
# Métricas entre modelos lineales y su aplicación al tratamiento de datos en medicina

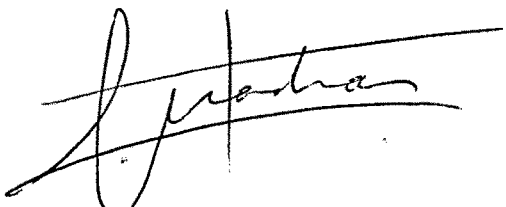
Martín Ríos Alcolea

# METRICAS ENTRE MODELOS LINEALES Y SU APLICACION AL TRATAMIENTO DE DATOS EN MEDICINA

Memoria presentada para
optar al título de
Doctor en Medicina, por

Martín Rios Alcolea

VºBº
EL DIRECTOR



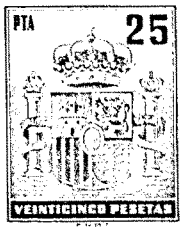
Prof. D. Carlos M. Cuadras Avellana,
Catedrático de Bioestadística.

Facultad de Biología

Universidad de Barcelona

Barcelona, 13 de Septiembre de 1.985

UNIVERSIDAD DE BARCELONA
FACULTAD DE BIOLOGIA
DPTO. BIOESTADISTICA




D. Carlos Cuadras Avellana, Catedrático numerario de la Facultad de Biología de la Universidad de Barcelona

CERTIFICA que la tesis presentada por D. Martín Ríos Alcolea bajo el título Métricas entre Modelos Lineales y su aplicación al tratamiento de datos en Medicina reune los requisitos exigidos de nivel científico y de nuevas contribuciones para ser considerada como tesis doctoral

Barcelona 15 de Octubre 1985

El director

A mi padre,

mi mejor maestro.

Ser consciente de la propia ignorancia

es un gran paso hacia el saber.

BENJAMIN DISRAELI

## AGRADECIMIENTOS

A mis compañeros, los profesores del Departamento de Bioestadística, que bajo la dirección del Prof. D. Carlos Cuadras, forman un equipo de jóvenes investigadores, cuya amistosa actitud de ayuda y colaboración ha hecho posible la realización de esta tesis. En este sentido, el Prof. José Mª Oller es objeto de especial reconocimiento por la cantidad de conocimientos que de él he aprendido.

A la Dra. Montse Millán por sus sugerencias, y al personal de Laboratorio de Hormonal del Hospital de San Juan de Dios por haberme proporcionado los datos que se usan en esta memoria.

Al personal del Centro de Cálculo de la U.B., a las mecanógrafas, Emilia, Mª Antonia, Ester y Pilar, y a las auxiliares del Hospital Clínico, Carmen e Isabel, por sus atenciones.

Finalmente, un cariñoso agradecimiento a Anna y a mis hijos Martín y Daniel, porque le dan sentido a las cosas.

INDICE GENERAL

# PROLOGO

Las numerosas consultas que he recibido de los profesionales de la medicina sobre comparación de curvas y sobre métodos estadísticos que permitieran clasificar a los individuos de una población afectados por un síndrome (Diabetes Melitus, Artritis Reumatoide, etc...), me motivaron a intentar desarrollar un método que ayudara a resolver satisfactoriamente estos problemas.

Para ello me plantee los siguientes objetivos:

a) Desarrollar una metodología estadística que me permitiera definir un índice (distancia) que expresara las diferencias y semejanzas entre las curvas.

b) Basándome en índices, como por ejemplo, el anteriormente definido, desarrollar una metodología que ofrezca al profesional de la medicina unas pautas a seguir ante el estudio de un síndrome del que sospecha que es la manifestación de distintas enfermedades, o variantes de una misma enfermedad que aun no están perfectamente definidas.

Así, siguiendo las investigaciones iniciadas por el Prof. Cuadras, en la comparación de modelos lineales con técnicas del Análisis de la Varianza y enlazando con la línea de investigación del Prof. Oller en el campo de la Geometría Riemanniana, aplicada al análisis de datos, he realizado este memoria en la que podemos distinguir dos partes:

En la primera parte (cap. 1 al 4) se define y estudia una distancia, entre modelos lineales univariantes y multivariantes, basada en el concepto de entropía o información de Shannon.

En la segunda parte, cap. 5 y cap. 6, se aplican los resultados obtenidos a la clasificación de una muestra de 171 niños, a los que se les practicó un Test de Tolerancia Oral a la Glucosa (TTOG) y se da un método diagnóstico, objetivo, de los grupos que aparecen en la clasificación.

# 1. GENERALIDADES

Página

Sumario:

## 1.1. INTRODUCCION

La actividad médica, tanto en la clínica como en la investigación, obliga a establecer analogías y diferencias entre individuos, poblaciones y objetos en general, caracterizados por un entorno social o familiar, unos hábitos, una determinada clínica, una pequeña muestra de un tejido, unas pruebas de laboratorio, etc. ...

El estudio de estas analogías y diferencias nos permite definir el concepto, la etiología, la patogenia, la patología y la clínica, específicas de una determinada enfermedad o síndrome. También nos permite establecer criterios para su diagnóstico, pronóstico y tratamiento.

Todo esto como sabemos es práctica habitual en el médico. Es por ello que resulta de interés estudiar de una forma cuantitativa estas diferencias y analogías, con el fin de dar criterios objetivos a la hora de definir, diagnosticar, aplicar tratamientos y dar un pronóstico sobre las enfermedades o poblaciones estudiadas.

Para estudiar las analogías y diferencias entre individuos u objetos en general, resulta de gran utilidad definir entre cada par de ellos un índice de desemejanza, de tal manera que valores grandes de dicho índice se correspondan a diferencias grandes entre individuos. Particularmente adecuado resulta el uso de distancias como índices de desemejanza. Esta cuantificación nos ayuda a poner de relieve características

difíciles de observar por la gran cantidad de información con que se nos presentan los resultados, tanto por los muchos individuos estudiados, como por la gran cantidad de variables observadas.

Una distancia definida entre objetos cualesquiera nos permite obtener una representación geométrica de los mismos en un espacio euclídeo de dimensión reducida, perceptible a nuestros sentidos, un plano generalmente. Esta representación se realiza siguiendo algún criterio de optimización, de forma que la distancia entre los puntos que representan a los individuos en el nuevo espacio sea la más parecida posible a la distancia originalmente definida. Esto lo podemos realizar mediante un Análisis de Coordenadas Principales o siguiendo diversas técnicas de "multidimensional sealing" (Cuadras (1981)). También podemos realizar una clasificación jerárquica de los individuos a partir de la distancia previamente definida, siguiendo los métodos propios de la Taxonomía Numérica (Benzecri (1973)).

Tras haber situado a los individuos en un espacio de dimensión reducida y una vez hechas las correspondientes clasificaciones, podemos buscar interpretaciones causales de estos resultados, relacionándolos con tipos distintos de enfermedades, con diferentes grados de una misma enfermedad o con todo aquello que pueda justificar la proximidad y la separación entre los distintos individuos.

Las distancias también nos permiten realizar un tipo de análisis conocido como análisis discriminante que consiste en asignar individuos a unos determinados grupos siguiendo algún criterio lógico, como sería incluir a un individuo estudiado en uno de los grupos establecidos siguiendo algún criterio de proximidad, por ejemplo asignándolo al grupo a cuyo individuo promedio sea más próximo.

## 1.2. CONCEPTO DE DISTANCIA. PROPIEDADES

Sea $\Omega = \{w_1, w_2, \ldots, w_n\}$ un conjunto cuyos elementos son objetos a distanciar. Decimos que d es una distancia definida en $\Omega$, si y sólo si d es una aplicación de $\Omega x \Omega$ en R que cumple las siguientes propiedades:

A) $d(w_i, w_j) \geq 0$ $\qquad w_1, w_j \in \Omega$

B) $d(w_i, w_j) = 0 \iff w_i = w_j$ (definida positiva)

C) $d(w_i, w_j) = d(w_j, w_i)$ $\qquad w_i, w_j \in \Omega$ (simétrica)

D) $d(w_i, w_k) \leq d(w_i, w_j) + d(w_j, w_k)$ $\qquad w_i, w_j, w_k \in \Omega$ (desigualdad triangular)

Cuando una aplicación cumple estas propiedades decimos que es una distancia métrica. Si en lugar de cumplir la propiedad B cumple la:

B') $w_i = w_j \Longrightarrow d(w_i, w_j) = 0$ (semidefinida positiva)

es decir, la implicación en un solo sentido, decimos que d es una distancia pseudométrica.

Si verifica además la propiedad,

E) $d(w_i,w_j) \leq \sup. \{d(w_i,w_k), d(w_k,w_j)\}$ $\quad w_i,w_j,w_k \in \Omega$

(desigualdad ultramétrica)

decimos que d es una distancia ultramétrica.

Si una distancia cumple además:

F) Que a cada par de elementos $w_i,w_j$ de $\Omega$ se le puede asociar un par de puntos $P_i,P_j$ de $R^m$, de coordenadas $(x_{i1},x_{12}\ldots, x_{im})$ y $(x_{j1},x_{j2}\ldots,x_{jm})$ tales que la distancia entre $w_i$ y $w_j$ coincide con la distancia euclídea ordinaria entre $P_i$ y $P_j$, es decir:

$$d(w_i,w_j) = d_{R^m}(P_i,P_j) = \sqrt{\sum_{h=1}^{m} (x_{ih} - x_{jh})^2} \qquad (1)$$

decimos que la distancia definida en $\Omega$ es euclídea.

Cuando se cumplen sólo las propiedades A, B', C, decimos que se trata de una disimilaridad o desemejanza.

Si tenemos una distancia pseudométrica puede haber individuos distintos cuya interdistancia sea cero, en este caso podemos definir una relación de equivalencia en $\Omega$, de forma que dos individuos estén relacionados si y sólo si la distancia entre ellos es cero. En el conjunto cociente resultante se induce de forma natural una distancia entre las clases de equivalencia. Es por esto que aunque en un conjunto tengamos

definida una pseudométrica podemos transformarla en una métrica mediante paso al conjunto cociente.

Destaquemos también que una distancia ultramétrica en $\Omega$ es equivalente a tener definida una jerarquía indexada en partes de $\Omega$, lo cual nos permite obtener clasificaciones jerárquicas.

## 1.3. CONSTRUCCION DE INDICES DE DESEMEJANZA

Como hemos visto en el punto anterior, cualquier aplicación que cumpla ciertas propiedades es una distancia. Se comprende pues que hay muchas formas de definir distancias y por consiguiente índices de desemejanza que sean menos restrictivos que una distancia, como son las disimilaridades. Veamos algunos métodos de construcción.

### 1.3.1. Construcción de una disimilaridad

Una disimilaridad es el índice más sencillo de los usados en el sentido de que sólo se cumplen las propiedades A, B', C, de tal manera, que no puede ser considerada como una distancia, sino sólo como un índice de desemejanza. Estos índices son útiles cuando queremos comparar individuos de una población que vienen caracterizados por variables dicotómicas, como pueden ser las variables que toman valores cero o uno según esté ausente o presente un carácter en el individuo en el cual está definida esta variable.

Las disimilaridades se definen en general a través de similaridades que no son más que una forma, como su nombre indica, de cuantificar la semejanza entre individuos en base a datos cualitativos. Veamos como podemos definir algunas similaridades.

Dados dos individuos de una población, $w_i, w_j$ y n variables aleatorias definidas sobre cada individuo que toman los valores cero o uno, según esté presente o ausente un determinado carácter, podemos formar una tabla de frecuencias:

| $w_i$ \ $w_j$ | 1 | 0 |
|---|---|---|
| 1 | $a_{ij}$ | $b_{ij}$ |
| 0 | $c_{ij}$ | $d_{ij}$ |

donde $a_{ij}$ es el número de caracteres presentes simultáneamente en los dos individuos, $d_{ij}$ el número de caracteres ausentes en los dos individuos, $b_{ij}$ el número de caracteres ausentes en $w_j$ y presentes en $w_i$, $c_{ij}$ el número de caracteres presentes en $w_j$ y ausentes en $w_i$, $n = a_{ij} + b_{ij} + c_{ij} + d_{ij}$ el número de variables aleatorias indicadores de los caracteres estudiados.

Una similaridad la podemos definir como una aplicación

$$s : \Omega \times \Omega \longrightarrow R$$

definida de la siguiente manera:

$$A)\ s_{ij} = s(w_i, w_j) = f_n(a_{ij}, b_{ij}, c_{ij}) \qquad w_i, w_j \in \Omega \tag{2}$$

con $s_{ij}$ no dependiendo de $d_{ij}$, pues añadiendo caracteres arbitrarios no comunes a $w_i$, $w_j$, podríamos falsear las similitudes entre los individuos, pues las dobles ausencias no suelen implicar analogías.

B) $f_n$ ha de ser monótona creciente en $a_{ij}$ pues al aumentar los caracteres comunes ha de aumentar el índice de similitud.

C) $f_n$ ha de ser monótona decreciente en $b_{ij}$ y $c_{ij}$ por razones obvias.

D) $f_n$ ha de ser simétrica en $b_{ij}$ y $c_{ij}$, es decir:

$$f_n\ (a_{ij},\ b_{ij},\ c_{ij}) = f_n\ (a_{ij},\ c_{ij},\ b_{ij})$$

A partir de una similaridad podemos definir siempre una disimilaridad como una aplicación

$$d : \Omega \times \Omega \longrightarrow R$$

de la siguiente manera:

$$d_{ij} = d(w_i,w_j) = \alpha_n - s\ (w_i,w_j) \qquad w_i,w_j \in \Omega \qquad (3)$$

donde $\alpha_n$ es el máximo de $f_n\ (\Omega)$ que existe, pues $f_n\ (\Omega)$ es finita. Podemos comprobar que está bien definida pues posee las propiedades A, B' y C.

La mayor parte de los coeficientes de similaridad están comprendidos entre 0 y 1, tomando el valor 0 cuando todos los caracteres de $w_i$ están ausentes en $w_j$ y el valor 1 cuando todos los caracteres de $w_i$ están en $w_j$ y viceversa.

A continuación se exponen algunos de los coeficientes de similaridad más usados:

$$s_1 = \frac{a+d}{a+b+c+d} \qquad \text{(Jokal y Michener)} \qquad (4)$$

$$s_2 = \frac{a+d}{a+2b+2c+d} \qquad \text{(Roger y Tanimoto)} \qquad (5)$$

$$s_3 = \frac{2a+2d}{2a+b+c+2d} \qquad \text{(Jokal y Sneath)} \qquad (6)$$

$$s_4 = \frac{a}{a+b+c} \qquad \text{(Jaccard)} \qquad (7)$$

$$s_5 = \frac{a}{a+b+c+d} \qquad \text{(Rusell y Rao)} \qquad (8)$$

Como podemos observar las disimilaridades obtenidas a partir de estas similaridades no tienen porqué tener la propiedad triangular.

Otra disimilaridad definida por Gower (1971) a través de las variables $X_1, X_2, \ldots, X_n$, cuantitativas y cualitativas, es:

$$d_{ij} = d(w_i,w_j) = 1/n \sum_{h=1}^{n} 1/R_h \mid x_{ih} - x_{jh} \mid \qquad (9)$$

donde $R_h$ es el recorrido de $X_h$.

### 1.3.2. Distancias definidas a través de una norma

Si existe una aplicación

$$f : \Omega \longrightarrow E$$

siendo E un espacio vectorial normado, podemos definir en $\Omega$ una distancia de la siguiente forma natural:

$$d(w_i, w_j) = \| f(w_i) - f(w_j) \|$$

$$w_i, w_j \in \Omega$$

donde $\| . \|$ simboliza la norma en E.

Si una norma cumple que

$$\|x+y\|^2 + \|x-y\|^2 = 2(\|x\|^2 + \|y\|^2) \qquad (10)$$

puede ser construida a partir de un producto escalar, es la conocida regla del paralelogramo (Collatz (1966)), por lo tanto toda distancia en $\Omega$ definida a través de dicha norma es euclídea en el sentido dado por la propiedad F. Esta es una condición suficiente para dicha propiedad, pero no necesaria pues es bien sabido que hay normas que no cumplen esta condición y sin embargo a través de ellas se pueden definir distancias euclídeas en $\Omega$, en el sentido de la propiedad D. Así, por ejemplo si tenemos un conjunto $\Omega$ con dos elementos, le podemos asociar dos puntos de un espacio normado que no cumplan la propiedad del paralelogramo y evidentemente esos dos puntos son encajables en el espacio euclídeo de dimensión 1, conservando la distancia primitiva.

Veamos algunas distancias definidas a través de una norma.

Si tenemos una población $\Omega$ y en ella tenemos definida n variables aleatorias $X_1$, $X_2$,....,$X_n$, podemos asociar a cada individuo $w_i$ de $\Omega$ un elemento de $R^n$ $x_i = (X_1(w_i), X_2(w_i),....,X_n(w_i)) = (x_{i1}, x_{i2},....,x_{in})$ y definir una distancia entre $w_i$ y $w_j$ de la siguiente forma:

$$d(w_i,w_j) = \|x_i - x_j\| = \sqrt{(x_i-x_j)^t (x_i-x_j)} \qquad (11)$$

que es la distancia euclídea normal entre dos puntos de $R^n$ de coordenadas $x_i$ y $x_j$.

Esta es la distancia que usamos en el Análisis de Componentes Principales (Pearson (1901)). Tiene dos inconvenientes fundamentales: primero que es sensible a cambios de escala de las variables, este problema se puede solucionar tipificandolas; segundo que no tiene en cuenta la posible dependencia entre ellas, es por ello que esta distancia se usa cuando es desconocida la matriz de varianzas-covarianzas.

Se han definido distancias que tratan de subsanar estos problemas, pero es Mahalanobis (1936) quien introduce una distancia que es casi siempre la más adecuada para distanciar poblaciones estadísticas en las que haya definidas $X_1$, $X_2$,....,$X_n$ variables aleatorias que se distribuyen según una normal conjunta, bajo la hipótesis de que todas las poblaciones tengan la matriz de varianzas-covarianzas, $\Sigma$ común.

Sean $w_i$ y $w_j$ dos poblaciones estadísticas y $M_i$ y $M_j$ los vectores columna n-dimensionales, cuyas componentes son las esperanzas de las variables aleatorias definidas en $w_i$ y $w_j$, si $\Sigma$ es la matriz de varianzas-covarianzas de las variables, común a ambas poblaciones, entonces la distancia de Mahalanobis entre $w_i$ y $w_j$ viene definida por:

$$d(w_i, w_j) = \sqrt{(M_i - M_j)^t \Sigma^{-1} (M_i - M_j)} \qquad (12)$$

En el caso de que la matriz $\Sigma$ sea singular $\Sigma^{-1}$ se sustituirá por una matriz inversa generalizada $\Sigma^-$, siendo la distancia independiente de la inversa generalizada usada. En el caso de que las medias teóricas de las variables definidas en $w_i$, $w_j$ sean desconocidas se sustituirán por sus estimaciones máximo-verosímiles y cuando $\Sigma$ sea desconocida la sustituiremos por una estimación insesgada de la misma.

Las propiedades más importantes de la distancia de Mahalanobis son:

1) Es invariante por transformaciones lineales no singulares de las variables, en particular es invariante por cambios de escala.

2) Si indicamos por $D_n$ la distancia referida a las variables $X_1, X_2, \ldots, X_n$ y $D_m$ la distancia referida a las variables $Y_1, Y_2, \ldots, Y_m$, entonces $D_{n+m}$, la distancia calculada con todas las variables $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$, verifica,

$$D^2_{n+m} \geq \max.(D^2_n, D^2_m) \qquad (13)$$

En el caso de que las $X_1, X_2, \ldots, X_n$ sean independientes de las $Y_1, Y_2, \ldots, Y_m$, se verifica que $D^2_{m+n} = D^2_n + D^2_m$.

3) Cuando la distribución de las variables es la de una normal multivariante, el estadístico

$$V = \frac{N_i\, N_j\, (N_i + N_j - n - 1)}{n\, (N_i + N_j)\, (N_i + N_j - 2)}\, \hat{D}^2 \qquad (14)$$

donde $\hat{D}^2$ es la estimación de la distancia al cuadrado, sigue cuando $M_i = M_j$, una distribución F con n y $N_i+N_j-n-1$ grados de libertad, donde $N_i$ y $N_j$ son los tamaños muestrales para las poblaciones $w_i$ y $w_j$ y n el número de variables observadas.

La distancia de Mahalanobis juega un papel fundamental en el Análisis Multivariante, principalmente en el Análisis Canónico de Poblaciones y Análisis Discriminante.

La distancia de Mahalanobis es la distancia más importante definida a través de una norma y que además está asociada a un producto escalar, es por lo tanto euclídea en el sentido de la propiedad F.

La distancia de Mahalanobis puede usarse también aunque las poblaciones no sean normales multivariantes o no sean comunes las matrices de varianzas-covarianzas, pero entonces,

generalmente son desconocidas las distribuciones de las estimaciones de dicha distancia.

Siguiendo la misma filosofía de definir distancias a través de normas, podemos definir la distancia de Minkowski, que está basada en la norma de Minkwoski

$$\|x\| = \left[\sum_{h=1}^{n} |x_h|^r\right]^{1/r} \qquad x \in R^n \tag{15}$$

y su expresión es:

$$d(w_i,w_j) = \left[\sum_{h=1}^{n} |x_{ih} - x_{jh}|^r\right]^{1/r} \qquad r \in N \tag{16}$$

Para el caso r = 1, la expresión de la distancia es:

$$d(w_i,w_j) = \sum_{h=1}^{n} |x_{ih} - x_{jh}| \tag{17}$$

que es la llamada distancia ciudad y fue usada por Prevosti et al (1975) en aplicaciones a la genética.

La distancia de Minkowski para el caso r = 2 es la distancia euclídea ordinaria. Para el caso r ≠ 2 la distancia no es euclídea puesto que la norma que la define no puede expresarse a través de un producto escalar, pues no cumple la regla del paralelogramo (Cuadras (1981)).

Otro ejemplo de distancia definida a través de normas es la distancia de Tchebychef, cuya expresión es:

$$d(w_i,w_j) = \max_h. \ |x_{ih} - x_{jh}| \qquad (18)$$

la norma usada para esta definición es:

$$\|x\| = \max_h. \ |x_h| \qquad x \in R^n \qquad (19)$$

que se puede considerar como un caso particular de la distancia o norma de Minkowski, cuando $r \to \infty$.

Hemos dicho que una distancia definida a través de una norma que venga definida por un producto escalar es una condición suficiente pero no necesaria para que la distancia definida en $\Omega$ verifique la propiedad F. Veamos seguidamente una condición necesaria y suficiente para que una distancia definida en $\Omega$ cumpla dicha propiedad.

Sea d una distancia definida en una población finita $\Omega = \{w_1,w_2,\ldots,w_n\}$; definamos una matriz $A = (a_{ij})$, con $a_{ij} = -\frac{1}{2} d_{ij}^2$ y $d_{ij} = d(w_i,w_j)$, evidentemente A será una matriz simétrica de orden n y con ceros en la diagonal principal, si además llamamos $I_n$ a la matriz identidad de orden n y a E un vector columna de n unos, $E = (1,1,\ldots,1)^t$, entonces la matriz $B = HAH$, con $H = I_n - 1/n \ EE^t$ es semidefinida positiva de rango $m \leq n-1$, si y solo si la distancia cumple la condición de euclideanidad en el sentido de la propiedad F, es decir es posible encontrar $P_i$ $(i=1,2,\ldots,n)$ puntos de $R^m$, de coordenadas conocidas, tal que $d_{R^m}(P_i,P_j) = d(w_i,w_j)$ con $d_{R^m}$ la distancia euclídea definida en $R^m$. Las coordenadas de

$P_i$ se llaman coordenadas principales para la distancia d, definida en $\Omega$ (Cuadras (1981)).

### 1.3.3. Identificación de individuos como puntos de una variedad riemanniana

Otra forma de definir una distancia consiste en identificar a los individuos de una población $\Omega$ con los puntos de una variedad riemanniana M, definiendo la distancia entre los individuos $w_i$ y $w_j$ de $\Omega$ como la distancia riemanniana entre los puntos a los que están asociados. En otras palabras, podemos construir una distancia en la población $\Omega$ definiendo una aplicación $\alpha$ inyectiva de $\Omega$ en M,

$$\alpha : \Omega \longrightarrow M$$

con

$$\alpha (w_i) = P_i, \; w_i \in \Omega \quad ; \quad P_i \in M$$

entonces la distancia en $\Omega$, d, se obtiene a partir de:

$$d(w_i,w_j) = d_M (P_i,P_j) \tag{20}$$

siendo $d_M$ la distancia definida en la variedad riemanniana.

Si la variedad es euclídea, esto es, si existe una transformación admisible de coordenadas tal que reduzca el campo tensorial métrico a un campo tensorial constante, entonces la

distancia es euclídea en el sentido de que viene definida por un producto escalar. Esto en general no se cumple si la variedad no es euclídea.

Una medida de la euclideanidad de la variedad nos la da la curvatura riemanniana que es nula cuando el espacio es euclídeo.

Como ejemplo de distancias definidas por este método están las distancias definidas a través del concepto de información o entropía. Este método fue introducido por Burbea y Rao (1982) que definen el funcional entropía, $H_\phi$, mediante la integral de una función $\phi$, real, de valores no negativos y de clase $C^{(2}$. El funcional entropía $H_\phi$ viene dado por la expresión:

$$H_\phi(f) = - \int_A \phi(f)\, d\mu \qquad (21)$$

donde $\mu$ es una medida $\sigma$-finita y aditiva definida sobre una $\sigma$-algebra de subconjuntos del espacio medible A, f es una función de densidad paramétrica definida sobre A.

A partir del funcional $H_\phi$ se puede definir una métrica riemanniana entre funciones $f = f(x,\theta)$, con $\theta = ( \theta_1, \theta_2, \ldots, \theta_n)$, cada una de las cuales representaría una población a distanciar y donde el tensor métrico vendría definido por:

$$g_{ij} = \int_A \phi''(f) \quad (D_i(f)) \quad (D_j(f)) \quad d\mu \qquad (22)$$

siendo $D_i(f)$ y $D_j(f)$ las derivadas parciales de f respecto de $\theta_i$ y $\theta_j$ respectivamente. El elemento de línea es:

$$ds = \sqrt{\sum_{i=1}^{h} \sum_{j=1}^{h} g_{ij} \; d\theta_i \; d\theta_j} \qquad (23)$$

En la actualidad se investigan métricas definidas a través de las funciones $\phi$

$$\phi_\alpha = \begin{cases} (\alpha - 1)^{-1}(x^\alpha - x) & \text{si } \alpha \neq 1 \\ x \ln x & \text{si } \alpha = 1 \end{cases} \qquad (24)$$

con ellas se trata de definir distancias entre funciones de densidad paramétrica.

Para el caso $\alpha = 1$ la función $\phi$ viene definida por $\phi(x) = x \ln x$, el funcional $H_\phi$ que obtenemos es el funcional de entropía o información de Shannon. En este caso, las componentes del tensor métrico para una determinada familia de funciones de densidad paramétrica que obtenemos a través de este funcional H, coinciden con las componentes de la matriz de información de Fisher.

Suponiendo la clase de funciones de densidad normales multivariantes con matriz de varianzas-covarianzas constantes, la matriz de información de Fisher coincide con la matriz inversa de varianzas-covarianzas y por consiguiente la distancia definida a través de la matriz de información de Fisher coincide con la distancia de Mahalanobis.

Con la matriz de información de Fisher se han obtenido distancias para la familia de distribuciones multinomiales (Bhattachanyya (1946)) y para la familia de distribuciones multinomiales negativas (Oller y Cuadras (1985)), entre otros casos.

## 1.4. CONCLUSIONES

Esta breve introducción pone de manifiesto la gran cantidad de índices y distancias definidas en la literatura científica, todas ellas correctas desde el punto de vista matemático, como ya quedó claro desde los trabajos de Lobachewsky y Boylai en el siglo pasado, donde se mostraba la posibilidad de construir varias geometrías distintas de la euclídea, todas ellas formalmente válidas.

Cuando nos planteamos la necesidad de definir una distancia para analizar datos concretos, se nos presenta el problema de qué distancia elegir. Al decidirnos por una distancia determinada corremos el riesgo de que ésta se elija por permitirnos interpretar los resultados de acuerdo con las ideas que se tenían ya preconcebidas, en cuyo caso la distancia no resulta de mucha utilidad pues no nos aporta, esencialmente, más información de la que ya poseíamos. También puede ocurrir que se defina una distancia por ser la que más conclusiones nos permite obtener. Este punto de vista, aunque quizás defendido por algunos desde un punto de vista pragmático, no parece demasiado satisfactorio, ya que esta forma de proceder puede introducir resultados que no se fundamenten con la realidad.

Mientras que el problema de la elección de la distancia más apropiada en la variedad espacio-tiempo, se puede resolver a nivel experimental, ya que el espacio físico es perceptible a través de nuestros sentidos, los datos con los que se trabajan en las Ciencias Biológicas, en particular en Medicina, al no poder ser observados directamente por aquellos, no nos es posible comprobar empíricamente si la geometría empleada es la adecuada.

Es ciertamente difícil dar un criterio para definir la distancia más adecuada, no obstante se pueden apuntar algunas propiedades que sería aconsejable poseyeran las distancias. Como ya hemos dicho una distancia sirve para estudiar las analogías y las diferencias entre los objetos distanciados, es por ello que una distancia debería estar basada en la información que sobre las analogías y diferencias entre los objetos poseyéramos. Esto significa que la distancia no solo dependería del objeto observado y del observador, sino que dependería fundamentalmente de las propiedades formales del proceso de observación, que vendrían caracterizadas por la función de densidad o distribución involucrados en dicho proceso. A su vez, la información viene definida matemáticamente a través de un funcional de las mismas. (Shannon (1948), Burbea (1982)).

También sería deseable que la distancia poseyera buenas propiedades matemáticas de invarianza; debería ser invariante frente a transformaciones admisibles de las variables aleatorias, ya que una transformación admisible de las mismas no de

be alterar la información que poseemos sobre el sistema observado, al no ser más que en el fondo un cambio de escala y debería ser invariante frente a transformaciones admisibles de parámetros, ya que la misma no altera la distribución de las variables, al ser sólo un cambio en la forma de referenciar la función de distribución. También puede ser interesante que la independencia estocástica fuera interpretable en términos geométricos, como ocurre en el caso de la distancia de Mahalanobis y distribución normal multivariante de las variables, donde la independencia estocástica es equivalente a la ortogonalidad.

Aún con todo, la elección de la distancia más apropiada resulta difícil, pues puede haber muchas distancias que encajen con las ideas anteriores, en este caso parece lógico decidirnos por aquella distancia que tenga un tratamiento matemático más sencillo.

En esta memoria se ha desarrollado una distancia entre modelos lineales normales, tanto para el caso univariante, como para el caso multivariante, con matriz de varianzas-covarianzas constante para el caso multivariante y constante y variable para el caso univariante, usando la distancia inducida por la matriz de información de Fisher. que puede ser definida a partir del concepto de información de Shannon. Además la distancia así obtenida, posee propiedades matemáticas que es conveniente satisfagan las distancias, como es, invarianza frente a transformaciones admisibles de variables aleatorias

y parámetros y relación entre independencia estocástica y ortogonalidad.

Los estimadores de estas distancias siguen distribuciones conocidas, por lo que se han empleado para comparar modelos lineales mediante contrastes de hipótesis de la forma:

$H_o$: Los modelos a comparar son iguales si y sólo si la suma de las interdistancias entre ellos es nula.

$H_1$: Los modelos a comparar son distintos si y sólo si la suma de las interdistancias entre ellos es mayor que cero.

Finalmente se han usado estos resultados para distanciar curvas de glucemia e insulina en niños a los que se les realizó una TTOG, explicando la utilidad de estos resultados en la clasificación y diagnóstico de diabetes y situaciones prediabéticas.

# 2. INTRODUCCION A LOS MODELOS LINEALES, GEOMETRIA RIEMANNIANA Y ANALISIS DE DATOS

Página

Sumario:

Este capítulo está dividido en tres partes, en la primera se describen los métodos estadísticos referidos a modelos lineales, en la segunda se hace una breve reseña de la geometría riemanniana y en la tercera se citan brevemente las técnicas de tratamiento de datos: Análisis de Coordenadas Principales y Taxonomía Numérica.

## 2.1. MODELOS LINEALES

A continuación se exponen definiciones y resultados de teoremas referidos a la teoría de modelos lineales, el tema puede ampliarse viendo Scheffé (1959), Searle (1971) o Seber (1977).

### 2.1.1. Introducción

Cuando tenemos un grupo de variables aleatorias y queremos encontrar una relación entre las mismas que nos permita predecir el valor de una de ellas conocidas las demás, se construyen modelos que, en general, contienen una parte determinista, controlable por el experimentador y una parte aleatoria que depende de otras causas no controlables.

Matemáticamente estos modelos vienen expresados por

$$y = f(x, \beta) + e \qquad (1)$$

donde y es la variable observada, $f(x, \beta)$ es una función que representa la parte determinista y e es una variable aleatoria que representa la parte no controlada o error del modelo.

## 2.1.2. Modelos lineales univariantes

Seguidamente, vamos a desarrollar el caso particular en el cual la variable observada es una variable aleatoria univariante.

### 2.1.2.1. Definiciones

Un modelo lineal es una ecuación matricial del tipo,

$$y = X\beta + e \qquad (2)$$

donde las cuatro matrices representan lo siguiente:

a) y es un vector columna $y = (y_1, y_2, \ldots, y_n)^t$, cuyas componentes son una muestra aleatoria simple de la variable aleatoria observada.

b) X es una matriz, $X = (x_{ij})$; $1 \leq i \leq n$; $1 \leq j \leq m$, cuyos elementos propone el investigador y que depende de las condiciones experimentales. A esta matriz la llamamos matriz de diseño.

c) $\beta$ es un vector columna $\beta = (\beta_1, \beta_2, \ldots, \beta_m)^t$, cuyas componentes son los parámetros desconocidos y que llamaremos parámetros de la regresión.

d) e, es un vector columna $e = (e_1, e_2, \ldots, e_n)^t$, cuyas componentes son variables aleatorias que representan los errores o desviaciones aleatorias de las observaciones.

