U N I V E R S I D A D N A C I O N A L

A U T O N O M A D E M E X I C O

UNIDAD ACADEMICA DE LOS CICLOS PROFESIONAL

Y DE POSGRADO DEL C C H

"MEZCLAS DE DISTRIBUCIONES NORMALES" MULTIVARIADAS

T E S I S

MAESTRIA EN ESTADISTICA E INVESTIGACION DE

OPERACIONES

JAVIER ALAGON CANO

JULIO 1983

I N D I C E

CAPITULO 1

INTRODUCCION

El término más usado para referirse a las técnicas estadísticas que tienen por objetivo separar un conjunto de datos y formar distintos grupos es el de "Análisis de Clasificación". Muchos autores usan este término para referirse también a técnicas que buscan agrupar variables.

Dentro del Análisis de Clasificación se encuentran el Análisis Discriminante y el Análisis de Conglomerados*. En el primero de éstos se conoce de antemano el tipo de grupos que se formarán, así como el número de éstos; en el segundo se desconocen a priori ambas características. El Análisis Discriminante es usado también para estudiar las características principales que hacen diferente a un grupo de individuos de otro; en base a esto, se puede asignar un nuevo individuo en alguno de los grupos. Por su parte, el Análisis de Conglomerados tiene como principal objetivo, simplemente el formar grupos de individuos de acuerdo a un conjunto de características medidas en éstos.

Este trabajo trata con una de las técnicas de Análisis de Conglomerados, específicamente la técnica que consis-

* En Inglés: Cluster Analysis.

te en ajustar mezclas de distribuciones multivariadas a un conjunto de datos multivariados. El objetivo de esta técnica es resolver el siguiente problema: dada una muestra de N objetos o individuos a cada uno de los cuales se han medido p variables, agrupar estos objetos o individuos en g grupos o clases distintas. El número de estos grupos no se conoce de antemano. Esta técnica de análisis de conglomerados es de tipo no jerárquico.

Clasificar gente en distintos tipos ha sido para el ser humano todo un pasatiempo desde épocas muy remotas. Los antiguos hindúes usaban características físicas y de comportamiento así como el sexo de las personas para clasificar a la gente en seis tipos distintos que ellos designaban con nombres de animales. Los antiguos fisiólogos griegos y romanos desarrollarion varias tipologías basadas en variaciones de características físicas resultantes de una mezcla de cuatro humores. La más destacada de estas clasificaciones fue la de Galeno (129-199 d.C.), que definió nueve tipos de personas de acuerdo a su temperamento y que se suponía se relacionaba con la susceptibilidad de la persona a distintas enfermedades y con las diferencias individuales en el comportamiento. En el siglo XVIII, Lineo hizo la primera clasificación científica dentro de los reinos animal y vegetal.

La mayoría del trabajo de clasificación se hizo

primero en los campos de la Botánica y la Zoología en donde se le conoció por Taxonomía. Esta fue en un principio, más un arte que una técnica científica, pero gradualmente se fueron desarrollando técnicas más objetivas hasta llegar a los métodos de la Taxonomía Numérica. Posteriormente, son particularmente importantes los intentos de Zubin en 1938 y de Thorndike en 1953 al usar técnicas de clasificación numérica en campos muy alejados de los de las ciencias naturales. Sin embargo, estas técnicas tuvieron que esperar hasta el desarrollo de las computadoras electrónicas puesto que incluyen una gran cantidad de cálculos numéricos.

Desde hace dos décadas, tanto estadísticos como matemáticos han estado trabajando en un enfoque más formal al análisis de conglomerados y se han hecho muchos intentos para formular modelos estadísticos y matemáticos acordes con los problemas de clasificación.

La mayoría de los procedimientos de análisis de conglomerados tratan de medir la similitud entre cualesquiera dos objetos y entonces agrupan los objetos maximizando la similitud dentro de grupos. Desafortunadamente, no existe una "medida apropiada" de similitud, pues ésta depende del problema bajo investigación, por lo que forzosamente se cae en esquemas subjetivos de clasificación. Para cierto tipo de datos puede ser muy conveniente una técnica de análisis de conglomerados sin suposiciones arbitrarias de similitud. La

técnica de ajustar mezclas de distribuciones a un conjunto de datos, es de este tipo, puesto que no incluye medidas de similitud subjetivas.

El principal objetivo de esta tesis consistió en resolver un problema particular en la técnica de mezclas. Específicamente, el problema consistía en encontrar los estimadores máximo verosímiles de los parámetros involucrados en una mezcla de distribuciones normales multivariadas (mediante los cuales podemos identificar los distintos grupos existentes en los datos) en el caso particular en que las matrices de dispersión de los grupos son del tipo $\Sigma_s = a_s \Sigma$ donde Σ es una matriz positiva definida y a_s es un número real positivo, $s=1,2,\ldots,g$ y g es el número de grupos a formar. Tal y como se demuestra en el Capítulo III de esta tesis, este caso es un caso particular de la situación en donde se tienen matrices de correlación iguales en los grupos. Este caso se presenta frecuentemente en problemas de Biología, en donde se pueden tener grupos en donde las varianzas y covarianzas de las variables medidas cambian proporcionalmente en los distintos grupos; por ejemplo, en Zoología podemos encontrarnos con una situación en donde al clasificar diferentes especies animales de acuerdo a un conjunto de medidas sobre varios órganos del cuerpo, éstas difieran en medias y varianzas para distintos grupos aunque se tengan las mismas correlaciones en los distintos grupos. Desde luego

que el caso en que se tienen matrices de correlación iguales ya estaba resuelto (Ver [1]). Sin embargo, simplificar el problema tiene muchas ventajas estadísticas y numéricas, puesto que se tiene un menor número de parámetros por estimar, lo cual no hace necesario un tamaño de muestra muy grande para calcular los estimadores de las varianzas y covarianzas en los grupos; además se simplifican los cálculos numéricos dentro de la computadora y la convergencia de los métodos iterativos usados es más rápida.

La simplificación anterior es semejante a la que se hace cuando en el caso general se suponen matrices de dispersión iguales. Estos fueron resueltos y presentados por J.H. Wolfe en 1967 (Ver [18]). El mismo Wolfe implantó estos casos en computadora en un programa al que tituló NORMIX (del Inglés NORmal MIXtures). Este programa se encuentra en IIMAS. En el último capítulo de este trabajo presentamos la manera de usar NORMIX.

CAPITULO II

MAXIMA VEROSIMILITUD EN MEZCLAS DE DISTRIBUCIONES MULTIVARIADAS

Tal y como lo señalamos en el capítulo anterior, este trabajo tratará con una técnica para el análisis de datos multivariados, cuyo objetivo es resolver el siguiente problema: Dada una muestra de N objetos o individuos a cada uno de los cuales se le han medido p variables, agrupar estos objetos o individuos en g grupos o clases distintas. El número de estos grupos no se conoce de antemano.

El objetivo central de este capítulo es presentar la técnica de máxima verosimilitud en mezclas de distribuciones multivariadas para resolver el problema anterior. Presentamos también algunos métodos iterativos para la resolución de las ecuaciones máximo verosímiles.

§ 1. DEFINICION Y CALCULO DE ESTIMADORES.

Dado que los objetos o individuos dentro de un grupo difieren uno de otro, es razonable suponer la existencia de una distribución de probabilidad de las características de una población que pertenece a un grupo. Así, los individuos

de grupos diferentes tendrán distintas distribuciones de probabilidad de sus características. La población combinada tomada de todos los grupos tendrá una distribución de probabilidad conocida como una mezcla de distribuciones. El problema de análisis de conglomerados consiste en identificar y describir las distribuciones componentes de una mezcla a partir de la muestra. Usualmente se supone que las distribuciones componentes son distribuciones estadísticas estándar con parámetros desconocidos. El problema de agrupación se resuelve entonces con técnicas estadísticas estándar de estimación paramétrica y asignando a cada individuo al grupo al grupo al cual tenga mayor probabilidad de pertenecer. Este trabajo trata única y exclusivamente, estimación por máxima verosimilitud.

Para ello, sean $\alpha_1(\underline{x}, \Theta_1), \ldots, \alpha_g(\underline{x}, \Theta_g)$, g distribuciones de probabilidad definidas en un espacio de dimensión p de vectores aleatorios

$$\underline{x} = (x_1, \ldots, x_p)$$

Supongamos además, que α_s es una función dos veces derivable en sus parámetros

$$\Theta_s = (\theta_{s1}, \ldots, \theta_{sq}), \quad s = 1, \ldots, g \; ; \; q \in \mathbb{N}$$

Formamos una mezcla de distribuciones tomando proporciones λ_s de la población del grupo s, es decir, la distribución de probabilidad de la mezcla está dada por

(2.1) $$f(\underline{x}) = \sum_{s=1}^{g} \lambda_s \alpha_s(\underline{x}, \textcircled{H}_s)$$

donde

(2.2) $$\sum_{s=1}^{g} \lambda_s = 1.$$

La probabilidad de que el vector $\underline{x}$ pertenezca al grupo s está dada por

(2.3) $$P(s/\underline{x}) = \frac{P(s)\,P(\underline{x}/s)}{P(\underline{x})} = \frac{\lambda_s\,\alpha_s(\underline{x}, \textcircled{H}_s)}{f(\underline{x})}$$

Supongamos que obtenemos una muestra de N vectores aleatorios. El vector k-ésimo lo representaremos por

$$\underline{X}_k = (x_{1k}, \ldots, x_{pk}), \quad k = 1, \ldots, N.$$

Los estimadores máximo verosímiles de los parámetros involucrados en la mezcla, $\lambda_s, \textcircled{H}_s$, $s = 1, \ldots, g$, son aquellos valores de $\lambda_s, \textcircled{H}_s$ que maximizan la verosimilitud de la muestra

$$\log L = \sum_{k=1}^{N} \log f(\underline{x}_k)$$

sujeta a la restricción

$$\sum_{s=1}^{g} \lambda_s = 1$$

Usando un multiplicador de Lagrange γ, formamos la función objetivo

$$L = \sum_{k=1}^{N} \log f(\underline{x}_k) - \gamma \left(\sum_{s=1}^{g} \lambda_s - 1 \right)$$

Las ecuaciones máximo verosímiles son

$$\frac{\partial L}{\partial \lambda_s} = \sum_{k=1}^{N} \frac{\kappa_s(\underline{x}_k, \Phi_s)}{f(\underline{x}_k)} - \gamma = 0$$

$$\frac{\partial L}{\partial \theta_{si}} = \sum_{k=1}^{N} \frac{\lambda_s}{f(\underline{x}_k)} \frac{\partial \kappa_s}{\partial \theta_{si}} = 0$$

Si multiplicamos la primera de estas ecuaciones por λ_s y sustituimos en (2.3) obtenemos que

$$\sum_{k=1}^{N} \frac{\lambda_s \kappa_s(\underline{x}_k, \Phi_s)}{f(\underline{x}_k)} - \gamma \lambda_s = 0$$

es decir

$$\sum_{k=1}^{N} P(s/\underline{x}_k) - \gamma \lambda_s = 0$$

Si ahora sumamos sobre s concluimos que

$$\sum_{s=1}^{g} \sum_{k=1}^{N} P(s/\underline{x}_k) - \gamma \sum_{s=1}^{g} \lambda_s = 0$$

i.e.,

$$\gamma = \sum_{s=1}^{g} \sum_{k=1}^{N} P(s/\underline{x}_k) = \sum_{k=1}^{N} \sum_{s=1}^{g} \frac{\lambda_s \kappa_s(\underline{x}_k, \Phi_s)}{f(\underline{x}_k)} = \sum_{k=1}^{N} \frac{f(\underline{x}_k)}{f(\underline{x}_k)} = N$$

Podemos obtener más información de estas ecuaciones máximo verosímiles, como presentamos en los siguientes teoremas.

Teorema 1. La estimación por máxima verosimilitud de las proporciones de una mezcla es igual a la media muestral de las probabilidades de pertenencia al grupo dado, es decir

(2.4) $$\hat{\lambda}_s = \frac{1}{N} \sum_{k=1}^{N} P(s/\underline{x}_k)$$

Demostración. Tenemos que

$$\sum_{k=1}^{N} P(s/\underline{x}_k) - \gamma \lambda_s = 0$$

Dado que $\gamma = N$, concluimos que

$$\hat{\lambda}_s = \sum_{k=1}^{N} \vec{P}(s/\underline{x}_k). \quad \square$$

Teorema 2. Las ecuaciones máximo verosímiles para estimar los parámetros de las distribuciones de una mezcla están dadas por

(2.5) $$\frac{\partial L}{\partial \theta_{si}} = \sum_{k=1}^{N} P(s/\underline{x}_k) \frac{\partial \log \alpha_s(\underline{x}_k, \textcircled{\#}_s)}{\partial \theta_{si}} = 0$$

Demostración. Tenemos que $P(s/\underline{x}_k) = \lambda_s \alpha_s(\underline{x}_k, \textcircled{\#}_s) / f(\underline{x}_k)$

Por lo tanto,

$$\frac{\partial L}{\partial \theta_{si}} = \sum_{k=1}^{N} \frac{\lambda_s}{f(\underline{x}_k)} \frac{\partial \alpha_s(\underline{x}_k, \textcircled{\#}_s)}{\partial \theta_{si}} = \sum_{k=1}^{N} \frac{P(s/\underline{x}_k)}{\alpha_s(\underline{x}_k, \textcircled{\#}_s)} \frac{\partial \alpha_s(\underline{x}_k, \textcircled{\#}_s)}{\partial \theta_{si}} =$$

$$= \sum_{k=1}^{N} P(s/\underline{x}_k) \frac{\partial \log \alpha_s(\underline{x}_k, \textcircled{\#}_s)}{\partial \theta_{si}} = 0. \quad \square$$

Notemos que si toda la población fuera obtenida de un solo grupo, las ecuaciones máximo verosímiles para Θ_s serían

$$\sum_{k=1}^{N} \frac{\partial \log \alpha_s}{\partial \theta_{si}} = 0$$

Por lo tanto, las ecuaciones máximo verosímiles para las estimaciones de los parámetros de una mezcla, son los promedios ponderados de las expresiones usadas al obtener las estimaciones máximo verosímiles para grupos puros, donde las ponderaciones son las probabilidades de pertenencia a los grupos.

Usualmente, el número de grupos g, es una hipótesis que debe probarse contra la alternativa de g' y esto se hace encontrando las estimaciones por máxima verosimilitud bajo ambas hipótesis y probando el cociente de verosimilitud con la fórmula

$$\chi^2 = -2 \log \left[\frac{L_g}{L_{g'}} \right]$$

con grados de libertad igual a la diferencia de los números de parámetros a estimar bajo las dos hipótesis. Por supuesto que la distribución del logaritmo del cociente de verosimilitud es aproximadamente χ^2 para muestras grandes solamente. Esto se verá con más detalle en el capítulo IV.

Notemos que las ecuaciones (2.4) y (2.5) no están en forma cerrada, por lo que tienen que ser resueltas numéricamente. Para ello, veremos algunos métodos iterativos de resolución de estas ecuaciones.

