la finalidad de ofrecer notas de apoyo para cursos en esta materia.

Se ejemplificó la teoría mediante ejercicios resueltos usando el paquete *Statistica,* porque es un software apropiado para usarse durante las sesiones de clase debido a su versatilidad, además de que los resultados del paquete tienen la base teórica utilizada en este trabajo. Sin embargo, existen otros paquetes como MINITAB o SPSS que también son utilizados en análisis multivariado. Cualquiera de ellos puede servir para los fines propuestos, lo importante es notar que el uso de las computadoras es indispensable para realizar los compejos cálculos.

También se hizo hincapié en las técnicas de graficación para datos multivariados, pues éstas son una herramienta para el investigador que le permite saber sobre la validación de los supuestos; así como tener una visión de qué técnica utilizar.

Finalmente, como la selección de las técnicas multivariadas está en función de la naturaleza de los datos y los objetivos de estudio, el análisis multivariado es muy útil en diferentes áreas, por ejemplo:

- El análisis de Componentes Principales es usado frecuentemente en relaciones internacionales, sociología, medicina.
- El análisis de Factores originalmente fué utilizado en los "cambios de población" como mortalidad, natalidad, migración.
- El análisis Discriminante es una técnica de clasificación de individuos, uno de sus usos es en medicina.
- El análisis de conglomerados puede ser útil para reducir los datos definiendo grupos homogéneos de objetos. Se utiliza en medicina, biología, arqueología, antropología.
- El análisis de escalamiento multidimensional busca la interpretación de coordenadas y puede ser utilizado en mercadotecnia.

Apéndice A

Algebra Lineal

Este apéndice tiene como finalidad proporcionar las definiciones y resultados de álgebra lineal que son utilizados en este trabajo.

A.1 Matrices

Definición A.1 *Una matriz $A_{n \times p}$ es un arreglo de números de n renglones y p columnas como:*

$$A_{n \times p} = [a_{ij}] = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{pmatrix}$$

Definición A.2 *Las matrices de una columna se denominan vectores columna y se denotan por:*

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

Definición A.3 *Si* $A = (a_{i\,j})$ *es una matriz de orden* $n \times n$, *se define la* $diag(A)$ *como el vector columna.*

$$Diag(A) = \begin{pmatrix} a_{1\,1} \\ a_{2\,2} \\ \vdots \\ a_{n\,n} \end{pmatrix}$$

La matriz A es diagonal si $diag(A) = (a_{i\,j})$ con $a_{i\,j} = 0$ para $i \neq j$.

Definición A.4 *El rango de una matriz* $A_{n \times p}$, *se define como el máximo número de renglones linealmente independientes; se denota por* $ran(A)$.

Definición A.5

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

Definición A.6 *Se dice que una matriz* $A_{n \times n}$ *es idempotente si* $A\ A = A$.

A.2 Transpuesta de una matriz

Definición A.7 *La matriz transpuesta de una matriz* $A_{n \times p}$ *se denota por* A^t *y se define como la matriz:*

$$A^t_{p \times n} = [a_{j\,i}] = \begin{pmatrix} a_{1\,1} & a_{2\,1} & \cdots & a_{n\,1} \\ a_{1\,2} & a_{2\,2} & \cdots & a_{n\,2} \\ \vdots & \vdots & \cdots & \vdots \\ a_{1\,p} & a_{2\,p} & \cdots & a_{n\,p} \end{pmatrix}$$

Definición A.8 *Una matriz A es simétrica si $A^t = A$.*

Teorema A.9 *La tranpuesta de una matriz satisface las siguinetes propiedades:*

1. $(A^t)^t = A$.
2. $(A + B)^t = A^t + B^t$.
3. $(A\ B)^t = B^t A^t$.

A.3 Traza

Definición A.10 *La traza de una matriz se denota por $tr\ (\ A\ )$ y se define como:*

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

Teorema A.11 *Sea A y B matrices cuadrada de orden n y $\underline{X}_1, \underline{X}_2, ..., \underline{X}_p$ vectores de dimensión n., entonces la traza de un matriz satisface las siguientes propiedades:*

a. $tr(A \pm B) = tr(A) \pm tr(B)$.

b. $tr(A\ B) = tr(B\ A)$.

c. $tr(a\ A) = a\ tr(A)$.

d. $tr\left[\sum_{i=1}^{p} \underline{X}_i^t\ A\ \underline{X}_i\right] = tr\left[A\ \left(\sum_{i=1}^{p} \underline{X}_i\ \underline{X}_i^t\right)\right]$.