Debemos suponer que $e_i$, $1 \leq i \leq n$ son variables aleatorias estocasticamente independientes de media 0 y varianza constante e igual a $\sigma^2$. Si además $e_i$ sigue una distribución normal, llamamos al modelo lineal, modelo lineal normal.

Supongamos que existen K condiciones experimentales distintas y que tenemos $n_i$ réplicas de la condición experimental i-ésima, entonces, para simplificar los cálculos, en lugar de usar la matriz de diseño X podemos usar la matriz $X_r$ que resulta de eliminar las filas repetidas y que llamaremos matriz de diseño reducida.

En caso de usar la matriz reducida el vector y de observaciones lo sustituiremos por el vector columna $\bar{y} = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k)^t$, cuyas componentes son las medias de las observaciones dentro de cada condición experimental.

A los números de réplicas $n_1, n_2, \ldots, n_k$ los disponemos en una matriz diagonal

$$D = \text{dig}\ (n_1, n_2, \ldots, n_k)$$

$$n = n_1 + n_2 + \ldots + n_k$$

Si $n_1 = n_2 = \ldots = n_k$, al modelo se le llama balanceado.

Al rango de la matriz de diseño le llamaremos rango del diseño. Si el rango del diseño es igual al número de parámetros, entonces diremos que el diseño es de rango máximo.

2.1.2.2. Estimación de parámetros

La estimación de los parámetros la realizamos por el método de los mínimos cuadrados, calculando los valores $\beta = (\beta_1, \beta_2, \ldots, \beta_m)^t$, que hagan mínima la función

$$L(\beta) = (y-X\beta)^t (y-X\beta) \tag{3}$$

La estimación mínima cuadrática de los parámetros del diseño es la solución del sistema de ecuaciones

$$X^t X \hat{\beta} = X^t y \tag{4}$$

este sistema es equivalente al

$$X_r^t D X_r \hat{\beta} = X_r^t D \bar{y} \tag{5}$$

A los sistemas (4) y (5) les llamamos sistemas de ecuaciones normales.

Si el diseño es de rango máximo, la estimación mínima cuadrática de los parámetros es única e igual a

$$\hat{\beta} = (X^t X)^{-1} X^t y \tag{6}$$

ó

$$\hat{\beta} = (X_r^t D X_r)^{-1} X_r^t D \bar{y} \tag{7}$$

Si el rango de la matriz de diseño es menor que el número de parámetros, las estimaciones son

$$\hat{\beta} = (X^t X)^- X^t y \tag{8}$$

ó

$$\hat{\beta} = (X_r^t \, D \, X_r)^- \, X_r^t \, D \, \bar{y} \qquad (9)$$

donde $(X^t X)^-$, $(X_r \, D \, X_r)^-$ son respectivamente las g-inversas de $(X^t X)$, $(X_r^t \, D \, X_r)$.

Cuando los modelos son lineales normales, la estimación mínimo cuadrática de β coincide con la estimación máximo verosímil.

2.1.2.3. Estimación de la varianza

Se llama suma de cuadrados residual, $R_o^2$, a la norma del vector $y - X\hat{\beta}$, siendo X la matriz de diseño, es decir,

$$R_o^2 = (y-X\hat{\beta})^t \, (y-X\hat{\beta}) \qquad (10)$$

A partir de la suma de cuadrado residual podemos obtener un estimador insesgado de la varianza del modelo.

$$\hat{\sigma}^2 = R_o^2 \, / \, (n-r) \qquad (11)$$

donde r es el rango del diseño y n el tamaño de la muestra.

Si el modelo es lineal normal, podemos demostrar que el estadístico $R_o^2 \, / \, \sigma^2$, sigue una distribución ji - cuadrado con n-r grados de libertad (Scheffé (1959)).

2.1.2.4. Hipótesis lineales. Condición de demostrabilidad

Una hipótesis lineal sobre los parámetros de un diseño $\beta = (\beta_1, \beta_2, \ldots, \beta_m)$ es una restricción lineal de los parámetros definida por una ecuación del tipo

$$H\beta = 0 \qquad (12)$$

donde $H = (h_{ij})$ $1 \leq i \leq t$; $1 \leq j \leq m$ y $\beta$ es el vector columna de los parámetros.

Decimos que una hipótesis lineal es demostrable cuando se cumple que el subespacio $V = \{u \in R^m / u = X\beta, \beta \in R^m, H\beta = 0\}$ está contenido estrictamente un $U = \{u \in R^m / u = X\beta\}$.

Una condición suficiente para que una hipótesis lineal sea demostrable es que el subespacio vectorial engendrado por las filas de H este contenido en el subespacio vectorial engendrado por las filas de X.

Una condición necesaria y suficiente para que una hipótesis lineal sea demostrable, es que alguna combinación lineal de las filas de H no nula sea combinación lineal de las filas de X (Scheffé, 1959).

2.1.2.5. Teorema fundamental del Análisis de la Varianza

Cuando proponemos una hipótesis lineal demostrable $H\beta = 0$, con rango H = t, el modelo (2) se puede reparametrizar

obteniéndose una nueva matriz de diseño X*, cuyos vectores columna engendran un subespacio vectorial propio del subespacio vectorial engendrado por los vectores columna de la matriz X. La expresión de este nuevo modelo es:

$$y = X^* \theta + e \qquad (13)$$

siendo,

$$\theta = (\theta_1, \theta_2, \ldots, \theta_p) \quad p = r-t \qquad (14)$$

los parámetros del modelo reparametrizado. Igual que en el modelo primitivo se pueden obtener las estimaciones de los parámetros $\hat{\theta}$ a través de las ecuaciones normales, cuyas soluciones son:

$$\hat{\theta} = (X^{*t} X^*)^{-1} X^{*t} y \quad \acute{o} \quad \hat{\theta} = (X_r^{*t} D X_r^*)^{-1} X_r^{*t} D \bar{y} \qquad (15)$$

siendo $X_r^*$ la matriz de diseño reducida del nuevo modelo. En este caso la matriz inversa de $X^{*t} X^*$ ; $X_r^{*t} D X_r$, siempre existe, ya que por construcción el nuevo diseño es de rango máximo.

La suma de cuadrados residual bajo $H_o$ será:

$$R_1^2 = (y - X^* \hat{\theta})^t (y - X^* \hat{\theta}) \qquad (16)$$

Si el modelo es lineal normal y se cumple $H_o$, los estadísticos $R_1^2 / \sigma^2$ y $(R_1^2 - R_o^2) / \sigma^2$, siguen una distribución ji-cuadrado con n-p y t grados de libertad respectivamente,

siendo p el rango de la matriz X* que como vemos en (14) es igual al número de parámetros del nuevo modelo.

Por otra parte, también bajo $H_o$, $R_1^2 - R_o^2$ y $R_o^2$ son independientes.

Como consecuencia de esto, se tiene que el estadístico,

$$F = \frac{(R_1^2 - R_o^2)/t}{R_o^2 / (n-r)} \qquad (17)$$

sigue una distribución F de Fisher-Snedecor con t y n-r grados de libertad.

2.1.2.6. Tabla general del Análisis de la Varianza

Si queremos contrastar la hipótesis demostrable $H_o$, con nivel de significación $\varepsilon$, seguiremos el criterio de decisión que nos proporciona la tabla general del análisis de la varianza (Searle, 1971).

Tabla

| | Grados de Libertad | Suma de Cuadrados | Cuadrados medios |
|---|---|---|---|
| Desviación de $H_o$ | t | $R_1^2 - R_o^2$ | $(R_1^2 - R_o^2)/t$ |
| Residuo | n-r | $R_o^2$ | $R_o^2/(n-r)$ |

Si el cociente (17) es mayor que K, siendo K tal que $P(F_\varepsilon > K) = \varepsilon$ y $F_\varepsilon$ una variable aleatoria F de Fisher-Snedecor

con t y n-r grados de libertad, entonces se rechaza la hipótesis $H_o$ y en caso contrario se acepta.

### 2.1.3. Modelos lineales multivariantes

En este punto vamos a desarrollar el caso en que la variable aleatoria observada, es multivariante.

#### 2.1.3.1. Definiciones

Un modelo lineal multivariante viene definido por una ecuación matricial del tipo

$$Y = XB + E \qquad (18)$$

donde las cuatro matrices representan lo siguiente:

a) Y es una matriz $Y = (y_{ij})$; $1 \leq i \leq n$; $1 \leq j \leq p$, cuyas columnas son muestras aleatorias simples de p variables aleatorias $Y_i$ que llamamos variables observables.

b) X es una matriz $X = (x_{ij})$; $1 \leq i \leq n$; $1 \leq j \leq n$, cuya construcción y significado es el mismo que en el caso univariante. También la llamamos matriz de diseño.

c) B es una matriz $B = (b_{ij})$; $1 \leq i \leq m$; $1 \leq j \leq p$, cuyos vectores columna $B_i$ tienen el mismo significado que en el caso univariante para la variable $Y_i$ observada.

d) E es una matriz $E = (e_{ij})$; $1 \leq i \leq n$; $1 \leq j \leq p$, cuyos vectores columna $E_i$ son variables aleatorias que representan los errores o desviaciones aleatorias de las observaciones correspondientes a la variable $Y_i$.

Se supone que las filas de E son variables aleatorias de esperanza $O = (o,o,\ldots,o)$ y con matriz de varianzas covarianzas $\Sigma$, común para todas las filas. Si además cada fila sigue una distribución normal multivariante $N(O, \Sigma)$, el modelo se llama lineal normal multivariante.

Si sólo existen K condiciones experimentales distintas con $n_i$ réplicas de la condición experimental i-ésima, entonces podemos, como en el caso univariante, usar en lugar de la matriz de diseño X, la matriz de diseño reducida $X_r$.

En su caso, la matriz Y de observaciones, la sustituiremos por la matriz $\overline{Y}$, cuyas filas son los vectores de componentes, las medias de las variables muestrales de cada condición experimental.

Los números de réplicas $n_1, n_2, \ldots, n_k$, se disponen como en el caso univariante en una matriz diagonal,

$$D = \text{diag.}\ (n_1, n_2, \ldots, n_k)\ ;\ n = n_1+n_2+\ldots+n_k$$

Cuando $n_1 = n_2 = \ldots = n_k$, al diseño también se le llama balanceado. Si el rango de X es igual al número de filas de B se le llama diseño de rango máximo.

2.1.3.2. Estimación de los parámetros

La estimación de los parámetros de la matriz B la realizamos estimando sus vectores columna $B_i$, que los podemos considerar como los vectores paramétricos de los modelos lineales univariantes

$$Y_i = X\, B_i + E_i \qquad (19)$$

que componen el modelo lineal multivariante. Las estimaciones de los parámetros son las soluciones de las ecuaciones normales

$$X^t X\, B = X^t Y$$

ó (20)

$$X_r^t D\, X_r\, B = X_r^t D\, \bar{Y}$$

en caso de usar la matriz de diseño reducida.

Si el rango del diseño es máximo, la estimación de la matriz de parámetros B es:

$$\hat{B} = (X^t X)^{-1} X^t Y$$

ó (21)

$$\hat{B} = (X_r^t D\, X_r)^{-1} X_r^t D\, \bar{Y}$$

si usamos la matriz de diseño reducida. En caso contrario, en las expresiones (21) pondremos, en lugar de las matrices inversas, las inversas generalizadas $(X^t X)^-$ ; $(X^t D\, X_r)^-$ en la primera y segunda ecuación respectivamente.

2.1.3.3. Estimación de la matriz de covarianzas

Si consideramos al modelo lineal multivariante

$$Y = X\,B + E \tag{22}$$

como p modelos univariantes

$$Y_i = X\,B_i + E_i \tag{23}$$

la suma de cuadrados residual para cada uno de estos es:

$$R_o^2\,(i) = R_o\,(i,i) = (Y_i - X\,\hat{B}_i)^t\,(Y_i - X\,\hat{B}_i) \tag{24}$$

Si además a la suma de productos cruzados lo denótamos por $R_o\,(i,j)$, tenemos que

$$R_o\,(i,j) = (Y_i - X\,\hat{B}_i)^t\,(Y_i - X\,\hat{B}_j) \tag{25}$$

A la matriz

$$\begin{aligned} &R_o = (r_{ij}) \\ &r_{ij} = R_o\,(i,j) \\ &1 \leq i,j \leq p \end{aligned} \tag{26}$$

la llamamos matriz residual del modelo lineal multivariante.

Por otra parte, como ya hemos visto en 2.1.2.3., $R_o\,(i,i)/(n-r)$ es un estimador de la varianza de $Y_i$. Además, es fácil ver que $R_o\,(i,j)/(n-r)$ es un estimador insesgado de la covarianza de $Y_i$, $Y_j$ y la matriz

$$\hat{\Sigma} = \frac{1}{n-r} \; R_o \qquad (27)$$

es una estimación insesgada de la matriz de covarianzas $\Sigma$. $R_o$ sigue una distribución de Wishart $W_p$ ($\Sigma$, n-r) (Rao (1973)).

Como en el caso univariante podemos formular hipótesis lineales, definidas por ecuaciones del tipo

$$H \; B = 0 \qquad (28)$$

con $H = (h_{ij})$; $1 \leq i \leq t$; $1 \leq j \leq m$, y B la matriz de parámetros. Las condiciones de demostrabilidad son las mismas que para el caso univariante.

Bajo la hipótesis lineal H B = 0 obtenemos una reparametrización del modelo original con una nueva matriz de diseño X*. La nueva expresión matrial es:

$$Y = X^* \; C + E \qquad (29)$$

siendo:

$$X^* = (x_{ij}) \qquad 1 \leq i \leq n; \qquad 1 \leq j \leq q$$

$$C = (c_{ij}) \qquad 1 \leq i \leq q; \qquad 1 \leq j \leq p$$

y el rango de la matriz X* máximo por construcción.

La estimación de la matriz de parámetros $\hat{C}$ se obtiene mediante las ecuaciones:

$$\hat{C} = (X^{*t} X^*)^{-1} X^{*t} Y$$

ó (30)

$$\hat{C} = (X_r^{*t} D X_r^*)^{-1} X_r^{*t} D \bar{Y}$$

si se usa la matriz de diseño reducida en el modelo (29).

Las sumas de cuadrados y productos cruzados del modelo (29) son:

$$R_1 (i,j) = (Y_i - X^* \hat{C}_i)^t (Y_j - X^* \hat{C}_j) \quad (31)$$

donde $\hat{C}_k$ son los vectores columna de la matriz $\hat{C}$

Sea $R_1$ la matriz

$$R_1 = (r_i{}^*{}_j) \quad (32)$$

$$r_i{}^*{}_j = R_1 (i,j)$$

$$1 \leq i, j \leq p$$

si la hipótesis lineal (28) es cierta y el modelo lineal multivariante es normal, $R_1$ sigue una distribución de Wishart $W_p$ $(\Sigma, n-q)$, siendo q el rando de X* e igual al número de filas de la matriz C. Además $R_1-R_o$, sigue una distribución de Wishart $W_p$ $(\Sigma,t)$, siendo t = rang (H).

Suponiendo la hipótesis lineal cierta, $R_1-R_o$ y $R_o$ son estocásticamente independientes.

Los contrastes de hipótesis lineales en el caso multivariante pueden resolverse de diferentes formas, una de ellas es el criterio de Hotelling basado en el estadístico.

$$V = \text{traza} \left[ (R_1 - R_o)\ R_o^{-1} \right] \tag{33}$$

Si tenemos dos modelos lineales multivariantes con la misma matriz de varianzas-covarianzas.

$$\begin{aligned} Y_1 &= 1.\mu_1 + E_1 \\ Y_2 &= 1.\mu_2 + E_2 \end{aligned} \tag{34}$$

siendo: 1 el vector columna de n unos

$$\begin{aligned} Y_1 &= (y_1{}^1{}_i, \ldots, y_1{}^q{}_i) \\ Y_2 &= (y_2{}^1{}_i, \ldots, y_2{}^q{}_i) \\ \mu_1 &= (\mu_1{}^1, \ldots, \mu_1{}^q) \\ \mu_2 &= (\mu_2{}^1, \ldots, \mu_2{}^q) \\ E_1 &= (e_1{}^1{}_i, \ldots, e_1{}^q{}_i) \\ E_2 &= (e_2{}^1{}_i, \ldots, e_2{}^q{}_i) \\ 1 &\leq i \leq n \end{aligned}$$

y queremos contrastar la hipótesis

$$H_o: \mu_1 = \mu_2$$

frente a

$$H_1: \mu_1 \neq \mu_2$$

usaremos el estadístico (33) que sigue, si $H_o$ es cierto, una distribución $T^2$ de Hotelling, cumpliéndose que,

$$F = T^2 \frac{(2\ n-q-1)}{(2\ n-2)\ q} \qquad (35)$$

sigue una distribución F de Fisher con q y 2n-p-1 grados de libertad (Anderson (1958)).

## 2.2. INTRODUCCION A LA GEOMETRIA RIEMANNIANA EN UNA VARIEDAD

En el presente punto se pretende hacer una breve reseña de la geometría riemanniana, dando un enfoque intuitivo de la misma y ofreciendo el planteo y resultado de algunos problemas.

### 2.2.1. Definiciones

Dados dos puntos P y P' de una variedad M, infinitesimalmente próximos y con coordenadas $x = (x^1, \ldots, x^n)$, $x + d\,x = (x^1 + d\,x^1, \ldots, x^n + d\,x^n)$, la distancia entre ellos al cuadrado viene dada por la forma cuadrática diferencial

$$ds^2 = \sum_{i,j} g_{ij}\, dx^i\, dx^j \qquad (36)$$

o abreviadamente

$$ds^2 = g_{ij}\, dx^i\, dx^j$$

en caso de usar el convenio de sumación de índices repetidos.

A ds la llamamos elemento de arco.

En cada punto de la variedad, $g_{ij}$ define un producto escalar en el espacio tangente a la misma en dicho punto.

La longitud entre dos puntos P yQ de una variedad, medida sobre una curva C de parámetro t, viene dada por

$$d_c = \int_{t_P}^{t_Q} \sqrt{g_{ij} \frac{dx^i}{dt} \frac{dx^j}{dt}}\, dt \qquad (37)$$

La distancia entre los puntos P y Q se define como el ínfimo del conjunto de las longitudes entre ellos, medidas sobre todas las posibles curvas que los unen.

Las curvas que hacen mínima la integral (37) se llaman geodésicas, las cuales no siempre existen globalmente (en toda la variedad) pero si localmente (Hicks (1974), Spirak (1979)).

La variedad M es euclídea sí y sólo sí es posible efectuar una transformación de coordenadas tal, que el campo tensorial métrico, quede reducido a un campo tensorial constante. En este caso, siempre será posible encontrar un sistema de referencia tal, que la distancia entre dos puntos de coordenadas $P = (x^1, x^2, \ldots, x^n)$ y $Q = (y^1, y^2, \ldots, y^n)$ viene dada por

$$d_{PQ} = \sqrt{\sum_{i=1}^{n} (x^i - y^i)^2} \qquad (38)$$

## 2.2.2. Símbolos de Christoffel

En este apartado introducimos algunas combinaciones de las derivadas parciales de $g_{ij}$, que son conocidas como los

símbolos de Christofeel.

Al conjunto de funciones

$$\Gamma_{ij,K} = \frac{1}{2} (D_j\, g_{ik} + D_i\, g_{jk} - D_k\, g_{ij}) \qquad (39)$$

$1 \leq i, j, k \leq n$, se denominan símbolos de Christoffel de primera especie y al conjunto de funciones definidas por

$$\Gamma_{ij}^{\ k} = g^{k\alpha}\ \Gamma_{ij,\alpha} \qquad (40)$$

$$1 \leq i, j, k \leq n$$

se denominan símbolos de Christoffel de segunda especie, siendo $g^{k\alpha}$ las componentes de un tensor contravariante de segundo orden, tal que

$$g_{ek}\ g^{k\alpha} = \delta_{ek} \qquad (41)$$

y

$$\delta_{ek} = \begin{cases} 1 \iff e = k \\ 0 \iff e \neq k \end{cases} \qquad (42)$$

Tanto en la expresión (39) como en la (40) hemos usado el criterio de sumación de índices repetidos.

## 2.2.3. Cálculo de las geodésicas

Como hemos dicho en 2.2.1. una geodésica es una curva sobre la cual la integral (37) se hace mínima localmente.

Utilizando las técnicas propias del cálculo de variaciones, mediante las ecuaciones de Euler, obtenemos el sistema de ecuaciones diferenciales

$$\ddot{x}^k + \Gamma_{ij}^{\ k}\, \dot{x}^i\, \dot{x}^j = 0 \qquad (43)$$

$$1 \leq k \leq n$$

donde

$$\dot{x}^i = \frac{dx^i}{dt}\ , \quad \dot{x}^j = \frac{dx^j}{dt}\ , \quad \ddot{x}^k = \frac{d^2 x^k}{dt^2}$$

que satisfacen dichas curvas.

Resolviendo el sistema (43), teniendo en cuenta unas determinadas condiciones contorno, que vienen fijadas por los puntos inicial y final de la curva, se obtiene la geodésica buscada. Hay que tener en cuenta que no siempre existe solución del sistema (43) según condiciones de contorno fijadas.

## 2.3. ANALISIS DE DATOS

En el presente punto, se hace una breve exposición de los métodos de representación de individuos de una población $\Omega$ respecto a unas variables observadas y que son usadas en esta memoria.

## 2.3.1. Análisis de Componentes Principales

Dada una población $\Omega$ de n individuos $w_1, w_2, \ldots, w_n$, caracterizado cada $w_i$ por unas medidas efectuadas sobre el mismo, $(x_{i1}, x_{i2}, \ldots, x_{ip})$, podemos identificar a cada $w_i$ con un punto $P_i$, del espacio euclídeo $R^p$ con el producto escalar usual cuyas coordenadas, respecto de un sistema de referencia cartesianas, sean precisamente $(x_{i1}, x_{i2}, \ldots, x_{ip})$.

Para poder representar geométricamente a estos individuos, situados en un espacio p-dimensional, procedemos a hallar una variedad lineal q-dimensional con $q < p$, generalmente en $q = 1, 2$ ó $3$, tal que al proyectar cada uno de los $P_i$ de $R^p$ sobre la misma, la configuración de puntos resultante sea lo más parecida posible a la configuración de puntos en $R^p$.

El criterio para obtener dicha variedad lineal consiste en minimizar la suma de los cuadrados de las distancias de los puntos en $R^p$ a la variedad buscada. Además se cumple simultáneamente que la suma de los cuadrados de las interdistancias de las proyecciones es máxima.

La ecuación que define la variedad lineal buscada viene dada por

$$y = \alpha + \beta_1 v_1 + \ldots + \beta_q v_q \qquad (44)$$

donde $\alpha$ es un vector cuyas componentes son las medias de las variables observadas,

$$\alpha = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p) \tag{45}$$

y donde

$$v_i = (v_{1i}, \ldots, v_{pi})^t \tag{46}$$

son los vectores propios correspondientes a los q primeros valores propios, ordenados de mayor a menor, de la matriz de varianzas-covarianzas de las variables utilizadas.

Las proyecciones de los puntos $P_i$ sobre la variedad lineal tienen por coordenadas, respecto la referencia determinada por $\alpha$, $v_1, \ldots, v_q$, $(y_{i1}, \ldots, y_{iq})$, que vienen dadas por el sumatorio,

$$y_{ij} = \sum_{u=1}^{p} x_{iu} v_{uj} \tag{47}$$

$$1 \leq j \leq q$$

Matricialmente las coordenadas de las proyecciones de todos los puntos $P_i$ vendrán dadas por las filas de la matriz Y.

$$Y = X V \tag{48}$$

siendo X la matriz de datos, matriz cuyas filas son las coordenadas de los puntos $P_i$ y V es la matriz de vectores propios, cuyos vectores columna son los q primeros vectores propios (46).

Para más detalles sobre este tema vean (Mardia (1979, Cuadras(1981), Hotelling (1933), Rao (1964)).

## 2.3.2. Análisis de Coordenadas Principales

El Análisis de Coordenadas Principales, es un método que nos permite representar gráficamente a una población $\Omega$ de individuos $w_1, w_2, \ldots, w_n$, sobre lo que tenemos definida una similaridad o una distancia.

Si tenemos definida una distancia d en $\Omega$

$$d: \Omega \times \Omega \longrightarrow R \qquad (49)$$

$$d(w_i, w_j) = d_{ij}$$

debemos, en primer lugar, asignar unas coordenadas a los individuos de la población $\Omega$ que los identifique con puntos de $R^p$ y tal que la distancia euclídea en $R^p$ coincida con la distancia primitiva d. El siguiente paso es realizar un Análisis de Componentes Principales que nos permitirá representar a los n individuos de la población $\Omega$, identificados ya a puntos $P_i$ de $R^p$ de coordenadas $(n_{i1}, \ldots, n_{ip})$, en una variedad lineal q-dimensional con $q < p$, generalmente $q = 1, 2$ ó $3$. A las coordenadas de $P_i$ en la nueva variedad las llamamos coordenadas Principales.

En la práctica para obtener las coordenadas principales, a partir de la distancia (49), se sigue el siguiente algoritmo:

1) Obtenemos la matriz,

$$S = (s_{ij}) \qquad (50)$$

con

$$s_{ij} = \frac{1}{2} d_{i}^{2}{}_{j}$$

2) Transformamos la matriz S en la matriz,

$$B = (b_{ij}) \qquad (51)$$

con $b_{ij} = s_{ij} - \bar{s}_i - \bar{s}_j + \bar{s}$, siendo $\bar{s}_i$ y $\bar{s}_j$ las medias de los valores de la filla i y la columna j de la matriz S respectivamente y $\bar{s}$ la media de todos los valores de la matriz S.

3) Diagonalizamos la matriz B,

$$B = T\, D_{\mu}\, T^{t} \qquad (52)$$

siendo $D_{\mu}$ la matriz diagonal $D_{\mu}$ = diag. $(\mu_1, \mu_2, \ldots, \mu_n)$ con $\mu_1, \mu_2, \ldots, \mu_n$ los valores propios de B ordenados de mayor a menor y T la matriz cuyos vectores columna son vectores propios de B asociados a los valores propios antes mencionados y $\mu_i$- normalizados.

4) Por último las coordenadas de los puntos $Q_i$ que representan a los individuos $w_i$ de $\Omega$ en la variedad q-dimensional, tienen por coordenadas $u_{i1}, \ldots, u_{iq}$, siendo $u_{ij}$ los elementos de la matriz

$$U = T\, D_{\mu}^{\frac{1}{2}} \qquad (53)$$

que tiene n filas y p columnas

$$U = (u_{ij})$$

$$1 \leq i \leq n; \quad 1 \leq j \leq p$$

La dispersión de los n puntos en la variedad q-dimensional viene dada por

$$\sum_{i,j} d_{ij}^2 = 2n\ (\mu_1 + \mu_2 + \ldots + \mu_q) \qquad (49)$$

que se puede comparar con la dispersión global a fin de obtener el porcentaje de la variedad explicada por las q variables. El valor de este porcentaje es

$$\Pi_q = 100\ \frac{\mu_1 + \ldots + \mu_q}{\mu_1 + \ldots + \mu_{n-1}} \qquad (55)$$

Para ver más detalles sobre este tema vease (Cuadras (1981), Lefebre (1976)).

### 2.3.3. Taxonomía Numérica

Si tenemos una similaridad o una distancia definida en una población $\Omega$, es posible construir una clasificación de los individuos de la misma, atendiendo a sus semejanzas que vienen medidas por la similaridad o distancia definida en la población, mediante los métodos de Taxonomía Numérica. Estos métodos nos permiten realizar una clasificación jerárquica de los individuos de $\Omega$, obteniendo sucesivas particiones de la población organizadas en diferentes niveles jerárquicos y constituídas a cada nivel, por grupos homogéneos y disjuntos de indivi-

duos llamados clusters. A la representación gráfica de esta clasificación, la llamamos dendograma.

El grado de homogeneidad entre los grupos obtenidos, lo valoramos mediante un índice llamado índice de jerarquía o distancia fenética.

A continuación vamos a dar las definiciones formales de estos conceptos.

Dada la población $\Omega$ de individuos $w_1$, $w_2 \ldots$, $w_n$, decimos que un subconjunto $\mathcal{H}$ de partes de $\Omega$ es una Jerarquía si se verifica:

a) Dados dos elementos de $\mathcal{H}$, o uno de ellos está contenido en el otro o son disjuntos.

b) Todo elemento de $\mathcal{H}$ es la unión de los elementos de $\mathcal{H}$ que contiene, o no contiene a ningún elemento de $\mathcal{H}$.

Si $\mathcal{H}$ tiene como elemento a $\Omega$ y a los conjuntos formados por un sólo individuo, se dice que $\mathcal{H}$ es una jerarquía total.

A los elementos de $\mathcal{H}$ se les llama clusters y si $\mathcal{H}$ es una partición de $\Omega$ a $\mathcal{H}$ se le llama clustering.

Una jerarquía $\mathcal{H}$ se dice que indexada si existe una aplicación

$$i: \mathcal{H} \longrightarrow R \qquad (56)$$

que cumple

a) $i(H) \geq 0, \quad H \in \mathcal{H}$

b) $i(H) = 0 \quad \text{si } H = \{w\} \text{ y } w \in \Omega$

c) $H' \subset H \Longrightarrow i(H') < i(H)$

A la aplicación i se le llama distancia fenética o índice de la jerarquía

La representación gráfica de una jerarquía indexada es el dendograma.

Mediante una jerarquía indexada se puede definir una distancia ultramétrica en $\Omega$ y recíprocamente con una distancia ultramétrica definida en $\Omega$, se puede construir una jerarquía indexada (Cuadras (1981)).

En la práctica no se tiene, en general, una distancia ultramétrica definida en la población $\Omega$, por lo tanto para hacer una clasificación jerárquica debemos transformar la distancia inicial en una distancia ultramétrica de una forma razonable y seguidamente construir la jerarquía. Los pasos para obtener una distancia ultramétrica a partir de una distancia cualquiera definida en $\Omega$ se le llama algorítmo de clasificación.

Si partimos de una distancia d definida en $\Omega$ y la deformamos en otra u, ultramétrica, puede ocurrir que haya una gran distorsión entre d y u y por consiguiente, la clasifica-

ción obtenida mediante u no sea representativa de la distancia d. Para medir esta distorsión se usa el llamado coeficiente de correlación cofenética $r_c$ que es el coeficiente de correlación entre los n (n-1)/2 pares de valores

$$(d\ (w_i, w_j),\ u\ (w_i, w_j)) \qquad w_i, w_j \in \Omega$$

Si $d\ (w_i, w_j) = u\ (w_i, w_j) \qquad w_i, w_j \in \Omega$, entonces $r_c = 1$. Valores próximos a 1 indican una clasificación jerárquica representativa de la distancia d, por el contrario valores bajos de $r_c$ indican una gran distorsión entre las dos distancias d y u, por lo tanto, en este caso, la clasificación no es fiable. Interesa pues, que los algoritmos de clasificación no deformen la distancia original.

Existen muchos algoritmos de clasificación, entre ellos está el Método del mínimo y Método del máximo (Johnson (1967)), Método del centroide y Método de la media (Sokal y Michener (1958)), Método de la mediana (Gower (1967)), pero posiblemente el más usado de los métodos es el Método UPGMA (Unweighted Pair Group Method Using Method Averages) descrito por Sokal en 1963.

Para ver otros resultados de Taxonomía y profundizar en lo expuesto vease entre otros (Cuadras (1980), Cuadras y Carmona (1984), Arcas y Salicrú (1984), Cuadras y Uson (1980)).

# 3. DISTANCIAS ENTRE MODELOS LINEALES UNIVARIANTES

Página

Sumario:

Este capítulo está dividido en dos partes; en la primera se define y estudia una distancia entre modelos lineales normales caso univariante con igual varianza y en la segunda parte se define y estudia una distancia entre modelos lineales normales caso univariante con varianzas distintas (la misma dentro de cada modelo, pero distinta de un modelo a otro).

En cada una de estas dos partes se hace una formalización de la introducción de una distancia entre clases de poblaciones estadísticas, representadas por modelos lineales, estudiando los aspectos geométricos y estadísticos de la misma.

## 3.1. DISTANCIAS ENTRE MODELOS LINEALES UNIVARIANTES Y CON IGUAL VARIANZA

Se trata de introducir una distancia entre clases de poblaciones estadísticas, asociadas a modelos lineales normales univariantes con igual varianza, fijados por unas condiciones experimentales que determinan la matriz de diseño, el número de parámetros, el número de réplicas por condición experimental y el número de variables observadas. Es decir, se parte del conjunto de modelos lineales cuya expresión matricial es:

$$y = X\beta + e \qquad (1)$$

donde $y = (y_1, y_2, \ldots, y_n)^t$ es un vector columna que representa a la muestra de tamaño n, $\beta = (\beta_1, \beta_2, \ldots, \beta_m)^t$ es el vector columna de los parámetros, $e = (e_1, e_2, \ldots, e_n)^t$ es

el vector columna de las desviaciones aleatorias del modelo y donde $X = (x_{ij})$ $1 \leq i \leq n$, $1 \leq j \leq m$, es la matriz de diseño que supondremos de rango m, rango máximo, de otra manera haríamos una reparametrización que transformara el modelo en un diseño de rango máximo.

Se supone además que $e_i$ $1 \leq i \leq n$, son variables aleatorias normales e independientes de media o y varianza $\sigma^2$ que en esta primera parte supondremos la misma para cada modelo.

Por todo lo dicho se deduce que las variables $y_1, y_2, \ldots, y_n$ son variables aleatorias normales e independientes y por consiguiente, y, es una variable aleatoria normal multivariante de vector de medias $\mu = X\beta$ y matriz de varianzas-covarianzas $\Sigma_o = \sigma^2 I$, con $\sigma^2$ positiva.

### 3.1.1. Formalización y Aspectos Geométricos

Cuando tenemos definidas sobre distintas poblaciones n variables aleatorias observadas $y_1, y_2, \ldots, y_n$ independientes, de distribución conjunta normal multivariante de media variable y matriz de varianzas-covarianzas constante $\Sigma_o$ y queremos introducir una distancia entre dichas poblaciones estadísticas, podemos caracterizar a cada población por el vector de medias $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$, considerándolo como las coordenadas de un punto de una variedad paramétrica E, en este caso $E = R^n$.

Definiremos la distancia entre dos poblaciones como la distancia entre dos puntos de la variedad paramétrica E que los representan.