2. METODOS ITERATIVOS

En la mayoría de los casos, las ecuaciones máximo verosímiles

(2.4) $$\frac{1}{\hat{\lambda}_s} \sum_{k=1}^{N} \hat{P}(s/\underline{x}_k) - N = 0$$

y

(2.5) $$\frac{\partial L}{\partial \theta_{si}} = \sum_{k=1}^{N} \hat{P}(s/\underline{x}_k) \frac{\partial \log \alpha_s(\underline{x}_k, \textcircled{H}_s)}{\partial \theta_{si}} = 0$$

tienen que ser resueltas numéricamente. La manera estándar de hacer esto es mediante un proceso Newton-Raphson. Kale [13], en 1962, estudió la convergencia de este método así como la de otro método de resolución al que llama el 'método de puntajes' (*). Estos métodos son parecidos. El primero usa la matriz J^{-1} donde

$$J = \left(\begin{array}{c|c} J_{\lambda\lambda} & J_{\lambda\theta} \\ \hline J_{\theta\lambda} & J_{\theta\theta} \end{array} \right)$$

y J se encuentra particionada en las submatrices $J_{\lambda\lambda}$, $J_{\lambda\theta}$, $J_{\theta\lambda}$, $J_{\theta\theta}$, las cuales están definidas por

$$J_{\lambda\lambda} = \left\{ \frac{\partial^2 \log L}{\partial \lambda_s \partial \lambda_p} \right\}$$

$$J_{\lambda\theta} = \left\{ \frac{\partial^2 \log L}{\partial \lambda_s \partial \theta_{pj}} \right\}$$

* En inglés: 'Method of scoring'

$$J_{\theta\theta} = \left\{ \frac{\partial^2 \log L}{\partial \theta_{si}\, \partial \theta_{pj}} \right\}$$

y $J_{\theta\lambda} = J_{\lambda\theta}^{t}$. Las ecuaciones máximo verosímiles pueden ser resueltas iterativamente usando las siguientes ecuaciones:

$$(2.6)\quad \begin{pmatrix} T_{r+1} \\ \\ S_{r+1} \end{pmatrix} = \begin{pmatrix} T_r \\ \\ S_r \end{pmatrix} - J^{-1} \begin{pmatrix} \left[\frac{1}{\lambda_s} \sum_{k=1}^{N} P(s/\underline{x}_k) - N \right]_{\underline{\lambda} = T_r} \\ \\ \left[\sum_{k=1}^{N} P(s/\underline{x}_k) \dfrac{\partial \log d_s(\underline{x}_k, \textcircled{H}_s)}{\partial \theta_{si}} \right]_{\textcircled{H} = S_r} \end{pmatrix}$$

donde $\underline{\lambda} = (\lambda_1, \ldots, \lambda_g)$, $\textcircled{H} = (\theta_{11}, \ldots, \theta_{1q}, \ldots, \theta_{g1}, \ldots, \theta_{gq})$, son los vectores de parámetros, T_r y S_r son las estimaciones r-ésimas para los λ_s y para los θ_s respectivamente. A su vez, el método de puntajes utiliza la inversa de la matriz de información:

$$\underline{I} = \left(\begin{array}{c|c} I_{\lambda\lambda} & I_{\lambda\theta} \\ \hline I_{\theta\lambda} & I_{\theta\theta} \end{array} \right)$$

donde I se encuentra particionada en las submatrices $I_{\lambda\lambda}$, $I_{\lambda\theta}$, $I_{\theta\lambda}$, $I_{\theta\theta}$, que se encuentran definidas por

$$(2.7)\quad I_{\lambda\lambda} = \left\{ E\left(\frac{\partial \log L}{\partial \lambda_s} \frac{\partial \log L}{\partial \lambda_p} \right) \right\} = N \left\{ \frac{1}{\lambda_s \lambda_p} E\left[P(s/\underline{x})\, P(p/\underline{x}) \right] \right\}$$

$$(2.8)\quad I_{\lambda\theta} = \left\{ E\left(\frac{\partial \log L}{\partial \lambda_s} \frac{\partial \log L}{\partial \theta_{pj}} \right) \right\} = N \left\{ \frac{1}{\lambda_s} E\left[P(s/\underline{x}) P(p/\underline{x}) \frac{\partial \log d_p}{\partial \theta_{pj}} \right] \right\}$$

$$(2.9)\qquad I_{\theta\theta} = \left\{ E\left(\frac{\partial \log L}{\partial \theta_{si}} \frac{\partial \log L}{\partial \theta_{pj}} \right) \right\} = N \left\{ E\left[P(s/\underline{x}) P(p/\underline{x}) \frac{\partial \log d_s}{\partial \theta_{si}} \frac{\partial \log d_p}{\partial \theta_{pj}} \right] \right\}$$

La submatriz $I_{\theta\lambda}$ es la transpuesta de $I_{\lambda\theta}$. Si utilizamos la misma notación que en (2.6), las ecuaciones máximo verosímiles pueden ser resueltas iterativamente con el método de puntajes usando las siguientes ecuaciones:

$$(2.10)\qquad \begin{pmatrix} T_{r+1} \\ S_{r+1} \end{pmatrix} = \begin{pmatrix} T_r \\ S_r \end{pmatrix} + \frac{1}{N} I^{-1} \begin{pmatrix} \left[\frac{1}{\lambda_s} \sum_{k=1}^{N} P(s/\underline{x}_k) - N \right]_{\lambda = T_r} \\ \left[\sum_{k=1}^{N} P(s/\underline{x}_k) \frac{\partial \log d_s(\underline{x}_k, \textcircled{H}_s)}{\partial \theta_{si}} \right]_{\textcircled{H} = S_r} \end{pmatrix}$$

Tanto la matriz I como la matriz J se evalúan en (T_1, S_1) que es la estimación inicial de los parámetros. Para que exista convergencia en ambos métodos es necesario que (T_1, S_1) sea un estimador consistente de $(\underline{\lambda}, \textcircled{H})$. En la mayoría de los casos, las esperanzas en la matriz de información contienen integrales imposibles de evaluar en forma cerrada y difícilmente aproximables por series. El enfoque usual ha sido estimar la matriz de información con la muestra reemplazando el símbolo E en (2.7), (2.8) y (2.10) por $\frac{1}{N} \sum_{k=1}^{N}$.

Varios investigadores, al llevar a cabo Análisis de conglomerados (por ejemplo Friedman y Rubin en 1967 [8] y Ball en 1967 [2]), han reportado que los métodos iterativos pueden producir una multiplicidad de soluciones diferentes,

dependiendo de los estimadores iniciales que se utilicen al comenzar el proceso iterativo. Los métodos aquí presentados también tienen esta inconveniencia. Desde un punto de vista puramente matemático, no es sorprendente que un conjunto de ecuaciones no lineales tenga varias raíces diferentes. Sin embargo, puede sorprender el hecho de que surjan soluciones múltiples en el contexto de estimación estadística.

Veamos por medio de un ejemplo, lo que intuitivamente está sucediendo. Supongamos que se tiene una colección heterogénea de personas y los queremos dividir exactamente en dos grupos. Desde luego que existe un gran número de formas en que podemos hacer esto: hombres y mujeres, niños y adultos, liberales y conservadores, ricos y pobres, etc. La división apropiada depende de las variables usadas en el análisis. Si la única variable usada en el análisis es la altura, cualquier método razonable de agrupamiento asignará a la mayoría de los niños en un grupo y a la mayoría de los adultos en el otro.Si la longitud del pelo es usada como única variable de agrupamiento, una buena separación de hombres y mujeres se obtendría para personas adultas, pero quizás esta separación no sería muy buena para niños y jóvenes. Supongamos ahora que en el análisis se incluyen una multitud de variables. Un método que diera iguales pesos o ponderaciones a todas las variables, no sería un buen método. La ventaja de los métodos iterativos en la formación de los

grupos, el que el mismo procedimiento va determinando los pesos aplicados a las variables. En cada iteración, estos pesos se revisan de tal forma que se produzca una mejor discriminación entre grupos. Precisamente esta virtud del método puede originar soluciones múltiples.

Así pues, supongamos que en este ejemplo se hace una partición aleatoria del conjunto de individuos para determinar estimadores iniciales. Supongamos que debido al azar existe una pequeña mayoría de liberales en un grupo y una mayoría de conservadores en el otro. Entonces, las variables relacionadas con la actitud política discriminarán mejor entre los grupos que cualquier otro conjunto de variables. Las actitudes políticas tendrán mayor peso en la siguiente iteración. Los liberales tendrán mayor probabilidad de pertenecer a un grupo que al otro. En la siguiente iteración, esta composición revisada de grupos generará pesos aún mayores para las variables de actitud política. De esta manera, el proceso convergerá a una tipología liberal-conservadora pura.

CAPITULO III

MEZCLAS DE DISTRIBUCIONES NORMALES

Este capítulo constituye la parte más importante de este trabajo. Aquí presentamos la estimación máxima verosímil de los parámetros involucrados en una mezcla, en el caso particular en que las distribuciones que forman la mezcla son normales multivariadas. Así pues, obtenemos los estimadores máximo verosímiles de las medias, varianzas y covarianzas en cada uno de los grupos, así como las proporciones de estos grupos en la mezcla.

El capítulo consta de cuatro secciones: en la primera se estudia el caso general en que no hay restricción alguna sobre las matrices de dispersión en los distintos grupos; en la segunda se supone que estas matrices son iguales para todos los grupos; en la tercera, que las matrices de dispersión son de la forma $\Sigma_s = a_s \Sigma$, donde a_s es un escalar, y Σ es una matriz común; finalmente, en la cuarta sección se estudia el caso en que se tenga la misma matriz de correlación en los distintos grupos.

1. CASO GENERAL.

En este caso tenemos que

$$(3.1) \qquad \alpha_s(\underline{x}, \Theta_s) = \alpha_s(\underline{x}, \underline{\mu}_s, \Sigma_s) = (2\pi)^{-p/2} |\Sigma^s|^{1/2} \exp\left\{-\frac{1}{2}(\underline{x}-\underline{\mu}_s)' \Sigma^s (\underline{x}-\underline{\mu}_s)\right\}$$

donde $\Sigma_s = \{\sigma_s^{ij}\}$ es la matriz $p \times p$ de dispersión de las p variables $x_1, x_2, \ldots, x_p$ en la s-ésima distribución de la mezcla y $\Sigma^s = \{\sigma_{ij}^s\} = \Sigma_s^{-1}$

Para obtener los estimadores máximo verosímiles de $\underline{\mu}_s = (\mu_{s1}, \ldots, \mu_{sp})$ y de $\Sigma_s = \{\sigma_s^{ij}\}$ utilizaremos la ecuación (2.5) y por lo tanto necesitamos calcular explícitamente $\frac{\partial \log \alpha_s}{\partial \mu_{si}}$ y $\frac{\partial \log \alpha_s}{\partial \sigma_{ij}^s}$.

Para ello notemos que

$$\log \alpha_s(\underline{x}, \underline{\mu}_s, \Sigma_s) = -\frac{p}{2}\log 2\pi + \frac{1}{2}\log|\Sigma^s| - \frac{1}{2}(\underline{x}-\underline{\mu}_s)'\Sigma^s(\underline{x}-\underline{\mu}_s)$$

de donde

$$\frac{\partial \log \alpha_s(\underline{x}_k, \underline{\mu}_s, \Sigma_s)}{\partial \mu_{si}} = -\frac{1}{2}\frac{\partial}{\partial \mu_{si}}\left[(\underline{x}_k - \underline{\mu}_s)'\Sigma^s(\underline{x}_k - \underline{\mu}_s)\right]$$

Es decir, tenemos que derivar la forma cuadrática

$$(\underline{x}_k - \underline{\mu}_s)'\Sigma^s(\underline{x}_k - \underline{\mu}_s)$$

Pero,

$$\frac{\partial}{\partial \underline{\mu}_s}(\underline{x}_k - \underline{\mu}_s)'\Sigma^s(\underline{x}_k - \underline{\mu}_s) = -2\Sigma^s(\underline{x}_k - \underline{\mu}_s),$$

de donde

$$\frac{\partial}{\partial \mu_{si}} (\underline{x}_k - \mu_s)' \Sigma^s (\underline{x}_k - \mu_s) = -2 \sum_{j=1}^{p} \nabla_{ij}^s (x_{jk} - \mu_{sj})$$

Entonces, obtenemos finalmente que

$$\frac{\partial \log \alpha_s (\underline{x}_k, \mu_s, \Sigma_s)}{\partial \mu_{si}} = \sum_{j=1}^{p} \nabla_{ij}^s (x_{jk} - \mu_{sj}). \tag{3.2}$$

Ahora obtendremos $\dfrac{\partial \log \alpha_s (\underline{x}_k, \mu_s, \Sigma_s)}{\partial \nabla_{ij}^s}$

Si $i=j$, entonces

$$\frac{\partial \log \alpha_s (\underline{x}_k, \mu_s, \Sigma_s)}{\partial \nabla_{ii}^s} = \frac{1}{2} \frac{\partial}{\partial \nabla_{ii}^s} \log |\Sigma^s| - \frac{1}{2} \frac{\partial}{\partial \nabla_{ii}^s} (\underline{x}_k - \mu_s)' \Sigma^s (\underline{x}_k - \mu_s)$$

Ahora bien, como

$$|\Sigma^s| = \nabla_{ii}^s [\text{cofactor de } \nabla_{ii}^s] + [\text{términos sin } \nabla_{ii}^s]$$

y

$$(\underline{x}_k - \mu_s)' \Sigma^s (\underline{x}_k - \mu_s) = (x_{ki} - \mu_{si}) \nabla_{ii}^s + [\text{términos sin } \nabla_{ii}^s]$$

entonces

$$\frac{\partial \log \alpha_s (\underline{x}_k, \mu_s, \Sigma_s)}{\partial \nabla_{ii}^s} = \frac{1}{2} \left\{ \frac{[\text{cofactor de } \nabla_{ii}^s]}{|\Sigma^s|} - (x_{ik} - \mu_{si})^2 \right\}$$

Pero,

$$\frac{[\text{cofactor de } \nabla_{ii}^s]}{|\Sigma^s|} = \nabla_s^{ii}$$

Por lo tanto

$$\frac{\partial \log \alpha_s(\underline{x}_k, \mu_s, \Sigma_s)}{\partial \sigma_{ij}^s} = \frac{1}{2}\left\{ \sigma_s^{ii} - (x_{ik} - \mu_{si})^2 \right\}.$$

Si $i \neq j$, entonces

$$\frac{\partial \log \alpha_s(\underline{x}_k, \mu_s, \Sigma_s)}{\partial \sigma_{ij}^s} = \frac{1}{2}\frac{\partial}{\partial \sigma_{ij}^s} \log |\Sigma^s| - \frac{1}{2}\frac{\partial}{\partial \sigma_{ij}^s} (\underline{x}_k - \mu_s)' \Sigma^s (\underline{x}_k - \mu_s)$$

Dado que

$$|\Sigma^s| = 2\,\sigma_{ij}^s [\text{cofactor de } \sigma_{ij}^s] + [\text{términos sin } \sigma_{ij}^s],$$

y

$$(\underline{x}_k - \mu_s)' \Sigma^s (\underline{x}_k - \mu_s) = 2(x_{ik} - \mu_{si})\sigma_{ij}^s (x_{jk} - \mu_{sj}) + [\text{términos sin } \sigma_{ij}^s],$$

tenemos que

$$\frac{\partial \log \alpha_s(\underline{x}_k, \mu_s, \Sigma_s)}{\partial \sigma_{ij}^s} = \frac{[\text{cofactor de } \sigma_{ij}^s]}{|\Sigma^s|} - (x_{ik} - \mu_{si})(x_{jk} - \mu_{sj})$$

y nuevamente, usando que

$$\frac{[\text{cofactor de } \sigma_{ij}^s]}{|\Sigma^s|} = \sigma_s^{ij},$$

obtenemos

$$\frac{\partial \log \alpha_s(\underline{x}_k, \mu_s, \Sigma_s)}{\partial \sigma_{ij}^s} = \sigma_s^{ij} - (x_{ik} - \mu_{si})(x_{jk} - \mu_{sj})$$

Podemos juntar ambos casos ($i = j$, $i \neq j$) para escribir una sola expresión:

$$[3.3] \quad \frac{\partial \log \alpha_s(\underline{x}_k, \mu_s, \Sigma_s)}{\partial \sigma_{ij}^s} = \left(1 - \frac{\delta_{ij}}{2}\right)\left\{ \sigma_s^{ij} - (x_{ik} - \mu_{si})(x_{jk} - \mu_{sj}) \right\}$$

en donde $\delta_{ij} = \begin{cases} 1 & si \quad i = j \\ 0 & si \quad i \neq j \end{cases}$

Podemos usar toda esta información para obtener el siguiente resultado.

Teorema 1. Los estimadores máximo verosímiles de los parámetros de una mezcla de distribuciones normales multivariadas con distintas matrices de dispersión están dadas por

(3.4) $$\hat{\lambda}_s = \frac{1}{N} \sum_{k=1}^{N} \hat{P}(s/\underline{x}_k)$$

(3.5) $$\hat{\mu}_{si} = \frac{1}{N\hat{\lambda}_s} \sum_{k=1}^{N} \hat{P}(s/\underline{x}_k)\, x_{ik}$$

(3.6) $$\hat{\sigma}_{s\,ij} = \frac{1}{N\hat{\lambda}_s} \sum_{k=1}^{N} \hat{P}(s/\underline{x}_k)(x_{ik}-\hat{\mu}_{si})(x_{jk}-\hat{\mu}_{sj})$$

Demostración. Si sustituimos (3.2) en (2.5) obtenemos que

$$\sum_{k=1}^{N} \hat{P}(s/\underline{x}_k) \sum_{j=1}^{p} \sigma_{ij}^{s}(x_{jk}-\hat{\mu}_{sj}) = 0$$

Entonces

$$\sum_{j=1}^{p} \hat{\sigma}_{ij}^{s} \left[\sum_{k=1}^{N} \hat{P}(s/\underline{x}_k)(x_{jk}-\hat{\mu}_{sj}) \right] =$$

$$= \sum_{j=1}^{p} \hat{\sigma}_{ij}^{s} \left[\sum_{k=1}^{N} \hat{P}(s/\underline{x}_k)\, x_{jk} - \hat{\mu}_{sj} \sum_{k=1}^{N} \hat{P}(s/\underline{x}_k) \right] =$$

$$= \sum_{j=1}^{p} \hat{\sigma}_{ij}^{s} \left[\sum_{k=1}^{N} \hat{P}(s/\underline{x}_k)\, x_{jk} - \hat{\mu}_{sj} N \hat{\lambda}_s \right] = 0.$$

Por lo tanto

$$\sum_{k=1}^{N} \hat{P}(s/\underline{x}_k)\, x_{jk} - \hat{\mu}_{sj}\, N\, \hat{\lambda}_s = 0,$$

de donde

$$\hat{\mu}_{sj} = \frac{1}{N\hat{\lambda}_s} \sum_{k=1}^{N} \hat{P}(s/\underline{x}_k)\, x_{jk},$$

que es la ecuación (3.5).