A.4 Determinante

Definición A.12 *El derterminate de una matriz cumple:*

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

A.5 Inversa

Definición A.13 *La inversa de una matriz cuadrada $A_{n \times n}$ es la matriz única denotada A^{-1}, que satisface $A\ A^{-1} = A^{-1}\ A = I$. Y existe si y solo si A es no singular; por lo que A es invertible.*

Teorema A.14 *Sea A y B matrices cuadradas e invertibles de orden n y c una constante, entonces se cumplen las siguientes propiedades:*

a. $A^{-1} = \dfrac{1}{|A|}\ (A_{ij})^{t}$

b. $(c\ A)^{-1} = c^{-1}\ A^{-1}$

c. $(A\ B)^{-1} = B^{-1}\ A^{-1}$

d. $|A| = |A_{11}|\ |A_{22} - A_{21}\ A^{11}\ A_{12}| = |A_{22}|\ |A_{11} - A_{12}\ A^{22}\ A_{21}|$

   i. La matriz A^{-1} en términos de submatrices corresponde a:

$$A^{-1} = \begin{pmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{pmatrix}$$

donde: $A^{11} = (A_{11} - A_{12}\ A^{22}\ A_{21})^{-1}$

$$A^{12} = -(A^{11}\, A_{12}\, A^{22})$$
$$A^{21} = -(A^{22}\, A_{21}\, A^{11})$$
$$A^{22} = (A_{22} - A_{21}\, A^{11}\, A_{12})^{-1}$$

Definición A.15 *Una matriz $Q_{n \times n}$ es ortogonal si es invertible, esto es $Q^t = Q^{-1}$; o bién $Q\, Q^t = I$ y se cumplen las siguientes propiedades:*

1. $Q^t = Q^{-1}$
2. $Q\, Q^t = I$
3. $|Q| = \pm 1$

A.6 Eigenvalores y Eigenvectores

Definición A.16 *Sea A una matriz cuadrada de orden n. El polinomio*

$$P(\lambda) = |A - \lambda I|$$

de grado n se conoce como la *ecuación característica* de A.

Definición A.17 *Sea A una matriz cuadrada de orden n, $P(\lambda)$ tiene n raíces $\lambda_1, \lambda_2, \ldots, \lambda_n$ que son los eigenvalores (valores propios) de la matriz A.*

Definición A.18 *Sea A una matriz cuadrada de orden n, los vectores $\underline{X}_1$, $\underline{X}_2, \ldots, \underline{X}_p$ que satisfacen $A\,\underline{X} = \lambda_i\,\underline{X}_i$ se denominan eigenvectores (vectores propios) de la matriz A, correspondientes al eigenvalor λ_i.*

La solución de $A\,\underline{X} = \lambda_i\,\underline{X}_i$ no es única por que el sistema es un conjunto homogéneo de ecuaciones.

Teorema A.19 *Sea A una matriz cuadrada de orden n, con eigenvalores $\lambda_1, \lambda_2, \ldots, \lambda_n$, se cumplen las siguientes propiedades:*

a. $tr(A) = \sum_{i=1}^{n} \lambda_i$

b. $|A| = \prod_{i=i}^{p} \lambda_i$

Teorema A.20 *Si A es una matriz cuadrada de orden n y $\lambda_1, \lambda_2, \ldots, \lambda_n$ los eigenvalores de A, entonces:*

i. $A > 0$ *si y solo si* $\lambda_i > 0, \qquad i = 1, \cdots, n$

ii. $A \geq 0$ *si y solo si* $\lambda_i \geq 0, \qquad i = 1, \cdots, n$

Teorema A.21 *(descomposición espectral)Cualquier matriz simétrica A puede escribirse como:*

$$A = U\Lambda U^{t} = \sum_{i=1}^{n} \lambda_i U_{(i)} U_{(i)}^{t}$$

donde $\Lambda = Diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$, y cada λ_i son los eigenvalores de la matriz A y donde U es una matriz ortogonal donde las columnas son los eigenvectores estandarizados de A.

Teorema A.22 *Si $A > 0$, existe una matriz simétrica y definida positiva $A^{1/2}$ tal que:*

$$A = A^{1/2} A^{1/2}$$

Definición A.23 *Una forma cuadrática en el vector $\underline{Y}$ es una función de la forma:*

$$\underline{Y}^{t} A \underline{Y} = \sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij} \underline{Y}_i \underline{Y}_j$$

donde A es una matriz simétrica, y $\underline{Y}$ es un vector de dimensión p.