La métrica en esta variedad se define a través del funcional entropía $H_\phi$, o información de Shannon, tomando $\phi(x) = x \ln x$, tal como mencionábamos en el primer capítulo. El campo tensorial vendrá dado por:

$$g_{ij} = \int_{R^n} \phi''(f)\,(D_i f)\,(D_j f)\,dy \qquad (2)$$

$$1 \leq i,j \leq n$$

siendo $D_k f$ la derivada parcial de f con respecto a $\mu_k$.

Puesto que,

$$\phi(x) = x \ln x \qquad (3)$$

se tiene

$$\phi''(x) = \frac{1}{x}$$

por consiguiente

$$g_{ij} = \int_{R^n} \frac{1}{f}\,(D_i f)\,(D_j f)\,dy =$$

$$= \int_{R^n} \frac{1}{f}\,(D_i f)\,\frac{1}{f}\,(D_j f)\,f\,dy = \qquad (4)$$

$$= E\left[\frac{1}{f}\,(D_i f)\,\frac{1}{f}\,(D_j f)\right] = E\left[(D_i(\ln f)\,D_j(\ln f)\right]$$

$$1 \leq i,j \leq n$$

luego la matriz $G = (g_{ij})$ que se obtiene coincide con la matriz de información de Fisher, bajo ciertas condiciones de regularidad que se dan en nuestro caso.

Alternativamente podemos escribir

$$g_{ij} = -E\ (D_{ij}\ \ln f) \tag{5}$$

siendo $D_{ij}\ (\ln f)$, la derivada segunda con respecto a $\mu_i$ y $\mu_j$ del logaritmo de la función de densidad.

Como $(y_1, y_2, \ldots, y_n)$, sigue una distribución normal multivariante de media $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ y matriz de varianzas-covarianzas $\Sigma_o = (\sigma_{ij})$ constante, entonces:

$$f\ (y,\mu,\Sigma_o) = (2\pi)^{-\frac{1}{2}}\ |\Sigma_o|^{-\frac{1}{2}}\ \exp.\ \left[\ -\tfrac{1}{2}\ (y-\mu)^t\ \Sigma_o^{-1}\ (y-\mu)\right] \tag{6}$$

Tomando logaritmos y teniendo en cuenta que $\Sigma_o^{-1} = (\sigma^{ij})$ resulta:

$$\ln f = k - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma^{ij}\ (y_i - \mu_i)\ (y_j - \mu_j) \tag{7}$$

La parcial del ln f con respecto a $\mu_\alpha$ será

$$D_\alpha\ \ln f = \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma^{ij} \left[(y_i - \mu_i)\ \delta_{j\alpha} + (y_j - \mu_j)\delta_{i\alpha}\right] \tag{8}$$

luego

$$D_\alpha\ \ln f = \tfrac{1}{2} \left[ \sum_{i=1}^{n} \sigma^{i\alpha}\ (y_i-\mu_i) + \sum_{j=1}^{n} \sigma^{\alpha j}\ (y_j-\mu_j) \right] =$$

$$= \sum_{i=1}^{n} \sigma^{i\alpha}\ (y_i-\mu_i) \tag{9}$$

Las derivadas segundas serán:

$$D_{\alpha\beta} \ln f = - \sigma^{\beta\alpha} \tag{10}$$

por consiguiente combinando (5) con (10), obtenemos finalmente

$$g_{\alpha\beta} = -E \left[ D_{\alpha\beta} \ln f) \right] = \sigma^{\beta\alpha} \tag{11}$$

por tanto

$$G = (g_{ij}) = (\sigma^{ij}) = \Sigma_o^{-1} \tag{12}$$

Nótese que en la demostración no se ha supuesto independencia estocástica entre las variables aleatorias $y_1$, $y_2$, .. .., $y_n$.

Para nuestro caso particular, modelo (1), $\Sigma_o = \sigma^2 I$ con $\sigma^2 > 0$, quedando:

$$g_{ij} = \frac{\delta_{ij}}{\sigma^2} \; ; \quad G = \frac{1}{\sigma^2} \; I \tag{13}$$

Hemos construido, a partir del concepto de información de Shannon, un campo tensorial covariante de segundo orden simétrico y definido positivo, por tanto E tiene estructura de variedad riemanniana. En nuestro caso particular el campo tensorial es constante, luego la variedad es euclídea y la distancia entre dos poblaciones vendrá dada en general por:

$$d = \sqrt{(\mu_A - \mu_B)^t \; \Sigma_o^{-1} \; (\mu_A - \mu_B)} \tag{14}$$

y en nuestro caso particular

$$d = \sqrt{\frac{1}{\sigma^2} (\mu_A - \mu_B)^t (\mu_A - \mu_B)} \qquad (15)$$

siendo $\mu_A$, $\mu_B$, dos puntos de la variedad E. Esta expresión coincide con la distancia de Mahalanobis, resultado éste que ya fue indicado en el primer capítulo.

Vamos a continuación a trasladar la distancia a las clases de poblaciones estadísticas representadas por modelos lineales.

Sea $\Omega$ el conjunto de poblaciones estadísticas, cada una de las cuales asociada a un modelo lineal de la forma (1) y por consiguiente asociada a un vector paramétrico $\beta$. De esta forma es posible definir una aplicación h de $\Omega$ en S, siendo S una subvariedad de E definida por:

$$S = \{ \mu = (\mu_1, \mu_2, \ldots, \mu_n) \in E \mid \mu = X\beta \} \qquad (16)$$

La aplicación h vendrá definida por:

$$h: \Omega \longrightarrow S$$

$$\omega \longrightarrow \beta$$

cumpliendo que:

$$E(y/\omega) = X\beta \qquad (17)$$

Al ser $S \subset E$, existe una aplicación inyectiva i de S en E tal que $i(\mu) = \mu$, $\mu \in S$. Existe pues una aplicación $\Psi = i \circ h$ que transforma $\Omega$ en E.

$$\begin{array}{ccccc} & & h & & i \\ \Omega & \longrightarrow & S & \longrightarrow & E \\ \omega & \longrightarrow & \beta & \longrightarrow & X\beta \end{array} \qquad (18)$$

Al ser i una aplicación inyectiva de S en E, ya que supondremos modelos de rango máximo, queda definida, de una forma natural, una métrica en S a través de la métrica en E.

$$\text{Si, } \beta_A, \beta_B \in S, \quad \text{entonces}$$

$$d_S(\beta_A, \beta_B) = d_E(X\beta_A, X\beta_B) \qquad (19)$$

donde $\beta_A$ y $\beta_B$ son los vectores paramétricos de los modelos lineales asociados mediante h a dos poblaciones estadísticas $\omega_A, \omega_B \in \Omega$ tales que $h(\omega_A) = \beta_A$, $h(\omega_B) = \beta_B$.

Podemos definir en $\Omega$ una pseudométrica s a través de la aplicación h:

$$\begin{aligned} s(\omega_A, \omega_B) &= d_S(h(\omega_A), h(\omega_B)) = \\ &= d_E(\Psi(\omega_A), \Psi(\omega_B)) = d_E(X\beta_A, X\beta_B) \end{aligned} \qquad (20)$$

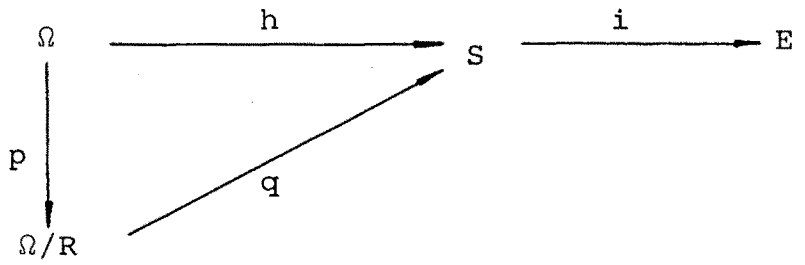
El problema, de tener definida en $\Omega$ una pseudométrica y no una métrica, se puede resolver definiendo en $\Omega$ una relación de equivalencia de la siguiente forma:

$$\omega_A \; R \; \omega_B \qquad s\ (\omega_A,\ \omega_B) = 0 \tag{21}$$

En el conjunto cociente $\Omega/R$ podemos definir una métrica

$$d_{\Omega/R}\ ([\omega_A],\ [\omega_B]) = d_E\ (X\ \beta_A,\ X\ \beta_B) \tag{22}$$

En esquema las aplicaciones entre los conjuntos quedarían así:




(23)

Veamos seguidamente cual sería la expresión de la distancia.

Si tenemos dos poblaciones estadísticas $\omega_A$ y $\omega_B$ tales $h\ (\omega_A) = \beta_A$, $h\ (\omega_B) = \beta_B$, la expresión de la distancia será:

$$d^2_{\Omega/R}\ ([\omega_A],\ [\omega_B]) = d^2_E\ (X\ \beta_A,\ X\ \beta_B) =$$

$$= (X\beta_A - X\ \beta_B)^t\ \Sigma_o^{-1}\ (X\ \beta_A - X\ \beta_B) \tag{24}$$

al ser $\Sigma_o^{-1} = \text{diag.}\ (1/\sigma^2,\ 1/\sigma^2,\ \ldots,\ 1/\sigma^2)$, resulta finalmente

$$d^2_{\Omega/R} = \frac{1}{\sigma 2}\ (\beta_A - \beta_B)^t\ X^t\ X\ (\beta_A - \beta_B) \tag{25}$$

### 3.1.2. Aspectos Estadísticos: estimación y contraste de hipotesis

Sean

$$y_A = X\,\beta_A + e_A$$
$$y_B = X\,\beta_B + e_B \qquad (26)$$

los modelos lineales expresados en forma matricial tales que

$$y_A = (y_1^A, y_2^A, \ldots, y_N^A)^t;\ y_B = (y_1^B, y_2^B, \ldots, y_N^B)^t$$

$$\beta_A = (\beta_A^1, \beta_A^2, \ldots, \beta_A^m)^t\ ;\ \beta_B = (\beta_B^1, \beta_B^2, \ldots, \beta_B^m)^t$$

$$e_A = (e_1^A, e_2^A, \ldots, e_N^A)^t\ ;\ e_B = (e_1^B, e_2^B, \ldots, e_N^B)^t$$

$$X = (x_{ij});\ 1 \leq i \leq N;\quad 1 \leq j \leq m$$

$$e_j^i \sim N\ (o,\sigma);\ i = A, B;\quad 1 \leq j \leq N$$

$e_i^A$, $e_j^B$, independientes para todo i, j.

Vamos a obtener un estimador de la distancia entre ambos modelos lineales, en función de la muestra observada.

En las aplicaciones prácticas conocemos la muestra observada, la matriz de diseño y desconocemos $\beta$ y generalmente también $\sigma$, por tanto, para el cálculo explícito de la distancia entre dos modelos lineales habrá que estimar tales parámetros.

La estimación mínimo cuadrática (L.S.) de los parámetros del modelo que coincide con la máximo verosímil, como es bien sabido, es:

$$\hat{\beta}_i = (X^t X)^{-1} X^t y_i \qquad (27)$$

llevando la estimación a (25) nos queda la estimación de la distancia como:

$$\hat{d}^2_{\Omega/R}([\omega_A], [\omega_B]) = \frac{1}{\sigma^2} (y_A - y_B)^t X (X^t X)^{-1} X^t (y_A - y_B) \qquad (28)$$

Nótese que al ser X de rango máximo, rango m, $X^t X$ también tiene de rango m y por ser de orden m x m siempre existirá su inversa.

En el caso de utilizar la matriz de diseño reducida, la expresión de la distancia estimada será:

$$\hat{d}^2_{\Omega/R}([\omega_A], [\omega_B]) = \frac{1}{\sigma^2} (\bar{y}_A - \bar{y}_B)^t \Delta X_r (X_r^t \Delta X_r)^{-1} X_r^t \Delta (\bar{y}_A - \bar{y}_B) \qquad (29)$$

donde, $\bar{y}_A = (\bar{y}_1^A, \bar{y}_2^A, \ldots, \bar{y}_k^A)$; $\bar{y}_B = (\bar{y}_1^B, \bar{y}_2^B, \ldots, \bar{y}_k^B)$ son los vectores de medias cuyas componentes son las medias de las réplicas correspondientes a una misma condición experimental y en donde $\Delta$ es igual a diag. $(n_1, n_2, \ldots, n_k)$ donde $n_i$ es el número de réplicas tomadas en la i-ésima condición experimental, k es el número de condiciones experimentales diferentes y $X_r$ es la matriz de diseño reducida (Cuadras (1974)). La utilización de $X_r$ y $\Delta$ permite abordar diseños no balanceados y con observaciones faltantes en alguna condición experimental (Cuadras (1982)).

Cuando la varianza no sea conocida tendremos que realizar una estimación de la misma que a continuación vamos a obtener.

Sea la expresión matricial del modelo conjunto

$$y = A\beta + e \tag{30}$$

con

$$y = \begin{pmatrix} y_A \\ y_B \end{pmatrix}; \quad A = \begin{pmatrix} X & 0 \\ 0 & X \end{pmatrix}; \quad \beta = \begin{pmatrix} \beta_A \\ \beta_B \end{pmatrix}; \quad e = \begin{pmatrix} e_A \\ e_B \end{pmatrix}$$

el rango del diseño es rang(A) = 2m, siendo m el número de parámetros de cada modelo, lo que quiere decir que la estimación insesgada de la varianza es $R_o^2/2(N-m)$, siendo $R_o^2$ la suma de cuadrados residuales del modelo conjunto

$$R_o^2 = (y - A\hat{\beta})^t (y - A\hat{\beta}) \tag{31}$$

al verificarse

$$\hat{\beta} = (A^tA)^{-1} A^t y \tag{32}$$

resulta finalmente

$$R_o^2 = y^t (I - A (A^tA)^{-1} A^t) y \tag{33}$$

Teniendo en cuenta (30), podemos escribir

$$R_o^2 = \sum_{i=A,B} (y_i^t (I - X (X^tX)^{-1} X^t) y_i \tag{34}$$

La expresión de la distancia estimada será pues:

$$\hat{d}^2 \left([\omega_A], [\omega_B]\right) = 2\,(N-m)\,\frac{(y_A-y_B)^t\, X\, (X^tX)^{-1}X^t\,(y_A-y_B)}{\sum_{i=A,B} (y_i^t\,(I-X(X^tX)^{-1}X^t)\,y_i)} \qquad (35)$$

Vamos a estudiar ahora contrastes de hipótesis del tipo

$$\begin{aligned} H_o&: d\left([\omega_A], [\omega_B]\right) = 0 \\ H_1&: d\left([\omega_A], [\omega_B]\right) > 0 \end{aligned} \qquad (36)$$

es decir, contrastes cuya hipótesis nula es que la distancia entre dos poblaciones es cero, frente a la hipótesis alternativa de que la distancia entre dos poblaciones estadísticas es mayor que cero. También se estudiará la relación de estos contrastes con los usuales del análisis de la varianza, estudiando la distribución del estimador de la distancia bajo la hipótesis nula. La región crítica de estos contrastes vendrá dada por:

$$W_\varepsilon = \left\{ z \in R^{2N} \;/\; \hat{d}\,(z) > A_\varepsilon \right\} \qquad (37)$$

Si suponemos que es cierta la hipótesis

$$d\left((\omega_A), (\omega_B)\right) = 0 \qquad (38)$$

esto es equivalente a considerar como hipótesis nula que

$$\beta_A = \beta_B = \beta_o \qquad (39)$$

por tanto: $\beta_i^A = \beta_i^B \quad 1 \leq i \leq m$. Matricialmente la hipótesis nula se expresaría así:

$$H\,\beta = 0 \tag{40}$$

con

$$H = (h_{ij}) \quad 1 \leq i \leq m; \quad 1 \leq j \leq m$$

$$h_{ij} = \begin{cases} 1 & \text{si } i = j \\ -1 & \text{si } j = m + 1 \\ 0 & \text{en caso contrario} \end{cases}$$

siendo m el rango de H.

Bajo la hipótesis nula, la expresión matricial del modelo conjunto sería, reparametrizado:

$$y = A\,\beta_o + e \tag{41}$$

con

$$y = \begin{pmatrix} y_A \\ y_B \end{pmatrix}; \quad A = \begin{pmatrix} X \\ X \end{pmatrix}; \quad e = \begin{pmatrix} e_A \\ e_B \end{pmatrix}$$

El resíduo bajo la hipótesis nula será:

$$R_1^2 = (y - A\,\hat{\beta}_o)^t\,(y - A\,\hat{\beta}_o) \tag{42}$$

donde $\hat{\beta}_o$ es la estimación L.S. de $\beta_o$

Teniendo en cuenta (41) obtenemos:

$$R_1^2 = (y_A - X\hat{\beta}_o)^t\,(y_A - X\hat{\beta}_o) + (y_B - X\hat{\beta}_o)^t\,(y_B - X\hat{\beta}_o) \tag{43}$$

y por otra parte la estimación L.S. de $\beta_o$ viene dada por:

$$\hat{\beta}_o = (A^t A)^{-1} A^t y = \frac{1}{2} (X^t X)^{-1} X^t (y_A + y_B) = \frac{1}{2}(\hat{\beta}_A + \hat{\beta}_B) \quad (44)$$

combinando (43) con (44) resulta:

$$R_1^2 = y_A^t y_A + y_B^t y_B - \frac{1}{2} (y_A+y_B)^t X (X^t X)^{-1} X^t (y_A+y_B) \quad (45)$$

La desviación de la hipótesis será teniendo en cuenta (34)

$$R_1^2 - R_o^2 = \frac{1}{2} \left[ (y_A - y_B)^t X (X^t X)^{-1} X^t (y_A - y_B) \right] \quad (46)$$

El cociente:

$$B = (R_1^2 - R_o^2) / \sigma^2 \quad (47)$$

si la hipótesis nula es cierta sabemos que sigue una distribución ji-cuadrado con m, rango de la matriz H, grados de libertad. En nuestro caso particular este cociente toma la expresión, cuando la varianza es conocida:

$$B = \frac{1}{2\sigma^2} \left[ (y_A - y_B)^t X (X^t X)^{-1} X^t (y_A - y_B) \right] = \frac{1}{2} \hat{d}^2 \left( [\omega_A], [\omega_B] \right) \quad (48)$$

Por consiguiente $\frac{1}{2} \hat{d}^2 \left( [\omega_A], [\omega_B] \right)$ sigue una distribución ji-cuadrado con m grados de libertad bajo la hipótesis nula $H_o$ (ambos modelos son iguales).

La región crítica de nivel de significación $\varepsilon$ de este test será:

$$W_\varepsilon = \left\{ z \in R^{2N} \;/\; \hat{d}^2 \left( [\omega_A], [\omega_B] \right) > 2K \right\} \qquad (49)$$

donde K es tal que $P(\chi^2 > K) = \varepsilon$ y $\chi^2$ una variable aleatoria que sigue una distribución ji-cuadrado con m grados de libertad. El cociente

$$F = \frac{(R_1^2 - R_o^2) \;/\; m}{R_o^2 \;/\; 2\,(N-m)} \qquad (50)$$

sigue bajo la hipótesis nula $H_o$, una distribución F de Fisher-Suedecor con m y 2(N-m) grados de libertad. Para nuestro caso tendremos, cuando la varianza es desconocida:

$$F = \frac{(R_1^2 - R_o^2) \;/\; m}{R_o^2 \;/\; 2\,(N-m)} = \frac{N-m}{m} = \frac{(y_A - y_B)^t\, X (X^t X)^{-1}\, X\, (y_A - y_B)}{\sum_{i=A,B} (y_i^t\, (I - X\,(X^t X)\, X^t)\, y_i)} =$$

$$= \frac{1}{2m}\; \hat{d}^2 \left( [\omega_A], [\omega_B] \right) \qquad (51)$$

Por consiguiente $\frac{1}{2m}\; \hat{d}^2 \left( [\omega_A], [\omega_B] \right)$ sigue, bajo la hipótesis nula $H_o$ (ambos modelos son iguales), una distribución F de Fisher-Sedecor con m y 2(N-m) grados de libertad.

La región crítica de nivel de significación $\varepsilon$ de este test será:

$$W_\varepsilon = \{ 2 \in R^{2N} \;/\; \hat{d}^2 \left( |\omega_A|, |\omega_B| \right) > 2mK \} \qquad (52)$$

donde K es tal que $P(F > K) = \varepsilon$ y F una variable aleatoria que sigue una distribución F de Fisher-Suedecor con m y 2 (N-m) grados de libertad.

Todo esto nos lleva al siguiente resultado: El test para comparar dos modelos a través del estadístico $\hat{d}^2$ ( $[\omega_A]$, $[\omega_B]$) es equivalente al test de comparación usado en el análisis de la varianza (Cuadras (1979)), ya que en ambos casos las regiones críticas coinciden. Concuerda además con la propiedad de que la distancia al cuadrado de Mahalanobis estimada, en el caso de que la distancia poblacional es nula, es proporcional a una distribución F de Fisher-Suedecor (Cuadras (1981)).

### 3.1.3. Comparación conjunta de p modelos

Este método de comparación de dos modelos lineales se puede generalizar a la comparación simultánea de p modelos lineales de rango máximo.

Sean p modelos lineales normales de rango máximo

$$y_i = X\,\beta_i + e_i \qquad (53)$$

$$1 \leq i \leq p$$

con

$$y_i = (y_1^i, y_2^i, \ldots, y_N^i)^t; \quad e_i = (e_1^i, e_2^i, \ldots, 2_N^i)^t$$

$$\beta_i = (\beta_1^i, \beta_2^i, \ldots, \beta_m^i)^t$$

siendo $X = (x_{ij})$ una matriz de orden N x m, siendo m el número de parámetros de cada modelo.

Según hemos visto en la sección anterior al estimador de la distancia entre dos poblaciones estadísticas (dos modelos lineales), $d_{ij} = d([\omega_i], [\omega_j])$, viene dada por:

$$\hat{d}^2_{ij} = \frac{1}{\sigma^2} \left[ (y_i - y_j)^t \, X \, (X^t X)^{-1} \, X^t \, (y_i - y_j) \right] \qquad (54)$$

La suma de los cuadrados de las interdistancias entre todas las poblaciones será:

$$D^2 = \sum_{i<j} d^2_{ij} \qquad (55)$$

y su estimación será:

$$\hat{D}^2 = \sum_{i<j} \hat{d}^2_{ij} = \frac{1}{\sigma^2} \sum_{i<j} \left[ (y_i - y_j)^t X \, (X^t X)^{-1} \, X^t \, (y_i - y_j) \right] \qquad (56)$$

o también

$$\hat{D}^2 = \frac{1}{2\sigma^2} \sum_{i=1}^{p} \sum_{j=1}^{p} \left[ (y_i - y_j)^t \, X \, (X^t X)^{-1} \, X^t \, (y_i - y_j) \right] \qquad (57)$$

Cuando la varianza no sea conocida tendremos que realizar una estimación de la misma, a continuación vamos a obtenerla.

Sea el modelo lineal conjunto de expresión matricial

$$y = A \, \beta + e \qquad (58)$$

siendo: $y = (y_1^t, y_2^t, \ldots, y_p^t)^t$; $A = I \times X$ (producto de Kronecker de la matriz I (p x p) por la matriz de diseño X); $\beta = (\beta_1^t, \beta_2^t, \ldots, \beta_p^t)^t$. El rango de la matriz A es m x p.

La estimación L.S. de $\beta$ es:

$$\hat{\beta} = (A^t A)^{-1} A^t y \tag{59}$$

sustituyendo (58) en (59) y desarrollando tendremos que:

$$\hat{\beta}_i = (X^t X)^{-1} X^t y_i \tag{60}$$

La estimación de la varianza será:

$$\hat{\sigma}^2 = R_o^2 / p\ (N - m) \tag{61}$$

siendo el valor de $R_o^2$:

$$R_o^2 = (y - A\ \hat{\beta})^t\ (y - A\ \hat{\beta}) \tag{62}$$

Por (58) y desarrollando se tiene que:

$$R_o^2 = \sum_{i=1}^{p} (y_i^t\ (I - X\ (X^t X)^{-1} X^t)\ y_i) \tag{63}$$

que es la suma de los resíduos de los modelos por separado, es decir:

$$R_o^2 = R_o^2\ (1) + R_o^2 + \ldots + R_o^2\ (p) \tag{64}$$

La estimación de la varianza será pues:

$$\hat{\sigma}^2 = \frac{1}{p(N-m)} \sum_{i=1}^{p} (y_i^t\ (I - X\ (X^t X)^{-1} X^t)\ y_i) \tag{65}$$

y la expresión de la estimación de la suma de las interdistancias será:

$$\hat{D}^2 = \frac{P(N-m)}{2} \frac{\sum_{i=1}^{p} \sum_{j=1}^{p} \left[ (y_i - y_j)^t X (X^t X)^{-1} X^t (y_i - y_j) \right]}{\sum_{i=1}^{p} (y_i^t (I - X (X^t X)^{-1} X^t) y_i)} \qquad (66)$$

Todo esto nos permite contrastar hipótesis del tipo:

$$H_o: D = 0$$
$$H_1: D > 0 \qquad (67)$$

La región crítica de nivel de significación $\varepsilon$ es:

$$W_\varepsilon = \{ z \in R^{pN} \ / \ \hat{D} > A_\varepsilon \} \qquad (68)$$

con ella podemos contrastar la hipótesis nula de que todos los modelos son iguales, es decir, todas las interdistancias $d_{ij}$ son nulas y por lo tanto su suma D es nula, frente a la hipótesis alternativa de que al menos alguno de los modelos es diferente de los demás.

La hipótesis D = 0 es equivalente a considerar como hipótesis nula que

$$\beta_1 = \beta_2 = \ldots = \beta_p = \beta_o \qquad (69)$$

o bien que

$$\beta_i^1 = \beta_i^2 = \ldots = \beta_i^p \quad 1 \leq i \leq m \qquad (70)$$

Matricialmente esta hipótesis la expresaríamos como

$$H\,\beta = 0 \tag{71}$$

con

$$H = (h_{ij}) \quad 1 \leq i \leq (p-1)m; \quad 1 \leq j \leq pm$$

siendo

$$h_{ij} = \begin{cases} 1 \text{ si } 1 \leq j \leq m;\ i = j + mk;\ k = 0,1,\ldots,p-2 \\ -1 \text{ si } m > j;\ j = i + m \\ 0 \text{ en los demás casos} \end{cases}$$

Bajo esta hipótesis la expresión matricial del modelo será:

$$y = A\beta_o + e \tag{72}$$

con

$$y = (y_1^t\ y_2^t \ldots y_p^t)^t;\ A = (X^t X^t \ldots X^t)^t$$

La estimación de $\beta_o$ será:

$$\hat{\beta}_o = (A^t A)^{-1} A^t y \tag{73}$$

sustituyendo (58) en (73) tenemos

$$\hat{\beta}_o = \frac{1}{p}\left[(X^tX)^{-1} X y_1 + (X^tX)^{-1} X y_2 + \ldots + (X^tX)^{-1} X y_p)\right] \tag{74}$$

teniendo en cuenta que

$$\hat{\beta}_i = (X^t X)^{-1} X y_i \tag{75}$$

se tendrá que

$$\hat{\beta}_o = \frac{1}{p} (\hat{\beta}_1 + \hat{\beta}_2 + \ldots + \hat{\beta}_p) \tag{76}$$

El resíduo bajo la hipótesis nula será:

$$R_1^2 = (y - A\hat{\beta}_o)^t (y - A\hat{\beta}_o) \tag{77}$$

sustituyendo (72) en (77) se obtiene

$$R_1^2 = \sum_{i=1}^{p} y_i^t y_i - \frac{1}{p} \left[ \left( \sum_{i=1}^{p} y_i^t \right) X (X^t X)^{-1} X^t \left( \sum_{i=1}^{p} y_i \right) \right] \tag{78}$$

La desviación de la hipótesis nula será:

$$R_1^2 - R_o^2 = \frac{1}{2p} \sum_{i=1}^{p} \sum_{j=1}^{p} \left[ (y_i - y_j)^t X (X^t X)^{-1} X^t (y_i - y_j) \right] \tag{79}$$

El cociente

$$B = \frac{R_1^2 - R_o^2}{\sigma^2} \tag{80}$$

si la hipótesis nula es cierta, sabemos que sigue una distribución ji-cuadrado con (p-1)m, rango de la matriz H, grados de libertad.

Para nuestro caso este cociente será:

$$B = \frac{R_1^2 - R_o^2}{\sigma^2} = \frac{1}{2p\sigma^2} \sum_{i=1}^{P} \sum_{j=1}^{P} \left[ (y_i - y_j)^t \; X \; (X^t X)^{-1} \; X^t \; (y_i - y_j) \right] = \tag{81}$$

$$= \frac{1}{p} \hat{D}^2$$

cuando la varianza es conocida. Por consiguiente $\frac{1}{p} \hat{D}^2$ sigue una distribución ji-cuadrado con (p-1)m grados de libertad cuando se cumple la hipótesis nula (todos los modelos son iguales).

La región crítica de este test, a nivel de significación $\varepsilon$ será:

$$W_\varepsilon = \{ z \in R^{pN} \; / \; \hat{D}^2 > pK \} \tag{82}$$

donde K es tal que

$$P \; (\chi^2 > K) = \varepsilon$$

y $\chi^2$ es una variable aleatoria que sigue una distribución ji-cuadrado con (p-1)m grados de libertad.

El cociente

$$F = \frac{(R_1^2 - R_o^2) \; / \; (p-1)m}{R_o^2 \; / \; p(N-m)} \tag{83}$$

sigue bajo la hipótesis nula una distribución F de Fisher-Suedecor con (p-1) y p(N-m) grados de libertad. Para nuestro caso este cociente será:

$$F = \frac{(N-m)}{2(p-1)\ m} \frac{\sum_{i=1}^{p} \sum_{j=1}^{p} \left[ (y_i - y_j)^t \ X \ (X^t X)^{-1} \ X^t \ (y_i - y_j) \right]}{\sum_{i=1}^{p} \left[ y_i^t \ (I - X \ (X^t X)^{-1} \ X^t) \ y_i \right]} =$$

$$= \frac{1}{p \ (p-1) \ m} \hat{D}^2 \qquad (84)$$

cuando la varianza es desconocida y ha de ser estimada. Por consiguiente $\hat{D}^2 \ / \ p \ (p-1) \ m$, sigue una distribución F de Fisher-Snedecor con $(p-1)\ m$ y $p \ (N-m)$ grados de libertad cuando se cumple la hipótesis nula (todos los modelos son iguales).

La región crítica de este contraste de hipótesis a nivel de significación $\varepsilon$ es:

$$W_\varepsilon = \{ z \in R^{pN} \ / \ \hat{D}^2 > p \ (p-1) \ mK \} \qquad (85)$$

siendo K tal que

$$P \ (F > K) = \varepsilon$$

y F una variable aleatoria que sigue una distribución F de Fisher-Snedecor con $(p-1)\ m$ y $p \ (N-m)$ grados de libertad.

Todo esto nos lleva al siguiente resultado: El test para comparar simultáneamente p modelos a través del estadístico $\hat{D}^2$ es equivalente al test de comparación usado en el análisis de la varianza ya que en ámbos casos las regiones críticas son iguales. (véase Cuadras, 1979).

## 3.2. DISTANCIAS ENTRE MODELOS LINEALES NORMALES UNIVARIANTES Y DE VARIANZA DISTINTA

Como en el apartado anterior se trata de introducir una distancia entre clases de poblaciones estadísticas, asociadas a modelos lineales normales univariantes con varianza distinta fijados por unas condiciones experimentales que determinan los elementos del modelo cuya expresión matricial es:

$$y = X\beta + e \qquad (86)$$

Las matrices y, X, $\beta$, e, tienen el mismo sentido que en el apartado anterior y las mismas características salvo que e es un vector columna $e = (e_1, e_2, \ldots, e_n)^t$, con $e_i$ $(1 \leq i \leq n)$ variables aleatorias normales e independientes de media 0 y varianza $\sigma^2$ que en esta segunda parte supondremos diferente en cada modelo. Por ello en este caso las variables $y_1, y_2, \ldots, y_n$ son variables aleatorias normales e independientes y por consiguiente y es una variable aleatoria normal multivariante de vector de medias $\mu = X\beta$ y matriz de varianzas-covarianzas $\Sigma = \sigma^2 I$, donde $\sigma^2$ toma diferentes valores para diferentes modelos. Al vector de parámetros lo denotaremos como $\beta = (\beta^1\ \beta^2\ \ldots\ \beta^m)^t$.

### 3.2.1. Formalización y aspectos geométricos

Cuando tenemos definidas sobre distintas poblaciones estadísticas n variables aleatorias observadas $y_1, y_2, \ldots, y_n$, independientes, con distribución conjunta normal multivariante

de media $\mu = (\mu^1, \mu^2, \ldots, \mu^n)$ y matriz de varianzas-covarianzas $\Sigma = \sigma^2 I$ variable con cada modelo, podemos caracterizar a cada población por el vector $\eta = (\mu^1, \mu^2, \ldots, \mu^n, \sigma^2)$ que se considera como las coordenadas de un punto de una variedad paramétrica, en este caso:

$$E = \{(\mu^1, \mu^2, \ldots, \mu^n, \sigma^2) \in R^{n+1} \ / \ \sigma^2 > 0\} \qquad (87)$$

Definimos la distancia entre dos poblaciones como la distancia entre dos puntos de la variedad paramétrica E que los representan.

Tal como se hizo en el punto anterior, la métrica en esta variedad se define a través del funcional entropía $H_\phi$ , tomando $\phi(x) = x \ln x$. El campo tensorial métrico vendrá dado por (5), $g_{ij} = -E(D_{ij} \ln f)$, $1 \leq i \leq j \leq n+1$.