Para demostrar (3.6) sustituyamos (3.3) en (2.5), con lo que obtenemos que

$$\sum_{k=1}^{N} \hat{P}(s/\underline{x}_k)\left(1 - \frac{\delta_{ij}}{2}\right)\left[\hat{V}_s^{ij} - (x_{ik} - \hat{\mu}_{si})(x_{jk} - \hat{\mu}_{sj})\right] = 0.$$

Entonces

$$\left(1 - \frac{\delta_{ij}}{2}\right) \sum_{k=1}^{N} \hat{P}(s/\underline{x}_k)\left[\hat{V}_s^{ij} - (x_{ik} - \mu_{si})(x_{jk} - \mu_{sj})\right] = 0.$$

De donde

$$\hat{V}_s^{ij} \sum_{k=1}^{N} \hat{P}(s/\underline{x}_k) = \sum_{k=1}^{N} \hat{P}(s/\underline{x}_k)(x_{ik} - \hat{\mu}_{si})(x_{jk} - \hat{\mu}_{sj}),$$

y despejando $\hat{V}_s^{ij}$ obtenemos finalmente que

$$\hat{V}_s^{ij} = \frac{1}{N\hat{\lambda}_s} \sum_{k=1}^{N} \hat{P}(s/\underline{x}_k)(x_{ik} - \hat{\mu}_{si})(x_{jk} - \hat{\mu}_{sj}).$$

La ecuación (3.4) es simplemente (2.4). □

Tal y como lo indicamos en el capítulo anterior, las ecuaciones (3.4) a (3.6) no están en forma cerrada, pues no se conocen las probabilidades de pertenencia a los distintos grupos para cada uno de los objetos o individuos en la muestra. En

el capítulo anterior presentamos algunos métodos iterativos de resolución de estas ecuaciones e indicamos que las soluciones dependían del agrupamiento inicial. Usualmente, este agrupamiento inicial se hace con técnicas jerárquicas de análisis de conglomerados o bien minimizando la traza o el determinante de la matriz W, que es la matriz de dispersión dentro de grupos. Esto se verá con mayor detalle en el capítulo V. Cuando ya se tienen los valores numéricos de los parámetros de la mezcla (después del proceso iterativo) se asigna a cada objeto o individuo al grupo al que tenga mayor probabilidad de pertenencia.

Finalmente, notemos que en el caso general de mezclas normales (distintas matrices de dispersión) se tienen que estimar $(g-1) + gp + \frac{1}{2} gp(p+1)$ parámetros (el primer término corresponde a las λ'_s , el segundo a las μ'_s y el tercero a las Σ'_s). Si g y p son relativamente grandes, se necesita un tamaño de muestra N muy grande para poder así estimar tantos parámetros -en particular, el número de individuos dentro de cada grupo debe ser suficientemente grande para obtener estimaciones satisfactorias de las matrices de dispersión. Debido a esto, resulta conveniente hacer algunas simplificaciones a las distribuciones componentes de la mezcla, principalmente a las matrices de dispersión de las distribuciones. Estas simplificaciones dependen de la naturaleza del problema que se tenga. Así pues, si se supone que las matrices de dis-

persión son iguales en los distintos grupos, se tienen que estimar $(g-1)+gp+\frac{1}{2}p(p+1)$ parámetros, lográndose una reducción de $\frac{1}{2}(g-1)p(p+1)$ parámetros. Otra posibilidad es suponer matrices de correlación iguales. En este caso se tienen que estimar $(g-1)+gp+\frac{1}{2}p(p-1)+gp$ parámetros, es decir, $\frac{1}{2}p(p-1)(g-1)$ parámetros menos que estimar. Aún otra posibilidad es suponer que las matrices de dispersión en los distintos grupos son iguales excepto por una constante multiplicativa, i.e., $\Sigma_s = a_s \Sigma$ $s = 1, 2, \ldots, g$. En este caso se tienen que estimar $(g-1)+gp+\frac{1}{2}p(p+1)+g$ parámetros (el cuarto término corresponde a las a_s's). Estas simplificaciones, aparte de reducir el número de parámetros por estimar, hacen que el problema de agrupación sea computacionalmente más simple. Desde luego, las suposiciones deben ser realistas, es decir, deben corresponder a la naturaleza del problema; de lo contrario las estimaciones no serán buenas.

A continuación presentamos las estimaciones máximo verosímiles para cada una de estas simplificaciones.

2. MATRICES DE DISPERSION IGUALES.

En este caso suponemos que $\Sigma_s = \Sigma = \{\sigma^{ij}\}$.

La ecuación (2.5) no es válida, pues fue derivada bajo la suposición de que los parámetros de distintos grupos no estaban relacionados funcionalmente. La ecuación de verosimilitud correcta es

$$\frac{\partial L}{\partial \nabla_{ij}} = \sum_{s=1}^{g} \frac{\partial L}{\partial \nabla_{ij}^{s}} \frac{\partial \nabla_{ij}^{s}}{\partial \nabla_{ij}} .$$

De hecho tenemos el siguiente resultado.

Teorema 2. Los estimadores máximo verosímiles de los parámetros de una mezcla de distribuciones normales multivariadas con matrices de dispersión iguales están dadas por

$$\hat{\lambda}_s = \frac{1}{N} \sum_{k=1}^{N} \hat{P}(s/\underline{x}_k) \tag{3.7}$$

$$\hat{\mu}_{si} = \frac{1}{N\hat{\lambda}_s} \sum_{k=1}^{N} \hat{P}(s/\underline{x}_k) x_{ik} \tag{3.8}$$

$$\hat{\nabla}_{ij} = \frac{1}{N} \sum_{k=1}^{N} x_{ik} x_{jk} - \sum_{s=1}^{g} \hat{\lambda}_s \hat{\mu}_{si} \hat{\mu}_{sj} . \tag{3.9}$$

Demostración. La ecuación (3.7) es simplemente (2.4). La ecuación (3.8) es clara si se observa la demostración de la ecuación (3.5) del Teorema 1 de este capítulo. Para demostrar (3.9) primero obtendremos $\frac{\partial L}{\partial \nabla_{ij}}$. Para esto, ya hicimos notar que

$$\frac{\partial L}{\partial \nabla_{ij}} = \sum_{s=1}^{g} \frac{\partial L}{\partial \nabla_{ij}^{s}} \frac{\partial \nabla_{ij}^{s}}{\partial \nabla_{ij}}$$

y dado que

$$\frac{\partial \nabla_{ij}^{s}}{\partial \nabla_{ij}} = 1$$

tenemos que

$$= (1-\delta_{ij}/2)\left[\sum_{s=1}^{g}\sum_{k=1}^{N} \frac{\lambda_s \alpha_s}{f(\underline{x}_k)} \nabla_s^{ij} - \sum_{s=1}^{g}\sum_{k=1}^{N} \frac{\lambda_s \alpha_s}{f(\underline{x}_k)} (x_{ik}-\mu_{si})(x_{jk}-\mu_{sj})\right]$$

$$= (1-\delta_{ij}/2)\left[N\nabla^{ij} - \sum_{s=1}^{g}\sum_{k=1}^{N} \frac{\lambda_s \alpha_s(\underline{x}_k, \mu_s, \Sigma_s)}{f(\underline{x}_k)} (x_{ik}-\mu_{si})(x_{jk}-\mu_{sj})\right]$$

Al igualar esto a cero obtenemos que

$$N\hat{\nabla}^{ij} - \sum_{s=1}^{g}\sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s(\underline{x}_k, \mu_s, \Sigma_s)}{f(\underline{x}_k)} (x_{ik}-\hat{\mu}_{si})(x_{jk}-\hat{\mu}_{sj}) = 0$$

Es decir,

$$N\hat{\nabla}^{ij} = \sum_{s=1}^{g}\sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s(\underline{x}_k, \mu_s, \Sigma_s)}{f(\underline{x}_k)} [x_{ik}x_{jk} - x_{ik}\hat{\mu}_{sj} - x_{jk}\hat{\mu}_{si} + \hat{\mu}_{si}\hat{\mu}_{sj}]$$

$$= \sum_{s=1}^{g}\sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s(\underline{x}_k, \mu_s, \Sigma_s)}{f(\underline{x}_k)} x_{ik}x_{jk} + \sum_{s=1}^{g}\sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s(\underline{x}_k, \mu_s, \Sigma_s)}{f(\underline{x}_k)} \hat{\mu}_{si}\hat{\mu}_{sj}$$

$$- \sum_{s=1}^{g}\sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s(\underline{x}_k, \mu_s, \Sigma_s)}{f(\underline{x}_k)} (x_{ik}\hat{\mu}_{sj} + x_{jk}\hat{\mu}_{si})$$

Ahora bien,

$$\sum_{s=1}^{g}\sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s(\underline{x}_k, \mu_s, \Sigma_s)}{f(\underline{x}_k)} \hat{\mu}_{si}\hat{\mu}_{sj} = \sum_{s=1}^{g} \hat{\lambda}_s \hat{\mu}_{si}\hat{\mu}_{sj} \sum_{k=1}^{N} \frac{\alpha_s(\underline{x}_k, \mu_s, \Sigma_s)}{f(\underline{x}_k)}$$

$$= \sum_{s=1}^{g} \hat{\lambda}_s \hat{\mu}_{si}\hat{\mu}_{sj} \left[\frac{1}{\hat{\lambda}_s}\sum_{k=1}^{N} \hat{P}(s|\underline{x}_k)\right] = N\sum_{s=1}^{g} \hat{\lambda}_s \hat{\mu}_{si}\hat{\mu}_{sj}.$$

$$\frac{\partial L}{\partial \bar{\sigma}_{ij}} = \sum_{s=1}^{g} \frac{\partial}{\partial \bar{\sigma}_{ij}^{s}} \left[\sum_{k=1}^{N} \log f(\underline{x}_k) \right] = \sum_{s=1}^{g} \sum_{k=1}^{N} \frac{\frac{\partial}{\partial \bar{\sigma}_{ij}^{s}} f(\underline{x}_k)}{f(\underline{x}_k)} .$$

Así pues, tenemos que encontrar $\frac{\partial}{\partial \bar{\sigma}_{ij}^{s}} f(\underline{x}_k)$:

$$\frac{\partial}{\partial \bar{\sigma}_{ij}^{s}} f(\underline{x}_k) = \frac{\partial}{\partial \bar{\sigma}_{ij}^{s}} \sum_{s=1}^{g} \lambda_s \alpha_s (\underline{x}_k, \mu_s, \Sigma_s) =$$

$$= \frac{\partial}{\partial \bar{\sigma}_{ij}^{s}} \lambda_s \alpha_s (\underline{x}_k, \mu_s, \Sigma_s) =$$

$$= \frac{\partial}{\partial \bar{\sigma}_{ij}^{s}} \left[e^{\log \lambda_s \alpha_s (\underline{x}_k, \mu_s, \Sigma_s)} \right] =$$

$$= \lambda_s \alpha_s (\underline{x}_k, \mu_s, \Sigma_s) \frac{\partial}{\partial \bar{\sigma}_{ij}^{s}} \log \left[\lambda_s \alpha_s (\underline{x}_k, \mu_s, \Sigma_s) \right] =$$

$$= \lambda_s \alpha_s (\underline{x}_k, \mu_s, \Sigma_s) \frac{\partial}{\partial \bar{\sigma}_{ij}^{s}} \log \alpha_s (\underline{x}_k, \mu_s, \Sigma_s) .$$

y usando la expresión (3.3) obtenemos que

$$\frac{\partial}{\partial \bar{\sigma}_{ij}^{s}} f(\underline{x}_k) = \lambda_s \alpha_s (\underline{x}_k, \mu_s, \Sigma_s) \left(1 - \frac{\delta_{ij}}{2}\right) \left[\bar{V}_s^{ij} - (x_{ik} - \mu_{si})(x_{jk} - \mu_{sj}) \right].$$

Por lo tanto

$$\frac{\partial L}{\partial \bar{\sigma}_{ij}} = \sum_{s=1}^{g} \sum_{k=1}^{N} \frac{\left(1 - \frac{\delta_{ij}}{2}\right) \lambda_s \alpha_s (\underline{x}_k, \mu_s, \Sigma_s)}{f(\underline{x}_k)} \left[\bar{V}_s^{ij} - (x_{ik} - \mu_{si})(x_{jk} - \mu_{sj}) \right] =$$

Y también,

$$\sum_{s=1}^{g}\sum_{k=1}^{N} \frac{\hat{\lambda}_s\, d_s(\underline{x}_k, \hat{\mu}_s, \Sigma_s)}{f(\underline{x}_k)} (x_{ik}\hat{\mu}_{sj} + x_{jk}\hat{\mu}_{si})$$

$$= \sum_{s=1}^{g}\sum_{k=1}^{N} \left[\hat{P}(s|\underline{x}_k)\, x_{ik}\hat{\mu}_{sj} + P(s|\underline{x}_k)\, x_{jk}\hat{\mu}_{si} \right]$$

$$= \sum_{s=1}^{g} \left[\hat{\mu}_{sj} \sum_{k=1}^{N} \hat{P}(s|\underline{x}_k)\, x_{ik} + \hat{\mu}_{si} \sum_{k=1}^{N} \hat{P}(s|\underline{x}_k)\, x_{jk} \right]$$

$$= \sum_{s=1}^{g} \left[\hat{\mu}_{sj}\hat{\mu}_{si} N \hat{\lambda}_s + \hat{\mu}_{si}\hat{\mu}_{sj} N \hat{\lambda}_s \right]$$

$$= 2N \sum_{s=1}^{g} \hat{\lambda}_s \hat{\mu}_{si}\hat{\mu}_{sj}.$$

Por lo tanto

$$N\hat{V}_{ij} = \sum_{s=1}^{g}\sum_{k=1}^{N} \frac{\lambda_s d_s(\underline{x}_k, \mu_s, \Sigma_s)}{f(\underline{x}_k)} x_{ik}x_{jk} - N\sum_{s=1}^{g} \hat{\lambda}_s\hat{\mu}_{si}\hat{\mu}_{sj}$$

$$= \sum_{k=1}^{N} x_{ik}x_{jk} - N\sum_{s=1}^{g} \hat{\lambda}_s \hat{\mu}_{si}\hat{\mu}_{sj}.$$

Es decir

$$\hat{V}_{ij} = \frac{1}{N}\sum_{k=1}^{N} x_{ik}x_{jk} - \sum_{s=1}^{g} \hat{\lambda}_s\hat{\mu}_{si}\hat{\mu}_{sj},$$

que es la ecuación (3.9). □

3. CASO $\Sigma_s = a_s \Sigma$

En ciertas ocasiones, el suponer que las matrices de dispersión son iguales para todos los grupos, puede no ser una suposición muy adecuada. Una suposición un poco más general y con muchos parámetros menos por estimar que en el caso general consiste en pensar que las matrices de dispersión son iguales excepto por una constante multiplicativa, i.e., $\Sigma_s = a_s \Sigma$ donde a_s es un número real positivo y Σ es una matriz positiva definida de tamaño $p \times p$, $s = 1, \cdots, g$

En este caso tenemos que

$$\phi_s(\underset{\sim}{x}, \mu_s, \Sigma_s) = \phi_s(\underset{\sim}{x}, \mu_s, a_s \Sigma) = (2\pi)^{-p/2} |\Sigma^s|^{1/2} \exp\left\{-\frac{1}{2}(\underset{\sim}{x} - \mu_s)' \Sigma^s (\underset{\sim}{x} - \mu_s)\right\}$$

donde

$$\Sigma_s = \{\sigma_s^{ij}\} = a_s \Sigma = \{a_s \sigma^{ij}\}$$

y

$$\Sigma^s = \{\sigma_{ij}^s\} = \{\Sigma_s\}^{-1} = a^s \Sigma^{-1} = \{a^s \sigma_{ij}\}$$

con

$$a^s = (a_s)^{-1}, \quad \Sigma = \{\sigma^{ij}\}, \quad \Sigma^{-1} = \{\sigma_{ij}\}$$

Tenemos que encontrar los estimadores máximo verosímiles de λ_s, μ_{si}, σ^{ij} y a_s, $s = 1, 2, \ldots, g$.
El siguiente teorema muestra estos resultados.