Teorema A.24 *Si* $f(G) = -n\ln(G) - tr(G^{-1}D)$ *donde* D *es una matriz definida positiva de orden* p, *el máximo de* G *existe y ocurre en* $G = \left(\frac{1}{n}\ D\right)$, *por lo que:*

$$f(\frac{1}{n}\ D) = p\ n\ \ln(n) - n\ \ln|D| - p\ n$$

Teorema A.25 *Sea* A *una matríz simétrica, se cumple:*

$$\frac{\partial \ln|A|}{\partial A} = 2\ A^{-1} - Diag(A^{-1})$$

A.7 Distancias

Definición A.26 *Sea* $\underline{x} = (\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n)$, *y* $\underline{y} = (\underline{y}_1, \underline{y}_2, \cdots, \underline{y}_n)$. *La distancia entre* i *y* j *está dada como:*

$$\begin{aligned} d(\underline{x}, \underline{y}) &= \sqrt{(\underline{x}_1 - \underline{y}_1)^2 + (\underline{x}_2 - \underline{y}_2)^2 + \cdots + (\underline{x}_n - \underline{y}_n)^2} \\ &= ||(\underline{x} - \underline{y})|| \end{aligned}$$

Definición A.27 *La distancia al cuadrado entre los centroides de dos grupos* p *y* q, *se define por:*

$$D^2_{pq} = \left(\underline{\mu}_p - \underline{\mu}_q\right)^t\ \Sigma^{-1}\left(\underline{\mu}_p - \underline{\mu}_q\right)$$

donde $\underline{\mu}_p$ *y* $\underline{\mu}_q$ *son los vectores columna que contienen las medias de las variables de los grupos respectivos y* Σ^{-1} *es la inversa de la matriz de covarianzas intragrupos conjuntamente de los dos grupos. Y se le conoce como la* `distancia de Mahalanobis`.

Frecuentemente la distancia de Mahalanobis es utilizada para medir la distancia de una única observación multivariada al centro de la población de la cual proviene.

Bibliografía

[1] Anderson, T. W. (1994). *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons. New York.

[2] Bisquerra, R. (1989). *Introduccion Conceptual al Análisis Multivariable*. Promociones y Publicaciones Universitarias. Barcelona.

[3] Cuadras, C. M. (1991). *Métodos de Análisis Multivariante*. Promociones y Publicaciones Universitarias. Barcelona.

[4] Dallas, E. (2000). *Métodos Multivariados Aplicados al Análisis de datos*. International Thomson Editores. México.

[5] Dillon, W., Goldstein, W.(1984). *Multivariate Analysis Methods and Applications*. John Wiley & Sons. New York.

[6] Everitt, B., Dunn, G. (1991). *Applied Multivariate Data Analysis*. John Wiley & Sons. New York.

[7] Hair, J., Anderson, R., Tatham, R., Black, W. (1984). *Multivariate Data Analysis*. Prentice Hall. New Jersey.

[8] Jhonson, R. A. (1982). *Applied Multivariate Statistical Analysis*. Prentice Hall. New Jersey.

[9] Lawrence, A. (1978). Graphical Representation of Multivariate Data. *On Chernoff Faces*, 93-114

[10] Manly, B. (1986). *Multivariate Statistical Methods*. Chapman and Hall. New Zealand.

[11] Mardia, K., Kent, J., Bibby, J. M. (1995). *Multivariate Statistical Analysis*. Academic Press. Great Britain.

[12] Mendoza B. J. (1987). *La distribución normal multivariada y su relación con otras distribuciones*. Tesis de Licenciatura.U.N.A.M.

[13] Moreno, H. (1992). *Técnicas de Graficación para datos multivariados*. Tesis de Licenciatura.U.N.A.M.

[14] Morrison, D. (1976). *Multivariate Statistical Methods*. McGraw-Hill. Nueva York.

[15] Rencher, A. C. (1995). *Methods of Multivariate Analysis*. John Wiley & Sons. New York.

[16] Seber, G. A. (1984). *Multivariate Observations*. John Wiley & Sons. New Zealand.

[17] Uriel, E. (1995). *Análisis de Datos de Series Temporales y Análisis Multivariante*. A. C. Madrid.