Como $(y_1, y_2, \ldots, y_n)$ sigue una distribución normal multivariante de media $\mu = (\mu^1, \mu^2, \ldots, \mu^n)$ y matriz de varianzas-covarianzas $\Sigma = \sigma^2 I$, entonces:

$$f(y,\mu,\Sigma) = (2\pi)^{-n/2} \ |\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(y-\mu)^t \ \Sigma^{-1} \ (y-\mu)\right]$$

Tomando logaritmos y teniendo en cuenta que $\Sigma^{-1} = (\sigma^{ij})$ con $\sigma^{ij} = \frac{\delta_{ij}}{\sigma^2}$ , resulta

$$\ln f = K - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu^i)^2 \qquad (88)$$

La derivada parcial de ln f con respecto a $\mu^i$ será:

$$D_i \ln f = \frac{1}{\sigma^2} (y_i - \mu^i) \qquad (89)$$

$$i \leq i \leq n$$

Las derivadas segundas de ln f con respecto a $\mu^i$ y $\mu^j$ serán:

$$D_{ij} \ln f = \frac{\delta ij}{\sigma^2} \qquad (90)$$

$$i \leq i,j \leq n$$

La derivada parcial de ln f con respecto a $\sigma^2$ es:

$$D_{n+1} \ln f = \frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \mu^i)^2 \qquad (91)$$

Las derivadas segundas con respecto a $\mu^i$ y $\sigma^2$ serán:

$$D_{i\ n+1} \ln f = - \frac{1}{\sigma^4} (y_i - \mu^i) \qquad (92)$$

La derivada segunda con respecto a $\sigma^2$ es:

$$D_{n+1\ n+1} \ln f = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^{n} (y_i - \mu^i)^2 \qquad (93)$$

Al ser

$$E (y_i - \mu^i)^2 = \sigma^2 \qquad (94)$$

se tiene que

$$-E (D_{ij} \ln f) = \frac{\delta ij}{\sigma^2} \qquad (95)$$

Además

$$-E\ (D_{i\ n+1}\ \ln f) = 0 \tag{96}$$

y finalmente

$$D_{n+1\ n+1}\ \ln f = \frac{n}{2\ \sigma^4} \tag{97}$$

Al ser $g_{ij} = -E\ (D_{ij}\ \ln f)$ se tendrá $G = (g_{ij})$; $1 \leq i,j \leq n+1$ con

$$g_{ij} = \frac{\delta ij}{\sigma^2}$$

$$g_{i\ n+1} = g_{n+1\ i} = 0 \tag{98}$$

$$g_{n+1\ n+1} = \frac{n}{2\ \sigma^4}$$

Hemos construído, como en el apartado anterior, a partir del concepto de información de Shannon, un campo tensorial covariante de segundo orden simétrico y definido positivo, por tanto E tiene ahora estructura de variedad riemanniana. En este caso, sin embargo, el campo tensorial no es constante.

La distancia entre dos puntos de E, la definimos a través del elemento de arco en E, cuya expresión es teniendo en cuenta (98):

$$ds^2 = \frac{1}{\sigma^2} \sum_{\gamma=1}^{m} (d\ \mu^\gamma)^2 + \frac{n}{2\ \sigma^4}\ (d\ \sigma^2)^2 \tag{99}$$

Como lo que buscamos es distanciar poblaciones estadísticas representadas por los modelos lineales (86), debemos definir el tensor métrico en la subvariedad S de E.

$$S = \left\{ (\mu^1, \mu^2, \ldots, \mu^n, \sigma^2) \;/\; \mu = (\mu^1, \mu^2, \ldots, \mu^n)^t = \right.$$
$$\left. = X\beta \;;\quad \sigma^2 > 0 \right\} \tag{100}$$

para definir la distancia entre dos puntos de S y una vez definida la distancia en S inducirla en el conjunto de las clases de poblaciones estadísticas representadas por los modelos lineales (véase esquema (23)).

En la subvariedad S, al ser

$$\mu^\gamma = \sum_{j=1}^{m} x_{\gamma j} \beta^j \tag{101}$$

$$1 \leq \gamma \leq n$$

tendremos que:

$$d\mu^\gamma = \sum_{j=1}^{m} x_{\gamma j} \, d\beta^j \tag{102}$$

$$1 \leq \gamma \leq n$$

por consiguiente el elemento de arco en la subvariedad S será:

$$d\bar{s}^2 = \sum_{j,k} \bar{g}_{jk} \, d\beta^j \, d\beta^k + \bar{g}_{m+1\; m+1} (d\sigma^2)^2 \tag{103}$$

siendo

$$\bar{g}_{jk} = \sum_{\gamma,\lambda} x_{\gamma j}\, x_{\lambda k}\, g_{\gamma\lambda} = \frac{1}{\sigma^2} \sum_{\lambda=1}^{n} x_{\lambda j}\, x_{\lambda k}$$

$$\bar{g}_{j\ m+1} = \bar{g}_{m+1\ j} = \sum_{\gamma=1}^{n} g_{n+1\ \gamma}\, x_{\gamma j} = 0 \qquad (104)$$

$$\bar{g}_{m+1\ m+1} = g_{n+1\ n+1} = \frac{n}{2\,\sigma^4}$$

$$1 \leq \gamma,\lambda \leq n, \quad 1 \leq j,k \leq m$$

La matriz del tensor métrico inducida en S será la matriz $G = (\bar{g}_{ij})$; $1 \leq i,j \leq m+1$

Busquemos un cambio del tipo

$$(\beta^1, \beta^2, \ldots, \beta^m, \sigma^2) \longrightarrow (\theta^1, \theta^2, \ldots, \theta^m, \sigma^2) \qquad (105)$$

que reduzca el tensor métrico a un tensor diagonal

$$(\theta^1, \theta^2, \ldots, \theta^m)^t = \tau D^{1/2}\, T^t\, (\beta^1, \beta^2, \ldots, \beta^m)^t \qquad (106)$$

donde $X^tX = T\ D\ T^t$, con T ortogonal y $\tau$ un escalar a determinar.

Entonces, llamando $A = T\ D^{-1/2}$, podemos escribir:

$$(\beta^1, \beta^2, \ldots, \beta^m)^t = \frac{1}{\tau} A\, (\theta^1, \theta^2, \ldots, \theta^m)^t \qquad (107)$$

o equivalentemente:

$$\beta^j = \frac{1}{\tau} \sum_{i=1}^{m} a_{ij}\, \theta^i; \quad \sigma^2 = \sigma^2 \qquad (108)$$

donde $a_{ij}$ son los elementos de A.

La matriz jacobiana de la transformación (108) es:

$$J = \begin{pmatrix} \frac{1}{\tau} T D^{-1/2} & 0 \\ 0 & 1 \end{pmatrix} \qquad (109)$$

por consiguiente el tensor métrico $\bar{G}$ queda transformado en:

$$\hat{G} = J^t \bar{G} J \qquad (110)$$

desarrollando este producto se tiene

$$\hat{G} = \begin{pmatrix} \frac{1}{2} D^{-1/2} T^t & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma^2} X^t X & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} \begin{pmatrix} \frac{1}{2} T D^{-1/2} & 0 \\ 0 & 1 \end{pmatrix}$$

Al ser $X^t X = T D T^t$ y tomando $\tau = \frac{1}{\sqrt{n}}$, resulta:

$$\hat{G} = \begin{pmatrix} \frac{1}{\tau^2 \sigma^2} I & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} = n \begin{pmatrix} \frac{1}{\sigma^2} I & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} \qquad (111)$$

Por lo tanto el tensor métrico bajo $(\theta^1, \theta^2, \ldots, \theta^m, \sigma^2)$ es $\hat{G} = (\hat{g}_{ij})$; $i,j = 1 \leq i,j \leq m+1$

con,

$$\hat{g}_{ij} = 0 \quad \text{si } i \neq j \qquad 1 \leq i,j \leq m+1$$

$$\hat{g}_{kk} = \frac{n}{\sigma^2} \qquad 1 \leq k \leq m \tag{112}$$

$$\hat{g}_{m+1\ m+1} = \frac{n}{2\ \sigma^4}$$

Vamos a hallar los símbolos de Cristoffel de primera especie, como:

$$\Gamma_{ij/k} = \frac{1}{2}\ (D_j \hat{g}_{ik} + D_i\ \hat{g}_{jk} - D_k\ \hat{g}_{ij}) \tag{113}$$

se tiene que

$$\Gamma_{ij/k} = 0; \quad \Gamma_{m+1\ j/k} = \Gamma_{j\ m+1/k} = \frac{n\ \delta_{ij}}{2\sigma^4} \tag{114}$$

$$\Gamma_{ij/m+1} = \frac{n\ \delta_{ij}}{2\sigma^4}\ ; \quad \Gamma_{m+1\ m+1} = 0; \quad \Gamma_{m+1\ j/m+1} = \Gamma_{i\ m+1/m+1} = 0$$

$$\Gamma_{m+1\ m+1/m+1} = -\frac{n}{2\sigma^6}$$

siendo

$$i \leq i,j,k \leq m$$

La matriz inversa del tensor métrico $G^{-1}$ será:

$$G^{-1} = (g^{k\alpha})$$

con

$$g^{k\alpha} = 0 \quad \text{si } k \neq \alpha$$

$$g^{hh} = \frac{\sigma^2}{n} \qquad 1 \leq h \leq m \tag{115}$$

$$g^{m+1\ m+1} = \frac{2\ \sigma^4}{n} \qquad 1 \leq k,d \leq m+1$$

Los símbolos de Cristoffel de segunda especie son:

$$\Gamma^{k}_{ij} = g^{k\alpha} \; \Gamma_{ij/\alpha} \tag{116}$$

En nuestro caso por ser $G^{-1}$ diagonal tenemos que:

$$\Gamma^{k}_{ij} = g^{kk} \; \Gamma_{ij/k} \tag{117}$$

los símbolos de Christoffel de segunda especie serán:

$$\Gamma^{k}_{ij} = 0; \; \Gamma^{k}_{m+1\ j} = \Gamma^{k}_{j\ m+1} = -\frac{1}{2}\frac{\delta jk}{\sigma 2} \; ; \; \Gamma^{m+1}_{ij} = \delta_{ij} \tag{118}$$

$$\Gamma^{k}_{m+1\ m+1} = 0; \; \Gamma^{m+1}_{m+1\ j} = \Gamma^{m+1}_{j\ m+1} = 0; \; \Gamma^{m+1}_{m+1\ m+1} = -\frac{1}{\sigma^2}$$

Pasemos seguidamente a calcular las ecuaciones de las geodésicas que, como sabemos, teniendo en cuenta el convenio de sumación de índices repetidos, son las soluciones del sistema

$$\left.\begin{array}{l} \dfrac{d^2\theta^{\gamma}}{ds^2} + \Gamma^{\gamma}_{\alpha\beta}\dfrac{d\,\theta^{\alpha}}{ds}\dfrac{d\,\theta^{\beta}}{ds} = 0 \\ \\ 1 \leq \gamma \leq m+1 \end{array}\right\} \tag{119}$$

siendo $\theta^{m+1} = \sigma^2$

Para $1 \leq \gamma \leq m$, las ecuaciones (119), serán en nuestro caso, teniendo en cuenta (118).

$$\frac{d^2\theta^{\gamma}}{ds^2} + 2\,\Gamma^{\gamma}_{m+1}\;\frac{d\,\sigma^2}{ds}\;\frac{d\,\theta^{\beta}}{ds} = 0$$

es decir

$$\frac{d^2\theta^\gamma}{ds^2} - \frac{d\,\sigma^2}{ds}\ \frac{d\,\theta^\beta}{ds}\ \frac{\delta_{\gamma\beta}}{\sigma^2} = 0 \iff$$

$$\iff \frac{d^2\theta^\gamma}{ds^2} - \frac{1}{\sigma^2}\ \frac{d\,\theta^\gamma}{ds}\ \frac{d\,\sigma^2}{ds} = 0 \qquad (120)$$

$$1 \leqq \gamma \leqq m$$

la ecuación que falta para completar el sistema (119) es:

$$\frac{d^2\,\sigma^2}{ds^2} + \Gamma^{m+1}_{\alpha\beta}\ \frac{d\,\theta^\alpha}{ds}\ \frac{d\,\theta^\beta}{ds} + \Gamma^{m+1}_{m+1\ m+1}\left(\frac{d\sigma^2}{ds}\right)^2 = 0 \qquad (121)$$

teniendo en cuenta (118) se tiene

$$\frac{d^2\,\sigma^2}{ds^2} + \delta_{\alpha\beta}\frac{d\,\theta^\alpha}{ds}\ \frac{d\,\theta^\beta}{ds} - \frac{1}{\sigma^2}\left(\frac{d\,\sigma^2}{ds}\right)^2 = 0 \qquad (122)$$

desarrollando

$$\frac{d^2\,\sigma^2}{ds^2} + \sum_{\alpha=1}^{m}\left(\frac{d\,\theta^\alpha}{ds}\right)^2 - \frac{1}{\sigma^2}\left(\frac{d\,\sigma^2}{ds}\right)^2 = 0 \qquad (123)$$

Tenemos pues que las ecuaciones de las geodésicas son las soluciones de:

$$\left.\begin{aligned} &\frac{d^2\theta^\gamma}{ds} - \frac{1}{\sigma^2}\ \frac{d\,\theta^\gamma}{ds}\ \frac{d\,\sigma^2}{ds} = 0 \\ &\gamma = 1 \leqq \gamma \leqq m \\ &\frac{d^2\,\sigma^2}{ds^2} + \sum_{\alpha=1}^{m}\left(\frac{d\,\theta^\alpha}{ds}\right)^2 - \frac{1}{\sigma^2}\left(\frac{d\,\sigma^2}{ds}\right)^2 = 0 \end{aligned}\right\} \qquad (124)$$

con la condición de que el vector tangente sea unitario, es decir:

$$1 = \frac{n}{2}\left(\frac{2}{\sigma^2}\sum_{\alpha=1}^{m}\left(\frac{d\,\theta^{\alpha}}{ds}\right)^2 + \frac{1}{\sigma^4}\left(\frac{d\,\sigma^2}{ds}\right)\right)^2 \qquad (125)$$

De las primeras ecuaciones del sistema (124) se obtiene

$$\ln\frac{d\,\theta^{\gamma}}{ds} = \ln\sigma^2 + k \Longleftrightarrow \ln\frac{d\,\theta^{\gamma}}{ds} = \ln A_{\gamma}\,\sigma^2 \qquad (126)$$

por consiguiente

$$\frac{d\,\theta^{\gamma}}{ds} = A_{\gamma}\,\sigma^2 \qquad 1 \leq \gamma \leq m \qquad (127)$$

con las condiciones de contorno para $s = 0$, $\theta^{\gamma} = \theta^{\gamma}_{A}$ y para $s = d, \theta^{\gamma} = \theta^{\gamma}_{B}$, $1 \leq \gamma \leq m$, resulta

$$A_{\gamma} = k\,(\theta^{\gamma}{}_{B} - \theta^{\gamma}{}_{A}) \qquad 1 \leq \gamma \leq m \qquad (128)$$

Definamos la transformación:

$$\left.\begin{aligned}(\mu^1, \mu^2, \ldots, \mu^m)^t &= V^t\,(\theta^1, \theta^2, \ldots, \theta^m)^t \\ \sigma^2 &= \sigma^2\end{aligned}\right\} \qquad (129)$$

con la matriz V ortonormal. De esta forma se tiene:

$$\mu^i = V_i^t\,\theta \qquad 1 \leq i \leq m \qquad (130)$$

siendo $V_i$ el vector columna de la matriz V y $\theta$ el vector columna $\theta = (\theta^1, \theta^2, \ldots, \theta^m)^t$.

Sin pérdida de generalidad se puede elegir $V_1$ como:

$$V_1^t = \frac{1}{\| \theta_B - \theta_A \|} (\theta_B - \theta_A)^t \qquad (131)$$

Teniendo en cuenta (128), (130) y (131) se obtiene

$$\frac{d\,\mu^1}{ds} = k\, V_1 \quad \sigma^2$$

$$\frac{d\,\mu^i}{ds} = 0 \qquad (132)$$

Al ser

$$\frac{d\mu}{ds} = V^t \quad \frac{d\sigma}{ds} \qquad (133)$$

resulta:

$$\left(\frac{d\,\mu}{ds}\right)^t \quad \frac{d\,\mu}{ds} = \left(\frac{d\,\theta}{ds}\right)^t \quad V\, V^t \frac{d\,\theta}{ds} = \left(\frac{d\,\theta}{ds}\right)^t \quad \frac{d\,\theta}{ds} \qquad (134)$$

como

$$\left(\frac{d\,\theta}{ds}\right)^t \quad \frac{d\,\theta}{ds} = \left(\frac{d\,\mu}{ds}\right)^t \quad \frac{d\,\mu}{ds} \qquad (135)$$

teniendo en cuenta (132) y (135) la última ecuación del sistema de ecuaciones (124) queda como

$$\frac{d^2\sigma^2}{ds^2} + \left(\frac{d\,\mu^1}{ds}\right)^t - \frac{1}{\sigma^2} \left(\frac{d\,\sigma^2}{ds}\right)^2 = 0 \qquad (136)$$

con la condición (125)

Las ecuaciones de las geodésicas son las soluciones de:

$$\left.\begin{aligned} \frac{d\,\mu^1}{ds} &= A\ \sigma^2 \\ \frac{d\,\mu^i}{ds} &= 0 \qquad 2 \leq i \leq n \\ \frac{d\,\sigma^2}{ds^2} + \left(\frac{d\,\mu^1}{ds}\right)^2 - \frac{1}{\sigma^2}\ \frac{d\,\sigma^2}{ds} &= 0 \end{aligned}\right\} \qquad (137)$$

con la condición de que el vector tangente sea unitario

$$1 = \frac{n}{2}\left(\frac{2}{\sigma^2}\left(\frac{d\,\mu^1}{ds}\right)^2 + \frac{1}{\sigma^4}\left(\frac{d\,\sigma^2}{ds}\right)^2\right) \qquad (138)$$

Estas ecuaciones coinciden con las planteadas por Burbea y Rao (1982), salvo que en el vector unidad está dividido por n, por lo que distancia al cuadrado entre dos puntos $(\mu^A, \sigma_A^2)$ y$(\mu^B, \sigma_B^2)$ que sería la distancia al cuadrado entre dos clases de poblaciones estadísticas $[\omega_A]$ y $[\omega_B]$ viene multiplicada por n. La expresión de la distancia será:

$$d_{AB} = \sqrt{2n}\ \ln \frac{1 + \sqrt{\Delta_{AB}}}{1 - \sqrt{\Delta_{AB}}} \qquad (139)$$

siendo

$$\Delta_{AB} = \frac{(\mu_1^B - \mu_1^A)^2 + 2\,(\sigma_B - \sigma_A)^2}{(\mu_1^B - \mu_1^A)^2 + 2\,(\sigma_B + \sigma_A)^2} \qquad (140)$$

Teniendo en cuenta la transformación (130) y las expresiones (131) tenemos que:

$$\mu_1^A = V_1^t \ \theta_A$$
$$\mu_1^B = V_1^t \ \theta_B \qquad (141)$$

por consiguiente

$$\mu_1^B - \mu_1^A = V_1^t \ (\theta_B - \theta_A) = \| \ \theta_B - \theta_A \ \| \qquad (142)$$

por otra parte teniendo en cuenta la transformación (106) se tiene que

$$\delta_{AB} = \| \theta_B - \theta_A \| = \frac{1}{\sqrt{n}} \left( (\beta_B - \beta_A)^t \ X^t \ X \ (\beta_B - \beta_A) \right)^{\frac{1}{2}} \qquad (143)$$

la expresión de la distancia será:

$$d_{AB} = \sqrt{2n} \ \ln \frac{1 + \sqrt{\Delta_{AB}}}{1 - \sqrt{\Delta_{AB}}} \qquad (144)$$

con

$$\Delta_{AB} = \frac{\delta_{AB}^2 + 2 \ (\sigma_B - \sigma_A)^2}{\delta_{AB}^2 + 2 \ (\sigma_B + \sigma_A)^2} \qquad (145)$$

### 3.2.2. Aspectos estadísticos

En el presente apartado se plantea un contraste entre dos modelos lineales normales univariantes de media y varianza no necesariamente iguales para cada modelo, usando la distancia obtenida en (144).

El problema de comparar las medias de dos poblaciones normales con distinta varianza fue tratado por Beherens i Fisher. En nuestro caso comparamos dos poblaciones estadísticas asociadas a dos modelos lineales normales, teniendo en cuenta las diferencias que presentan el vector de medias y la matriz de varianzas-covarianzas simultáneamente.

Sean

$$\begin{aligned} y_A &= X\,\beta_A + e_A \\ y_B &= X\,\beta_B + e_B \end{aligned} \qquad (146)$$

dos modelos lineales como los descritos en (86) y tal que:

$$y_A = (y_1^A, y_2^A, \ldots, y_N^A)^t; \quad y_B = (y_1^B, y_2^B, \ldots, y_N^B)^t$$

$$\beta_A = (\beta_A^1, \beta_A^2, \ldots, \beta_A^m)^t; \quad \beta_B = (\beta_B^1, \beta_B^2, \ldots, \beta_B^m)^t$$

$$e_A = (e_1^A, e_2^A, \ldots, e_N^A)^t; \quad e_B = (e_1^B, e_2^B, \ldots, e_N^B)^t$$

$$X = (x_{ij}) \qquad 1 \leq i \leq N; \qquad 1 \leq j \leq m$$

$$e_j^A \sim N(o, \sigma_A); \quad e_j^B \sim N(o, \sigma_B); \quad 1 \leq j \leq N$$

$e_j^A$, $e_j^B$ independientes para todo j

Vamos a obtener un estimador de la distancia entre ambos modelos lineales. El estimador de esta distancia tendrá por expresión:

$$\hat{d}_{AB} = \sqrt{2n}\ \ln \frac{1 + \sqrt{\hat{\Delta}_{AB}}}{1 - \sqrt{\hat{\Delta}_{AB}}} \qquad (147)$$

siendo

$$\hat{\Delta}_{AB} = \frac{\hat{\delta}_{AB} + 2\ (\hat{\sigma}_B - \hat{\sigma}_A)^2}{\hat{\delta}_{AB} + 2\ (\hat{\sigma}_B + \hat{\sigma}_A)^2} \qquad (148)$$

y con

$$\hat{\delta}^2_{AB} = \frac{1}{n}\ (\hat{\beta}_B - \hat{\beta}_A)^t\ X^t\ X\ (\hat{\beta}_B - \hat{\beta}_A) \qquad (149)$$

teniendo en cuenta (27) nos queda

$$\delta^2_{AB} = \frac{1}{n}\ \ (y_B - y_A)^t\ X\ (X^tX)^{-1}\ X^t\ (y_B - y_A) \qquad (150)$$

por otra parte, la estimación máximo verosímil de $\hat{\sigma}_i$ (i=A,B) es:

$$\hat{\sigma}_i = R^2_{oi}\ /N \qquad (151)$$

Todo esto nos permite realizar contrastes de hipótesis del tipo

$$\begin{aligned} H_o&: d_{AB} = 0 \\ H_1&: d_{AB} > 0 \end{aligned} \qquad (152)$$

la hipótesis nula es que la distancia entre dos poblaciones estadísticas es cero, frente a la hipótesis alternativa de que la distancia entre dichas poblaciones es mayor de cero, puesto que bajo la hipótesis nula y asintóticamente, el estadístico,

$$U = \frac{N}{2} \hat{d}^2_{AB} \tag{153}$$

converge a una distribución ji-cuadrado con m+1, número de parámetros, grados de libertad. Oller (1982).

Por lo tanto la región crítica asociada al contraste (152) será a nivel de significación $\varepsilon$,

$$W_\varepsilon = \{z \in R^{2N} \ / \ \hat{d}^2_{AB} > \frac{2k}{N}\}$$

donde k es tal que

$$P\ (\chi^2 > k) = \varepsilon \tag{155}$$

y $\chi^2$ es una variable aleatoria que sigue una distribución ji-cuadrado con m+1 grados de libertad.

# 4. DISTANCIAS ENTRE MODELOS LINEALES MULTIVARIANTES

Página

Sumario:

En este capítulo se define y estudia una distancia entre modelos lineales normales, caso multivariante con matriz de varianzas-covarianzas iguales.

También aquí, en primer lugar se ha introducido una distancia entre clases de poblaciones estadísticas, representadas por modelos lineales normales multivariantes, efectuando después un estudio de sus aspectos estadísticos.

## 4.1. DISTANCIAS ENTRE MODELOS LINEALES NORMALES MULTIVARIANTES Y DE IGUAL MATRIZ DE VARIANZAS-COVARIANZAS

Se trata de introducir una distancia entre clases de poblaciones estadísticas, asociadas a modelos lineales normales multivariantes con la misma matriz de varianzas-covarianzas. Cada modelo viene fijado por unas condiciones experimentales que determinan la matriz de diseño, el número de parámetros, el número de réplicas por condición experimental y el número de variables observadas. Es decir, se parte del conjunto de modelos lineales cuya expresión matricial es:

$$Y = XB + E \qquad (1)$$

donde Y es una matriz cuyos vectores fila representan a la muestra de tamaño n p-dimensional y cuyas componentes son la p variables $Y_1$, $Y_2$, ..., $Y_p$, medidas en las distintas condiciones experimentales determinadas por las filas de X, siendo $Y = (y_{ij})$ ; $1 \leq i \leq n$; $1 \leq j \leq p$. B es la matriz de paráme-

tros $B = (b_{ij})$; $1 \leq i \leq m$; $1 \leq j \leq p$ cuyos vectores columna $\beta_i = (b_{1i}, b_{2i}, \ldots, b_{mi})^t$, son los parámetros que determinan los valores medios de $Y_i$ por condición experimental. $E = (e_{ij})$; $1 \leq i \leq n$; $1 \leq j \leq p$; es la matriz de las desviaciones aleatorias del modelo cuyas filas son variables aleatorias normales multivariantes de media 0 con matriz de varianzas-covarianzas $S = (\sigma_{ij})$; $1 \leq i,j \leq p$ que supondremos igual para todos los modelos a distanciar. $X = (x_{ij})$; $1 \leq i \leq n$; $1 \leq j \leq m$; es la matriz de diseño que supondremos de rango máximo rang $(X) = m$, de otra forma se haría una reparametrización que transformaría el modelo en un diseño de rango máximo.

El modelo lineal multivariante (1) lo podemos considerar constituído por p modelos univariantes para las variables $Y_i$

$$Y_i = X \beta_i + e_i \qquad (2)$$

siendo $e_i$ los vectores columna de la matriz E.

Para facilitar la construcción de la distancia, la matriz de observaciones Y la podemos identificar con un vector columna cuyos componentes son

$$Y = (y_1^t, y_2^t, \ldots, y_p^t)^t$$

Análogamente la matriz B de parámetros la podemos identificar como un vector columna

$$\beta = (\beta_1^t, \beta_2^t, \ldots, \beta_p^t)^t$$

y la matriz E de desviaciones aleatorias del modelo la podemos identificar con el vector columna

$$e = (e_1^t, e_2^t, \ldots, e_p^t)^t$$

El modelo (1) quedaría como

$$y = A\beta + e \qquad (3)$$

donde la nueva matriz de diseño será: A = I x X (producto de Kronecker de I (p x p) por X (matriz de diseño del modelo (1)).

Con esta nueva presentación del modelo (1) en la forma (3), se tiene que el vector columna de las desviaciones aleatorias es una variable aleatoria normal multivariante de media 0 y matriz de varianzas-covarianzas.

$$\Sigma_o = S \times I \qquad (4)$$

(producto de Kronecker de la matriz S por la matriz identidad I (nxn)).

Por todo esto se puede considerar que la variable aleatoria y es una variable aleatoria normal multivariante de media $\mu = A\beta$ y matriz de varianzas-covarianzas $\Sigma_o$, igual para todos los modelos.

4.1.1. Formalización y Aspectos Geométricos

Vamos a construir a continuación una distancia entre dos poblaciones estadísticas representadas por dos modelos lineales multivariantes del tipo (1).

$$Y_a = XB_a + E_a \qquad Y_b = XB_b + E_b \tag{5}$$

Para formalizar la construcción de la distancia, consideraremos a los modelos (5) en la forma (3) quedando ambos modelos como

$$y_a = A\,\beta_a + e_a \qquad y_b = A\,\beta_b + e_b \tag{6}$$

Siguiendo el mismo proceso que en el punto 3.1.1., si tenemos dos poblaciones estadísticas $\omega_a$ y $\omega_b$ tales que $h(\omega_a) = \beta_a$, $h(\omega_b) = \beta_b$, la expresión al cuadrado de la distancia será:

$$L^2 = d^2((\omega_a), (\omega_b)) = (A\,\beta_a - A\,\beta_b)^t\ \Sigma_o^{-1}\ (A\,\beta_a - A\,\beta_b) = \tag{7}$$

$$= (\beta_a - \beta_b)^t\, A^t\, \Sigma_o^{-1}\, A\, (\beta_a - \beta_b)$$

siendo

$$\Sigma_o^{-1} = S^{-1} \times I \tag{8}$$

con $S^{-1} = (\sigma^{ij})$; $1 \leq i,j \leq p$ (inversa de la matriz de varianzas-covarianzas de las filas p-dimensionales de las matrices de los errores de los modelos (6) que supondremos igual para los dos modelos) y donde I es la matriz identidad de orden n x n. Por consiguiente $\Sigma_o^{-1}$ será una matriz de cajas

$$S_{ij}^{-1} = \text{diag.}\ (\sigma^{ij}, \sigma^{ij}, \ldots, \sigma^{ij}) \qquad (9)$$

$$1 \leq i,j \leq p$$

Desarrollando la expresión (7) tenemos que

$$L^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} (\beta_j^a - \beta_j^b)^t X^t S_{ij}^{-1} X (\beta_i^a - \beta_i^b) \qquad (10)$$

siendo $\beta_k^a$, $\beta_k^b$ los vectores columna de las matrices de parámetros $\beta_a, \beta_b$ de los modelos (6).

Pero $S_{ij}^{-1}$ es una matriz escalar, por consiguiente

$$L^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} (\sigma^{ij}) (\beta_j^a - \beta_j^b)^t X^t X (\beta_i^a - \beta_i^b) \qquad (11)$$

o también

$$\begin{aligned} L^2 &= \text{tr}\left[S^{-1} (B_a - B_b)^t X^t X (B_a - B_b)\right] = \\ &= \text{tr}\left[(B_a - B_b)^t X^t X (B_a - B_b) S^{-1}\right] \end{aligned} \qquad (12)$$

Esta traza como sabemos es la suma de los valores propios $\lambda_1, \lambda_2, \ldots, \lambda_p$ de la matriz

$$V = (B_a - B_b)^t (X^t X) (B_a - B_b) S^{-1} \qquad (13)$$

respecto de la matriz I (p x p), o la suma de los valores propios de la matriz

$$W = (B_a - B_b)^t (X^t X) (B_a - B_b) \qquad (14)$$

respecto de S.

La suma, por consiguiente, de los valores propios de W respecto de S es la distancia al cuadrado $L^2$.

Obsérvese en (11) que en caso de que las variables componentes de la variable p-dimensional sean independientes, la distancia $L^2$ es la suma de los cuadrados de las distancias entre los modelos lineales univariantes que forman parte de los modelos multivariantes (5).

### 4.1.2. Aspectos Estadísticos: Estimación y contrastes de hipótesis

Sean dos modelos lineales multivariantes, expresados en forma matricial como:

$$\begin{aligned} Y_a &= XB_a + E_a \\ Y_b &= XB_b + E_b \end{aligned} \qquad (15)$$

donde

$$Y_a = (y^a_{ij}); \quad Y_b = (y^b_{ij}); \quad 1 \leq i \leq N; \quad 1 \leq j \leq p$$

$$B_a = (b^a_{ij}); \quad B_b = (b^b_{ij}); \quad 1 \leq i \leq m; \quad 1 \leq j \leq p$$

$$E_a = (e^a_{ij}); \quad E_b = (e^b_{ij}); \quad 1 \leq i \leq N; \quad 1 \leq j \leq p$$

y $e^a_i$, $e^b_i$, los vectores fila de las matrices $E_a$ y $E_b$ respectivamente, con $e^a_i$, $e^b_i \sim N(0,S)$; $S = (\sigma_{ij})$, $1 \leq i,j \leq p$; $X = (x_{ij})$ $1 \leq i \leq N$; $1 \leq j \leq m$.

Vamos a obtener un estimador de la distancia entre ambos modelos lineales en función de la muestra observada, para ello vamos a expresar los modelo (15) en la forma (6), es decir:

$$\begin{aligned} y_a &= A\,\beta_a + e_a \\ y_b &= A\,\beta_b + e_b \end{aligned} \qquad (16)$$

donde

$$y_a = \left((y_1^a)^t\ (y_2^a)^t \ldots (y_p^a)^t\right)^t ;\quad y_b = \left((y_1^b)^t\ (y_2^b)^t \ldots (y_p^b)^t\right)^t$$

$$\beta_a = \left((\beta_1^a)^t\ (\beta_2^a)^t \ldots (\beta_p^a)^t\right)^t ;\quad \beta = \left((\beta_1^b)^t\ (\beta_2^b)^t \ldots (\beta_p^b)^t\right)^t$$

$$e_a = \left((e_1^a)^t\ (e_2^a)^t \ldots (e_p^a)^t\right)^t ;\quad e = \left((e_1^b)^t\ (e_2^b)^t \ldots (e_p^b)^t\right)^t$$

$A = I \times X$ (producto de Kronecker de I (pxp) por la matriz X)

por otra parte, $e_i^a,\ e_i^b \sim N(0,\Sigma_o)$; $1 \leq i \leq p$, siendo $\Sigma_o = S \times I$ (producto de Kronecker de S por la matriz I (n x n).

En las aplicaciones prácticas, como hemos dicho en el capítuio anterior, conocemos la muestra observada, la matriz de diseño y desconocemos en general $\beta$ y $\Sigma_o$, por lo tanto para el cálculo explícito de la distancia entre los dos modelos lineales habrá que estimarlos.