Teorema 3. Los estimadores máximo verosímiles de los parámetros de una mezcla de distribuciones normales multivariadas con matrices de dispersión de la forma $\Sigma_s = a_s \Sigma$, $s = 1, 2, \ldots, g$, están dados por

$$\hat{\lambda}_s = \frac{1}{N} \sum_{k=1}^{N} \hat{P}(s/\underline{x}_k) \tag{3.10}$$

$$\hat{\mu}_{si} = \frac{1}{N\hat{\lambda}_s} \sum_{k=1}^{N} \hat{P}(s/\underline{x}_k)\, x_{ik} \tag{3.11}$$

$$\hat{\sigma}^{ij} = \frac{1}{N} \sum_{s=1}^{g} \sum_{k=1}^{N} \frac{\hat{P}(s/\underline{x}_k)\, x_{ik} x_{jk}}{\hat{a}_s} - \sum_{s=1}^{g} \frac{\hat{\lambda}_s \hat{\mu}_{si} \hat{\mu}_{sj}}{\hat{a}_s} \tag{3.12}$$

$$\hat{a}_s = \frac{1}{pN\hat{\lambda}_s} \sum_{k=1}^{N} \hat{P}(s/\underline{x}_k)(\underline{x}_k - \hat{\mu}_s)' \hat{\Sigma}^{-1} (\underline{x}_k - \hat{\mu}_s). \tag{3.13}$$

Demostración. Es obvio que las ecuaciones (3.10) y (3.11) siguen siendo válidas en este caso. Demostraremos ahora la ecuación (3.12). La ecuación de verosimilitud en este caso es

$$\frac{\partial L}{\partial \sigma^{ij}} = \sum_{s=1}^{g} \frac{\partial L}{\partial \sigma_s^{ij}} \frac{\partial \sigma_s^{ij}}{\partial \sigma^{ij}} = 0$$

Ya tenemos en el caso general que

$$\frac{\partial L}{\partial \sigma_s^{ij}} = \sum_{k=1}^{N} \frac{\lambda_s f_s(\underline{x}_k, \mu_s, \Sigma_s)}{f(\underline{x}_k)} \left(1 - \frac{\delta_{ij}}{2}\right) \left[V_s^{ij} - (x_{ik} - \mu_{si})(x_{jk} - \mu_{sj}) \right],$$

Y además en este caso

$$\frac{\partial \sigma_s^{ij}}{\partial \sigma^{ij}} = \frac{\partial}{\partial \sigma^{ij}} \left(a^s \sigma^{ij} \right) = a^s.$$

Por lo tanto

$$\frac{\partial L}{\partial \sigma_{ij}} = \sum_{s=1}^{g} \sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s}{f(\underline{x}_k)} \left(1 - \frac{\delta_{ij}}{2}\right) \left[-\hat{\Gamma}_s^{ij} - (x_{ik} - \hat{\mu}_{si})(x_{jk} - \hat{\mu}_{sj}) \right] \hat{a}^s = 0$$

y como $\Gamma_s^{ij} = a_s \Gamma^{ij}$, tenemos

$$\left(1 - \frac{\delta_{ij}}{2}\right) \sum_{s=1}^{g} \sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s}{f(\underline{x}_k)} \left[\hat{a}_s \hat{\Gamma}^{ij} - (x_{ik} - \hat{\mu}_{si})(x_{jk} - \hat{\mu}_{sj}) \right] \hat{a}^s = 0$$

i.e.,

$$\sum_{s=1}^{g} \sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s}{f(\underline{x}_k)} \left[-\hat{\Gamma}^{ij} - \hat{a}^s (x_{ik} - \hat{\mu}_{si})(x_{jk} - \hat{\mu}_{sj}) \right] = 0.$$

Entonces

$$\hat{\Gamma}^{ij} \sum_{s=1}^{g} \sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s}{f(\underline{x}_k)} = \sum_{s=1}^{g} \sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s}{f(\underline{x}_k)} \hat{a}^s (x_{ik} - \hat{\mu}_{si})(x_{jk} - \hat{\mu}_{sj})$$

$$= \sum_{s=1}^{g} \sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s}{f(\underline{x}_k)} \hat{a}^s x_{ik} x_{jk} - \sum_{s=1}^{g} \sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s \hat{a}^s}{f(\underline{x}_k)} (x_{ik} \hat{\mu}_{sj} + x_{jk} \hat{\mu}_{si})$$

$$+ \sum_{s=1}^{g} \sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s}{f(\underline{x}_k)} \hat{a}^s \hat{\mu}_{si} \hat{\mu}_{sj}$$

$$= \sum_{s=1}^{g} \sum_{k=1}^{N} \hat{a}^s \hat{p}(s/\underline{x}_k) x_{ik} x_{jk} - \sum_{s=1}^{g} \hat{a}^s \left[\sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s}{f(\underline{x}_k)} (x_{ik} \hat{\mu}_{sj} + x_{jk} \hat{\mu}_{si}) \right]$$

$$+ \sum_{s=1}^{g} \hat{a}^s \hat{\mu}_{si} \hat{\mu}_{sj} \sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s(\underline{x}_k, \mu_s, \Sigma_s)}{f(\underline{x}_k)}$$

$$= \sum_{s=1}^{g} \sum_{k=1}^{N} \hat{a}^s \hat{p}(s/\underline{x}_k)(x_{ik} x_{jk}) - \sum_{s=1}^{g} \hat{a}^s \left[\hat{\mu}_{sj} \sum_{k=1}^{N} \hat{p}(s/\underline{x}_k) x_{ik} + \hat{\mu}_{si} \sum_{k=1}^{N} \hat{p}(s/\underline{x}_k) x_{jk} \right]$$

$$+ \sum_{s=1}^{g} \hat{a}^s \hat{\mu}_{si} \hat{\mu}_{sj} N \hat{\lambda}_s =$$

$$= \sum_{s=1}^{g} \sum_{k=1}^{N} \hat{a}^s \hat{P}(s/\underline{x}_k) x_{ik} x_{jk} - \sum_{s=1}^{g} \hat{a}^s \left[\hat{\mu}_{sj} N \hat{\lambda}_s \hat{\mu}_{si} + \hat{\mu}_{si} N \hat{\lambda}_s \hat{\mu}_{sj} \right]$$

$$+ N \sum_{s=1}^{g} \hat{\lambda}_s \hat{a}^s \hat{\mu}_{si} \hat{\mu}_{sj}$$

$$= \sum_{s=1}^{g} \sum_{k=1}^{N} \hat{a}^s \hat{P}(s/\underline{x}_k) x_{ik} x_{jk} - 2N \sum_{s=1}^{g} \hat{a}^s \hat{\lambda}_s \hat{\mu}_{si} \hat{\mu}_{sj} + N \sum_{s=1}^{g} \hat{a}^s \hat{\lambda}_s \hat{\mu}_{si} \hat{\mu}_{sj}$$

$$= \sum_{s=1}^{g} \sum_{k=1}^{N} \hat{a}^s \hat{P}(s/\underline{x}_k) x_{ik} x_{jk} - N \sum_{s=1}^{g} \hat{a}^s \hat{\lambda}_s \hat{\mu}_{si} \hat{\mu}_{sj}$$

De donde obtenemos finalmente que

$$\hat{T}^{ij} \cdot N = \sum_{s=1}^{g} \sum_{k=1}^{N} \hat{a}^s \hat{P}(s/\underline{x}_k) x_{ik} x_{jk} - N \sum_{s=1}^{g} \hat{a}^s \hat{\lambda}_s \hat{\mu}_{si} \hat{\mu}_{sj}.$$

Al despejar de esta ecuación obtenemos (3.12).

Finalmente, demostremos la ecuación. (3.13). Para ello:

$$\frac{\partial L}{\partial a^s} = \frac{\partial}{\partial a^s} \left[\sum_{k=1}^{N} \log \sum_{s=1}^{g} \lambda_s \alpha_s (\underline{x}_k, \mu_s, \Sigma_s) \right]$$

$$= \sum_{k=1}^{N} \frac{\partial}{\partial a^s} \log \sum_{s=1}^{g} \lambda_s \alpha_s (\underline{x}_k, \mu_s, \Sigma_s)$$

$$= \sum_{k=1}^{N} \lambda_s \frac{\frac{\partial}{\partial a_s} \alpha_s (\underline{x}_k, \mu_s, \Sigma_s)}{f(\underline{x}_k)}.$$

Ahora bien,

$$\frac{\partial}{\partial a^s}\alpha_s = \frac{\partial}{\partial a^s}\exp(\log\alpha_s) = \alpha_s \frac{\partial}{\partial a^s}\log\alpha_s =$$

$$= \alpha_s \frac{\partial}{\partial a^s}\left[\frac{1}{2}\log|a^s\Sigma^{-1}| - \frac{1}{2}(\underline{x}_k-\mu_s)' a^s \Sigma^{-1}(\underline{x}_k-\mu_s)\right]$$

$$= \alpha_s\left[\frac{1}{2}\frac{\partial}{\partial a^s}\log\left((a^s)^p|\Sigma^{-1}|\right) - \frac{1}{2}\frac{\partial}{\partial a^s}\left(a^s(\underline{x}_k-\mu_s)'\Sigma^{-1}(\underline{x}_k-\mu_s)\right)\right]$$

$$= \alpha_s\left[\frac{1}{2}\frac{p(a^s)^{p-1}|\Sigma^{-1}|}{(a^s)^p|\Sigma^{-1}|} - \frac{1}{2}(\underline{x}_k-\mu_s)\Sigma^{-1}(\underline{x}_k-\mu_s)\right]$$

$$= \frac{1}{2}\alpha_s(\underline{x}_k,\mu_s,\Sigma_s)\left[\frac{p}{a^s} - (\underline{x}_k-\mu_s)\Sigma^{-1}(\underline{x}_k-\mu_s)\right]$$

Por lo tanto

$$\frac{\partial L}{\partial a^s} = \frac{1}{2}\sum_{k=1}^{N}\frac{\hat{\lambda}_s\alpha_s}{f(\underline{x}_k)}\left[p\hat{a}_s - (\underline{x}_k-\hat{\mu}_s)'\hat{\Sigma}^{-1}(\underline{x}_k-\hat{\mu}_s)\right] = 0$$

De donde

$$\sum_{k=1}^{N}\hat{P}(s|\underline{x}_k)\,p\,\hat{a}_s = \sum_{k=1}^{N}\hat{P}(s|\underline{x}_k)(\underline{x}_k-\hat{\mu}_s)'\hat{\Sigma}^{-1}(\underline{x}_k-\hat{\mu}_s).$$

Despejando $\hat{a}_s$ obtenemos finalmente

$$\hat{a}_s = \frac{1}{pN\hat{\lambda}_s}\sum_{k=1}^{N}\hat{P}(s|\underline{x}_k)(\underline{x}_k-\hat{\mu}_s)'\hat{\Sigma}^{-1}(\underline{x}_k-\hat{\mu}_s)$$

que es la ecuación (3.13). □

Notemos que si $Q_s = 1$, $s = 1, 2, \ldots, g$, la ecuación (3.12) es la ecuación (3.9), que es el caso en que se tienen matrices de dispersión iguales.

Observemos también que la estimación máximo verosímil de los Q_s' está en términos de la distancia de cada objeto a la media del grupo en cuestión con la inversa de la matriz estimada de dispersión como métrica, esto es, la distancia de Mahalanobis, y de las probabilidades de pertenencia al grupo en cuestión para cada uno de los objetos.

Finalmente notemos que si los grupos estuvieran muy separados tendríamos que

$$\hat{Q}_s = \frac{1}{pN\hat{\lambda}_s} \sum_{j \in S'} (x_j - \hat{\mu}_s)\, \hat{\Sigma}^{-1} (x_j - \hat{\mu}_s)$$

donde

$$S' = \{ i : x_i \text{ está en el grupo } s \}$$

4. MATRICES DE CORRELACION IGUALES.

Existen situaciones en donde la suposición de iguales matrices de dispersión en los grupos no es válida, porque las variables en cada grupo tienen distintas varianzas, pero podemos suponer que tienen la misma matriz de correlación; por ejemplo, en Taxonomía nos podemos encontrar con una situación donde al clasificar diferentes especies animales de acuerdo a un conjunto de medidas sobre varios órganos del

cuerpo, estas medidas estén muy correlacionadas en cada grupo pero difieren en medias y varianzas para distintos grupos; o en Medicina, tratando de agrupar pacientes de acuerdo a un conjunto de variables como temperatura, presión, etc., las variables en distintos grupos pueden tener las mismas correlaciones pero diferentes medias y varianzas. Tal y como ya indicamos, en estas situaciones tenemos $(g-1)p(p-1)/2$ parámetros menos por estimar que en el caso general, lo cual es una gran ventaja.

Desde luego que los estimadores máximo verosímiles de los λ_s' y de los μ_s' siguen siendo los mismos que en el caso general. Los estimadores de las varianzas se obtienen de la ecuación (3.6) y están dados por

$$\hat{v}_s^{ii} = \frac{1}{N \hat{\lambda}_s} \sum_{k=1}^{N} \hat{P}(s/\underline{x}_k)(x_{ik} - \hat{\mu}_{si})^2.$$

Ahora bien, como estamos suponiendo que los distintos grupos tienen una matriz de correlación común $C = \{c^{ij}\}$ entonces para la población s, tenemos que la matriz de dispersión se puede escribir de la manera siguiente:

$$\Sigma_s = D_s C D_s$$

donde $C = \{c^{ij}\}$ y D_s es la matriz diagonal dada por

$$D_s = \begin{bmatrix} d_s^1 & & 0 \\ & d_s^2 & \\ 0 & & \ddots \\ & & & d_s^p \end{bmatrix} = \begin{bmatrix} \sqrt{v_s^{11}} & & 0 \\ & \sqrt{v_s^{22}} & \\ 0 & & \ddots \\ & & & \sqrt{v_s^{pp}} \end{bmatrix}$$

Si $\{C_{ij}\} = \{C^{ij}\}^{-1}$ y $\{\alpha_i^s\} = \{\alpha_s^i\}^{-1}$, podemos expresar a $f(\underline{x})$ de la manera siguiente

$$f(\underline{x}) = \sum_{s=1}^{g} \lambda_s (2\pi)^{-p/2} |D_s^{-1} C^{-1} D_s^{-1}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\underline{x}-\underline{\mu}_s)' D_s^{-1} C D_s^{-1} (\underline{x}-\underline{\mu}_s)\right\}.$$

En forma análoga a las dos secciones anteriores las ecuaciones máximo verosímiles son

$$\frac{\partial L}{\partial C_{ij}} = \sum_{s=1}^{g} \frac{\partial L}{\partial \overline{v}_{ij}^{s}} \frac{\partial \overline{v}_{ij}^{s}}{\partial C_{ij}} = 0$$

y dado que

$$C_{ij} = \frac{\overline{v}_{ij}^{s}}{\sqrt{\overline{v}_{ii}^{s}\,\overline{v}_{jj}^{s}}} = \frac{\overline{v}_{ij}^{s}}{\alpha_i^s \alpha_j^s}$$

entonces

$$\frac{\partial \overline{v}_{ij}^{s}}{\partial C_{ij}} = \alpha_i^s \alpha_j^s$$

Luego, las ecuaciones de máxima verosimilitud son

$$\frac{\partial L}{\partial C_{ij}} = \sum_{s=1}^{g} \frac{\partial L}{\partial \overline{v}_{ij}^{s}} \hat{\alpha}_i^s \hat{\alpha}_j^s = 0$$

Procediendo de manera idéntica que en la sección 2 de este capítulo, obtenemos

$$\sum_{s=1}^{g} \sum_{k=1}^{N} \frac{\frac{\partial}{\partial \overline{v}_{ij}^{s}} f(\underline{x}_k)}{f(\underline{x}_k)} \hat{\alpha}_i^s \hat{\alpha}_j^s = 0$$

i.e.,

$$\sum_{s=1}^{g} \sum_{k=1}^{N} \frac{(1-\frac{\delta_{ij}}{2})\,\hat{\lambda}_s \alpha_s}{f(\underline{x}_k)} \left[\hat{\overline{v}}_s^{ij} - (x_{ik}-\hat{\mu}_{si})(x_{jk}-\mu_{sj})\right] \hat{\alpha}_i^s \hat{\alpha}_j^s =$$

$$= (1-\tfrac{\delta_{ij}}{2}) \sum_{s=1}^{g} \sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s}{f(\underline{x}_k)} \left[\hat{c}^{ij} \hat{\alpha}_s^i \hat{\alpha}_s^j - (x_{ik}-\hat{\mu}_{si})(x_{jk}-\hat{\mu}_{sj}) \right] \hat{\alpha}_i^s \hat{\alpha}_j^s$$

$$= (1-\tfrac{\delta_{ij}}{2}) \sum_{s=1}^{g} \sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s}{f(\underline{x}_k)} \left[\hat{c}^{ij} - (x_{ik}-\hat{\mu}_{si})(x_{jk}-\hat{\mu}_{sj}) \hat{\alpha}_i^s \hat{\alpha}_j^s \right] = 0.$$

De donde

$$N \hat{c}^{ij} = \sum_{s=1}^{g} \sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s}{f(\underline{x}_k)} (x_{ik}-\hat{\mu}_{si})(x_{jk}-\hat{\mu}_{sj}) \hat{\alpha}_i^s \hat{\alpha}_j^s =$$

$$= \sum_{s=1}^{g} \sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s}{f(\underline{x}_k)} x_{ik} x_{jk} \hat{\alpha}_i^s \hat{\alpha}_j^s + \sum_{s=1}^{g} \sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s}{f(\underline{x}_k)} \hat{\mu}_{si} \hat{\mu}_{sj} \hat{\alpha}_i^s \hat{\alpha}_j^s$$

$$- \sum_{s=1}^{g} \sum_{k=1}^{N} \frac{\hat{\lambda}_s \alpha_s}{f(\underline{x}_k)} (x_{ik} \hat{\mu}_{sj} + x_{jk} \hat{\mu}_{si}) \hat{\alpha}_i^s \hat{\alpha}_j^s.$$

$$= \sum_{s=1}^{g} \sum_{k=1}^{N} \hat{P}(s/\underline{x}_k) x_{ik} x_{jk} \hat{\alpha}_i^s \hat{\alpha}_j^s + N \sum_{s=1}^{g} \hat{\lambda}_s \hat{\mu}_{si} \hat{\mu}_{sj} \hat{\alpha}_i^s \hat{\alpha}_j^s$$

$$- 2N \sum_{s=1}^{g} \hat{\lambda}_s \hat{\mu}_{si} \hat{\mu}_{sj} \hat{\alpha}_i^s \hat{\alpha}_j^s$$

$$= \sum_{s=1}^{g} \sum_{k=1}^{N} \hat{P}(s/\underline{x}_k) x_{ik} x_{jk} \hat{\alpha}_i^s \hat{\alpha}_j^s - N \sum_{s=1}^{g} \hat{\lambda}_s \hat{\mu}_{si} \hat{\mu}_{sj} \hat{\alpha}_i^s \hat{\alpha}_j^s$$

Por lo tanto

$$\hat{c}^{ij} = \frac{1}{N}\sum_{s=1}^{g}\sum_{k=1}^{N}\hat{P}(s/\underline{x}_k)\,x_{ik}x_{jk}\hat{d}_i^{s}\hat{d}_j^{s} - \sum_{s=1}^{g}\hat{\lambda}_s\hat{\mu}_{si}\hat{\mu}_{sj}\hat{d}_i^{s}\hat{d}_j^{s}$$

que es la estimación máxima verosímil de las correlaciones.

Estos resultados los podemos condensar en el siguiente teorema.

Teorema 4. Los estimadores máximo verosímiles de los parámetros de una mezcla de distribuciones normales multivariadas con matriz de correlación común a todos los grupos son

(3.14) $$\hat{\lambda}_s = \frac{1}{N}\sum_{k=1}^{N}\hat{P}(s/\underline{x}_k)$$

(3.15) $$\hat{\mu}_{si} = \frac{1}{N\hat{\lambda}_s}\sum_{k=1}^{N}\hat{P}(s/\underline{x}_k)\,x_{ik}$$

(3.16) $$\hat{\Gamma}_s^{ii} = \frac{1}{N\hat{\lambda}_s}\sum_{k=1}^{N}\hat{P}(s/\underline{x}_k)(x_{ik}-\mu_{si})^2$$

(3.17) $$\hat{c}^{ij} = \frac{1}{N}\sum_{s=1}^{g}\sum_{k=1}^{N}\hat{P}(s/\underline{x}_k)x_{ik}x_{jk}\hat{d}_i^{s}\hat{d}_j^{s} - \sum_{s=1}^{g}\hat{\lambda}_s\hat{\mu}_{si}\hat{\mu}_{sj}\hat{d}_i^{s}\hat{d}_j^{s}$$

donde $\hat{d}_i^{s} = (d_s^{i})^{-1} = \left(\sqrt{\hat{\Gamma}_s^{ii}}\right)^{-1} = \sqrt{\hat{\Gamma}_{ii}^{s}}$; c^{ij} son las correlaciones.

Es interesante hacer notar que si los grupos en la mezcla tienen varianzas comunes Γ^{ii}, $i = 1, 2, \ldots, p$, entonces

$$\hat{c}^{ij} = \left\{ \frac{1}{N} \sum_{k=1}^{N} x_{ik} x_{jk} - \sum_{s=1}^{g} \hat{\lambda}_s \hat{\mu}_{si} \hat{\mu}_{sj} \right\} \Big/ \sqrt{\hat{\nabla}_*^{ii} \hat{\nabla}_*^{jj}} = \frac{\hat{\nabla}_*^{ij}}{\sqrt{\hat{\nabla}_*^{ii} \hat{\nabla}_*^{jj}}}$$

donde $\hat{\nabla}_*^{ij}$ son las estimaciones máximo verosímiles de ∇^{ij} suponiendo que las poblaciones en la mezcla tienen una matriz de dispersión común $\Sigma = \{\nabla^{ij}\}$. Los estimadores $\hat{\alpha}_i^s$ no son estrictamente los estimadores máximo verosímiles para este modelo; sin embargo, son buenos estimadores.

Es importante señalar que el tercer caso ($\Sigma_s = a_s \Sigma$) es un caso particular del cuarto caso (matrices de correlación iguales). Para ver esto, notemos que si las matrices de dispersión son de la forma $\Sigma_s = a_s \Sigma$, entonces

$$\Sigma_s = \{\nabla_s^{ij}\} = a_s \Sigma = a_s \{\nabla^{ij}\} = \{a_s \nabla^{ij}\}$$

y las correlaciones son

$$C_s^{ij} = \frac{\nabla_s^{ij}}{\sqrt{\nabla_s^{ii} \nabla_s^{jj}}} = \frac{a_s \nabla^{ij}}{\sqrt{a_s \nabla^{ii} a_s \nabla^{jj}}} = \frac{\nabla^{ij}}{\sqrt{\nabla^{ii} \nabla^{jj}}}$$

para todo $s = 1, 2, \ldots g$; es decir, que las correlaciones entre las variables en los distintos grupos son iguales.

En el tercer caso ($\Sigma_s = a_s \Sigma$) se estiman menos parámetros que en el caso de matrices de correlación iguales siempre y cuando $g(p-1) \geq p$. Para ver la justificación de esto recordemos que el número de parámetros a estimar cuando se tienen matrices de correlación iguales es $g - 1 + gp + \frac{1}{2} p(p-1) + gp$, mientras que para el caso $\Sigma_s = a_s \Sigma$, se tienen que estimar

$g-1 + gp + \frac{1}{2}p(p+1) + g$ y restando ambas expresiones obtenemos $g(p-1) - p$; para que esto sea positivo es necesario que $g(p-1) \geq p$, como queríamos verificar.

CAPITULO IV

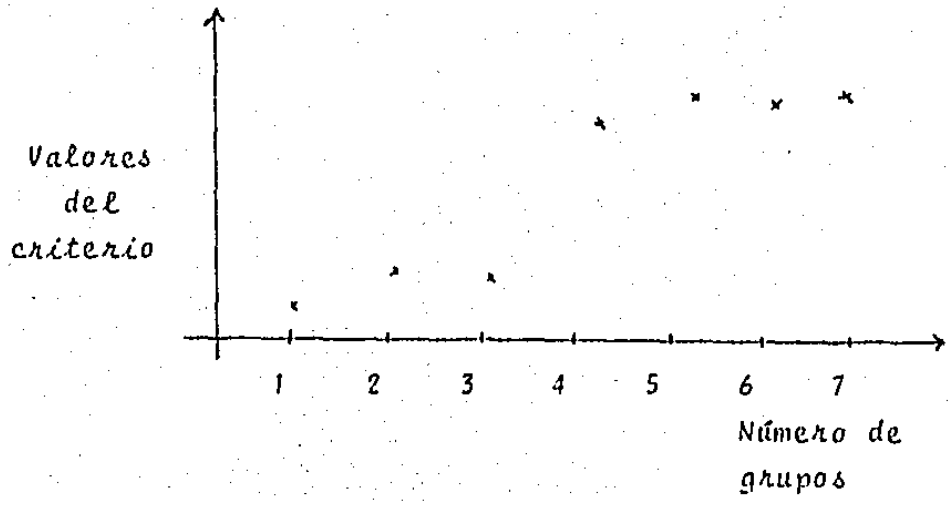
PRUEBAS SOBRE EL NUMERO DE GRUPOS EN UNA MEZCLA

Un problema común a todas las técnicas de análisis de conglomerados consiste en decidir el número de grupos a formarse. Este problema será tratado en este capítulo.

1. REVISION DE ALGUNOS METODOS.

Para decidir cuántos grupos formar al hacer un agrupamiento de objetos o individuos, varios autores han sugerido diversos métodos que dependen de la técnica utilizada al hacer dicho agrupamiento.

Para aquellas técnicas que buscan optimizar un criterio, Friedman y Rubin (1967) [8], sugieren que una gráfica del valor del criterio contra el número de grupos indicará el número más adecuado de éstos en donde haya un cambio muy pronunciado de dicho valor. Por ejemplo, la siguiente gráfica sugeriría que el número más adecuado de grupos a formar es 4:



Sin embargo, este procedimiento no ha sido satisfactorio por completo. Otros intentos se han hecho para resolver este problema. Por ejemplo, Beale (1969) [2] sugiere una estadística F definida de la manera siguiente:

$$F(C_2, C_1) = \frac{R_{C_2} - R_{C_1}}{R_{C_2}} \Big/ \left[\frac{N - C_1}{N - C_2} \left(\frac{C_2}{C_1} \right)^{2/P} - 1 \right]$$

basada en $p(C_2 - C_1)$ y $p(N - C_2)$ grados de libertad. En esta fórmula $R_C = (N - C)S_C^2$ y S_C^2 es la desviación cuadrática media respecto a los centros de los grupos en la muestra. Un resultado significativo sería indicador de que una división en C_2 grupos es significativamente mejor que una división en C_1 grupos (C_2 C_1). Calinski y Harabasz (1971)

, sugirieron usar el cociente R dado por

$$R = \frac{\text{traza}(B) / (g - 1)}{\text{traza}(W) / N - g}$$

donde B y W son las matrices de dispersión entre grupos y dentro de grupos respectivamente. Si en esta expresión el valor de R crece monótonamente con el de g, entonces se sugiere que no existe estructura de grupos en los datos; si por el contrario, R decrece monótonamente con g, se sugiere la presencia de una estructura jerárquica en los grupos con g^* grupos, donde g^* es aquel valor de g donde se produce el cambio más pronunciado conforme variamos g. Tanto el método de Beale, como el de Calinski y Harabasz son útiles, pero sólo en situaciones donde los grupos son aproximadamente esféricos y de igual tamaño.

En 1971, Marriot investigó en detalle las propiedades del 'Criterio $|W|$'. El propone el uso de $g^2|W|$ donde g es el número correcto de grupos, como un indicador de la estructura de datos, tomando como el número correcto de grupos aquel valor de g para el que $g^2|W|$ es mínimo. Asimismo, Marriot investigó las propiedades muestrales de la estadística asociada, es decir de $g^2|W|/T$, donde la población tiene una distribución uniforme, y concluye que a partir de ésta se puede saber si hay alguna estructura de grupos en los datos, o bien, si se está tratando con un solo grupo. Este método puede aplicarse a conglomerados más generales que en los casos anteriores, puesto que no hay suposición alguna sobre la forma de los grupos, aunque sí se requiere que éstos sean aproximadamente del mismo tamaño.

Para las técnicas jerárquicas de análisis de conglomerados no existen indicadores claros del número más adecuado de grupos, y puede ser, que en los tipos de datos para los que estas técnicas sean las más apropiadas, el problema del número de grupos no sea de gran interés. Pero en aquellos casos en donde sí necesitamos tener algún indicador del número correcto de grupos, el simple análisis de los dendogramas no es muy satisfactorio.

En el caso de ajustar mezclas de distribuciones multivariadas a un conjunto de datos, existe un criterio, mucho más razonable que los anteriores, para investigar el número correcto de grupos en la muestra; este criterio es la Prueba del Cociente de Verosimilitud, el cual estudiaremos a continuación.

2. PRUEBA DEL COCIENTE DE VEROSIMILITUD.

Para probar la hipótesis de r grupos contra s grupos $(r < s)$, podemos usar la estadística

$$\Lambda = -2 \log \lambda$$

donde λ es el cociente de las verosimilitudes al tomar

s y r grupos, es decir

$$\lambda = L_s / L_r$$

donde L_s y L_r son las funciones de verosimilitud de la muestra bajo las hipótesis de s y r grupos, respectivamente.

Tanto Wilks (1938) [16] como Silvey (1959) [17], demostraron que bajo ciertas condiciones de regularidad, la estadística Λ se distribuye como una χ^2 con grados de libertad igual a la diferencia de los números de parámetros bajo las dos hipótesis. Bartlett (1939) [1], al el investigar el problema de probar la igualdad de r medias en análisis de varianza multivariado, mejoró el resultado de Wilks, para muestras pequeñas usando la estadística modificada

$$\Lambda' = -2C \log \lambda$$

con $p(s - r)$ grados de libertad y $C=(N - 1 -(p+s)/2)/N$, donde N es el tamaño de la muestra y p es el número de variables.

En 1970, Wolfe [19] notó que la fórmula de Wilks no funciona correctamente. Tanto para los datos de "Fisher-Iris" como para unos conglomerados generados artificialmente, la prueba de Wilks rechazaba la hipótesis nula cuando era verdadera. Un año después, el mismo Wolfe

señaló que la causa de que hubiera ocurrido esto, es que algunas de las condiciones de regularidad pedidas en el Teorema de Wilks no son satisfechas por el cociente de verosimilitud en el problema de mezclas.

Para probar la hipótesis de r grupos contra s grupos, Wolfe recomienda una modificación al cociente de verosimilitud. Esta modificación consiste en usar la estadística

$$\Lambda^* = -\frac{2}{N}\left(N - 1 - p - \frac{s}{2}\right)\log \lambda$$

como una distribución χ^2 con $2p(s-r)$ grados de libertad si se hace la suposición de matrices de dispersión iguales, $2\left[\frac{p(p+1)}{2}+p\right](s-r)$ grados de libertad en el caso general y $2(2p)(s-r)$ grados de libertad para el caso de matrices de correlación iguales. Esta recomendación es sólo heurística, no tiene justificación teórica o numérica.

CAPITULO V

USO DE NORMIX

Tal y como lo señalamos en los capítulos anteriores, en este capítulo presentamos la manera como debe hacerse un agrupamiento de un conjunto de datos usando la técnica de mezclas. Para ello usamos un programa cuya versión original se debe a J.H. Wolfe (1967) (Ver 18) titulado NORMIX. La versión utilizada en el IIMAS incorpora algunas modificaciones hechas por A. Hernández.

Este último capítulo consta de dos partes: en la primera presentamos las características generales de este programa así como un manual de usuario para el mismo; en la segunda parte presentamos un ejemplo de un agrupamiento hecho en los datos de Iris de Fisher.

1. CARACTERISTICAS GENERALES Y MANUAL DE USUARIO.

El objetivo de este programa consiste en encontrar los estimadores máximo verosímiles de una mezcla de distribuciones normales multivariadas; una vez teniendo éstos se

lleva a cabo el agrupamiento de los datos de acuerdo a las probabilidades de pertenencia de cada uno de los objetos o individuos a cada uno de los grupos. Existe una opción que permite al usuario especificar si las matrices de dispersión son iguales o diferentes (aún no han sido incluidas las opciones para los casos 3 y 4 del capítulo anterior).