La estimación mínimo cuadrática (L.S.) de los parámetros del modelo (15) es:

$$\hat{\beta}_i = (A^t A)^{-1} A^t y_i \qquad (17)$$

si

$$\beta_j^i = (b_{1j}^i, b_{2j}^i, \ldots, b_{mj}^i)^t; \quad y_j^i = (y_{1j}^i, y_{2j}^i, \ldots, y_N^i)^t$$

$$\hat{\beta}_j^i = (X^tX)^{-1} X^t y_j^i \qquad (18)$$

$$i = a, b$$

La estimación de la distancia, $L^2$, cuando la matriz de varianzas-covarianzas es conocida, se obtiene sustituyendo en (11) los parámetros por su estimación L.S.

$$\hat{L}^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} (\sigma^{ij}) \; (y_i^a - y_i^b)^t X \; (X^tX)^{-1} X^t \; (y_j^a - y_j^b) \qquad (19)$$

En el caso de usar las ecuaciones normales reducidas, véase Cuadras (1979), la expresión de la distancia estimada será:

$$\hat{L}^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} (\sigma^{ij}) \; (\bar{y}_i^a - \bar{y}_i^b)^t \; \Delta \; X_r (X_r^t \Delta X_r) \; X_r^t \Delta \; (\bar{y}_j^a - \bar{y}_j^b) \qquad (20)$$

donde $\bar{y}_j^k = (\bar{y}_{1j}^k, \bar{y}_{2j}^k, \ldots, \bar{y}_{hj}^k)^t$; $k = 1,b$; $1 \leq j \leq p$, son los vectores de medias de la variable j observada correspondiente a una misma condición experimental, $\Delta = \text{diag}\,(n_1, n_2, \ldots, n_h)$, con $n_i$ número de réplicas de la i-ésima condición experimental, h es el número de condiciones experimentales diferentes y $X_r$ es la matriz de diseño reducida.

Por otra parte, la expresión (19) y (20), de la estimación de la distancia al cuadrado, $\hat{L}^2$, la podemos expresar como:

$$\hat{L}^2 = \text{tr}\left((Y_a-Y_b)^t \; X \; (X^tX)^{-1} \; X^t \; (Y_a-Y_b) \; S^{-1}\right) \qquad (21)$$

que como sabemos es la suma de los valores propios de la matriz

$$V = (Y_a-Y_b)^t \; X \; (X^tX)^{-1} \; X^t \; (Y_a-Y_b) \; S^{-1} \qquad (22)$$

respecto a la matriz I (p x p) o bien la suma de los valores propios de la matriz

$$W = (Y_a-Y_b)^t \; X \; (X^tX)^{-1} \; X^t \; (Y_a-Y_b) \qquad (23)$$

respecto a la matriz S.

También puede expresarse en términos de $\bar{Y}_k, k = a,b$; $X_r$, $\Delta$, siendo $\bar{Y}_k$ una matriz cuyos vectores columna son $\bar{y}^k_j$, $1 \leq j \leq p$, y queda

$$\hat{L}^2 = \text{tr}\left((\bar{Y}_a-\bar{Y}_b)^t \Delta \; X_r \; (X_r^t \; \Delta X_r) \; X_r^t \Delta \; (\bar{Y}_a-\bar{Y}_b) \, S^{-1}\right) \qquad (24)$$

Cuando la matriz de varianzas-covarianzas no sea conocida tendremos que sustituirla por una estimación de la misma que a continuación vamos a obtener.

Sea la expresión matricial del modelo lineal multivariante conjunto

$$Y = ZB + E \qquad (25)$$

con

$$Y = \begin{pmatrix} Y_a \\ Y_b \end{pmatrix} \qquad Z = \begin{pmatrix} X & 0 \\ 0 & X \end{pmatrix} \qquad B = \begin{pmatrix} B_a \\ B_b \end{pmatrix} \qquad E = \begin{pmatrix} E_a \\ E_b \end{pmatrix}$$

El rango del diseño es rang (Z) = 2m.

La estimación de la matriz de varianzas-covarianzas es:

$$\hat{S} = (2\ (N - m))^{-1}\ R_o \tag{26}$$

siendo $R_o$ la matriz de dispersión del modelo (25), por consiguiente $R_o$ será:

$$R_o = (Y - Z\hat{B})^t\ (Y - Z\hat{B}) \tag{27}$$

Al verificarse

$$\hat{B} = (Z^tZ)^{-1}\ Z^t\ Y \tag{28}$$

resulta que

$$R_o = Y^t\ (1 - Z\ (Z^tZ)^{-1}\ Z^t)\ Y \tag{29}$$

La expresión de la distancia estimada será:

$$\hat{L}^2 = tr\ \left(\hat{S}^{-1}\ (Y_a - Y_b)^t\ X\ (X^tX)^{-1}\ X^t\ (Y_a - Y_b)\right) \tag{30}$$

sustituyendo (26) en (30) nos queda:

$$\hat{L}^2 = 2\ (N-m)\ tr\ \left(R_o^{-1}\ (Y_a - Y_b)^t\ X\ (X^tX)^{-1}\ (Y_a - Y_b)\right) \tag{31}$$

A continuación vamos a estudiar contrastes de hipótesis del tipo

$$H_o: L = 0$$
$$H_1: L > 0 \qquad (32)$$

es decir, contrastes cuya hipótesis nula es que la distancia entre dos poblaciones es cero, frente a la hipótesis alternativa de que la distancia entre las dos poblaciones estadísticas es mayor que cero. También se estudiará la relación de estos contrastes con los contrastes usuales del análisis multivariante a la varianza, estudiando la distribución del estimador de la distancia bajo $H_o$. Este problema fue propuesto por Cuadras (1979).

La región crítica de estos contrastes vendrá dada por:

$$W_\varepsilon = \{T \in M_{2N \times p} / \hat{L}(T) > A_\varepsilon\} \qquad (33)$$

Si suponemos que es cierta la hipótesis $H_o$, esto es equivalente a considerar como hipótesis nula

$$B_a = B_b = B_o \qquad (34)$$

Matricialmente, la hipótesis nula la podemos expresar como:

$$HB = 0 \qquad (35)$$

con H definida igual que en (41) del punto 3.1.2. y con el rango de H igual a m.

Bajo la hipótesis nula la expresión matricial del modelo conjunto sería, reparametrizando

$$Y = Z\ B_o + E \tag{36}$$

con

$$Y = \begin{pmatrix} Y_a \\ Y_b \end{pmatrix} \qquad Z = \begin{pmatrix} X \\ X \end{pmatrix} \qquad E = \begin{pmatrix} E_a \\ E_b \end{pmatrix}$$

El resíduo bajo la hipótesis nula será:

$$R_1 = (Y - Z\hat{B}_o)^t\ (y - Z\hat{B}_o) \tag{37}$$

donde $\hat{B}_o$ es la estimación L.S. de $B_o$.

Teniendo en cuenta (36), obtenemos

$$R_1 = (Y_a - X\hat{B}_o)^t\ (Y_a - X\hat{B}_o) + (Y_b - X\hat{B}_o)^t\ (Y_b - X\hat{B}_o) \tag{38}$$

y por otra parte, la estimación L.S. de $\hat{B}_o$ viene dada por:

$$\hat{B}_o = (Z^tZ)^{-1}\ Z^t\ Y = \frac{1}{2}\ (X^tX)^{-1}\ X^t\ (Y_a + Y_b) = \frac{1}{2}\ (\hat{B}_a + \hat{B}_b) \tag{39}$$

combinando (38) y (39), resulta:

$$R_1 = Y_a^t\ Y_a + Y_b^t\ Y_b - \frac{1}{2}\ /Y_a + Y_b)^t\ X\ (X^tX)^{-1}\ X^t\ (Y_a + Y_b) \tag{40}$$

La desviación de la hipótesis será, teniendo en cuenta la expresión (29),

$$R_1 - R_o = \frac{1}{2} (Y_a - Y_b)^t X (X^t X)^{-1} X^t (Y_a - Y_b) \qquad (41)$$

Por otra parte,

$$U = \frac{1}{p} tr \left[ (R_1 - R_o) S^{-1} \right] \qquad (42)$$

sigue una distribución $\chi^2$ con m grados de libertad, ya que U cumple

$$U = \frac{1}{p} \sum_{i=1}^{p} \lambda_i \qquad (43)$$

siendo $\lambda_i$ los valores propios de $(R_1-R_o)S^{-1}$ respecto a I (pxp).

En efecto, sean $v_i$ y $\lambda_i$ vector y valor propio respectivamente, por consiguiente

$$v_i^t (R_1-R_o) S^{-1} = \lambda_i v_i^t \iff v_i^t (R_1-R_o) v_i = \lambda_i v_i^t S v_i \qquad (44)$$

despejando $\lambda_i$, se obtiene

$$\lambda_i = \frac{v_i^t (R_1-R_o) v_i}{v_i^t S v_i} \qquad (45)$$

Al ser $R_1 - R_o$ semidefinida positiva y S definida positiva podemos elegir los vectores propios, tales que

$$\begin{aligned} v_i^t (R_1-R_o) v_j &= \delta_{ij} \lambda_i \\ v_i^t S v_j &= \delta_{ij} \end{aligned} \qquad (46)$$

Teniendo en cuenta (43) y (45), tenemos:

$$U = \frac{1}{p} \operatorname{tr} (R_1 - R_o) S^{-1} = \frac{1}{p} \sum_{i=1}^{p} \frac{v_i^t (R_1 - R_o) v_i}{v_s^t S v_i} =$$

$$= \frac{\left(\sum_{i=1}^{p} v_i\right)^t (R_1 - R_o) \left(\sum_{i=1}^{p} v_i\right)}{\left(\sum_{i=1}^{p} v_i\right)^t S \left(\sum_{i=1}^{p} v_i\right)} \qquad (47)$$

que como sabemos, véase Cuadras (1982), sigue una distribución ji-cuadrado con m grados de libertad.

Al ser

$$U = \frac{1}{p} \operatorname{tr} \left((R_1 - R_o) S^{-1}\right) = \frac{1}{2p} \operatorname{tr} \left((Y_a - Y_b)^t X (X^t X)^{-1} X^t (Y_a - Y_b) S^{-1}\right) = \frac{1}{2p} \hat{L}^2 \qquad (48)$$

resulta que cuando la matriz de varianzas-covarianzas es conocida, el estadístico $U = \frac{1}{2p} \hat{L}^2$ sigue una distribución ji-cuadrado con m grados de libertad bajo la hipótesis nula $H_o$ (ambos modelos son iguales). Por tanto la constante $A_\varepsilon$ de (33) vendrá determinada de forma que

$$P\left(\chi^2 > \frac{A_\varepsilon^2}{2p}\right) = \varepsilon \qquad (49)$$

donde $\chi^2$ es una v.a. que sigue una distribución ji-cuadrado con m grados de libertad.

También tenemos que:

$$W = tr\ (R_1 - R_o)\ R_o^{-1} \tag{50}$$

sigue una distribución $T^2$ de Hotelling con 2 N-2 grados de libertad, que como sabemos, véase capítulo II, cumple

$$F = T^2\ \frac{2N - p - 1}{p\ (2N - 2)} \tag{51}$$

sigue una distribución F de Fisher-Suedecor con p y (2N-p-1) grados de libertad.

Por tanto,

$$W = \frac{1}{2}\ tr\ \left[(Y_a - Y_b)^t\ X\ (X^tX)^{-1}\ (Y_a - Y_b)\ R_o^{-1}\right] = \frac{\hat{L}^2}{4\ (N-m)} \tag{52}$$

cuando la matriz de varianzas-covarianzas es desconocida y ha de ser estimada.

Por consiguiente,

$$W = \frac{\hat{L}^2}{4\ (N-m)} \tag{53}$$

sigue, bajo la hipótesis nula $H_o$ (ambos modelos son iguales), una distribución $T^2$ de Hotelling. Por tanto, teniendo en cuenta (51) tenemos que

$$F = \frac{2N - p - 1}{4p\ (2N-2)\ (N-m)}\ \hat{L}^2 \tag{54}$$

sigue una distribución F de Fisher-Suedecor con p y (2N-1) grados de libertad.

La región crítica de nivel de significación $\varepsilon$, de este test, será:

$$W_\varepsilon = \left\{ T \in M_{2N \times p} \;/\; \hat{L}^2 (T) > \frac{4p\ (2N - 2)\ (N - m)}{2N - P} k \right\} \qquad (55)$$

donde k es tal que

$$P\ (F > K) = \varepsilon \qquad (56)$$

y donde F es una variable aleatoria que sigue una distribución F de Fisher-Snedecor con p y 2N - p-1 grados de libertad.

Todo esto nos lleva al siguiente resultado: El test para comparar dos modelos a través del estadístico $\hat{L}^2$ es equivalente al test de comparación usado en el análisis multivariante de varianza, Cuadras (1982), ya que en ambos casos las regiones críticas son equivalentes.

### 4.1.3. Comparación conjunta de q modelos

Este método de comparación de dos modelos lineales multivariantes se puede generalizar como en el capítulo anterior a la comparación simultánea de q modelos lineales multivariantes de rango máximo. Este problema fue propuesto por Cuadras (1979).

Sean q modelos lineales normales multivariantes de rango máximo:

$$Y_i = X\ B_i + E_i \qquad (57)$$

con

$$Y_i = (h^i_{hj}); \quad E_i = (e^i_{hj})$$

$i = 1,2,\ldots,q$; $j = 1,2,\ldots,p$; $h = 1,2,\ldots,N$, siendo $X = (x_{ij})$ la matriz de diseño de rango N x m.

Según hemos visto en el punto anterior el estimador de la distancia entre dos poblaciones estadísticas (dos modelos lineales) $L_{ij} = d((\omega_i), (\omega_i))$, viene dada por:

$$\hat{L}^2_{ij} = tr\left((Y_i-Y_j)^t X (X^tX)^{-1} X^t (Y_i-Y_j) S^{-1}\right) \qquad (58)$$

La suma de los cuadrados de las interdistancias entre todas las poblaciones será:

$$L^2 = \sum_{i<j} L^2_{ij} \qquad (59)$$

y su estimación será:

$$\hat{L}^2 = \sum_{i<j} \hat{L}_{ij} = \sum_{i<j} tr\left((Y_i-Y_j)^t X (X^tX)^{-1} (Y_i-Y_j) S^{-1}\right) \qquad (60)$$

Cuando la matriz de varianza-covarianzas no sea conocida tendremos que realizar una estimación de la misma. A continuación vamos a obtenerla.

Sea el modelo lineal multivariante conjunto de expresión matricial

$$Y = ZB + E \qquad (61)$$

siendo

$$Y = (Y_1^t, Y_2^t, \ldots, Y_q^t)$$

Z = I x X (producto de Kronecker de I (p x p) por X (matriz de diseño)). El rango de Z es rang (Z) = mq.

$$B = (B_1^t, B_2^t, \ldots, B_q^t)^t; \quad E = (E_1^t, E_2^t, \ldots, E_q^t)^t$$

La estimación L.S. de B es:

$$\hat{B} = (Z^t Z)^{-1} Z^t Y \tag{62}$$

desarrollando (62) tenemos que:

$$\hat{B}_i = (X^t X)^{-1} X^t Y_i \quad i = 1,2,\ldots, q \tag{63}$$

Una estimación insesgada de la matriz de varianzas-covarianzas vendrá dada por:

$$\hat{S} = \frac{1}{q\ (N - m)}\ R_o \tag{64}$$

teniendo en cuenta que

$$R_o = (Y - Z\hat{B})^t\ (Y - Z\hat{B}) \tag{65}$$

por (63) y desarrollando se obtiene

$$R_o = \sum_{i=1}^{q} (Y_i^t\ (I - X)\ (X^t X)^{-1} X^t)\ Y_i) \tag{66}$$

dicha expresión es la suma de los resíduos de cada uno de los q modelos por separado, es decir:

$$R_o = R_o(1) + R_o(2) + \ldots + R_o(q) \qquad (67)$$

La estimación de la matriz de varianzas-covarianzas será finalmente

$$\hat{S} = \frac{1}{q\,(N - m)} \sum_{i=1}^{q} \left(Y_i^t \left(I - X\,(X^tX)^{-1}\,X^t\right)\,Y_i\right) \qquad (68)$$

y la estimación de la suma de los cuadrados de las interdistancias será:

$$\hat{L}^2 = \frac{q\,(N - m)}{2} \sum_{i=1}^{q} \sum_{j=1}^{q} tr \left| (Y_i - Y_j)^t\, X\, (X^tX)^{-1}\, X^t\, (Y_i - Yj)\, R_o^{-1} \right| \qquad (69)$$

Todo esto nos permite contrastar hipótesis del tipo:

$$H_o: L = 0 \qquad (70)$$

$$H_1: L > 0$$

la región crítica natural de nivel de significación $\varepsilon$ es:

$$W_\varepsilon = \{T \in M_{qN \times p} \;/\; \hat{L}(T) > A_\varepsilon\} \qquad (71)$$

con esto podemos contrastar la hipótesis nula de que todos los modelos son iguales, es decir, todas las interdistancias son nulas, frente a la hipótesis alternativa de que al menos alguno de los modelos es diferente de los demás.

La hipótesis L = 0 es equivalente a considerar como hipótesis nula que

$$B_1 = B_2 = \ldots = B_q = B_o \tag{72}$$

Matricialmente esta hipótesis la expresaríamos como

$$HB = 0 \tag{73}$$

siendo H la misma matriz que la definida en (71) del punto 3.1.3., que tiene por rango (q - 1) m número de modelos a comparar menos uno, multiplicado por el número de parámetros.

Bajo esta hipótesis la expresión matricial del modelo será:

$$Y = ZB_o + E \tag{74}$$

con $Z = (X^t\ X^t\ \ldots\ X^t)^t$.

La estimación de $B_o$ será:

$$\hat{B}_o = (Z^tZ)^{-1}\ Z\ Y \tag{75}$$

teniendo en cuenta (74) y desarrollando (75) tenemos que

$$\hat{B}_o = \frac{1}{q}\ \left((X^tX)^{-1}\ X^t\ Y_1 + \ldots + (X^tX)^{-1}\ X^t\ Y_q\right) \tag{76}$$

como por otra parte

$$B_i = (X^tX)^{-1}\ X^t\ Y_i \tag{77}$$

se tendrá que

$$\hat{B}_o = \frac{1}{q}\ (\hat{B}_1 + \hat{B}_2 + \ldots + \hat{B}_q) \tag{78}$$

El residuo bajo la hipótesis nula será:

$$R_1 = (Y - Z \hat{B}_o)^t (Y - Z \hat{B}_o) \qquad (79)$$

sustituyendo (76) en (79) se obtiene:

$$R_1 = \sum_{i=1}^{q} Y_i^t Y_i - \frac{1}{q} \left( \sum_{i=1}^{q} Y_i^t \right) X (X^t X)^{-1} X^t \left( \sum_{i=1}^{q} Y_i \right) \qquad (80)$$

la desviación de la hipótesis nula será:

$$R_1 - R_o = \frac{1}{2q} \sum_{i=1}^{q} \sum_{j=1}^{q} \left( (Y_i - Y_j)^t X (X^t X)^{-1} X^t (Y_i - Y_j) \right) \qquad (81)$$

Por otra parte

$$U = \frac{1}{p} \operatorname{tr} \left[ (R_1 - R_o) S^{-1} \right]$$

sigue una distribución ji-cuadrado con (q-1) m grados de libertad, véase (42).

Al ser,

$$U = \frac{1}{p} \operatorname{tr} \left( (R_i - R_o) S^{-1} \right) = \frac{1}{2qp} \sum_{i=1}^{q} \sum_{j=1}^{q} \operatorname{tr} \left[ (Y_i - Y_j) X (X^t X)^{-1} X^t \right.$$

$$\left. (Y_i - Y_j) S^{-1} \right] = \frac{1}{qp} \hat{L}^2$$

resulta que cuando la matriz de varianzas-covarianzas es conocida el estadístico $U = \frac{1}{qp} \hat{L}^2$ sigue una distribución ji-cuadrado con (q-1) m grados de libertad bajo la hipótesis nula $H_o$ (todos los modelos son iguales).

La región crítica de este test, a nivel de significación $\varepsilon$ será:

$$W_\varepsilon = \{T \in M_{q\ N \times p} \ / \ L^2\ (T) > qk\}$$

donde k es tal que

$$P\ (\chi^2 > K) = \varepsilon$$

y $\chi^2$ es una variable aleatoria que sigue una distribución ji-cuadrado con $(q - 1)m$ grados de libertad.

Por otra parte

$$W = tr\ \left((R_1 - R_o)\ R_o^{-1}\right)$$

sigue una distribución $T^2$ de Hotelling con q N-q grados de libertad.

Además se tiene que

$$W = \frac{1}{2q} \sum_{i=1}^{q} \sum_{j=1}^{q} tr\ \left((Y_i - Y_j)\ X\ (X^tX)^{-1}\ X^t\ (Y_i - Y_j)\ R_o^{-1}\right) =$$

$$= \frac{\hat{L}^2}{q^2\ (N - m)}$$

cuando la matriz de varianzas-covarianzas es desconocida y ha de ser estimada.

Por

$$W = \frac{\hat{L}^2}{q^2\ (N - m)}$$

sigue bajo la hipótesis nula $H_o$ (todos los modelos son iguales), una distribución $T^2$ de Hotelling con q N-q grados de libertad.

La región crítica de nivel de significación $\varepsilon$, de este test será:

$$W_\varepsilon = \{T \in M_{q\ Nxp} \ / \ \hat{L}^2\ (T) > q^2\ (N - m)K\}$$

donde k es tal que

$$P\ (T^2 > K) = \varepsilon$$

y $T^2$, siguiendo una distribución $T^2$ de Hotelling con q N-q grados de libertad.

# 5. UNA APLICACION A LA CLASIFICACION DE LOS TEST DE TOLERANCIA ORAL A LA GLUCOSA

Página

Sumario:

## 5.1. INTRODUCCION

Con el término de Diabetes Melitus, la Organización Mundial de la Salud designa a un estado de hiperglucemia crónica producida por la incapacidad del páncreas para segregar insulina o por un exceso de los factores que se oponen a su acción.

En muchos casos para estudiar la etiología, realizar el diagnóstico y hacer un pronóstico de la enfermedad resulta de gran interés hacer determinaciones de glucosa e insulina en plasma, bajo el estímulo de una dosis de glucosa. La prueba de estimulación más usada es el test de tolerancia oral a la glucosa, TTOG.

Todo diagnóstico de la diabetes exige un TTOG anormal, considerándose anormal un test cuando los valores de glucosa observados en el tiempo que dura la prueba, son mayores que unos standard establecidos, véase National Diabetes Data Group (1979).

Por otra parte, es interesante constatar que tanto en individuos catalogados de diabéticos, como en los individuos catalogados de normales, los valores de insulina varían dentro de un amplio rango. Así hay individuos con diabetes que presentan valores de insulina más altos que la media de los normales e individuos de glucemia normal que presentan valores subnormales y supranormales de insulina (Ellemberg (1982)). Esta variabilidad en la respuesta del páncreas podría justificar, junto con la de los valores de glucosa las diferentes formas de

evolucionar la enfermedad, su etiología y sus manifestaciones clínicas.

Todo ello hace que el TTOG, valorando los valores de glucosa e insulina, sea cada vez más usado en el estudio de la diabetes.

No obstante, creemos que los métodos usuales de valoración de las curvas de glucosa / insulina en tiempos puntuales dan menos información que una valoración conjunta de las mismas, tanto cuantitativamente, considerando los valores de las curvas, como cualitativamente, considerando su forma que depende de la absorción y regulación.

Por ello, en este capítulo se hace un estudio de 171 TTOG realizados en el Hospital San Juan de Dios de Barcelona, a una población de niños sospechosos de ser diabéticos. Este estudio se realizó construyendo un modelo matemático de las mismas, definiendo una distancia entre ellos y realizando una clasificación jerárquica, utilizando los métodos habituales de la taxonomía numérica. El método taxonómico usado es el UPGMA (UNWEIGHTED PAIR GROUP METHOD USING METHOD AVERAGES), ya que proporciona correlaciones cofenéticas elevadas (Arcas (1982)). La distancia usada es la definida en los capítulos 3 y 4, trás la linealización del modelo.

Se ha completado esta clasificación haciendo una representación gráfica de las curvas de glucemia/insulina como puntos del plano, usando un análisis de coordenadas principales y to

mando como disimilaridad la distancia definida en los capítulos 3 y 4.

Con estas clasificaciones se pretende establecer una relación entre tipos diferentes de curvas de glucemia/insulina, y situaciones clínicas u otras características de los individuos, así como establecer un patrón de curva de cada grupo formado.

En una segunda etapa, como veremos en el capítulo siguiente, esta clasificación nos servirá para proponer un algoritmo con el que se pueda realizar diagnósticos automatizados de tipos de curvas glucemia/insulina.

Por último se propondrá un método para el estudio pronóstico de cualquier paciente que no ha podido ser desarrollado por razones de tiempo para seguir a los pacientes.

Dado que después se hará un estudio del TTOG en una población de niños, digamos que este test se realiza administrando por vía oral al niño 1,75 gr. de glucosa por kilo de peso hasta un máximo de 75 gr. Si se observan valores de glucosa en plasma venoso de al menos 140 mg/100 c.c. en ayunas, 200 mg/100 c.c. o más a las 2 horas de la ingestión de la glucosa y otro valor de más de 200 mg/100 c.c. en alguno de los otros tiempos de la prueba, diremos que el niño presenta un TTOG patológico (Ellemberg (1982)).

## 5.2. METODOLOGIA

### 5.2.1. Datos

Los datos usados en este trabajo corresponden a 171 curvas de glucemia/insulina. Los valores de glucosa e insulina se observaron a los 0 min. (basal), 30 min., 60 min., 90 min., 120 min., 150 min., 180 min.

La precisión de las determinaciones, tanto en insulina como en glucosa, no era constante, es decir, las variables aleatorias observadas no tenían igual varianza, por lo que hubo que hacer una transformación de las variables glucosa e insulina para que los modelos construídos tuvieran varianza constante. Para ello, se ha usado el resultado experimental que el cociente entre las medias muestrales de las variables observadas y los errores típicos eran aproximadamente iguales, tanto para la glucosa como para la insulina.

La función de transformación se eligió de la siguiente manera; sea

$$Z = \varphi(X) \qquad (1)$$

la función que transforma la variable glucosa o insulina en una nueva variable, cuya varianza queremos que sea constante. Desarrollando por Taylor en la media de X, m, y cogiendo los dos primeros términos del desarrollo tendremos:

$$Z = \varphi(X) = \varphi(m) + (X-m)\,\varphi'(m) \qquad (2)$$

por consiguiente

$$\text{Var } (Z) = \left(\varphi' \ (m)\right)^2 \ \text{var } (X)$$

imponiendo que la varianza de Z sea constante e igual a K, tendremos:

$$\frac{K}{\text{var } (X)} = \left(\varphi' \ (m)^{\ 2}\right) \tag{4}$$

además como

$$\frac{m}{\sigma \ (X)} = \alpha \quad \text{(cte.)} \tag{5}$$

la expresión (4) nos queda como:

$$\frac{\alpha \ \sqrt{K}}{m} = \varphi' \quad (m) \tag{6}$$

integrando (7) resulta

$$\varphi(m) = \alpha \ \sqrt{K} \ \ln m + C \tag{7}$$

tomando

$$K = \frac{1}{\alpha^2} \ \wedge \ C = - \ln H \tag{8}$$

siendo H arbitraria, la transformación elegida será:

$$\varphi(X) = \ln \frac{X}{H} \tag{9}$$

5.2.2. Modelo empleado

Para estudiar los cambios en las concentraciones de glucosa e insulina en sangre, al efectuar un TTOG, supondremos que los valores medios de dichas variables, en un determinado individuo a lo largo del tiempo, pueden ser obtenidos a través de un sistema de ecuaciones diferenciales de primer orden. Si llamamos x a la concentración de glucosa en sangre (mg/100 ml) e y a la concentración de insulina en sangre (μu/ml), admitiremos que se verifica el sistema:

$$\begin{aligned} \dot{x} &= \frac{dx}{dt} = f\ (x,y) + P\ (t) \\ \dot{y} &= \frac{dy}{dt} = g\ (x,y) \end{aligned} \qquad (10)$$

donde f y g son funciones de las concentraciones de glucosa e insulina solamente y que supondremos diferenciables con continuidad y P es una función del tiempo, que describe la perturbación a que sometemos al sistema (organismo), debido a la ingestión de glucosa.

Admitiremos además que:

$$x\ (o) = x_o \qquad y\ (o) = y_o \qquad (11)$$

que el punto $(x_o, y_o)$ verifica:

$$f\ (x_o, y_o) = o \qquad g\ (x_o, y_o) = o$$

y que dicho punto es un punto de reposo, asintoticamente estable,

$$\lim_{t \to \infty} x(t) = x_o \quad , \quad \lim_{t \to \infty} y(t) = y_o \tag{12}$$

del sistema

$$\dot{x} = f(x,y) \quad , \quad \dot{y} = g(x,y) \tag{13}$$

Estudiemos a continuación, cualitativamente, las propiedades de las funciones f, g y P. Para que se conserve la homeostasis, parece natural suponer que f es una función monótona decreciente, respecto sus dos argumentos, al menos en un entorno del punto de reposo $(x_o, y_o)$. Por tanto se verificará:

$$\left(\frac{\partial f}{\partial x}\right)(x_o, y_o) = a_{11} < o, \quad \left(\frac{\partial f}{\partial y}\right)(x_o, y_o) = a_{12} < o \tag{14}$$

Por otra parte, g debe de ser una función monótona creciente respecto la concentración de glucosa y decreciente respecto la concentración de insulina, en un entorno del punto de reposo. Por consiguiente:

$$\left(\frac{\partial g}{\partial x}\right)(x_o, y_o) = a_{21} > o, \quad \left(\frac{\partial g}{\partial y}\right)(x_o, y_o) = a_{22} < o \tag{15}$$

Todo ello encaja bien con las habituales descripciones fisiológicas del proceso de regulación del par glucosa/insulina en sangre: al sobrepasarse la concentración de glucosa del equilibrio, el organismo responde aumentando la concentración de insulina, mecanismo activo para restaurar el equilibrio, a la vez que tiende a restaurarse el mismo de forma no activa, debido al transporte pasivo de glucosa a los tejidos.

En cuanto a la función P (t), que describe la perturbación del sistema al ingerir glucosa, es razonable aceptar que es una función monótona decreciente que verifica:

$$\lim_{t \to \infty} P(t) = o \quad y \quad \int_o^{+\infty} P(t)\, dt = \beta G, \ \beta > o \tag{16}$$

G sería la cantidad de glucosa ingerida y βG la concentración de glucosa en sangre. En ausencia del proceso de regulación

$$\dot{x} = P(t) \tag{17}$$

De lo que antecede se deduce que, en un entorno del punto de reposo, las funciones f y g pueden aproximarse por:

$$\begin{aligned} f(x,y) &\cong a_{11}(x-x_o) + a_{12}(y-y_o) \\ g(x,y) &\cong a_{21}(x-x_o) + a_{22}(y-y_o) \end{aligned} \tag{18}$$

Por otra parte, podemos escoger como función P (t), a la sencilla expresión

$$P(t) = \beta G\, \alpha\, e^{-\alpha t}, \ t \geq o \quad \alpha > o \quad \beta > o \tag{19}$$

que cumple con las condiciones antes requeridas.