Tal y como vimos en el Capítulo II, para hacer un agrupamiento en un conjunto de datos por medio de la técnica de mezclas, es necesario contar con estimaciones iniciales para todos los parámetros involucrados en la mezcla. En la versión de Wolfe se obtenían estas estimaciones usando técnicas jerárquicas de agrupamiento. La versión existente en el IIMAS, incluye la opción de obtener estas estimaciones minimizando la traza de W o bien minimizando el determinante de W, donde W es la matriz de dispersión dentro de grupos. Esta modificación provoca que la velocidad de convergencia de los métodos iterativos al resolver las ecuaciones máximo verosímiles sea mucho más rápida.

Si el número de variables es muy grande, los cálculos dentro de la computadora son muy pesados. Sin embargo, esta no es una dificultad tan seria, puesto que dado que los resultados son independientes de transformaciones lineales, las variables observadas pueden ser reemplazadas por un número menor de componentes principales, y el programa da los mismos resultados, modificados solamente por la pér-

dida de información en los componentes principales descartados.

Tal y como ya lo indicamos, existe una opción que permite especificar si las matrices de dispersión son iguales o distintas. Las ventajas de los dos enfoques son similares a aquéllas de usar funciones discriminantes lineales o cuadráticas; el análisis de conglomerados trata con diferencias en la localización y en la práctica la versión con matrices de dispersión distintas puede producir estimaciones muy diferentes a las obtenidas con matices de dispersión iguales, aunque las reglas de colocación o asignación al formar los grupos sean muy similares para ambos casos.

La salida del programa nos da para cada valor de g (el número de grupos), la verosimilitud, la estimación de las proporciones y de las probabilidades de pertenencia para cada uno de los grupos. Si los grupos están muy separados la convergencia es muy rápida y la mayoría de los puntos son asignados a un grupo con probabilidad cercana a uno. Si no lo están, la convergencia es muy lenta, además de que pueden resultar varios máximos locales; las estimaciones dependen de los valores iniciales escogidos, por lo que es muy recomendable hacer varias corridas para un mismo conjunto de datos con diferentes estimaciones iniciales. El hecho de que haya varios máximos locales, no es una dificultad muy seria, desde el punto de vista de Análisis de Conglomerados,

pues la mayoría de las veces se trata de dividir distribuciones multimodales más que de dividir distribuciones platicúrticas.

Para correr NORMIX en el IIMAS, la instrucción para hacerlo es:

R$BINNORMIX4;FILE FILE5(DISK, TITLE=nombre del programa, FILETYPE=7), FILE1(DISK, TITLE=nombre del archivo de datos, FILETYPE=7).

Las instrucciones para cada uno de los renglones del programa son:

RENGLON	ESPECIFICACIONES
100	Los primeros cuatro renglones se usan para algún
200	título del programa, comentarios u observaciones
300	que se dee aparezcan al principio de cada corrida.
400	Si no desean utilizarse tienen que dejarse en blanco.

500

Columnas 34-35: Número de variables medidas a cada individuo.

Columnas 50-53: Número de individuos u observaciones; tiene que ser un número del 1 al 9999.

Columnas 69-70: Número del archivo de datos. Generalmente se el elige el 01.

---

600 Columnas 32 en adelante: número de grupos a formar; se escriben en forma ascendiente, separados por una coma; deben ser números del 01 al 99.

---

700 Columna 44: se escribe un 0 si se desea la opción de matrices de dispersión iguales; se escribe un 1 si se desea la opción de matrices de dispersión diferentes.

Columnas 68-70: mínimo número de puntos que debe tener un grupo para que pueda ser estimada su matriz de dispersión. Si el tamaño de algún grupo resulta menor que este número se toma a la matriz de dispersión entre grupos como matriz de dispersión para dicho grupo.

---

800 Columnas 70-72: el programa formará un número mayor de grupos si la prueba resulta significativa. Esta prueba consiste en ir probando un grupo más que el último especificado. En estas columas se pone el valor de α, la significancia de la prueba.

---

900 Columnas 36-38: Se usa solamente en el caso en que

se desee un agrupamiento inicial de tipo jerárquico. Se especifica el tamaño de la submuestra que se usa para generar estimaciones iniciales. Debe ser un número del 040 al 250.

Columna 58: si se escribe un 1 entonces en la salida del programa se imprimen los resultados de cada iteración. Desde luego esto origina que se tenga una salida muy grande.

Columna 68: si se escribe un 1 entonces el usuario puede identificar cada observación. Para esto las etiquetas deben estar en las primeras n columnas ($1 \leq n \leq 8$) del archivo de datos. Por lo tanto, el formato en los renglones 13 y/o 14 debe ser de la forma (A_n, formato de las observaciones) y la etiqueta debe ocupar n caracteres en el archivo de datos. si se escribe un 0 no hay alteraciones y las observaciones son numeradas e identificadas de acuerdo al orden en que se leen. Desde luego que es posible modificar el orden en que se leen las etiquetas en el archivo de datos alterando la instrucción de lectura del archivo de datos.

1000 Columnas 25-16: En estas columnas se especifica un límite de tiempo (en minutos) para que que el programa tenga tiempo de imprimir los resultados. Este tiempo debe ser alrededor de un minuto menos que el límite de tiempo del JOB correspondiente, con el objeto de que se puedan imprimir los resultados antes de que el sistema operativo concluya el JOB. Si se dejan en blanco, es probable que el listado de resultados no esté completo.

Columnas 57-59: Si se dejan en blanco el programa iterará hasta que se obtenga convegencia. Esto toma de 1 a 100 iteraciones. Si no se dispone de tiempo máquina suficiente se pueden indicar en estas columnas un número menor de iteraciones. Generalmente con 010 iteraciones es más que suficiente. Si se escribe aún un número menor de iteraciones es posible que ya no se tengan los estimadores máximo verosímiles, aunque los obtenidos serán siempre mejores que los que se obtienen del agrupamiento inicial.

1100 Columna 34: Si se deja en blanco se hace un agrupamiento inicial jerárquico. Si se escribe un 1 se hará un agrupamiento inicial no jerárquico: si se escribe un 1 en la columna 46 y un 0 en la 58 el método es minimizando la traza de W; si se escribe un 2 en la columna 46 y un 1 en la 58 el método es minimizando el determinante de W.

1200 Se deja en blanco

1300 Columna 1 en adelante: Formato de entrada; o sea el formato con el que se leerá el archivo de datos. Va escrito entre paréntesis.

1400 Sólo se usa en caso de que el usuario dé estimaciones iniciales para los parámetros de la mezcla. En este caso, en las columnas 26-27 se escribe el número de grupos del conjunto de estimaciones iniciales. Además en las columnas 44, 55 y 70 se escribe un 1 si medias iniciales, desviaciones estándar iniciales o correlaciones iniciales van a ser leídas respectivamente. En este caso el formato de

entrada es 8f10.4 y ya no se tiene que especificar. Las estimaciones iniciales se escriben los renglones sucesivos con este formato, escribiendo primero todas las medias, después desviaciones estándar y finalmente correlaciones. En caso de que no se den estimaciones iniciales, este renglón se deja en blanco.

1500 Columna 1-2: Se escribe lo siguiente: /* (diagonal asterisco) que indica el fin de los comandos. Se debe tener un renglón en blanco antes de éste.

Después de esto ya se puede correr el programa con la instrucción dada para hacerlo.

En la siguiente sección de este capítulo presentamos un ejemplo de como hacer un agrupamiento en un conjunto de datos usando este programa. Se presentan también los resultados obtenidos.

2. ANALISIS DE LOS DATOS DE IRIS DE FISHER.

En esta sección se analiza un conjunto de datos publicados por Fisher en 1936, correspondientes a tres varie-

dades de Iris, a saber, Iris Setosa, Iris Versicolor e Iris Virginica. Se tienen 50 observaciones para cada una de las variedades y para facilitar el análisis las hemos numerado de la manera siguiente: con los números 1 al 50 las observaciones correspondientes a Iris setosa; con los números 51 al 100 las correspondientes a Iris versicolor y con los números 101 al 150 las correspondientes a Iris virginica. Cada observación consta de 4 mediciones hechas a cada uno de los ejemplares de Iris: longitud del sépalo (LS), ancho del sépalo (AS), longitud del pétalo (LP) y ancho del pétalo (AP). Dos de las especies, Iris setosa e Iris Versicolor fueron encontradas en la misma colonia por el Botánico E. Anderson. La especie Iris virginica fue encontrada en una colonia distinta. El mismo Fisher (1936) aplicó su procedimiento de análisis discriminante a los tres grupos anteriores y encontró que la Iris setosa podía separarse muy bien de las otras especies, sin embargo, las especies Iris Versicolor e Iris Virginica tenían un traslape considerable y podían confundirse algunos elementos de una especie con los de la otra.

En la siguiente tabla se muestran este conjunto de datos.

DATOS DE LAS TRES VARIEDADES DE IRIS

IRIS SETOSA

	LS	AS	LP	AP
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
	5.0	3.4	1.5	0.2
	4.4	2.9	1.4	0.2
10	4.9	3.1	1.5	0.1
11	5.4	3.7	1.5	0.2
12	4.8	3.4	1.6	0.2
	4.8	3.0	1.4	0.1
	4.3	3.0	1.1	0.1
15	5.8	4.0	1.2	0.2
16	5.7	4.4	1.5	0.4
17	5.4	3.9	1.3	0.4
18	5.1	3.5	1.4	0.3
19	5.7	3.8	1.7	0.3
20	5.1	3.8	1.5	0.3
21	5.4	3.4	1.7	0.2
22	5.1	3.7	1.5	0.4
23	4.6	3.6	1.0	0.2
	5.1	3.3	1.7	0.5
25	4.8	3.4	1.9	0.2
26	5.0	3.0	1.6	0.2
27	5.0	3.4	1.6	0.4
28	5.2	3.5	1.5	0.2
29	5.2	3.4	1.4	0.2
30	4.7	3.2	1.6	0.2
31	4.8	3.1	1.6	0.2
32	5.4	3.4	1.5	0.4
33	5.2	4.1	1.5	0.1
34	5.5	4.2	1.4	0.2
35	4.9	3.1	1.5	0.2
36	5.0	3.2	1.2	0.2
37	5.5	3.5	1.3	0.2
38	4.9	3.6	1.4	0.1
39	4.4	3.0	1.3	0.2
40	5.1	3.4	1.5	0.2
41	5.0	3.5	1.3	0.3
42	4.5	2.3	1.3	0.3
43	4.4	3.2	1.3	0.2
44	5.0	3.5	1.6	0.6
45	5.1	3.8	1.9	0.4
46	4.8	3.0	1.4	0.3
47	5.1	3.8	1.6	0.2
48	4.6	3.2	1.4	0.2
49	5.3	3.7	1.5	0.2
50	5.0	3.3	1.4	0.2

IRIS VERSICOLOR

	LS	AS	LP	AP
51	7.0	3.2	4.7	1.4
52	6.4	3.2	4.5	1.5
53	6.9	3.1	4.9	1.5
54	5.5	2.3	4.0	1.3
55	6.5	2.8	4.6	1.5
56	5.7	2.8	4.5	1.3
57	6.3	3.3	4.7	1.6
58	4.9	2.4	3.3	1.0
59	6.6	2.9	4.6	1.3
60	5.2	2.7	3.9	1.4
61	5.0	2.0	3.5	1.0
62	5.9	3.0	4.2	1.5
63	6.0	2.2	4.0	1.0
64	6.1	2.9	4.7	1.4
65	5.6	2.9	3.6	1.3
66	6.7	3.1	4.4	1.4
67	5.6	3.0	4.5	1.5
68	5.8	2.7	4.1	1.0
69	6.2	2.2	4.5	1.5
70	5.6	2.5	3.9	1.1
71	5.9	3.2	4.8	1.8
72	6.1	2.8	4.0	1.3
73	6.3	2.5	4.9	1.5
74	6.1	2.8	4.7	1.2
75	6.4	2.9	4.3	1.3
76	6.6	3.0	4.4	1.4
77	6.8	2.8	4.8	1.4
78	6.7	3.0	5.0	1.7
79	6.0	2.9	4.5	1.5
80	5.7	2.6	3.5	1.0
81	5.5	2.4	3.8	1.1
82	5.5	2.4	3.7	1.0
83	5.8	2.7	3.9	1.2
84	6.0	2.7	5.1	1.6
85	5.4	3.0	4.5	1.5
86	6.0	3.4	4.5	1.6
87	6.7	3.1	4.7	1.5
88	6.3	2.3	4.4	1.3
89	5.6	3.0	4.1	1.3
90	5.5	2.5	4.0	1.3
91	5.5	2.6	4.4	1.2
92	6.1	3.0	4.6	1.4
93	5.8	2.6	4.0	1.2
94	5.0	2.3	3.3	1.0
95	5.6	2.7	4.2	1.3
96	5.7	3.0	4.2	1.2
97	5.7	2.9	4.2	1.3
98	6.2	2.9	4.3	1.3
99	5.1	2.5	3.0	1.1
100	5.7	2.8	4.1	1.3

IRIS VIRGINICA

	LS	AS	LP	AP
101	6.3	3.3	6.0	2.5
102	5.8	2.7	5.1	1.9
103	7.1	3.0	5.9	2.1
104	6.3	2.9	5.6	1.8
105	6.5	3.0	5.8	2.2
106	7.6	3.0	6.6	2.1
107	4.9	2.5	4.5	1.7
108	7.3	2.9	6.3	1.8
109	6.7	2.5	5.8	1.8
110	7.2	3.6	6.1	2.5
111	6.5	3.2	5.1	2.0
112	6.4	2.7	5.3	1.9
113	6.8	3.0	5.5	2.1
114	5.7	2.5	5.0	2.0
115	5.8	2.8	5.1	2.4
116	6.4	3.2	5.3	2.3
117	6.5	3.0	5.5	1.8
118	7.7	3.8	6.7	2.2
119	7.7	2.6	6.9	2.3
120	6.0	2.2	5.0	1.5
121	6.9	3.2	5.7	2.3
122	5.6	2.8	4.9	2.0
123	7.7	2.8	6.7	2.0
124	6.3	2.7	4.9	1.8
125	6.7	3.3	5.7	2.1
126	7.2	3.2	6.0	1.8
127	6.2	2.8	4.8	1.8
128	6.1	3.0	4.9	1.8
129	6.4	2.8	5.6	2.1
130	7.2	3.0	5.8	1.6
131	7.4	2.8	6.1	1.9
132	7.9	3.8	6.4	2.0
133	6.4	2.8	5.6	2.2
134	6.3	2.8	5.1	1.5
135	6.1	2.6	5.6	1.4
136	7.7	3.0	6.1	2.3
137	6.3	3.4	5.6	2.4
138	6.4	3.1	5.5	1.8
139	6.0	3.0	4.8	1.8
140	6.9	3.1	5.4	2.1
141	6.7	3.1	5.6	2.4
142	6.9	3.1	5.1	2.3
143	5.8	2.7	5.1	1.9
144	6.8	3.2	5.9	2.3
145	6.7	3.3	5.7	2.5
146	6.7	3.0	5.2	2.3
147	6.3	2.5	5.0	1.9
148	6.5	3.0	5.2	2.0
149	6.2	3.4	5.4	2.3
150	5.9	3.0	5.1	1.8

Para este conjunto de datos hicimos varias corridas de NORMIX. Estas corridas fueron hechas con las opciones de matrices de dispersión iguales y diferentes; también variamos los métodos de agrupamiento inicial: jerárquico, minimización del determinante de w y minimización de la traza de W. A continuación presentamos los resultados de estas corridas. En la tabla 1 presentamos las pruebas sobre el número de grupos, usando hipótesis de un grupo, dos grupos, tres grupos y cuatro grupos. Cada hipótesis se prueba contra la anterior usando la modificación propuesta por Wolfe al cociente de verosimilitud, a saber

$$\chi^2 = -\frac{2}{N}\left(N-1-p-\frac{s}{2}\right)\log\frac{L_s}{L_r}$$

con $2p(s-r)$ grados de libertad si se hace la suposición de matrices de dispersión iguales y $2\left(\frac{p(p+1)}{2}+p\right)(s-r)$ grados de libertad si se hace la suposición de matrices de dispersión diferentes. Desde luego que en este caso tenemos $p=4$ y las pruebas de hipótesis fueron hechas con $s-r=1$, por lo que tenemos 8 grados de libertad en el caso de matrices de dispersión iguales y 28 grados de libertad en el caso de matrices de dispersión diferentes. En la tabla 2 presentamos las estimaciones para algunos parámetros de la mezcla (de las proporciones y las medias) y los comparamos con los valores reales que el mismo Fisher publicó en su citado artículo. Esta tabla incluye también el número de individuos mal clasificados con cada uno

de los métodos y cuales fueron estos individuos mal clasificados en cada caso para poder hacer una comparación posterior de los métodos.

TABLA 1

Pruebas de significancia sobre el número de grupos

MATRICES DE DISPERSION IGUALES

Número de grupos: H_o/H_a	JERARQUICO	P	MIN DET W	P	MIN TR W	P
1/2	-	-	151.61	10^{-8}	151.61	10^{-8}
2/3	40.71	.000002	59.75	10^{-8}	59.75	10^{-8}
3/4	95.93	10^{-8}	12.04	.034	12.04	.034

MATRICES DE DISPERSION DIFERENTES

Número de grupos: H_o/H_a	JERARQUICO	P	MIN DET W	P	MIN TR W	P
1/2	-	-	-	-	-	-
2/3	38.18	0.09505	49.98	.00640	49.25	.00783
3/4	58.64	.000606	20.89	.82992	16.92	.95010

Notas: 1) La P es la probabilidad de la hipótesis nula ante la alternativa.

2) Los valores faltantes no los dio la máquina; son muy altos los valores de χ^2.

TABLA 2
Estimación de parámetros de Iris

I. IRIS SETOSA

		MATRICES DE DISPERSION IGUALES			MATRICES DE DISPERSION DIFERENTES		
PARAMETRO	FISHER	Jerárquico	Min det W	Min tr W	Jerárquico	Min det W	Min tr W
Proporción	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333
Media (LS)	5.0060	5.0060	5.0060	5.0060	5.0060	5.0060	5.0060
Media (AS)	3.4280	3.4280	3.4280	3.4280	3.4280	3.4280	3.4280
Media (LP)	1.4620	1.4620	1.4620	1.4620	1.4620	1.4620	1.4620
Media (AP)	0.2460	0.2460	0.2460	0.2460	0.2460	0.2460	0.2460
# mal clasificados	0	0	0	0	0	0	0

II. IRIS VERSICOLOR

		MATRICES DE DISPERSION IGUALES			MATRICES DE DISPERSION DIFERENTES		
PARAMETRO	FISHER	Jerárquico	Min det W	Min tr W	Jerárquico	Min det W	Min tr W
Proporción	0.3333	0.5200	0.3153	0.2973	0.1410	0.2850	0.2810
Media (LS)	6.5880	7.0045	6.5982	6.6454	6.9706	6.6733	6.6815
Media (AS)	2.9740	2.9318	3.0021	3.0120	2.9216	2.9980	3.0005
Media (LP)	5.5520	5.3591	5.5587	5.5926	5.3929	5.6413	5.6500
Media (AP)	2.0260	1.6591	2.0567	2.0757	1.6607	2.0679	2.0697
# mal clasificados	0	3	1	0	12	0	0

III. IRIS VIRGINICA

		MATRICES DE DISPERSION IGUALES			MATRICES DE DISPERSION DIFERENTES		
PARAMETRO	FISHER	Jerárquico	Min det W	Min tr W	Jerárquico	Min det W	Min tr W
Proporción	0.3333	0.1467	0.3514	0.3694	0.5260	0.3810	0.3850
Media (LS)	5.9360	6.0526	5.9607	5.9535	6.0721	5.9544	5.9550
Media (AS)	2.7700	2.8531	2.7555	2.7594	2.8587	2.7788	2.7781
Media (LP)	4.2600	4.7782	4.3244	4.3545	4.7755	4.3561	4.3625
Media (AP)	1.3260	1.6808	1.3349	1.3544	1.6801	1.3829	1.3884
# mal clasificados	0	36	4	4	40	6	6
Total de mal clasificados	0	39	5	4	52	6	6

Las pruebas de significancia sobre el número de grupos indican que la hipótesis de un grupo debe ser rechazada ante la alternativa de dos grupos; también indican que la hipotesis de dos grupos debe ser rechazada ante la alternativa de tres grupos; si acaso el único método donde se presenta una duda es el de agrupamiento jerárquico con matrices de dispersión diferentes donde la P tiene un valor de 0.09505 por lo que no podríamos rechazar al 95 % de confianza. Asimismo, estas pruebas indican (con excepción de dos métodos) que no debe rechazarse la hipótesis de 3 grupos ante la alternativa de 4 grupos. Los dos métodos en donde esto no es tan claro son los de agrupamiento por min det W y min tr W con matrices de dispersión iguales, en donde la P tiene un valor de .034.

La tabla 2 presenta los estimadores máximo verosímiles de algunos parámetros de la mezcla (proporciones y medias) que se obtienen con NORMIX y se comparan con los valores reales de dichos parámetros que el mismo Fisher proporcionó. El renglón de individuos mal clasificados da el número de ejemplares de Iris cuyas probabilidades de pertenencia fueron más altas para otros grupos que para el que pertenecían en realidad por lo que quedaron mal clasificados. Excepto para los métodos en donde se realizó un agrupamiento inicial jerárquico, las estimaciones que se obtienen con NORMIX fueron muy parecidas a las reales; con la opción de matrices iguales se obtienen

resultados ligeramente mejores que con la opción de matrices de dispersión diferentes. El método que hizo el mejor agrupamiento fue el de un agrupamiento inicial por min tr W con matrices de dispersión iguales en donde solamente se malclasificaron 4 ejemplares de Iris, a saber 120, 130, 134 y 135 que siendo de la especie Iris Virginica los incluyó el programa en la especie Iris Versicolor. El método de min det W inicial con matrices de dispersión iguales, clasificó mal a los mismos individuos además del individuo 71. Los métodos min det W y min tr W inicial con matrices de dispersión diferentes clasificaron mal a seis ejemplares, a saber, los marcados con los números 124, 127, 128, 134, 139 y 150 que siendo de la especie Iris Virginica fueron incluidos en la especie Iris Versicolor. Como un comentario final, debe indicarse que los métodos con agrupamiento jerárquico inicial dieron tan malos agrupamientos debido a que los datos no presentan para nada un esquema jerárquico de agrupamiento.

Con el objeto de averiguar que sucedía con estos mismos datos si primero se obtenían componentes principales, se corrió también NORMIX para los primeros dos componentes principales que resultan de estos datos y que explican el 96 % de la variabilidad total existente en estos datos, resultando que el método que dio mejores resultados fue min tr W inicial y matrices de dispersión diferentes. Con este método se mal clasi-

ficaron 13 individuos, los marcados con los números: 78, 102, 107, 114, 122, 124, 127, 128, 134, 135, 139, 143, 150. Es decir, se volvieron a traslapar algunos ejemplares de Iris Versicolor e Iris Virginica, mientras que los de Iris Setosa fueron agrupados perfectamente.

Finalmente y con propósitos de ilustración, presentamos una corrida de NORMIX así como sus resultados. Elegimos la opción de matrices de dispersión diferentes y un agrupamiento inicial por minimización de la traza de W. Limitamos también el número de iteraciones a 5 para que no se tuviera una salida muy grande.

Las instrucciones de entrada quedan de la manera siguiente:

```
ANALISIS DE LOS DATOS DE IRIS DE FISHER
AGRUPAMIENTO NO JERARQUICO INICIAL;MATRICES DE DISPERSION DIFERENTES
5 ITERACIONES COMO MAXIMO
SEGUNDA CORRIDA
                                04             0150               01
                              02,03,04
                                          1                       005
                                                                   .90

                      01                                005
                                 1           1           0

(2X,4F4.1)

/*
```

La salida del programa queda así: (con esto concluimos el capítulo)

```
                    PROGRAM NORMIX FOR CLUSTER ANALYSIS
     MAXIMUM LIKELIHOOD ESTIMATION OF A MIXTURE OF MULTIVARIATE NORMAL POPULATIONS

ANALISIS DE LOS DATOS DE IRIS DE FISHER
AGRUPAMIENTO NO JERARQUICO INICIAL; MATRICES DE DISPERSION DIFERENTES
SITERACIONES COMO MAXIMO
SEGUNDA CORRIDA

STORAGE REQUIREMENT=  14[illegible]24 DECIMAL BYTES
DIMENSION A( 143C7)

     1      7     31     91    141    187    211    215    231    247    263
   413    563    713    863   1464   1613   1745  13[illegible]91  13241  13391  134[illegible]7

 13717

        A NON HIERARCHICAL INITIAL GROUPING WAS PERFORMED

        EUCLIDEAN DISTANCE WILL BE USED
        THE CRITERION IS TO MINIMIZE TRACE W

                    ************************************

THE CRITERION VALUE FOR THE INITIAL GROUPING FOR   2 TYPES IS [illegible]

                    ************************************

                    ************************************

      SAMPLE SIZE =          15[illegible]
      NUMBER OF VARIABLES =   4
      NUMBER OF TYPES =       2
                  ITERATION NUMBER    1
  LOGLIKELIHOOD OF  2 TYPES IN THIS SAMPLE =   [illegible]
            CHARACTERISTICS OF THE WHOLE SAMPLE
                           MEANS
      1           2           3           4
   5.8433      3.0573      3.7381      1.1993
                    STANDARD DEVIATIONS
      1           2           3           4
   0.8281      0.4359      1.7653      0.7622
                       CORRELATIONS
      1           2           3           4
   1.0000     -0.1176      0.8718      0.8179
  -0.1176      1.0000     -0.4284     -0.3661
   0.8718     -0.4284      1.0000      0.9629
   0.8179     -0.3661      0.9629      1.0000
            CHARACTERISTICS OF TYPE    1
       THE PROPORTION OF THE POPULATION FROM THIS TYPE=  0.647
                           MEANS
      1           2           3           4
   6.3310      2.8866      4.9588      1.6959
                    STANDARD DEVIATIONS
      1           2           3           4
   0.6337      0.3266      0.73[illegible]1      0.4156
                       CORRELATIONS
      1           2           3           4
   1.0000      0.5151     [illegible]      [illegible]
   0.5151      1.0000      0.4771      0.5353
   [illegible]      0.4771      1.0000      0.8122
   0.5553      0.5353      0.8122      1.0000
            CHARACTERISTICS OF TYPE    2
       THE PROPORTION OF THE POPULATION FROM THIS TYPE=  0.353
                           MEANS
      1           2           3           4
   5.0057      3.3698      1.56[illegible]4      0.2716
                    STANDARD DEVIATIONS
      1           2           3           4
   0.3427      0.4397      0.44[illegible]4      0.21[illegible]6
                       CORRELATIONS
      1           2           3           4
   1.0000      0.6239      0.0945      [illegible]
   0.6239      1.0000     -0.4182     -0.3791
   0.0945     -0.42[illegible]      1.0000      0.8617
   0.134[illegible]     -0.3791      0.86[illegible]7      1.0000

                    ************************************

ITERATION   1     LOG LIKELIHOOD OF  2 TYPES IN THIS SAMPLE =   [illegible]
ITERATION   2     LOG LIKELIHOOD OF  2 TYPES IN THIS SAMPLE =   [illegible]
ITERATION   3     LOG LIKELIHOOD OF  2 TYPES IN THIS SAMPLE =   [illegible]

                    ITERATION TIME =       1.17 SECONDS.
                    TOTAL TIME USED =      2.55 SECONDS.
```

SAMPLE SIZE = 151
NUMBER OF VARIABLES = 4
NUMBER OF TYPES = 2

ITERATION NUMBER 4

LOG LIKELIHOOD OF 2 TYPES IN THIS SAMPLE = -337.[illegible]

CHARACTERISTICS OF THE WHOLE SAMPLE

MEANS

1	2	3	4
5.8423	3.0573	3.7584	1.1993

STANDARD DEVIATIONS

1	2	3	4
.8281	.4359	1.7653	.7622

CORRELATIONS

1	2	3	4
1.0000	-.1176	-.8718	-.8179
-.1176	1.0000	.4284	.3661
.8718	-.4284	1.0000	.9629
.8179	-.3661	.9629	1.0000

CHARACTERISTICS OF TYPE 1

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .667

MEANS

1	2	3	4
6.2624	2.8720	4.9161	1.6761

STANDARD DEVIATIONS

1	2	3	4
.6595	.3311	.8215	.4227

CORRELATIONS

1	2	3	4
1.0000	.5539	.8265	.5937
.5539	1.0000	.5198	.5662
.8265	.5198	1.0000	.8234
.5937	.5662	.8234	1.0000

CHARACTERISTICS OF TYPE 2

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .333

MEANS

1	2	3	4
5.0163	3.4281	1.4607	.2460

STANDARD DEVIATIONS

1	2	3	4
.3469	.3752	.1719	.1143

CORRELATIONS

1	2	3	4
1.0000	.7425	.2672	.2781
.7425	1.0000	.1777	.2326
.2672	.1777	1.0000	.3316
.2781	.2326	.3316	1.0000

PROBABILITIES OF TYPE MEMBERSHIP

1 2

PROBABILITIES OF TYPE MEMBERSHIP

1 2

PROBABILITIES OF TYPE MEMBERSHIP

1 2

PROBABILITIES OF TYPE MEMBERSHIP

1 2

INDIVIDUALS IN GROUP NUMBER 1

51	52	53	54	55	56	57	58	59	60	61	62	63	64	
65	66	67	68	69	70	71	72	73	74	75	76	77	78	
79	80	81	82	83	84	85	86	87	88	89	90	91	92	
93	94	95	96	97	98	99	100	101	102	103	104	105	106	
107	108	109	110	111	112	113	114	115	116	117	118	119	120	
121	122	123	124	125	126	127	128	129	130	131	132	133	134	
135	136	137	138	139	140	141	142	143	144	145	146	147	148	
149	150													

INDIVIDUALS IN GROUP NUMBER 2

1	2	3	4	5	6	7	8	9	10	11	12	13	14	
15	16	17	18	19	20	21	22	23	24	25	26	27	28	
29	30	31	32	33	34	35	36	37	38	39	40	41	42	
43	44	45	46	47	48	49	50	51						

THE CRITERION VALUE FOR THE INITIAL GROUPING FOR 3 TYPES IS .7815[illegible]

****

****

SAMPLE SIZE = 151
NUMBER OF VARIABLES = 4
NUMBER OF TYPES = 3

ITERATION NUMBER 4

LOGLIKELIHOOD OF 3 TYPES IN THIS SAMPLE = .327.[illegible]

CHARACTERISTICS OF THE WHOLE SAMPLE

MEANS

1	2	3	4
5.8433	3.0573	3.7581	1.1993

STANDARD DEVIATIONS

1	2	3	4
0.8281	0.4359	1.7655	0.7622

CORRELATIONS

1	2	3	4
1.0000	-0.1176	0.8713	0.8179
-0.1176	1.0000	-0.4284	-0.3661
0.8713	-0.4284	1.0000	0.9629
0.8179	-0.3661	0.9629	1.0000

CHARACTERISTICS OF TYPE 1

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .333

MEANS

1	2	3	4
5.0[illegible]	3.42[illegible]	1.463[illegible]	.246[illegible]

STANDARD DEVIATIONS

1	2	3	4
0.3525	0.3791	0.1737	0.1[illegible]54

CORRELATIONS

1	2	3	4
1.0[illegible]	0.7425	0.2677	0.2741
0.7425	1.0000	0.1777	0.[illegible]312
0.2677	0.1777	0.[illegible]	0.3312
0.2741	0.2328	0.3310	1.[illegible]

CHARACTERISTICS OF TYPE 2

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .412

MEANS

1	2	3	4
5.9316	2.7484	4.2935	1.4139

STANDARD DEVIATIONS

1	2	3	4
0.4664	0.2963	0.[illegible]99	0.2975

CORRELATIONS

1	2	3	4
1.0[illegible]	0.4619	0.3457	0.3[illegible]19
0.4619	1.0[illegible]	0.3[illegible]57	0.3[illegible]39
0.3457	0.3[illegible]57	1.0[illegible]	0.[illegible]117
0.2489	0.3289	0.8117	1.0[illegible]

CHARACTERISTICS OF TYPE 3

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .293

MEANS

1	2	3	4
6.85[illegible]	3.[illegible]737	5.7421	2.[illegible]711

STANDARD DEVIATIONS

1	2	3	4
0.4942	.3[illegible]291	0.4086	0.2799

CORRELATIONS

1	2	3	4
1.0[illegible]	0.1722	0.7363	0.[illegible]31
0.1722	1.0[illegible]	0.[illegible]462	0.[illegible]331
0.7363	0.[illegible]462	1.0[illegible]	0.1861
0.6[illegible]1	0.4231	0.1921	1.0[illegible]

****

ITERATION 1 LOG LIKELIHOOD OF 3 TYPES IN THIS SAMPLE = .[illegible]
ITERATION 2 LOG LIKELIHOOD OF 3 TYPES IN THIS SAMPLE = .[illegible]
ITERATION 3 LOG LIKELIHOOD OF 3 TYPES IN THIS SAMPLE = .[illegible]
ITERATION 4 LOG LIKELIHOOD OF 3 TYPES IN THIS SAMPLE = .[illegible]
ITERATION 5 LOG LIKELIHOOD OF 3 TYPES IN THIS SAMPLE = .[illegible]

1

ITERATION TIME = 1.57 SECONDS.
TOTAL TIME USED = 5.45 SECONDS.