Por tanto, el modelo que utilizaremos para describir aproximadamente el test de tolerancia oral a la glucosa, es

$$\dot{x} = a_{11}\,(x-x_o) + a_{12}\,(y-y_o) + \beta G\alpha\, e^{-\alpha t}$$

$$\dot{y} = a_{21}\,(x-x_o) + a_{22}\,(y-y_o) \qquad (20)$$

Dicho sistema se simplifica con el cambio $u = x - x_o$, $v = y - y_o$, resultando:

$$\dot{u} = a_{11}\,u + a_{12}\,v + \beta\, G\, \alpha\, e^{-\alpha t}$$

$$\dot{v} = a_{21}\,u + a_{22}\,v \qquad (21)$$

Resolvamos en primer lugar el sistema homogéneo

$$\dot{u} = a_{11}\,u + a_{12}\,v$$

$$\dot{v} = a_{21}\,u + a_{22}\,v \qquad (22)$$

Llamando $\lambda_1$ y $\lambda_2$ a las soluciones de la ecuación característica:

$$\begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = o \qquad (23)$$

que sin gran pérdida de generalidad podemos suponer distintas, nos permiten obtener la solución general del sistema homogéneo, como

$$u = c_1\, e^{\lambda_1 t} + e_2\, e^{\lambda_2 t}$$

$$v = e_3\, e^{\lambda_1 t} + e_4\, e^{\lambda_2 t} \qquad (24)$$

Por otra parte si A es la matriz

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \tag{25}$$

y $\alpha$ y $\beta$ son la parte real e imaginaria de las raíces $\lambda_1$, $\lambda_2$ de la ecuación características (23)

$$\begin{aligned} \lambda_1 &= \alpha + \beta_i \\ \lambda_2 &= \alpha - \beta_i \end{aligned} \tag{26}$$

siendo además $\alpha < o$, $\beta > o$, se tiene (véase Apostol 1982)

$$e^{At} = \frac{e^{\alpha t}}{\beta} \left[ (\beta \cos \beta t - \alpha \operatorname{sen} \beta t)\ I + \operatorname{sen} \beta t\ A \right] \tag{27}$$

Si llamamos Q(s) a la matriz

$$Q(s) = \begin{pmatrix} \gamma\ G\ \mu\ e^{-\mu s} \\ 0 \end{pmatrix} \tag{28}$$

la solución $\omega(t)$ del sistema (24) será:

$$\omega(t) = \int_o^t e^{(t-s)\ A}\ Q(s)\ ds \tag{29}$$

y desarrollando nos queda:

$$\omega(t) = \int_o^t \frac{e^{\alpha(t-s)}}{\beta} \left[(\beta \cos \beta t - \alpha \operatorname{sen} \beta t)\ I + \operatorname{sen} \beta t A\right] \begin{pmatrix} \gamma\ G\ \mu\ e^{-\mu t} \\ 0 \end{pmatrix} ds \tag{30}$$

volviendo a desarrollar obtenemos:

$$\omega(t) = \frac{\gamma G \mu}{\beta} \begin{pmatrix} \int_o^t e^{\alpha t - (\alpha+\mu)s} (\beta \cos \beta(t-s) + (a_{11} - ) \operatorname{sen} \beta(t-s) ds \\ \int_o^t e^{\alpha t-(\alpha+\mu)s} a_{12} \operatorname{sen} \beta\ (t-s)\ d.s \end{pmatrix} \tag{31}$$

si llamamos

$$\begin{aligned} I_A &= \int_o^t e^{\alpha t - (\alpha+\mu)\ s} \cos \beta(t-s)\ d\ s \\ I_B &= \int_o^t e^{\alpha t - (\alpha+\mu)\ s} \operatorname{sen} \beta(t-s)\ d\ s \end{aligned} \tag{32}$$

la expresión (31) queda como:

$$\omega(t) = \frac{\gamma G \mu}{\beta} \begin{pmatrix} \beta I_A + (a_{11} - \alpha) I_B \\ a_{12}\ \ I_B \end{pmatrix}, \tag{33}$$

integrando las expresiones (32) obtendremos:

$$\begin{aligned} I_A &= -\frac{e^{-\mu t}}{\alpha+\mu} + \frac{e^{\alpha t}}{\alpha+\mu} \cos \beta t + \frac{\beta}{\alpha+\mu} I_B \\ I_B &= \frac{e^{\alpha t}}{\alpha+\mu} \operatorname{sen} \beta t - \frac{\beta}{\alpha+\mu} I_A \end{aligned} \tag{34}$$

por lo tanto

$$I_A = \frac{-(\alpha+\mu)e^{-\mu t} + e^{\alpha t}\left((\alpha+\mu)\cos \beta t + \beta \operatorname{sen} \beta t\right)}{(\alpha+\mu)^2 + \beta^2}$$

$$I_B = \frac{e^{\alpha t}\left((\alpha+\mu) \operatorname{sen} \beta t - \beta \cos \beta t\right) + \beta e^{-\mu t}}{(\alpha+\mu)^2 + \beta^2} \qquad (35)$$

Las soluciones del sistema (21) son pues:

$$u(t) = \frac{\gamma G \mu}{\beta\left((\alpha+\mu)^2 + \beta^2\right)}\left[e^{-\mu t}\left(-\beta(\alpha+\mu) + \beta(a_{11} - \alpha)\right) + \right.$$

$$\left. + e^{\alpha t}\left(\cos \beta t\left[\beta(\alpha+\mu) - \beta(a_{11} - \alpha)\right] + \operatorname{sen}\beta t\left[\beta^2 + (a_{11} - \alpha)(\alpha+\mu)\right]\right)\right]$$

$$u(t) = a_{21}\frac{e^{\alpha t}\left((\alpha+\mu) \operatorname{sen} \beta t - \beta \cos \beta t\right) + \beta e^{-\mu t}}{(\alpha+\mu)^2 + \beta^2}$$

pero al ser

$$\alpha = \frac{1}{2}(a_{11} + a_{22}) \qquad (36)$$

se tiene que

$$u(t) = \frac{\gamma G \mu}{(\alpha+\mu)^2 + \beta^2}\left[-e^{-\mu t}(a_{22}+\mu) + e^{\alpha t}\left[(a_{22}+\mu)\cos \beta t + \right.\right.$$

$$\left.\left. + \left(\beta + \frac{(a_{11} - a_{22})}{2\beta}\left(\frac{a_{11} + a_{22}}{2} + \mu\right)\right)\right] \operatorname{sen} \beta t\right] \qquad (37)$$

Llamando por otra parte:

$$C = \frac{-\gamma G \mu\ (a_{22} + \mu)}{(\alpha+\mu)^2 + \beta^2}$$

$$D = \frac{\gamma G \mu \quad \beta + \dfrac{(a_{11} - a_{22})\ (a_{21} + a_{22} - 2\mu)}{4\ \beta}}{(\alpha+\mu)^2 + \beta^2} \qquad (38)$$

$$E = \frac{a_{21}\ (\mu+\alpha)}{(\alpha+\mu)^2 + \beta^2} \qquad F = \frac{a_{12}\ \beta}{(\alpha+\mu)^2 + \beta^2}$$

$$E = \frac{\mu + \alpha}{\beta}\ F$$

Las ecuaciones soluciones de (10) quedan como:

$$x(t) = x_o + C\ e^{-\mu t} - C\ e^{\alpha t} \cos \beta t + D\ e^{\alpha t} \operatorname{sen} \beta t$$

$$y(t) = y_o + E\ e^{\alpha t} \operatorname{sen} \beta t - \frac{E\ \beta}{\mu+\alpha}\ e^{\alpha t} \cos \beta t + \frac{E\ \beta}{\mu+\alpha}\ e^{-\mu t} \qquad (39)$$

siendo

$$\alpha < o \qquad \beta > o \qquad \mu > o$$

se puede observar que

$$x(o) = x_o$$

$$y(o) = y_o \qquad (40)$$

y que además

$$\lim_{t \to \infty} x\ (t) = x_o$$

$$\lim_{t \to \infty} y\ (t) = y_o \qquad (41)$$

Una vez obtenido el modelo teórico de las curvas de glucemia e insulina en el TTOG, se determinaron las ecuaciones de los modelos para un individuo normal obtenido de Ellernber (1982). Los valores de los parámetros fueron en este caso.

$$\begin{aligned} x_o &= 95 \quad C = -7.1404.10^{1};\ D = -5.720.10^{2};\ \mu = 3.4318.10^{-3}; \\ &\qquad \beta = 0.5626.10^{-1} \\ y_o &= 25 \quad E = -6.2988.10^{3};\ \alpha = -0.9476 \end{aligned} \tag{42}$$

El error típico fue de $\sqrt{44.2}$ por lo que el modelo fue aceptado como válido.

Por razones expuestas en el punto 5.2.1., las variables observadas de cada individuo hay que transformarlas tomando logaritmos neperianos de ellas partido por una constante que nosotros elegimos como los valores de la curva teórica del individuo normal en cada tiempo, es decir, las nuevas variables serán al tiempo t

$$\omega(t) = \ln \frac{X(t)}{X_N(t)}$$

$$\omega'(t) = \ln \frac{y(t)}{y_N(t)} \tag{43}$$

Como por otra parte,

$$\ln x \cong x - 1 \tag{44}$$

cuando $x \cong 1$, se tiene que las variables $\omega(t)$, $\omega'(t)$, las podemos considerar como:

$$\omega(t) = \frac{X(t)}{X_N(t)} - 1$$

$$\omega'(t) = \frac{y(t)}{y_N(t)} - 1 \qquad (45)$$

siendo W y W' variables aleatorias normales de varianza $\sigma_1^2$ y $\sigma_2^2$ respectivamente.

Por otra parte, la medida de la glucosa es más fiable que la medida de la insulina, cumpliéndose que el cociente entre la media de las determinaciones de glucosa y desviación típica es el doble que el mismo cociente para la insulina, es decir:

$$\frac{m}{\sigma(X)} = 2\alpha \quad \hat{} \quad \frac{m}{\sigma(Y)} = \alpha \qquad (46)$$

siendo X la variable aleatoria, determinación de glucosa e Y la variable aleatoria, determinación de insulina. La varianza de Y es pues 4 veces la varianza de X. De la variable insulina se obtenían dos réplicas por tiempo, por lo que la variable observada de la insulina es la media de las dos réplicas $\bar{y}(t)$, por tanto, las variables (45) quedan como:

$$\omega(t) = \frac{X(t)}{X_N(t)} - 1$$

$$\bar{\omega}'(t) = \frac{\bar{y}(t)}{y_N(t)} - 1 \qquad (47)$$

En este caso las varianzas de $W, \bar{W}'$, son respectivamente $\sigma_1^2$ y $2\sigma_1^2$. La varianza de la variable correspondiente a la insulina

ha quedado dividida por 2, al ser la media de dos réplicas. Con el fin de que las dos variables, la correspondiente a la glucosa y la correspondiente a la insulina tengan la misma varianza, se eligió en lugar de la variable $\bar{\omega}(t)$ la variable

$$\bar{\omega}^*(t) = \frac{1}{\sqrt{2}} \left( \frac{y(t)}{y_N(t)} - 1 \right) \tag{48}$$

Como puede apreciarse las variables $\omega(t)$ y $\bar{\omega}^*(t)$ representan la variación en tanto por uno de un individuo respecto del normal. Esta perturbación se supuso que en el intervalo de tiempo de observación se ajustaba a un polinomio de tercer grado, es decir, que

$$\begin{aligned} \omega(t) &= \alpha_o + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + e \\ \bar{\omega}^*(t) &= \beta_o + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + e^* \end{aligned} \tag{49}$$

Estas ecuaciones las podemos expresar como un modelo lineal multivariante

$$Y = X B + E \tag{50}$$

siendo

$$Y = \begin{pmatrix} \omega_o & \bar{\omega}_o^* \\ \omega_1 & \bar{\omega}_1^* \\ \cdot & \\ \cdot & \\ \omega_n & \bar{\omega}_n^* \end{pmatrix} \qquad X = \begin{pmatrix} 1 & t_o & t_o^2 & t_o^3 \\ 1 & t_1 & t_1^2 & t_1^3 \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 1 & t_n & t_n^2 & t_n^3 \end{pmatrix}$$

$$B = \begin{pmatrix} \alpha_o & \beta_o \\ \alpha_1 & \beta_1 \\ \alpha_2 & \beta_2 \\ \alpha_3 & \beta_3 \end{pmatrix} \qquad E = \begin{pmatrix} e_o & e_o^* \\ e_1 & e_1^* \\ \cdot & \cdot \\ \cdot & \cdot \\ e_n & e_n^* \end{pmatrix}$$

En nuestro caso n=6, $t_i = 3\,oi$ $\quad (o \leq i \leq 6)$

En rango del diseño es fácil comprobar que es 4, es decir, el diseño es de rango máximo.

5.2.3. Cálculo de la distancia

Si tenemos dos modelos como los definidos en (50)

$$Y_a = B_a + E_a \qquad Y_b = B_b + E_b \tag{51}$$

la estimación de la distancia al cuadrado entre ellos es, según el punto 4.1.2.

$$\hat{L}^2 = \text{traza}\left((Y_a - Y_b)^t \times (X^tX)^{-1} (Y_a - Y_b)\, S^{-1}\right) \tag{52}$$

donde S es la estimación insesgada de la matriz de varianzas-covarianzas común para ambos modelos.

Si suponemos que las variables $\omega$, $\bar{\omega}^*$ son independientes la estimación de la distancia entre ellos es:

$$\hat{L}^2 = \hat{L}_1^2 + \hat{L}_2^2 \tag{53}$$

donde $\hat{L}_1^2$ es la estimación de la distancia al cuadrado de los modelos univariantes para la glucosa:

$$\omega_a = X\,\alpha_a + e_a$$
$$\omega_b = X\,\alpha_b + e_b \qquad (54)$$

siendo $\alpha_a$ y $\alpha_b$ los primeros vectores columna de la matriz $B_a$ y $B_b$ respectivamente y $\hat{L}_2^2$ es la estimación de la distancia al cuadrado de los dos modelos univariantes para la insulina

$$\bar{\omega}_a^* = X\,\beta_a + e_a^*$$
$$\bar{\omega}_b^* = X\,\beta_b + e_b^* \qquad (55)$$

siendo $\beta_a$, $\beta_b$ los segundos vectores columna de $B_a$ y $B_b$ respectivamente.

Las expresiones de $\hat{L}_1^2$ y $\hat{L}_2^2$ son

$$\hat{L}_1^2 = \frac{1}{\sigma^2}\ (\omega_a-\omega_b)^t\ X\ (X^tX)^{-1}\ X^t\ (\omega_a-\omega_b)$$

$$\hat{L}_2 = \frac{1}{\sigma^2}\ (\bar{\omega}_a^*-\bar{\omega}_b^*)^t\ X\ (X^tX)^{-1}\ X^t\ (\bar{\omega}_a^*-\bar{\omega}_b^*)$$

Por otra parte se sabe, según las experiencias del laboratorio, que:

$$\frac{m}{\sigma(x)} = 5 \qquad (57)$$

por consiguiente:

$$\sigma^2 = \frac{1}{25} \qquad (58)$$

### 5.2.4. Métodos de clasificación

Los métodos de clasificación usados a partir de la distancia (52) fueron los de Taxonomía Numérica completada con un Análisis de Coordenadas Principales.

La Taxonomía Numérica nos sirvió para realizar una clasificación jerárquica. El método para construir la jerarquía fue el UPGMA (véase capítulo 2). Los cálculos para realizar la clasificación taxonómica se hicieron con el programa Clustan.

El dendograma viene dado en la fig. 1.

Mediante el Análisis de Coordenadas Principales se hizo una representación gráfica de los 172 individuos como puntos de un plano. Esto nos permite visualizar las proximidades entre los individuos.

Los cálculos para realizar el Análisis de Coordenadas Principales, se hicieron con el programa ANCPR (Cuadras 1979).

## 5.3. RESULTADOS

Como ya hemos dicho en el punto anterior, se ha construído un dendograma por el método UPGMA. La correlación cofenética resultó ser de $r_c = 0.997$, lo que indica que la jerarquía indexada obtenida, reproduce muy bien la distancia inicial y que la población está muy jerarquizada.

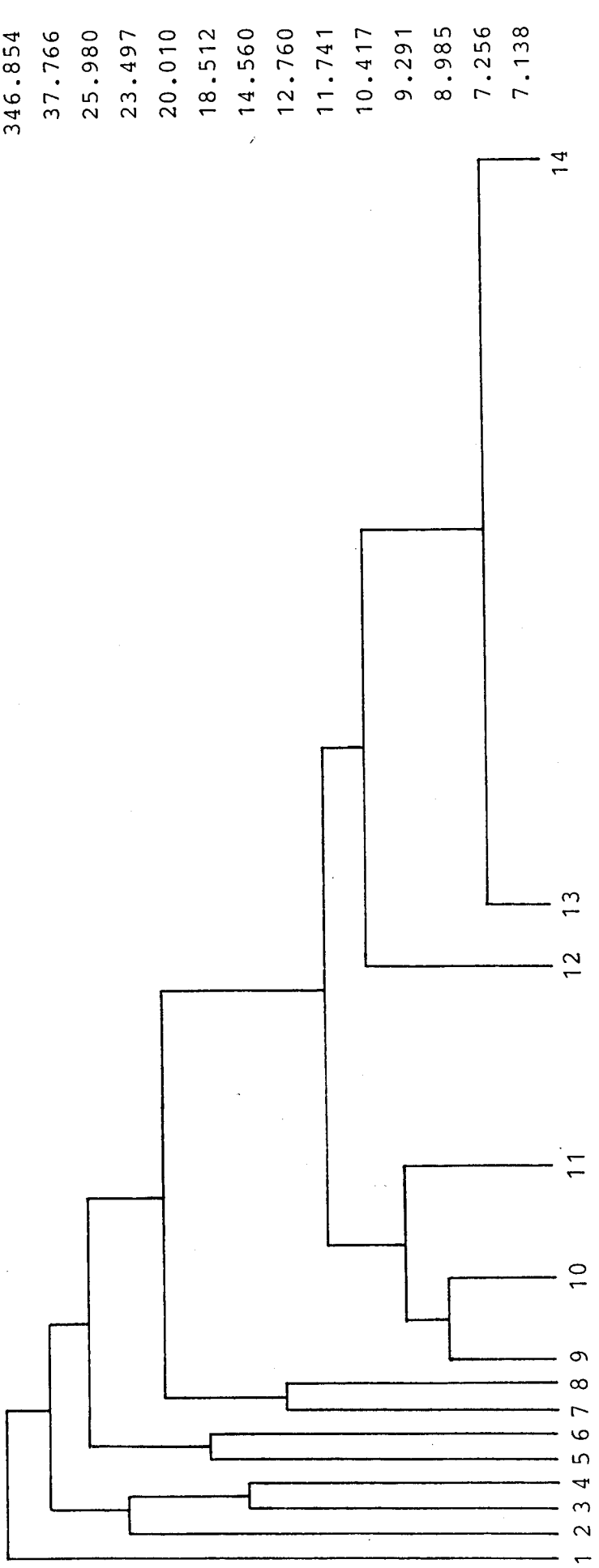

346.854
37.766
25.980
23.497
20.010
18.512
14.560
12.760
11.741
10.417
9.291
8.985
7.256
7.138
1 2 3 4 5 6 7 8 9 10 11 12 13 14

Fig. 1

Basándonos en el capítulo IV, se obtiene que dos TTOG son distintos, a nivel de significación 0.05, cuando la distancia entre ellos es mayor de 7. Por consiguiente y puesto que el coeficiente de correlación cofenético es tan alto, se puede cortar en el dendograma a nivel de distancia 7, obteniéndose 14 grupos de niños de los cuales, aunque todos ellos son estadísticamente diferentes, no todos son necesariamente clínicamente distintos con los criterios actuales.

El análisis de coordenadas principales, tomando las dos primeros coordenadas principales que explican más del 98% de la dispersión total, nos permitió hacer una representación gráfica de los individuos (figs. 2 y 2 bis). En ellas se observan perfectamente diferenciados 5 grupos de individuos que poseen un TTOG patológico, según los criterios del National Diabetes Data Group (1979), el resto de los individuos quedan más agrupados correspondiendo éstos, posiblemente, a variantes menos marcadas de la normalidad.

A continuación pasamos a describir los 14 grupos que aparecen en la clasificación taxonómica, comparando los 13 primeros con el grupo XIV al que pertenece el individuo teórico.

Los grupos I, II, III, IV, XII tienen un TTOG patológico, mientras que el resto de los grupos tienen un TTOG dentro de los límites de la normalidad.

Los grupos II y IV ya presentan a nivel basal valores patológicamente elevados de glucosa, también en el resto de los

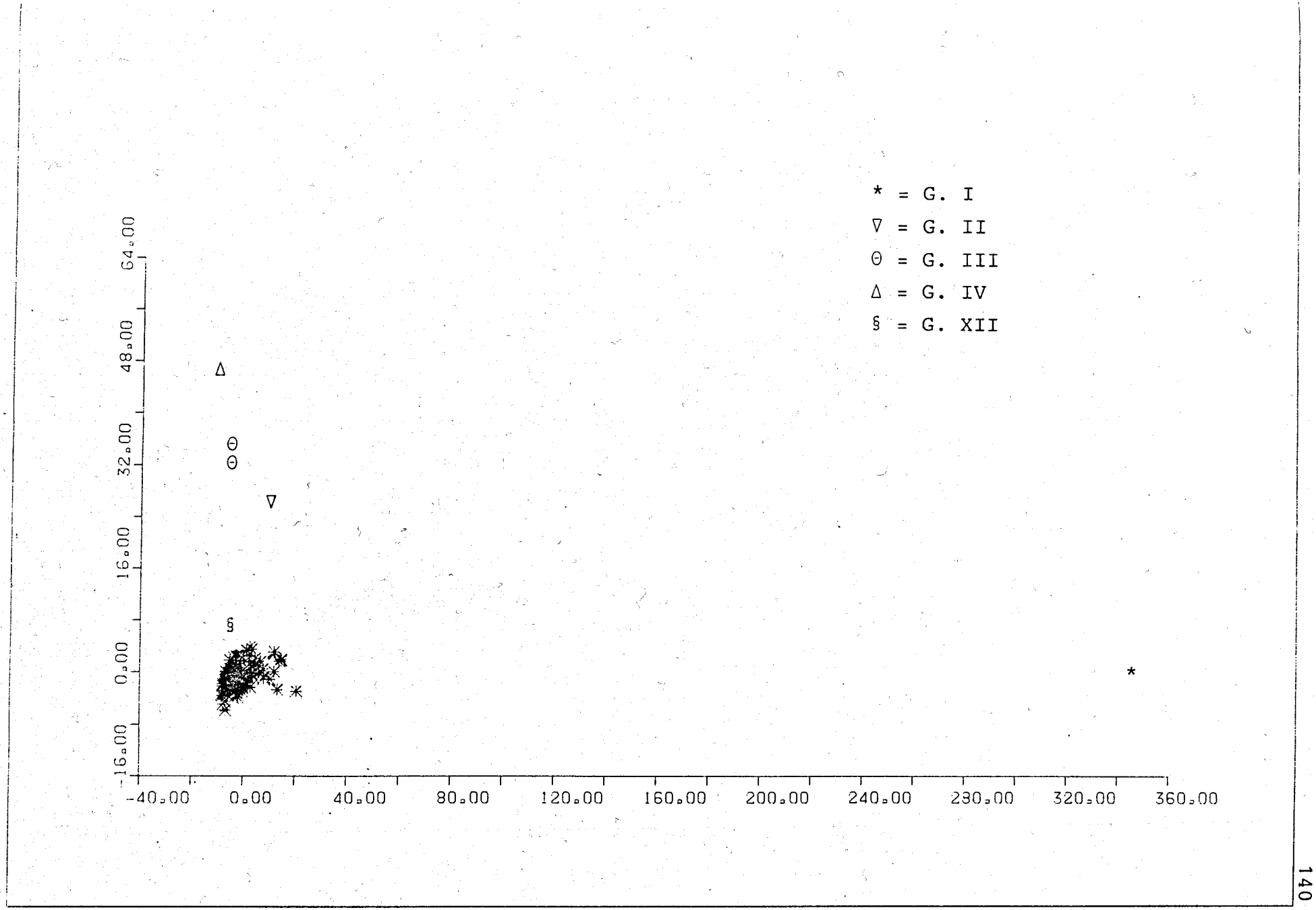

* = G. I
∇ = G. II
Θ = G. III
Δ = G. IV
§ = G. XII
64.00
48.00
32.00
16.00
0.00
-16.00
-40.00
0.00
40.00
80.00
120.00
160.00
200.00
240.00
280.00
320.00
360.00

Fig. 2

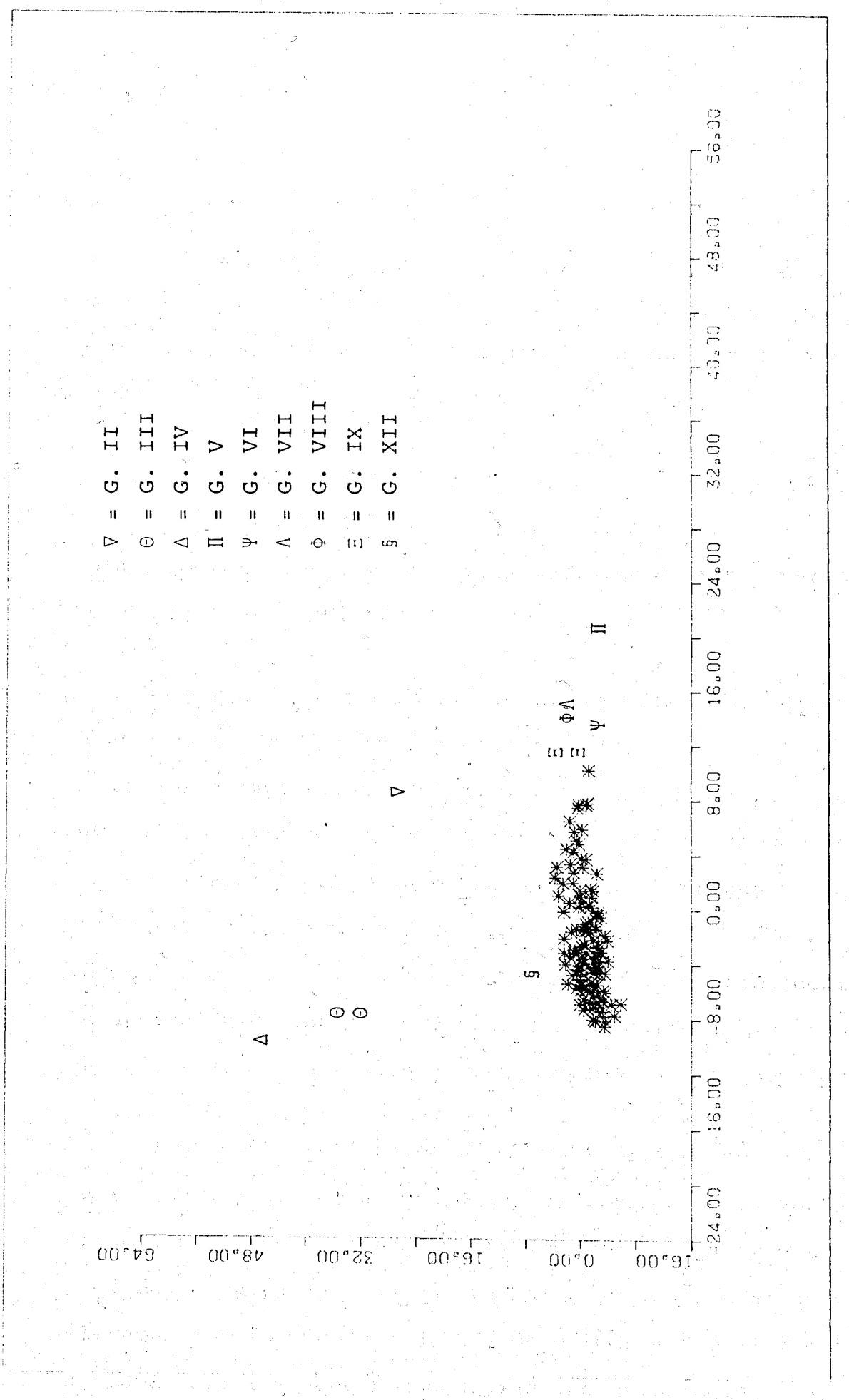

∇ = G. II
Θ = G. III
Δ = G. IV
Π = G. V
Ψ = G. VI
Λ = G. VII
Φ = G. VIII
Ξ = G. IX
§ = G. XII
64.00
48.00
32.00
16.00
0.00
-16.00
-24.00
-16.00
-8.00
0.00
8.00
16.00
24.00
32.00
40.00
48.00
56.00

Fig. 2 bis.

tiempos, siendo las hiperglucemias más marcadas en el grupo IV.

En el grupo III no se estimula prácticamente la secreción de insulina y en el grupo IV la insulina, tanto a nivel basal como en el resto de los tiempos que dura la prueba, es inapreciable. En ambos casos, se trataría de grupos de diabéticos insulinopénicos, con un grado mayor de afectación en el grupo IV.

Los grupos I, II, XII, tienen unos niveles basales de glucosa dentro de los límites de la normalidad, sin embargo presentan un TTOG patológico.

El grupo I está formado por un sólo individuo diagnosticado de Acantosis Nigrigans y presenta, tanto a nivel basal, como en los tiempos de la prueba, niveles altísimos de insulina, que no se corresponden con los valores de glucosa que incluso en los tiempos 30 m. y 60 m., alcanzan valores superiores a 200 mg/100 c.c. Esto confirma que en esta enfermedad se produce una insulinresistencia acompañada de una hiperespuesta en la secreción de insulina que resulta eficaz en el sentido de que a los 180 m. se alcanzan los niveles basales de glucosa.

El grupo II está formado por un individuo que a nivel basal tiene una glucemia dentro de los límites de la normalidad, presentando, en los restantes tiempos de la prueba, valores elevados de glucosa. Los valores de insulina son sin embargo elevados por lo que se trataría de un diabético diagnosticado a través del TTOG que presenta insulinresistencia.

Se podría especular con que estos individuos podrían evolucionar a diabéticos insulinopénicos , como los del grupo III y IV, por agotamiento de las células β.

El grupo XII está formado por un solo individuo obeso, con glucemia basal normal y con TTOG patológica. La respuesta en la secreción de insulina es deficiente en los tiempos 30 m. y 60 m., pero en los tiempos 120 m. y 150 m., alcanza los niveles de los tiempos 30 m. y 60 m. de un individuo normal. Se podría estudiar si la respuesta es buena pero lenta.

Las curvas promedio de los grupos descritos vienen dadas en las figs: 3, 4, 4 (bis), 5, 6 y 6 (bis).

Los grupos V, VI, VII, VIII, IX, X, XI y XII tienen curva de glucosa normal, aunque todos ellos presentan niveles de insulina elevados, ver figs: 6, 7, 8, 9, 10, 11 y 12.

Los grupos V, VII, IX y XI tienen una curva de glucosa dentro de la normalidad y curva de insulina elevada en diferentes grados; el grupo V, formado por un solo individuo diagnosticado de Distrofia Miotónica, presenta niveles de insulina que alcanzan hasta los 700 μu/ml.; el grupo VII está formado por dos individuos obesos, con curvas de glucemia dentro de la normalidad y con curva de insulina elevado pero sin llegar a los valores del grupo V; el grupo IX está formado por dos individuos obesos que tienen curva de glucosa normal y con niveles de insulina elevados, pero menos que el grupo VII; el grupo XI está formado por 15 individuos de los cuales 6

eran obesos, 2 estaban diagnosticados de Distrofia Miotónica y 1 de Polimiopatía, las características del TTOG eran similares a las del grupo IX, salvo que los niveles de insulina eran ligeramente más bajos.

Todos estos grupos V, VII, IX y XI tienen una característica importante y es que los valores de glucosa a los 180 m., no alcanzan los niveles basales, son más elevados, esto se podría interpretar como que estos grupos presentan una insulinresistencia en diferentes grados que los incapacita para la regulación.

Los grupos VI, VIII, X y XIII tienen una curva de glucosa dentro de la normalidad y una curva de insulina elevada en diferentes grados que de mayor a menor sería la del grupo VI, VIII, X y XIII, sin embargo es interesante ver que resulta eficaz en el sentido de que a los 180 m. la glucosa alcanza los valores basales que en ninguno de los grupos presentaba diferencias con los niveles basales del grupoXIV. Esto vendría a decir, que aunque el gasto de insulina es mayor durante la prueba que en el grupo XIV, la regulación es buena.

El grupo XIV, al que pertenece el individuo teórico, lo consideramos como grupo standard.

## 5.4. DISCUSION Y CONCLUSIONES

En el presente capítulo se han obtenido 14 grupos de individuos cuyos TTOG eran estadísticamente diferentes. El esquema de la clasificación viene dada en la tabla I.

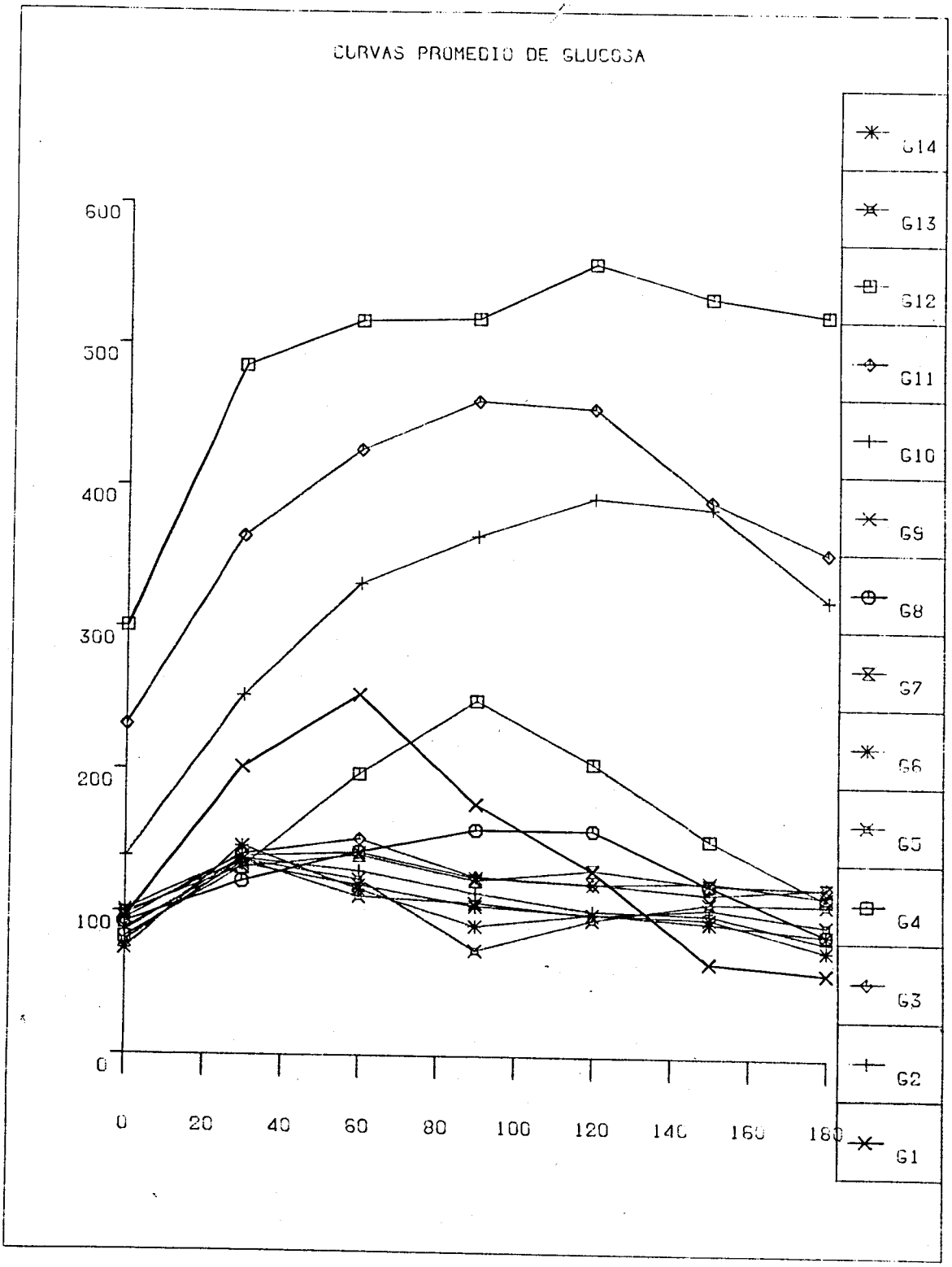

CURVAS PROMEDIO DE GLUCOSA
600
500
400
300
200
100
0
0
20
40
60
80
100
120
140
160
180
G14
G13
G12
G11
G10
G9
G8
G7
G6
G5
G4
G3
G2
G1