****

SAMPLE SIZE = 151
NUMBER OF VARIABLES = 4
NUMBER OF TYPES = 3

ITERATION NUMBER 6

LOGLIKELIHOOD OF 3 TYPES IN THIS SAMPLE = .3622[illegible]

CHARACTERISTICS OF THE WHOLE SAMPLE

MEANS

1	2	3	4
5.8433	3.0573	3.7580	1.1993

STANDARD DEVIATIONS

1	2	3	4
0.8281	0.4359	1.7653	0.7632

CORRELATIONS

1	2	3	4
1.0000	-0.1176	0.8717	0.8179
-0.1176	1.0000	-0.4284	-0.3661
0.8717	-0.4284	1.0000	0.9629
0.8179	-0.3661	0.9629	1.0000

CHARACTERISTICS OF TYPE 1

THE PROPORTION OF THE POPULATION FROM THIS TYPE = 0.333

MEANS

1	2	3	4
5.0160	3.4280	1.4620	0.2460

STANDARD DEVIATIONS

1	2	3	4
0.3489	0.3753	0.1719	0.1163

CORRELATIONS

1	2	3	4
1.0000	0.7425	0.2672	0.2781
0.7425	1.0000	0.1777	0.2328
0.2672	0.1777	1.0000	0.3316
0.2781	0.2328	0.3316	1.0000

CHARACTERISTICS OF TYPE 2

THE PROPORTION OF THE POPULATION FROM THIS TYPE = 0.284

MEANS

1	2	3	4
5.9555	2.7741	4.3629	1.3954

STANDARD DEVIATIONS

1	2	3	4
0.5051	0.3132	0.5163	0.2472

CORRELATIONS

1	2	3	4
1.0000	0.5256	0.8054	0.6315
0.5256	1.0000	0.5165	0.5412
0.8054	0.5165	1.0000	0.7867
0.6315	0.5412	0.7867	1.0000

CHARACTERISTICS OF TYPE 3

THE PROPORTION OF THE POPULATION FROM THIS TYPE = 0.284

MEANS

1	2	3	4
6.6815	3.0685	5.6503	2.0697

STANDARD DEVIATIONS

1	2	3	4
0.6135	0.3244	0.5251	0.2659

CORRELATIONS

1	2	3	4
1.0000	0.4138	0.8644	0.7719
0.4138	1.0000	0.3698	0.5124
0.8644	0.3698	1.0000	0.9429
0.7719	0.5124	0.9429	1.0000

****

PROBABILITIES OF TYPE MEMBERSHIP

1 2 3

PROBABILITIES OF TYPE MEMBERSHIP

1 2 3

PROBABILITIES OF TYPE MEMBERSHIP

1 2 3

PROBABILITIES OF TYPE MEMBERSHIP

1 2 3

****

INDIVIDUALS IN GROUP NUMBER 1

1	2	3	4	5	6	7	8	9	10	11	12	13	14
15	16	17	18	19	20	21	22	23	24	25	26	27	28
43	44	45	46	47	48	49	50	51	52	53	54	55	56

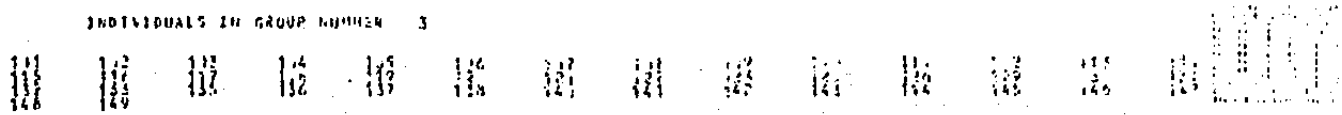
****

INDIVIDUALS IN GROUP NUMBER 2

57	58	59	60	61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80	81	82	83	84
99	100	101	102	103	104	105	106	107	108	109	110	111	112

****

INDIVIDUALS IN GROUP NUMBER 3



****

LOGARITHM OF LIKELIHOOD RATIO OF 3 TO 2 TYPES = .2565 732[illegible]2
CHI-SQUARE WITH 23 DEGREES OF FREEDOM= 49.25
PROBABILITY OF NULL HYPOTHESIS= 3.[illegible]733345

****

****

THE CRITERION VALUE FOR THE INITIAL GROUPING FOR 4 TYPES IS .5726[illegible]

****

****

SAMPLE SIZE = 151
NUMBER OF VARIABLES = 4
NUMBER OF TYPES = 4

ITERATION NUMBER 7

LOGLIKELIHOOD OF 4 TYPES IN THIS SAMPLE = .5[illegible]

CHARACTERISTICS OF THE WHOLE SAMPLE

MEANS

1	2	3	4
5.8433	3.0573	3.7581	1.1993

STANDARD DEVIATIONS

1	2	3	4
0.8281	1.4359	1.7651	.7622

CORRELATIONS

1	2	3	4
[illegible]	-.1176	-.8718	-.8179
-.1176	[illegible]	-.4284	-.3661
-.8718	-.4284	[illegible]	.9629
-.8179	-.3661	.9629	[illegible]

CHARACTERISTICS OF TYPE 1

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .[illegible]1

MEANS

1	2	3	4
7.0148	3.0963	5.9185	2.1556

STANDARD DEVIATIONS

1	2	3	4
.4809	.2131	.4541	.2342

CORRELATIONS

1	2	3	4
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]

CHARACTERISTICS OF TYPE 2

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .[illegible]

MEANS

1	2	3	4
[illegible]	[illegible]	[illegible]	[illegible]

STANDARD DEVIATIONS

1	2	3	4
[illegible]	[illegible]	[illegible]	[illegible]

CORRELATIONS

1	2	3	4
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]

CHARACTERISTICS OF TYPE 3

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .[illegible]

MEANS

1	2	3	4
[illegible]	[illegible]	[illegible]	[illegible]

STANDARD DEVIATIONS

1	2	3	4
[illegible]	[illegible]	[illegible]	[illegible]

CORRELATIONS

1	2	3	4
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]

[illegible]

THE PROPORTION OF THE POPULATION FROM THIS TYPE= 0.2[illegible]

MEANS

1	2	3	4
6.2644	2.8844	4.[illegible]67	1.6667

STANDARD DEVIATIONS

1	2	3	4
0.3206	[illegible].2878	[illegible].3468	[illegible].2492

CORRELATIONS

1	2	3	4
1.[illegible]	[illegible].297[illegible]	-[illegible].1331	-[illegible].35[illegible]3
[illegible].297[illegible]	1.[illegible]	-[illegible].[illegible]22	-[illegible].16[illegible]
-[illegible].1331	-[illegible].[illegible]22	1.[illegible]	[illegible].6232
-[illegible].3513	[illegible].1696	[illegible].6232	1.[illegible]

****

ITERATION 1 LOG LIKELIHOOD OF 4 TYPES IN THIS SAMPLE = .3[illegible]
ITERATION 2 LOG LIKELIHOOD OF 4 TYPES IN THIS SAMPLE = .36[illegible]
ITERATION 3 LOG LIKELIHOOD OF 4 TYPES IN THIS SAMPLE = .3[illegible]
ITERATION 4 LOG LIKELIHOOD OF 4 TYPES IN THIS SAMPLE = .37 [illegible]
ITERATION 5 LOG LIKELIHOOD OF 4 TYPES IN THIS SAMPLE = .37[illegible]

ITERATION TIME = 1.54 SECONDS.
TOTAL TIME USED = 1[illegible] SECONDS.

****

SAMPLE SIZE = 15[illegible]
NUMBER OF VARIABLES = 4
NUMBER OF TYPES = 4

ITERATION NUMBER 6

LOGLIKELIHOOD OF 4 TYPES IN THIS SAMPLE = .37[illegible]

CHARACTERISTICS OF THE WHOLE SAMPLE

MEANS

1	2	3	4
5.8423	3.[illegible]573	3.7[illegible]	1.1993

STANDARD DEVIATIONS

1	2	3	4
0.8281	0.4359	1.7653	[illegible].7622

CORRELATIONS

1	2	3	4
1.[illegible]	-[illegible].1176	[illegible].8718	[illegible].8179
-[illegible].1176	1.[illegible]	-[illegible].4284	-[illegible].3661
[illegible].8718	-[illegible].4284	1.[illegible]	[illegible].9629
[illegible].8179	-[illegible].3661	[illegible].9629	1.[illegible]

CHARACTERISTICS OF TYPE 1

THE PROPORTION OF THE POPULATION FROM THIS TYPE= 0.216

MEANS

1	2	3	4
6.[illegible]822	3.[illegible]666	5.8[illegible]28	2.151[illegible]

STANDARD DEVIATIONS

1	2	3	4
[illegible].5547	[illegible].30[illegible]2	[illegible].4859	[illegible].2299

CORRELATIONS

1	2	3	4
1.[illegible]	[illegible].1665	[illegible].8162	-[illegible].2483
[illegible].1665	1.[illegible]	[illegible].1242	[illegible].3432
[illegible].8162	[illegible].1242	1.[illegible]	-[illegible].31[illegible]
-[illegible].2483	[illegible].3432	-[illegible].31[illegible]6	1.[illegible]

CHARACTERISTICS OF TYPE 2

THE PROPORTION OF THE POPULATION FROM THIS TYPE= 0.333

MEANS

1	2	3	4
5.[illegible]	[illegible].426[illegible]	1.4621	[illegible].246[illegible]

STANDARD DEVIATIONS

1	2	3	4
[illegible].3489	[illegible].3753	[illegible].1719	[illegible].1[illegible]43

CORRELATIONS

1	2	3	4
1.[illegible]	[illegible].7425	[illegible].2[illegible]	[illegible].278[illegible]
[illegible].7425	1.[illegible]	[illegible].1777	[illegible].2[illegible]
[illegible].2[illegible]	[illegible].1777	1.[illegible]	[illegible].3[illegible]
0.27[illegible]	[illegible].2328	[illegible].3316	1.[illegible]

CHARACTERISTICS OF TYPE 3

THE PROPORTION OF THE POPULATION FROM THIS TYPE= 0.2[illegible]9

MEANS

1	2	3	4
5.[illegible]253	2.[illegible]976	4.[illegible]44	1.2[illegible]03

STANDARD DEVIATIONS

1	2	3	4
[illegible].3636	[illegible].2944	[illegible].42[illegible]8	[illegible].1929

CORRELATIONS

1	2	3	4
1.[illegible]	[illegible].[illegible]771	[illegible]	[illegible].2395
[illegible].5771	1.[illegible]	[illegible].6[illegible]7	[illegible].6115
[illegible].58[illegible]	[illegible].6[illegible]7	1.[illegible]	[illegible].71[illegible]
0.2395	[illegible].6614	[illegible].71[illegible]5	1.[illegible]

CHARACTERISTICS OF TYPE 4

THE PROPORTION OF THE POPULATION FROM THIS TYPE= 0.231

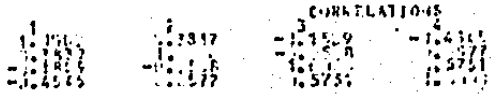
MEANS

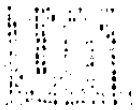
1	2	3	4
6.28[illegible]2	2.8369	4.8756	1.6273

STANDARD DEVIATIONS

1	2	3	4
[illegible]	[illegible]	[illegible]	[illegible]

CORRELATIONS





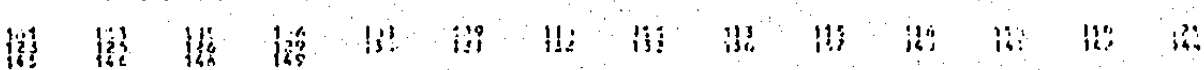
PROBABILITIES OF TYPE MEMBERSHIP
1 2 3 4

PROBABILITIES OF TYPE MEMBERSHIP
1 2 3 4

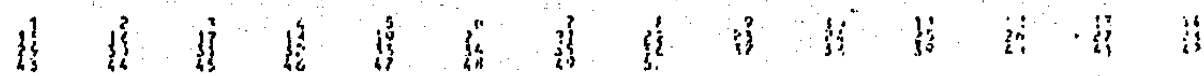
PROBABILITIES OF TYPE MEMBERSHIP
1 2 3 4

PROBABILITIES OF TYPE MEMBERSHIP
1 2 3 4

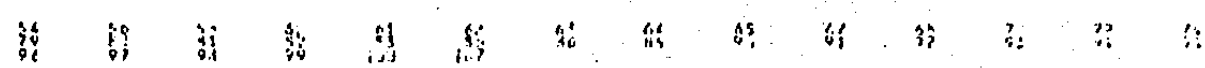
INDIVIDUALS IN GROUP NUMBER 1

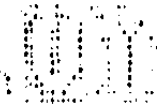


INDIVIDUALS IN GROUP NUMBER 2

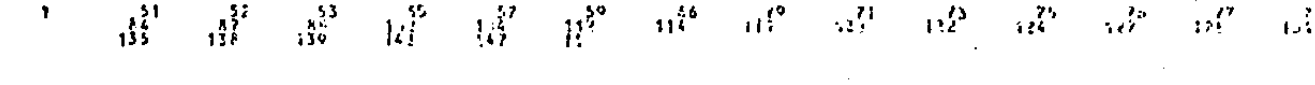


INDIVIDUALS IN GROUP NUMBER 3





INDIVIDUALS IN GROUP NUMBER 4



LOGARITHM OF LIKELIHOOD RATIO OF 4 TO 3 TYPES = -.4843892754E+1
CHI-SQUARE WITH 24 DEGREES OF FREEDOM = 19.992
PROBABILITY OF NULL HYPOTHESIS= 1.9013127E-1

BIBLIOGRAFIA

1. Bartlett M. S. (1939), "A note on tests of significance in multivariate Analysis", Proc. Camb. Vol 8, pp 376-86.
2. Beale E.M.L. (1969), Cluster Analysis. London: Scientific Control Systems.
3. Cohen A. (1967), "Estimation in mixtures of two normal distributions", Technometrics, Vol. 9, No. 1, pp 15-28.
4. Day E. (1969) "Estimating the components of a mixture of normal distributions". Biometrika, Vol 56, pp 463-72.
5. Dempster et Al. (1976), "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, pp 3 -40.
6. Everitt B., (1974), Cluster Analysis, SSRC, Heinemann, London.
7. Fisher R. A. (1936), "Multiple measurements in taxonomic problems", Annals of Eugenics, Vol VII, pp 179-88.
8. Friedman y Rubin (1967), "On some invariant criteria for grouping data", JASA, Vol. 62, pp 1159-78.
9. Gnanadesikan R. (1977), Methods for Statistical data analysis of multivariate observations" John Wiley & Sons New York, pp 82-120.
10. Hernández A. y Marriot F.H.C. (1981), "The theory of distributions mixtures", Comunicaciones técnicas del IIMAS, No. 228, Serie Naranja.

11. Hernández A. (1979), "Problems in Cluster Analysis" Ph. D. Thesis. Oxford University.

12. Hasselblad, V. (1966), "Estimation of parameters for a mixture of normal distributions, Technometrics, Vol. 8, No. 3, pp 431-46.

13. Kale B.K. (1962), "On the solution of likelihood equations by iteration processes. The multiparametric case", Biometrika, Vol. 49, pp 479-86.

14. Marriot F.H.C. (1971), "Practical problems in a method of cluster analysis". Biometrics, Vol. 27, pp 501-14.

15. Marriot F.H.C. (1975), "Separating mixtures of normal distributions", Biometrics, 31, pp 767-69.

16. Silvey S.D. (1959), "Lagrangian Multiple tests", Annals of Mathematical Statistics, Vol. 30, pp 389-407.

17. Wilks S.S. (1938), "A large sample distribution of the likelihood ratio for testing composite hypothesis", Annals of Mathematical Statistics, Vol. 9, pp 60-62.

18. Wolfe J.H. (1967). "Normix; computational methods for estimating the parameters of multivariate normal mixtures of distributions". Research memorandum, SRM 68-2, U.S. Naval Personnel Research Activity, San Diego.

19. Wolfe J.H. (1970), "Pattern clustering by multivariate mixture analysis", U.S. Naval Personnel Research Laboratory, pp 329-50.