Fig. 3

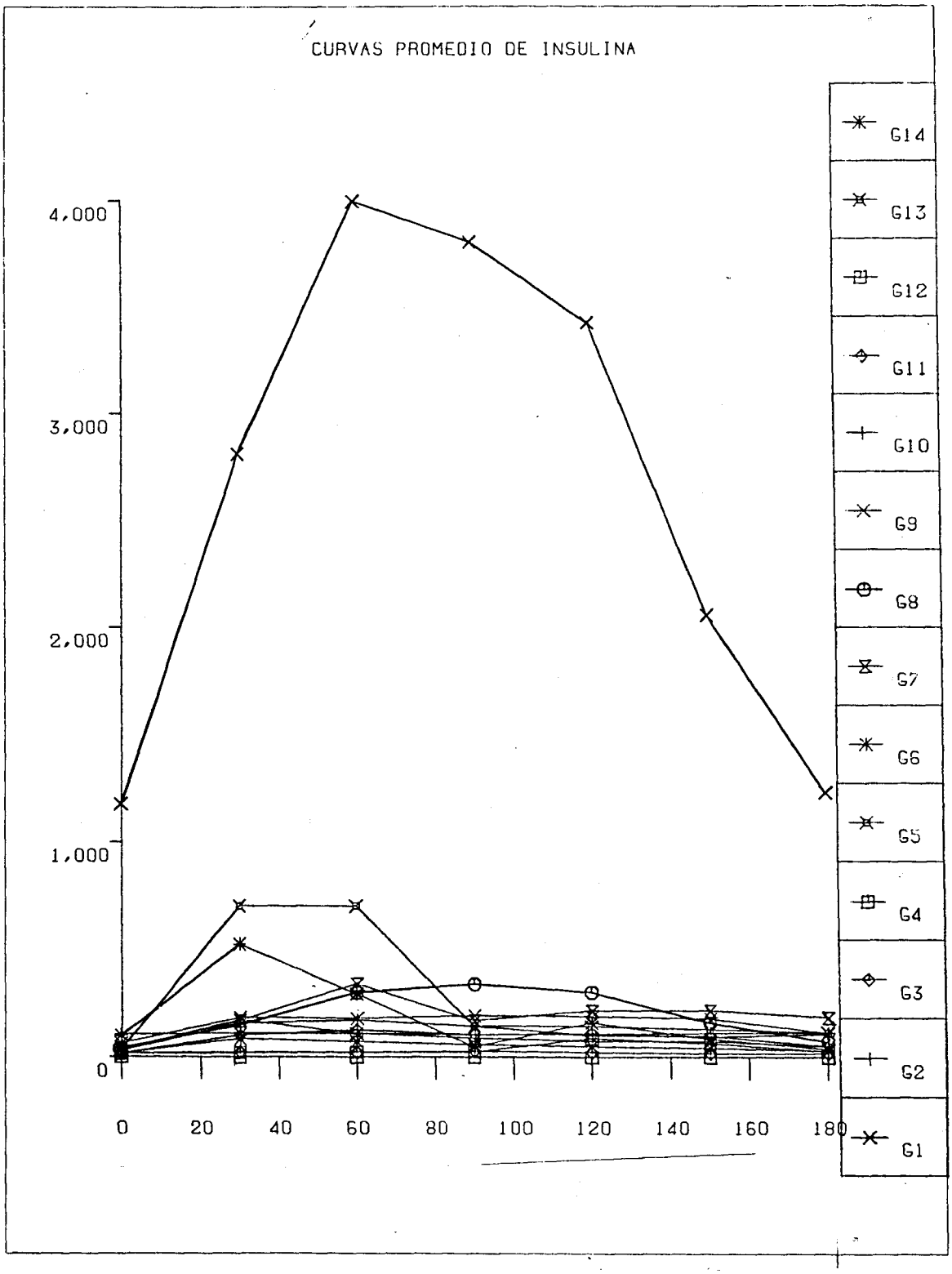

CURVAS PROMEDIO DE INSULINA
4,000
3,000
2,000
1,000
0
0
20
40
60
80
100
120
140
160
180
G14
G13
G12
G11
G10
G9
G8
G7
G6
G5
G4
G3
G2
G1

Fig. 4

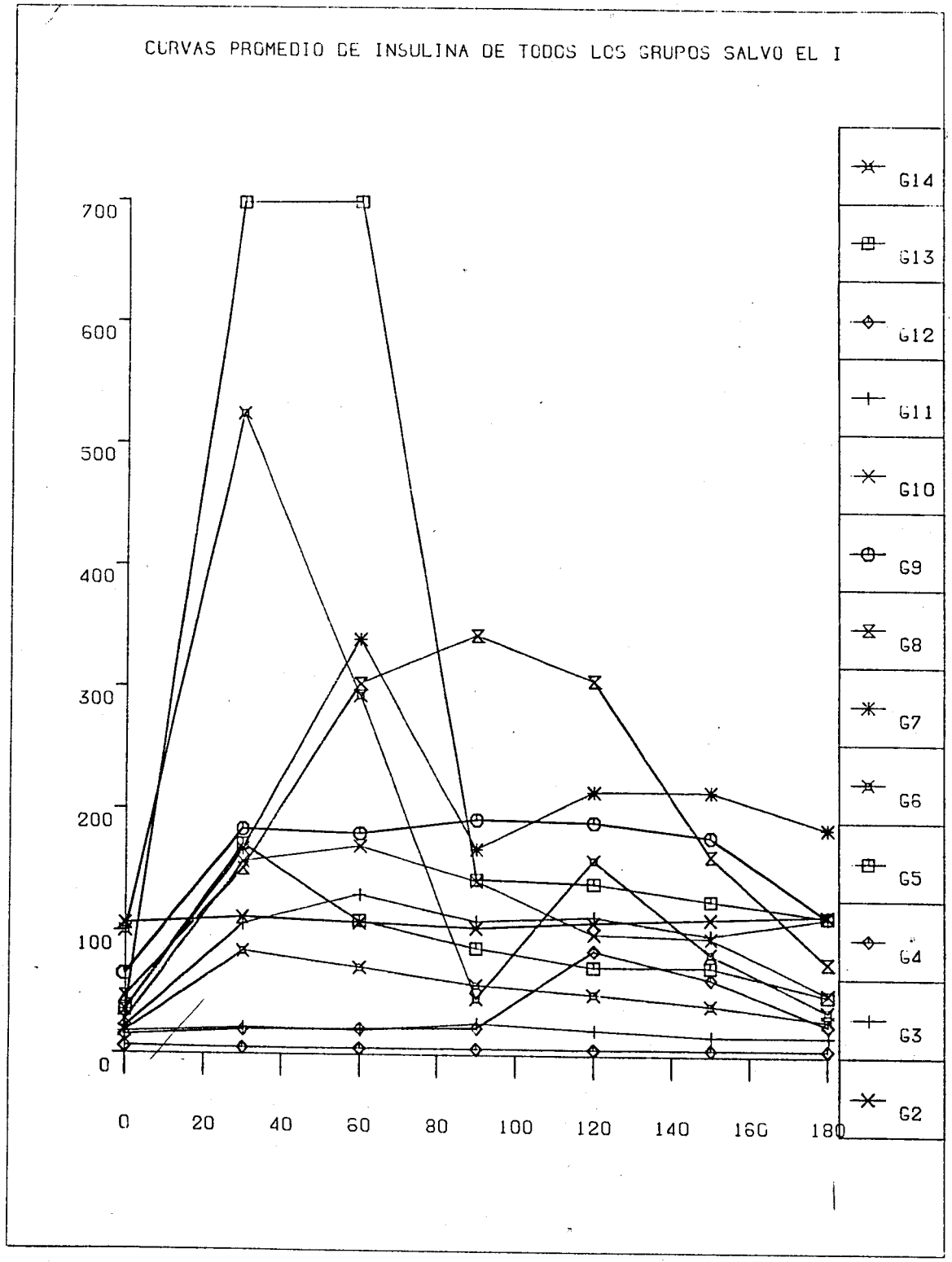

CURVAS PROMEDIO DE INSULINA DE TODOS LOS GRUPOS SALVO EL I
700
600
500
400
300
200
100
0
0
20
40
60
80
100
120
140
160
180
G14
G13
G12
G11
G10
G9
G8
G7
G6
G5
G4
G3
G2

Fig. 4 (bis)

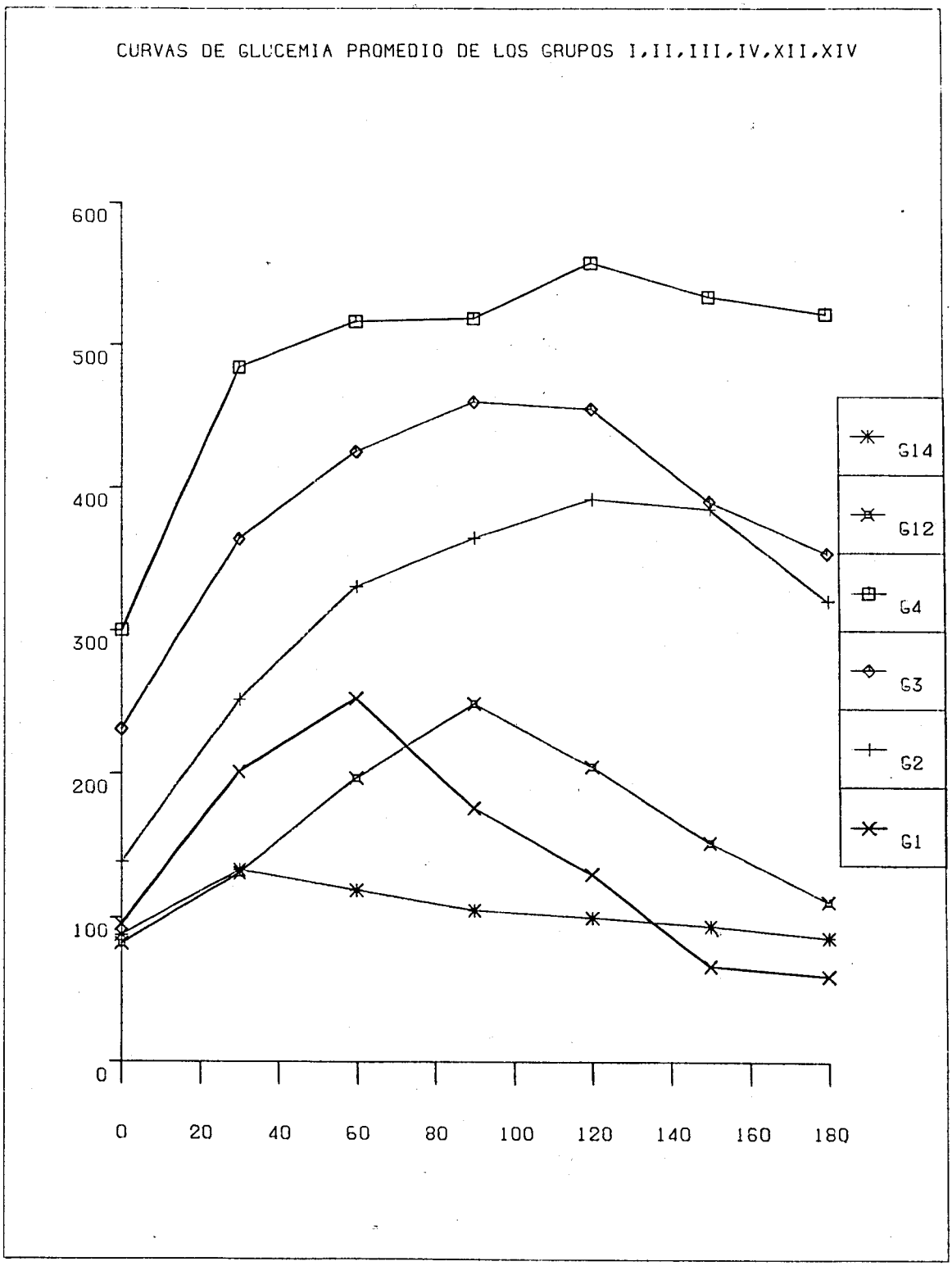

CURVAS DE GLUCEMIA PROMEDIO DE LOS GRUPOS I,II,III,IV,XII,XIV
600
500
400
300
200
100
0
0
20
40
60
80
100
120
140
160
180
G14
G12
G4
G3
G2
G1

Fig. 5

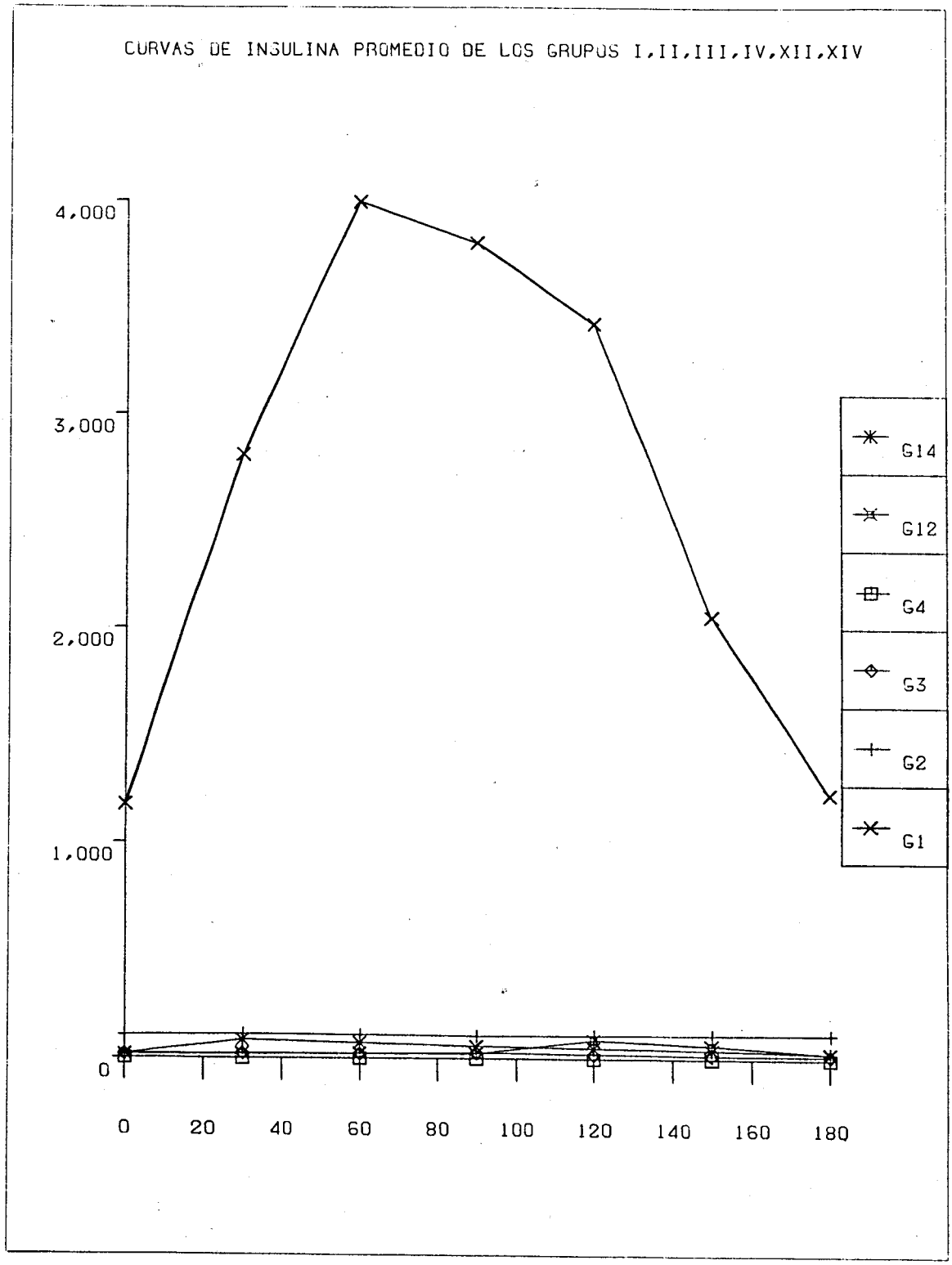

CURVAS DE INSULINA PROMEDIO DE LOS GRUPOS I,II,III,IV,XII,XIV
4,000
3,000
2,000
1,000
0
0
20
40
60
80
100
120
140
160
180
G14
G12
G4
G3
G2
G1

Fig. 6

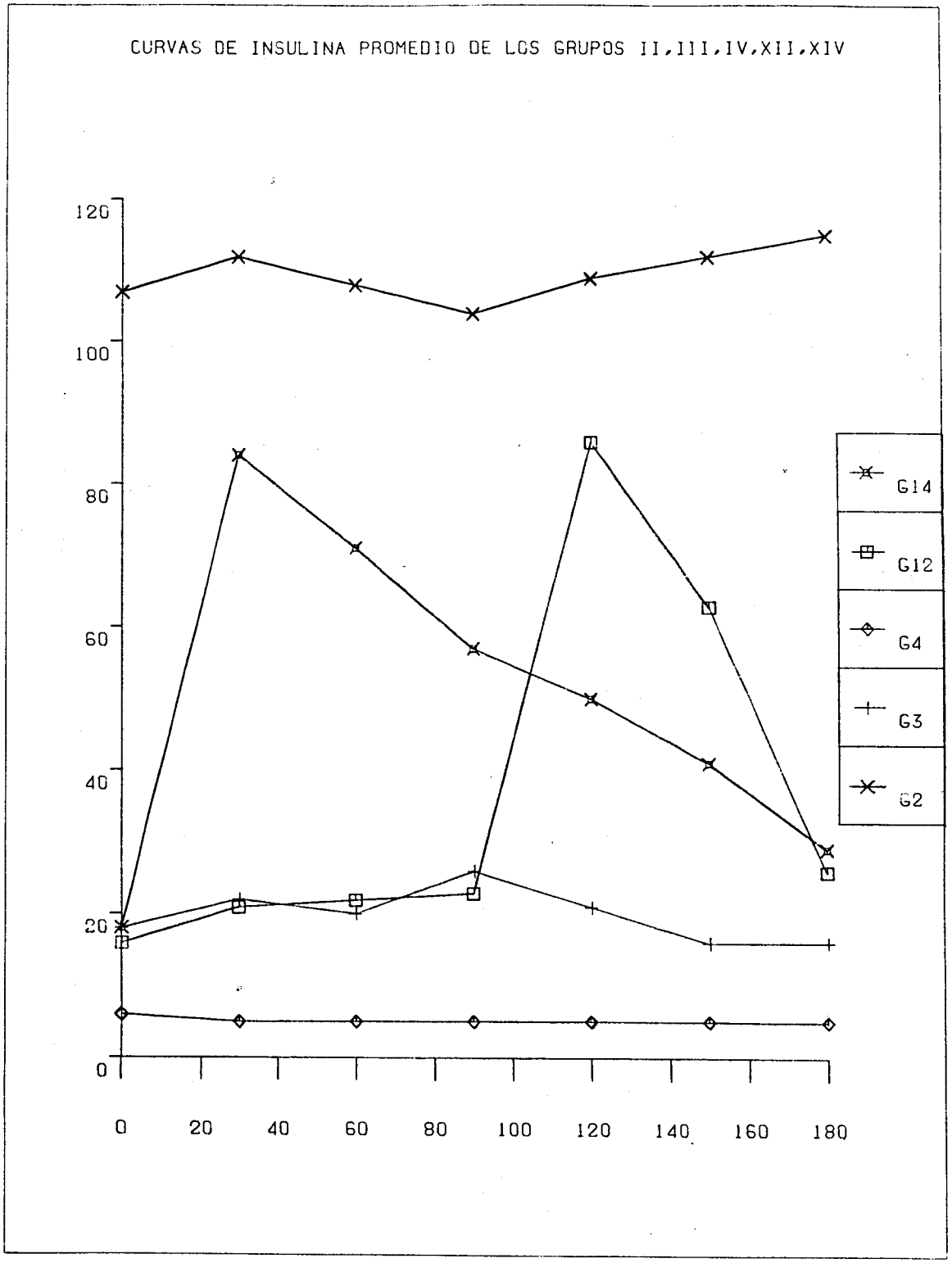

CURVAS DE INSULINA PROMEDIO DE LOS GRUPOS II,III,IV,XII,XIV
120
100
80
60
40
20
0
0
20
40
60
80
100
120
140
160
180
G14
G12
G4
G3
G2

Fig. 6 (bis)

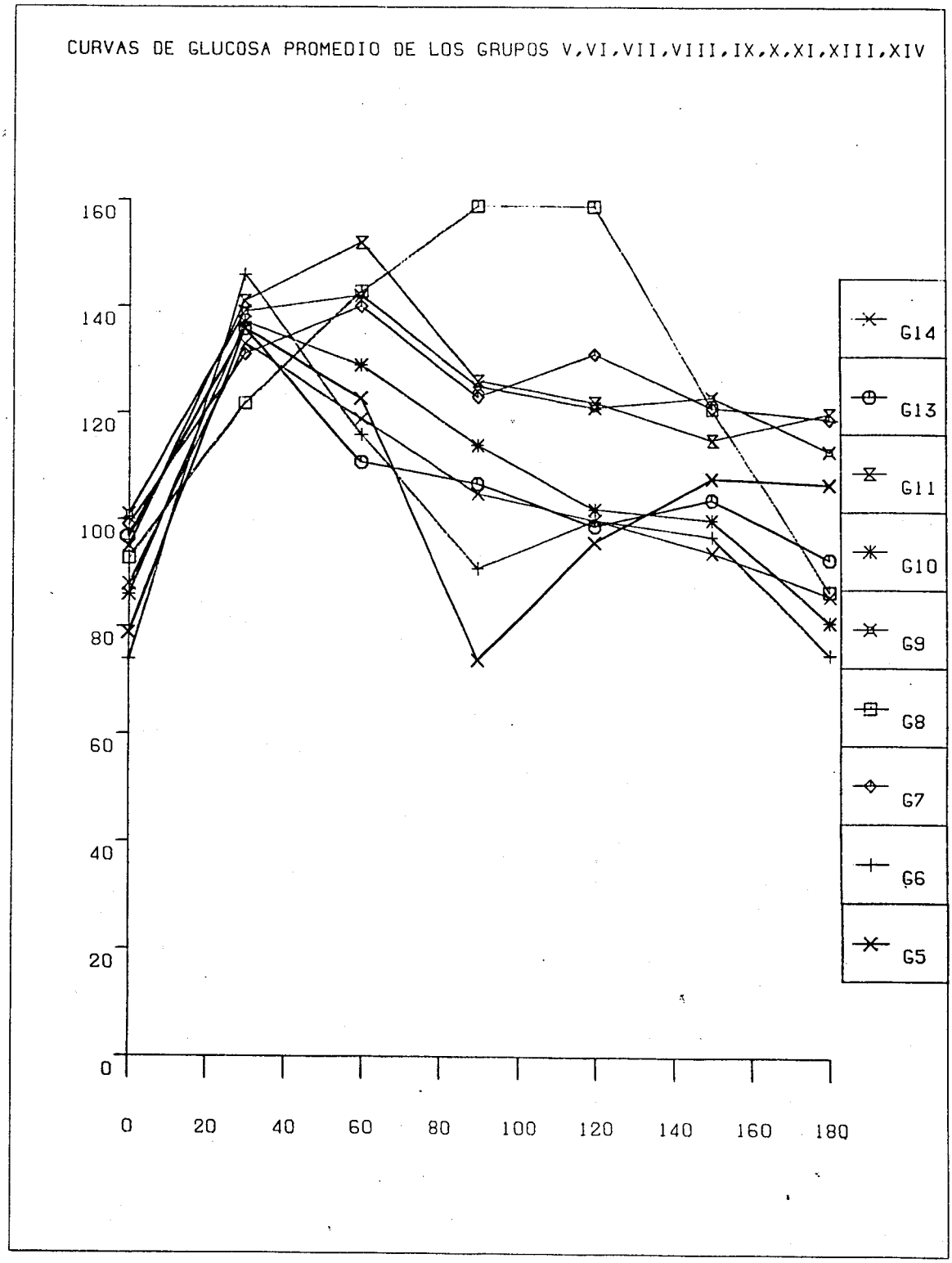

CURVAS DE GLUCOSA PROMEDIO DE LOS GRUPOS V,VI,VII,VIII,IX,X,XI,XIII,XIV
160
140
120
100
80
60
40
20
0
0
20
40
60
80
100
120
140
160
180
G14
G13
G11
G10
G9
G8
G7
G6
G5

Fig. 7

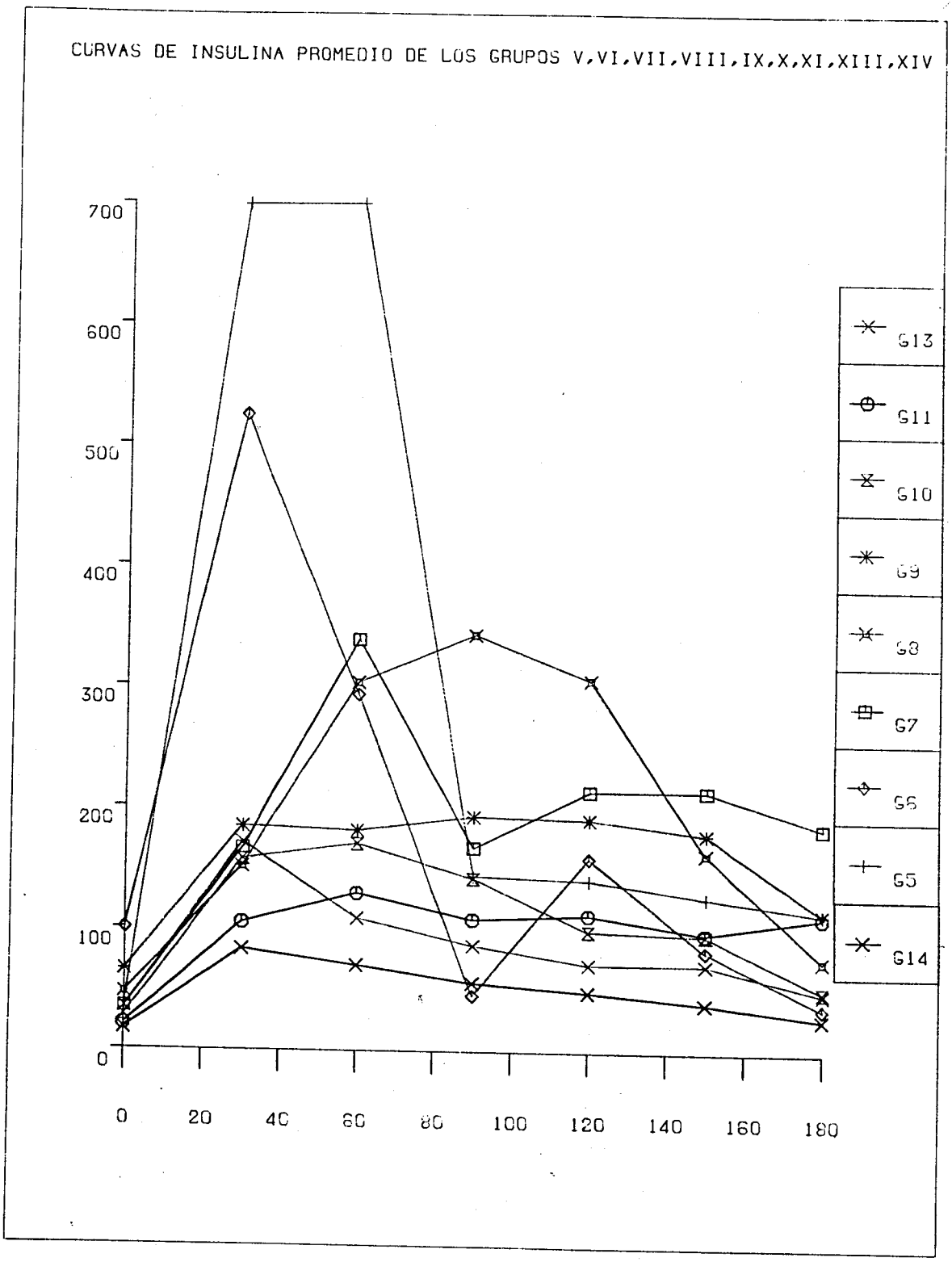

CURVAS DE INSULINA PROMEDIO DE LOS GRUPOS V,VI,VII,VIII,IX,X,XI,XIII,XIV
700
600
500
400
300
200
100
0
0
20
40
60
80
100
120
140
160
180
G13
G11
G10
G9
G8
G7
G6
G5
G14

Fig. 8

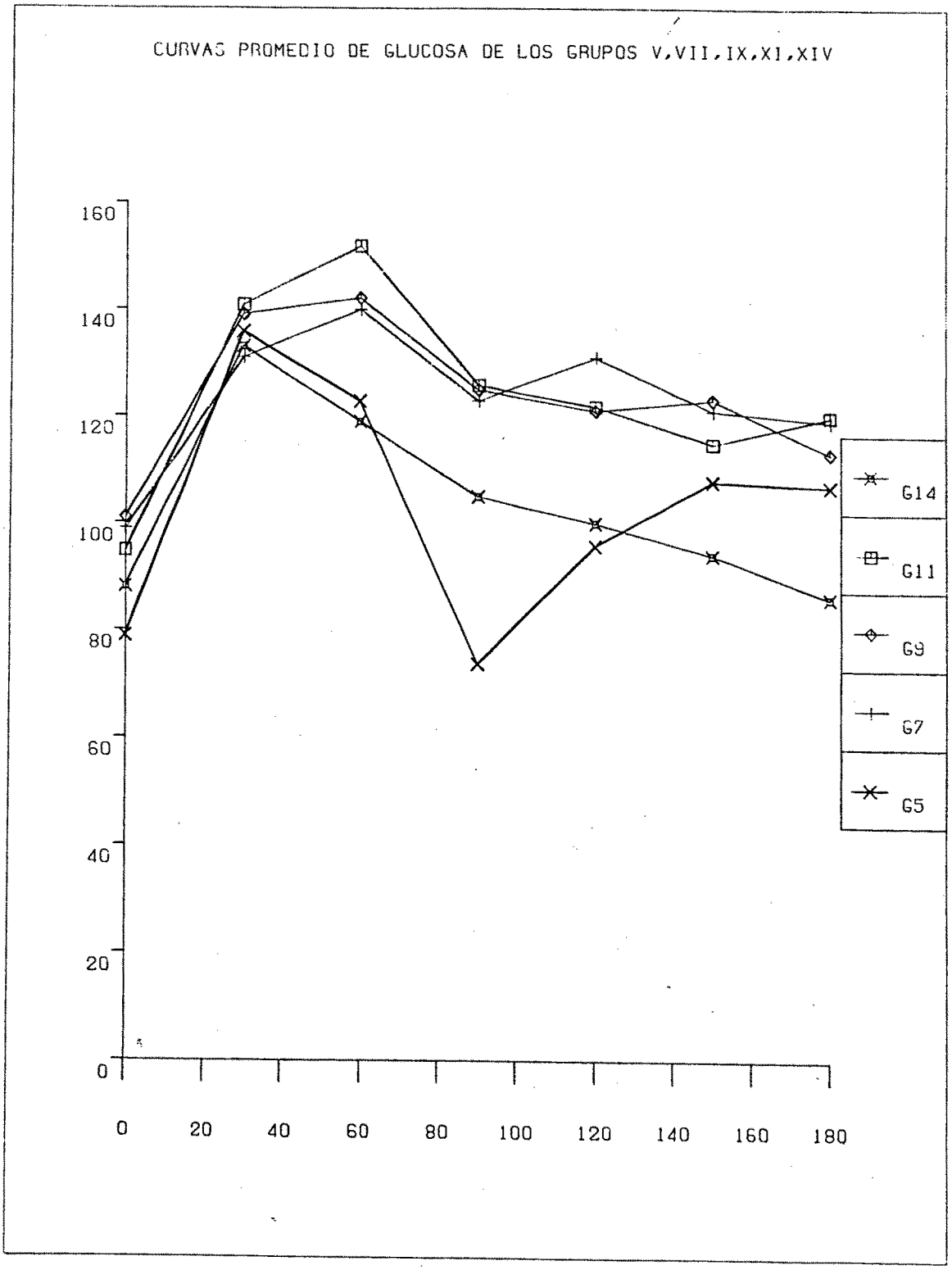

CURVAS PROMEDIO DE GLUCOSA DE LOS GRUPOS V,VII,IX,XI,XIV
160
140
120
100
80
60
40
20
0
0
20
40
60
80
100
120
140
160
180
G14
G11
G9
G7
G5

Fig. 9

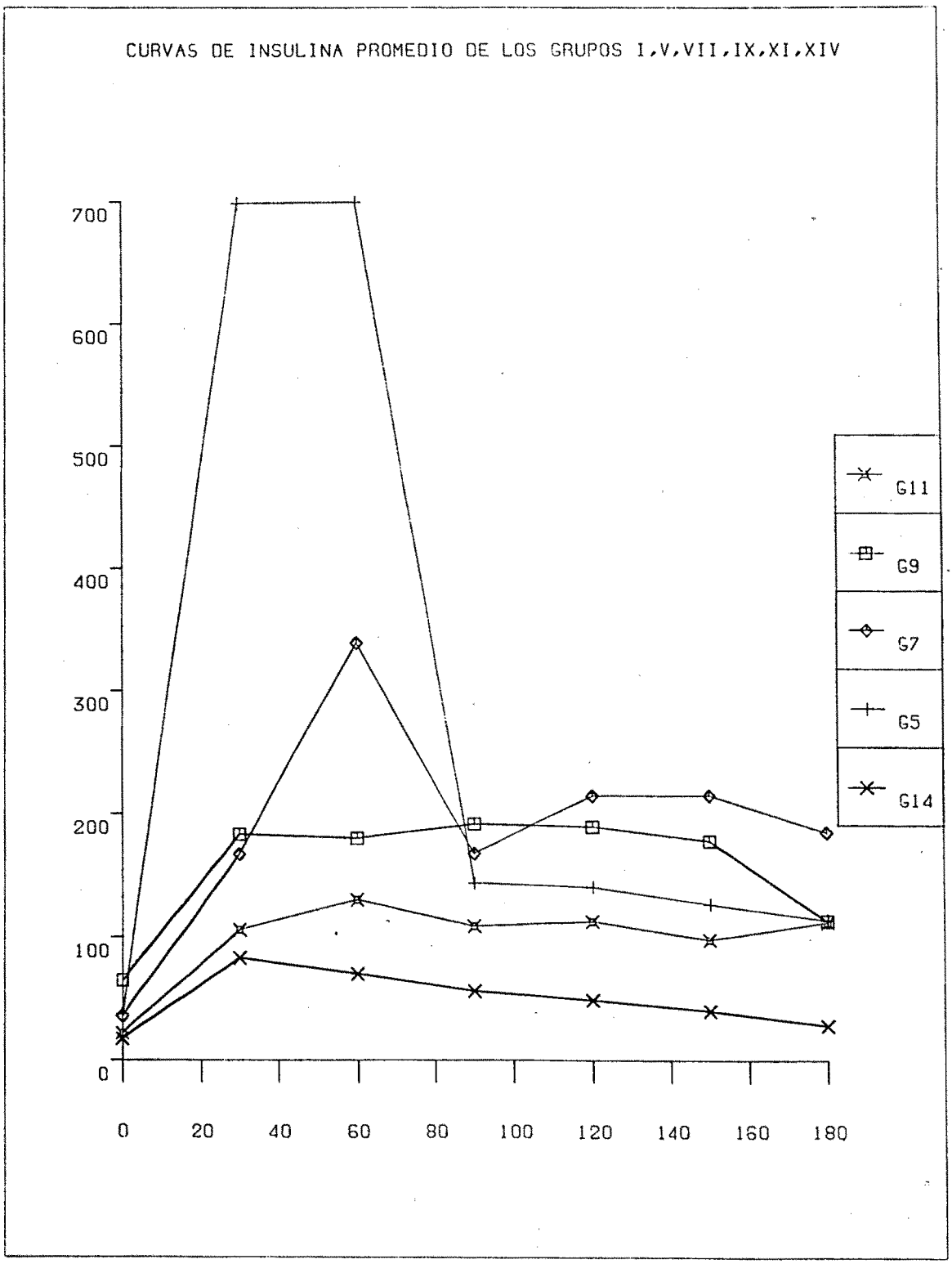

CURVAS DE INSULINA PROMEDIO DE LOS GRUPOS I,V,VII,IX,XI,XIV
700
600
500
400
300
200
100
0
0
20
40
60
80
100
120
140
160
180
G11
G9
G7
G5
G14

Fig. 10

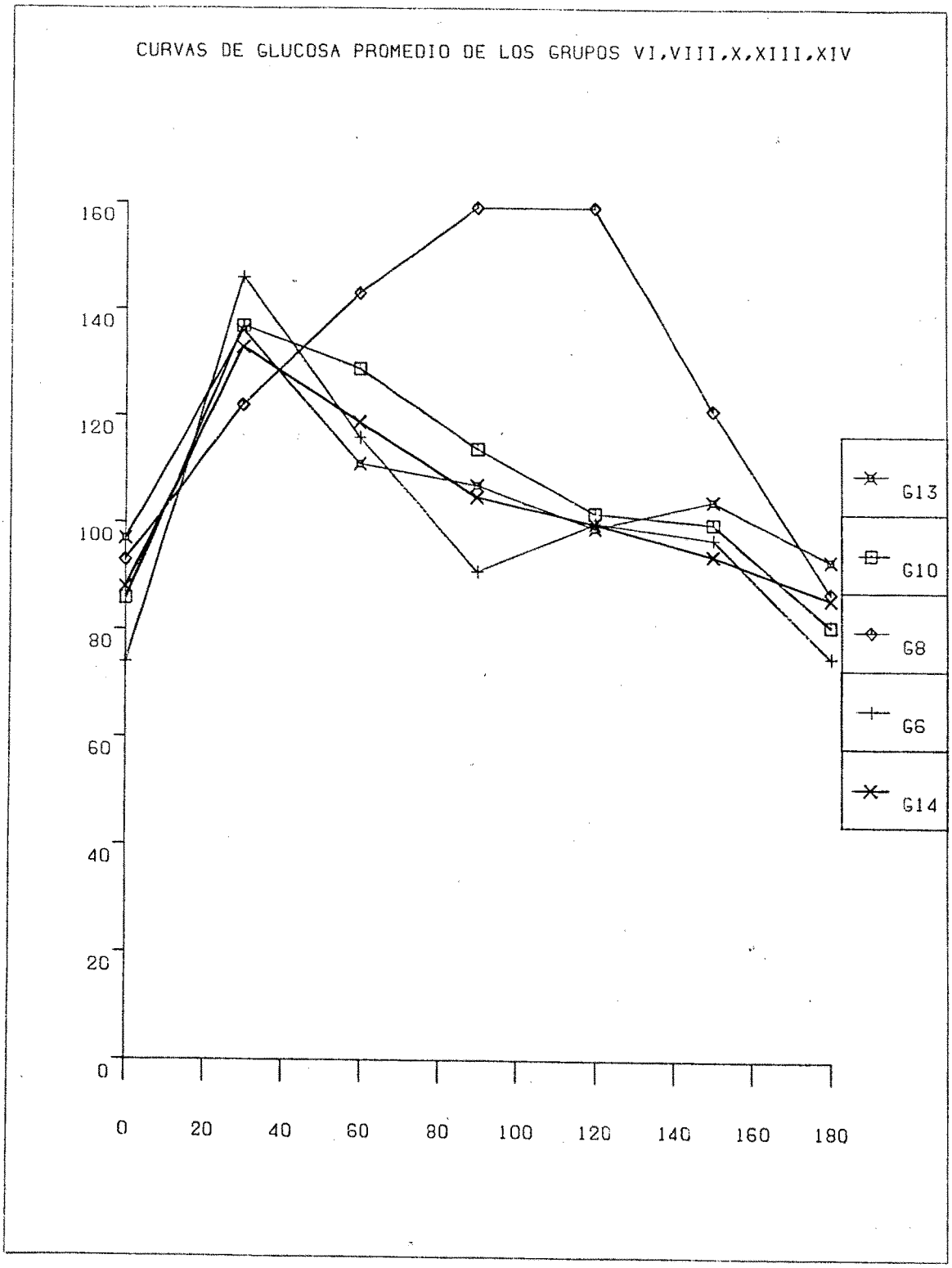

CURVAS DE GLUCOSA PROMEDIO DE LOS GRUPOS VI,VIII,X,XIII,XIV
160
140
120
100
80
60
40
20
0
0
20
40
60
80
100
120
140
160
180
G13
G10
G8
G6
G14

Fig. 11

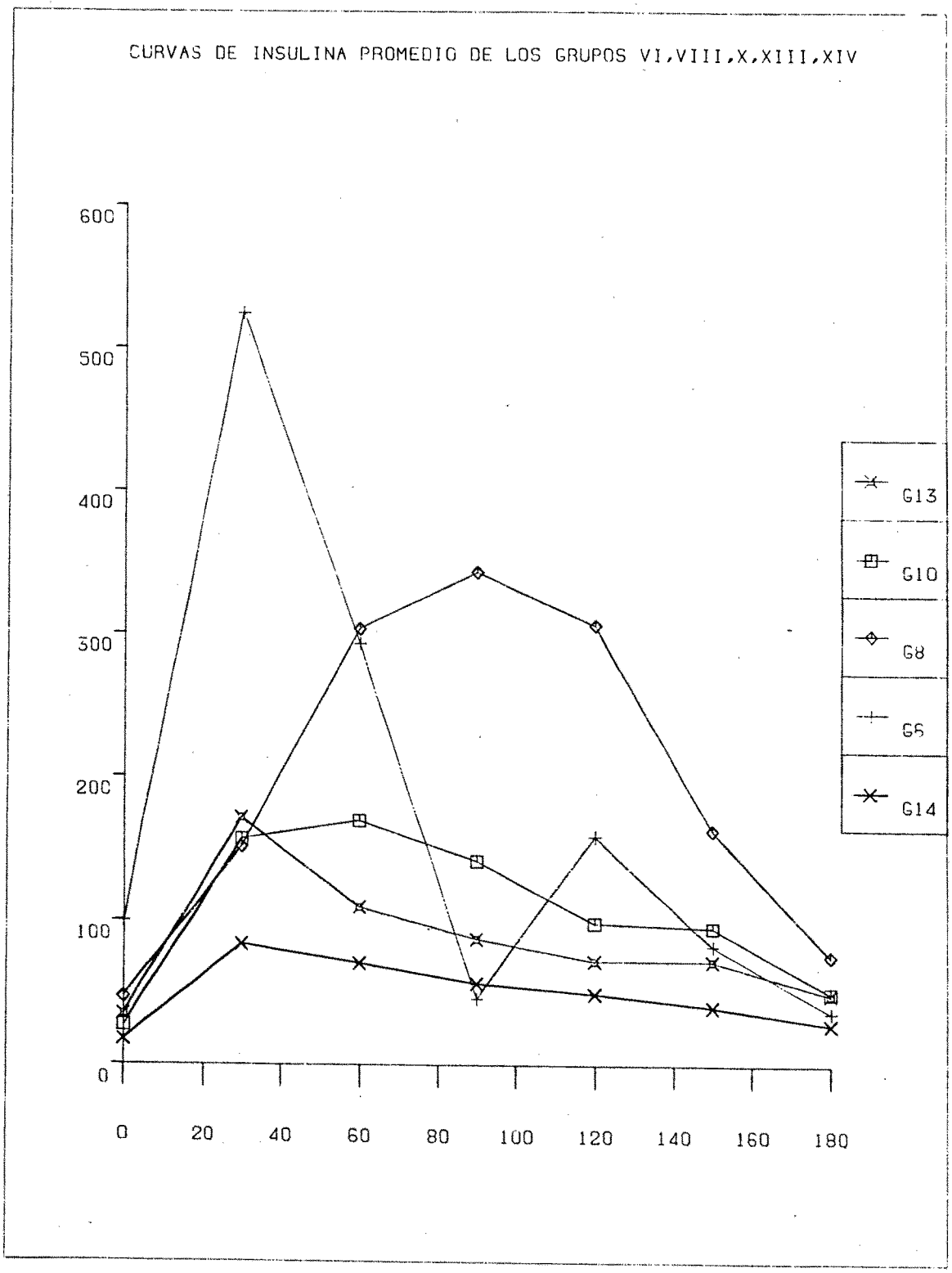

CURVAS DE INSULINA PROMEDIO DE LOS GRUPOS VI,VIII,X,XIII,XIV
600
500
400
300
200
100
0
0
20
40
60
80
100
120
140
160
180
G13
G10
G8
G6
G14

Fig. 12

| | |
|---|---|
| TTOG PATOLOGICA | INSULINOPENICOS: GRUPOS III, IV |
| | INSULINRESISTENCIA: GRUPOS I, II, XII |
| TTOG DENTRO DE LA NORMALIDAD | HIPERINSULINEMIA INEFICAZ: GRUPOS V, VII, IX, XI |
| | HIPERINSULINEMIA EFICAZ: GRUPOS VI, VIII, X, XIII |
| | NORMOINSULINEMIA: GRUPOS XIV |

TABLA I

Aunque los grupos son estadísticamente diferentes, no lo han de ser clínicamente, no obstante, es interesante tenerlos en cuenta, pues distintas evoluciones clínicas podrían estar relacionadas con algunos de los grupos descritos. Así, por ejemplo, los grupos con insulinresistencia podrían evolucionar a insulinopenia y dentro de los grupos que tuvieran esta evolución, algunos lo podrían hacer más o menos rápidamente.

También resulta interesante observar que en algunos grupos los individuos que los componen, responden mayoritariamente a una determinada constitución (obesos en los grupos X, VIII, IX y XII), una determinada enfermedad (grupos I y V con Acantosis Nigrigans y Distrofia Miotónica respectivamente).

Con las historias clínicas completas, se podría investigar si existe alguna relación con otras variables como edad, sexo, etc...

Todo esto ayudaría a aclarar algunas paradojas que existen en esta enfermedad.

# 6. DIAGNOSTICO AUTOMATIZADO DE LOS TEST DE TOLERANCIA ORAL A LA GLUCOSA

Página

Sumario:

## 6.1. INTRODUCCION

En el presente capítulo se pretende establecer un algoritmo basado en una función distancia y en la aplicación de técnicas de bayesianas que permita asignar a un niño, sobre el que se ha realizado un TTOG, a uno de los grupos establecidos en el capítulo V de una forma objetiva, automatizada y con ciertas propiedades de aprendizaje, siguiendo la línea de trabajo que sobre diagnóstico automatizado de enfermedades hematológicas realizaron Oller y Rios (1985).

## 6.2. METODO DIAGNOSTICO

Sea una población $\Omega$ formada por todos los niños de una determinada región geográfica que acuden a un centro hospitalario para realizar un TTOG, en ella podemos distinguir k grupos distintos $E_1$, $E_2$, ..., $E_k$, que supondremos abarcan a toda la población y que podemos identificar, al menos provisionalmente, con los grupos establecidos en el capítulo anterior.

Dado un individuo de la población, si no disponemos de más información del mismo, hay "a priori" una determinada probabilidad $p(E_i)$ de que pertenezca al grupo $E_i$. Cuando efectuamos sobre este individuo un TTOG, la información que nos da éste, debe determinar un cambio en la distribución de probabilidad de los distintos grupos formados. Si pretendemos efectuar un diagnóstico, una regla de decisión razonable consiste en asignar a este individuo al grupo más probable. El problema

consiste pues en cuantificar dichas probabilidades "a posteriori", siendo razonable la utilización del teorema de Bayes.

Una primera dificultad reside en el hecho de disponer de un número no numerable (contínuo) de resultados posibles, r, por lo que deberíamos escribir

$$p\ (E_i/r) = \frac{p(E_i)\ f(r/E_i)}{\sum_{\mu=1}^{k} p\ (E_\mu)\ f\ (r/E_\mu)} \qquad (1)$$

donde $f(r/E_\mu)$ es la función de densidad de los resultados, condicionados al grupo $E_\mu$.

La obtención de una buena estimación de la función de densidad es dificultosa (Duda and Hart (1973)) por lo que se propone una alternativa razonable para soslayar este problema.

Si A es el conjunto de todos los posibles resultados de los parámetros del modelo (50) descrito en el capítulo V que queda determinado por los resultados del TTOG y además se tiene definida una distancia en el conjunto como la descrita en el capítulo IV, se puede establecer una partición $P = \{R_1, R_2, \ldots, R_k\}$ de A definida de la siguiente forma

$$a \in R_j \iff \begin{cases} \bar{d}\ (a, E_j) < \bar{d}\ (a, E_i) & 1 \leq i \leq j-1 \\ \\ \bar{d}\ (a, E_j) < \bar{d}\ (a, E_i) & j \leq 1 \leq i \leq k \end{cases} \qquad (2)$$

siendo $\bar{d}\ (a, E_j)$ la distancia promedio del resultado a, al grupo $E_j$ y cuyo valor lo podemos definir como:

$$\bar{d}\ (a,\ E_j) = E\ (d\ (a,r)\ /\ E_j) = \int_A d\ (a,r)\ f\ (r/E_j)\ dr \qquad (3)$$

el cual si tenemos una muestra M suficientemente grande de individuos de la población, es fácilmente estimable a partir de los resultados de los análisis de los niños de cada uno de los grupos y que vendrá dada por:

$$\bar{d}\ (a,\ E_j) = \frac{1}{n_j} \sum_{e \in M \cap E_j} d\ (a,e) \qquad (4)$$

siendo $n_j$ el número de individuos de M que pertenecen a $E_j$.

Para la aplicación de este método, es necesario que en A haya definida una distancia.

La estimación de las probabilidades $p\ (R_j/E_i)$ se efectúa a partir de la proporción de individuos de la muestra M, perteneciente al grupo $E_i$ que presentan el resultado $R_j$.

Si queremos asignar un nuevo individuo $\omega$, sobre el que se ha obtenido un resultado $a \in A$, a una de las clases $E_1$, $E_2$, ... ..., $E_k$, debemos determinar a que elemento de la partición $R_j$ pertenece dicho resultado y posteriormente aplicamos el teorema de Bayes.

$$p\ (E_i/R_j) = \frac{p\ (R_j/E_i)\ p\ (E_i)}{\sum_{\mu=1}^{k} p\ (R_j/E_\mu).p(E_\mu)} \qquad (5)$$

siendo la estimación de $p\ (E_i)$ la frecuencia relativa del suceso $E_i$ en la muestra M.

Por último asignamos al individuo $\omega$ que tuvo el resultado a, al grupo más probable. Es decir, decidimos

$$\omega \in E_e \iff p\,(E_e/R_j) = \max.\ \{p(E_i/R_j) \quad 1 \leq i \leq k\} \tag{6}$$

siendo la probabilidad de error de esta decisión

$$p\,(\text{error}) = 1 - p\,(E_e/R_j) \tag{7}$$

Si se ha podido determinar con certeza, mediante algún otro procedimiento el grupo al que pertenece el individuo $\omega$ que queríamos diagnosticar, puede ser utilizado éste como un elemento más de la muestra M, mejorando así la definición de la partición P y las estimaciones de las probabilidades $p\,(E_i)$ y $p\,(R_j/E_i)$.

## 6.3. DISTANCIA EN EL CONJUNTO DE RESULTADOS

El conjunto de resultados de los TTOG puede identificarse con la variedad paramétrica

$$E = \{\beta \in R^8 \ / \ \beta = (\alpha_o, \alpha_1, \alpha_2, \alpha_3, \beta_o, \beta_1, \beta_2, \beta_3);\} \tag{8}$$

donde las componentes de $\beta$ son los parámetros del modelo (49) descrito en el capítulo V.

La distancia entre dos puntos $\beta_A$ y $\beta_B$ de la variedad E, viene dada por

$$d^2 = \frac{1}{\sigma^2}\,(\beta_A - \beta_B)^t\, X^t X\, (\beta_A - \beta_A) \tag{9}$$

donde X es la matriz de diseño del modelo (50) y $\sigma^2 = \frac{1}{25}$ según el capítulo V.

## 6.4. RESULTADOS

Como hemos dicho en el punto 6.2., partimos de una población formada por los niños que acuden a un centro hospitalario para que se les realice un TTOG y que la consideramos dividida en 14 grupos $E_i$, los cuales tienen unas probabilidades "a priori", estimadas a través de la muestra M de 171 niños, iguales a:

$$p(E_1)=p(E_2)=p(E_4)=p(E_5)=p(E_6)=p(E_8)=p(E_{12})=0.5848.10^{-2}$$

$$p\ (E_3) = p\ (E_7) = p\ (E_9) = 0.1169.10^{-1}$$

$$p\ (E_{11}) = 0.7602.10^{-1} \qquad (10)$$

$$p\ (E_{13}) = 0.8772.10^{-1}$$

$$p\ (E_{14}) = 0.7135$$

Las probabilidades condicionadas a los grupos, estimadas a partir de la muestra controlada, fueron las siguientes

$$p\ (R_j/E_i) = \begin{cases} 1 & i=j \\ 0 & 1\neq j \end{cases} \qquad (11)$$

$$1 \leq j \leq 14$$

$$1 \leq i \leq 13$$

esto indica que los TTOG dan mucha información para diagnosticar los 13 primeros grupos, por otra parte:

$$p\ (R_j/E_{14}) = 0; \quad j \neq 11,13,14$$

$$p\ (R_{12}/E_{14}) = 0.1639.10^{-1} \qquad (12)$$

$$p\ (R_{13}/E_{14}) = 0.9016.10^{-1}$$

$$p\ (R_{14}/E_{14}) = 0.8934$$

Seguidamente, por el método descrito en los puntos anteriores, se trató de asignar al grupo más probable, 16 nuevos TTOG que se realizaron en el mismo hospital, elegidos aleatoriamente. Los resultados $R_j$ de los casos j, fueron los siguientes:

| CASO | TIEMPOS | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0 m. | 30 m. | 60 m. | 90 m. | 120 m. | 150 m. | 180 m. |
| 1 | | | | | | | |
| Glucosa | 77 | 150 | 112 | 113 | 84 | 93 | 75 |
| Insulina | 19 | 196 | 66 | 26 | 52 | 79 | 34 |
| 2 | | | | | | | |
| Glucosa | 68 | 89 | 110 | 122 | 95 | 77 | 87 |
| Insulina | 85 | 28 | 43 | 47 | 34 | 27 | 39 |
| 3 | | | | | | | |
| Glucosa | 94 | 152 | 100 | 98 | 109 | 104 | 127 |
| Insulina | 18 | 152 | 52 | 32 | 24 | 26 | 40 |
| 4 | | | | | | | |
| Glucosa | 108 | 151 | 121 | 100 | 97 | 103 | 99 |
| Insulina | 21 | 96 | 110 | 52 | 50 | 40 | 18 |

| CASO | TIEMPOS | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0 m. | 30 m. | 60 m. | 90 m. | 120 m. | 150 m. | 180 m. |
| 5 | | | | | | | |
| Glucosa | 88 | 182 | 134 | 100 | 109 | 95 | 85 |
| Insulina | 12 | 88 | 80 | 64 | 30 | 22 | 18 |
| 6 | | | | | | | |
| Glucosa | 101 | 127 | 92 | 84 | 94 | 84 | 97 |
| Insulina | 18 | 86 | 50 | 32 | 36 | 14 | 10 |
| 7 | | | | | | | |
| Glucosa | 83 | 103 | 160 | 147 | 125 | 118 | 110 |
| Insulina | 16 | 32 | 122 | 106 | 52 | 30 | 36 |
| 8 | | | | | | | |
| Glucosa | 91 | 147 | 93 | 71 | 93 | 100 | 104 |
| Insulina | 11 | 55 | 57 | 16 | 25 | 17 | 22 |
| 9 | | | | | | | |
| Glucosa | 103 | 117 | 161 | 156 | 141 | 119 | 100 |
| Insulina | 25 | 48 | 115 | 109 | 118 | 70 | 34 |
| 10 | | | | | | | |
| Glucosa | 113 | 187 | 136 | 118 | 116 | 118 | 120 |
| Insulina | 24 | 190 | 148 | 122 | 86 | 92 | 92 |
| 11 | | | | | | | |
| Glucosa | 92 | 173 | 119 | 101 | 111 | 89 | 94 |
| Insulina | 19 | 188 | 94 | 70 | 84 | 34 | 34 |
| 12 | | | | | | | |
| Glucosa | 76 | 117 | 98 | 59 | 98 | 89 | 98 |
| Insulina | 10 | 24 | 34 | 16 | 18 | 18 | 18 |
| 13 | | | | | | | |
| Glucosa | 86 | 179 | 185 | 107 | 125 | 93 | 103 |
| Insulina | 12 | 54 | 78 | 48 | 70 | 30 | 32 |
| 14 | | | | | | | |
| Glucosa | 82 | 124 | 197 | 205 | 153 | 71 | 105 |
| Insulina | 128 | 36 | 68.8 | 93.3 | 62.9 | 53,6 | 35.4 |
| 15 | | | | | | | |
| Glucosa | 111 | 207 | 155 | 121 | 100 | 123 | 95 |
| Insulina | 27.4 | 212 | 176 | 120 | 61.9 | 103 | 51.2 |
| 16 | | | | | | | |
| Glucosa | 111 | 189 | 195 | 153 | 141 | 102 | 123 |
| Insulina | 29 | 130 | 142 | 110 | 104 | 62 | 122 |

Las clasificaciones fueron las siguientes:

| Número del caso | Grupo clasificador | Probabilidad estimada de error |
|---|---|---|
| 1 | 14 | 0 |
| 2 | 14 | 0 |
| 3 | 14 | 0 |
| 4 | 14 | 0 |
| 5 | 14 | 0 |
| 6 | 14 | 0 |
| 7 | 14 | 0 |
| 8 | 14 | 0 |
| 9 | 14 | 0 |
| 10 | 13 | 0.4231 |
| 11 | 14 | 0 |
| 12 | 14 | 0 |
| 13 | 14 | 0 |
| 14 | 12 | 0 |
| 15 | 13 | 0.4231 |
| 16 | 11 | 0.1333 |

## 6.5. DISCUSION

Hemos visto un método automático de diagnóstico que además permite cuantificar la probabilidad de error. Este método también tiene propiedades de aprendizaje, pues una vez diagnosticado un individuo con certeza, se introduce en la muestra M para de esta forma mejorar las estimaciones de las probabilidades (10), (11) y (12).

No obstante, este método admitiría la siguiente crítica; por las características de las variables, valores de glucemia

e insulinemia, podrían existir más grupos de los que se han descrito en el capítulo V, por lo que en lugar de realizar el proceso diagnóstico del capítulo VI, en donde se suponía que sólo podrían existir 14 grupos, se podría haber aumentado la muestra M de 171 niños, con los 16 nuevos casos y haber realizado de nuevo la clasificación, siguiendo el capítulo V.

# 7. RESUMEN DE LOS RESULTADOS

Página

Sumario:

## 7.1. RESULTADOS

Los resultados más relevantes que se pueden extraer de esta memoria son los siguientes:

Partiendo del concepto de información o entropia de Shannon se han definido y obtenido expresiones algebraicas de distancias entre modelos lineales normales univariantes, de igual y distinta varianza, y entre modelos lineales multivariantes de igual varianza, que gozan de buenas propiedades matemáticas entre ellas, propiedades de invarianza.

Se han obtenido explícitamente expresiones algebraicas de los estimadores de estas distancias que nos sirvieron para diseñar contrastes de hipótesis que comparan dos o más modelos lineales normales univariantes y multivariantes de varianzas iguales y dos modelos lineales normales univariantes de varianzas distintas, estableciendo su relación con los contrastes de hipótesis propios del Análisis de la Varianza en el caso de varianzas iguales.

Las ventajas del uso de las distancias en la realización de estos contrastes, radica en que se pueden hacer representaciones gráficas de los resultados, bien sea representando a los modelos lineales comparados en un grafo (dendograma o árbol aditivo), o bien en un espacio euclídeo de dimensión reducida (usualmente 2).

Definir los contrastes de hipótesis del Análisis de la Varianza a través de una distancia tendría interés para el

investigador experimental que en su necesidad de cuantificar diferencias, hace en muchos casos mal uso metodológico del nivel de significación en los contrastes de hipótesis.

En cuanto a las aplicaciones prácticas, se hace un análisis y clasificación de los distintos tipos de respuesta al TTOG que presentan una muestra de niños. A cada niño se le asocia un modelo lineal, en concreto un polinomio de tercer grado, que representa la desviación de un niño "teórico". A través de la distancia definida anteriormente se comparan entre sí. Posteriormente se utilizan las técnicas del Análisis de Coordenadas Principales y de la Taxanomía Numérica para establecer grupos que sean significativamente distintos estadísticamente, aunque estas diferencias en principio no tengan interés clínico.

Los grupos formados son caracterizados por su curva promedio y se hace una descripción de los mismos en términos médicos.

Posteriormente se procede a establecer un método de diagnóstico automático, que permite asignar a un niño al que se le realiza un TTOG a uno de los grupos establecidos anteriormente. Esta asignación se realiza de una forma objetiva, que no admite ambigüedades, cuantificando el error de la clasificación y con propiedades de aprendizaje.

Es de destacar que la metodología expuesta es aplicable al estudio de cualquier enfermedad o síndrome siempre que los

individuos estudiados sean identificables a elementos de un espacio métrico, por ejemplo siempre que se asocie cada individuo a un modelo lineal, a una distribución de probabilidad, etc...

En un estudio posterior, no desarrollado en esta memoria, se podría calcular, aunque fuera de una forma numérica, distancias entre modelos lineales multivariantes de distinta varianza y en cuanto al apartado de las aplicaciones se podría realizar un seguimiento de los individuos de los grupos evaluando tiempo hasta que aparecen complicaciones, gravedad de las mismas, tratamientos, etc... También se podría hacer un estudio más formal de las curvas promedio de los grupos.

# BIBLIOGRAFIA

ANDERSON, T.W. (1958). An Introduction to Multivariable Statistical Analysis. John Wiley & Sons, Inc.

APOSTOL, T. (1978). Calculus. Ed. Reverté.

ARCAS, A. (1982). Contribuciones a las clasificaciones estratificadas. Tesina. Universidad de Barcelona (inédita).

ARCAS, A. y SALICRU, M. (1984). Qüestió V.8 nº 3. pp. 113-120.

ATKINSON, C. & MITCHELL, A.F.A. (1981). Rao's distance measure. Sankhyâ, 43, A, 345-365.

BENZECRI, J.P. (1973). L'Analyse des Données. I. La taxonomie. L'Analyse des Donnés. II. L'Analyse des correspondances. Dunod, Paris.

BHATTACHARYYA, A. (1946). On a measure of divergence between two multinomial populations. Sankhyâ, 7, 401-406.

BURBEA, J. & RAO, C.R. (1982 a). Entropy differential metrie, distance and divergence measures in probability spaces: a unified approach. Jour. Multivariate Analysis, 12, 575-596.

BURBEA, J. & RAO, C.R. (1982 b). Differential metrics in probability spaces. Probability Math. Statist., 12, 115-132.

COLLATZ, L. (1966). Functional Analysis and Numerical Mathematics. Academic Press.

CUADRAS, C.M. (1974 a). Análisis Estadístico Multivariante y Representación Canónica de Funciones Paramétricas Estimables. Universidad de Barcelona, Secretariado de Publicaciones.

CUADRAS, C.M. (1974 b). Análisis discriminante de funciones paramétricas estimables. Trabajos de Estadística y de Investigación Operativa, vol. XXV, (1974), 3-31.

CUADRAS, C.M. (1979 ). Sobre la comparació estadística de corbes experimentals. Qüestió, 3 (1), 1-9.

CUADRAS, C.M. ( 1980 ). Métodes de Representació de Dades i la seva aplicació a la Biología. Col.loquis de la Societat Catalana de Biología (1979). (Matemática i Biología) 96-133.

CUADRAS, C.M. (1981). Métodos de Análisis Multivariante. Eunibar. Barcelona.

CUADRAS, C.M. y USON, T. (1980). Sobre la propiedad euclídea de las distancias ultramétricas. Actas XII Reunión Nac. Estad. Inv. Op. Inform.

ELLENBERG, M & RIFKIN, M. (1982). Diabetes Mellitus. Medical Examination Publishing Co., INC. New York.

ELSGOLTZ, L. (1977). Ecuaciones diferenciales y cálculo variacional. Ed. Mir.

GORDON L. ATKINS. Investigation of some Theorétical Models Relating the Concentrations of Glucosa and Insulin in Plasma. J. theor, Biol. 32, 471-494.

GOWER, J.C. (1967). A comparison of some methods of cluster analysis. Biometrics, 23, 623-637.

GOWER, J.C. (1971). A general coefficient of similarity and some of its properties. Biometrics, 27, 857-874.

HICKS, N.J. (1965). Notes on Differential Geometry, Van Nostran, Princeton.

HICKS, N.J. (1974). Notas sobre geometría diferencial. Hispano Europea.

HOTELLING, H. (1931). The generalization of Student's ratio. Annuals of Math. Stat. 2, 360-378.

HOTTELLING, H. (1933). Analysis of a complex of Statistical variables into principal components. J. Educ. Psychol., 24 (6), 417-441.

JOHNSON, S.C. (1967). Hierarchical clustering schemes. Psychometrika, 32, 241-254.

LAWLEY, D.N. (1938). A generalization of Fisher's z-test. Biometrika, 30, 180-187.

LEFEBVRE, J. (1976). Introduction aux Analyses Statistiques Multidimensionnelles. Masson.

LINDGREN, B.W. (1976). Statistical Theory. Macmillan Publishing Co., New York.

MAHALANOBIS, P.C. (1936). On the generalized distance in statistics. Proc. Natl. Inst. Sci. India, 2 (1), 49-55.

MARDIA, K.V.; KENT, J.T. y BIRBY, J.M. (1979). Multivariate Analysis. Academic Press, London.

MATUSITA, K. (1966). A distance and Related statistics in Multivariate Analysis. En: Multivariate Analysis I (P.R. Krishnaiah, ed.) Academic Press, New York.

MOOD y GRAYBILL (1969). Introducción a la teoría de la Estadística. Aguilar.

NATIONAL DIABETES DATA GROUP: "Clasification and diagnosis of diabetes mellitus and other categories of glucose intolerance" "Diabetes", 28: 1039, 1979.

NORWICH, K.H.: Mathematical Models of the kinetics of glucose and insulin in plasma. Bulletin of Mathematical Biophysics. Volume 31, 1969.

OLLER, J.M. (1982). Utilización de Métricas Riemannianas en Análisis de Datos Multidimensionales y su Aplicación a la Biología. Tesis Doctoral.

OLLER, J.M. & CUADRAS, C.M. (1982). On a defined distance for negative multinomial distribution. Abstracts of XIth Inter. Biometrie Conf., Toulouse, 69.

OLLER, J.M. y CUADRAS, C.M. (1985). Rao's distance for Negative Multinomial Distributions, Sanukyà, V 47, A, 75-83.

OLLER, J.M. y RIOS, M.: Description of an algorithm for automized diagnosis. Pendiente de publicación.

PEARSON, K. (1901). On lines and Planes of Closest Fit to Systems of Points in Space. Phil Mag., Ser. 6, 2 (11), 559-572.

PUIG ADAM, P. (1975). Curso teórico-práctico de Ecuaciones diferenciales aplicado a la física y técnica. Biblioteca Matemática, S.L.

RAO, C.R. (1945). Information and the accuracy attainable in the estimation os statistical parameters. Bull. Calcutta Math. Soc., 27, 81-91.

RAO, C.R. (1948 a). The utilization of Multiple Measurements in Problems of Biological Classification. J. Roy, Stat. Soc. B10 (2), 159-203.

RAO, C.R. (1948 b). On the distance between two populations. Sankhyà, 9, 246-248.

RAO, C.R. (1949). On the distance between two populations. Sankhyà, 9, 246-248.

RAO, C.R. (1952). Advanced statistical methods in Biometrics Research (1952). Hafner Publishing Company.

RAO, C.R. (1964). The use and interpretation of principal component analysis in applied research. Sankhya, Serv. A, 26, 329-358 (1964).

ROUVIER, R. (1966). L'Analyse en Composantes principales: son utilisation en Genetique et ses rapports avec l'analyse discriminatoire. Biometrics 22 (2), 343-357.

ROZMAN, C. (1979). Medicina Interna. Marin

SALICRU, M. (1983). Consideraciones sobre desemejanzas y clasificaciones asociadas. Tesina.

SCHEFFE, H. (1959). The Analysis of Variance. J. Wiley, New York.

SEARLE, S.R. (1971). Linear Models, John Wiley, New York.

SEBER, G.A.F. (1977). Linear Regression Analysis. J. Wiley, New York.

SHANNON, C.E. (1948). A mathematical theory of communications, Bell System Tech. J. 27, 379-423, 623-656.

SHEPARD, R.N. (1962). The analysis of proximities. Multidimensional scaling with an unknown distance function. I. Psychometrika, 27, 219-246.

SOKAL, R.R. & ROHLF, F.J. (1969). Biometría. Blume. Madrid.

SOKOLNIKOFF, I.S. (1971). Análisis Tensorial (1971). Index.

SPIEGEL, M.R. (1970). Análisis vectorial. McGraw-Hill.

SPIVAK, M. (1970). Cálculo en variedades. Reverté.

SPIVAK, M. (1979). A comprehensive Introduction to Differential Geometry. Publish or Perish, Inc. Berkeley.

STRUIK, D.J. (1966). Geometría diferencial clásica. Aguilar.

TUCKER, H.G. (1973). Introducción a la teoría matemática de las probabilidades y a la estadística. Vicens Vives.

WHO Expert Committée on Diabetes Mellitus. Ginebra, 1980.