Acta Geológica Hispánica, t. VII (1972), n.º 1, págs. 22-25

El análisis canónico y su aplicación en Geología*

por C. Cuadras-Avellana **, J. A. Campá-Viñeta *** y Joaquín Montoriol-Pous ***,

Resumen

Se exponen los fundamentos del análisis canónico, combinación del análisis de la varianza y del factorial, así como sus posibles aplicaciones a los problemas geológicos.

Résumé

On expose les bases de l'analyse canonique, qui est la combinaison de l'analyse de la variance et l'analyse factoriel. On expose aussi les posibles applications de l'analyse canonique aux problèmes géologiques.

Introducción

Si el método de trabajo geológico se funda en la repetida observación de un fenómeno, para deducir las leyes que lo caracterizan, es evidente que está fundado en una base estadística. La reiterada medición de una o varias variables nos conducirá a la obtención de unas funciones en virtud de las cuales explicaremos un hecho.

Con estas consideraciones previas, establecemos un nexo de unión entre los modelos estadísticos y matemáticos y las ciencias experimentales, concretamente con la geología. Es decir: la geología debe apoyarse en modelos abstractos para explicar hechos concretos.

El análisis canónico se engloba dentro del multivariante como técnica estadística que combina el análisis factorial y de la covarianza. Si el primero permite explicar los factores que influyen en un conjunto de variables dentro de un mismo grupo o población y el segundo refleja la diferencia entre varios grupos o poblaciones, el análisis canónico detecta los factores que influyen en las diferencias de las variables a lo largo de tales grupos (1).

En su aspecto formal se plantea así:

Sean X_1, X_2, ..., X_m variables aleatorias (que miden unos determinados caracteres); de cada una de estas variables se conocen K muestras de valores tomados al azar e independientemente, correspondientes a K grupos (poblaciones, estratos, etc.), de tamaño n_1, n_2, ..., n_k respectivamente.

Sea $\sum_i = (\sigma_j K)$ la matriz de varianzas-covarianzas de las m variables dentro del grupo i-ésimo.

Si σ_{jj} es la varianza de la variable j, nos indica el grado de variabilidad de esta variable.

Y σ_{je} es la covarianza entre las variables j y e, que mide el grado de dependencia entre ambas.

Podemos representar las muestras en un espacio de dimensión m, como una nube de puntos alrededor del vector medias de las variables (fig. 1).

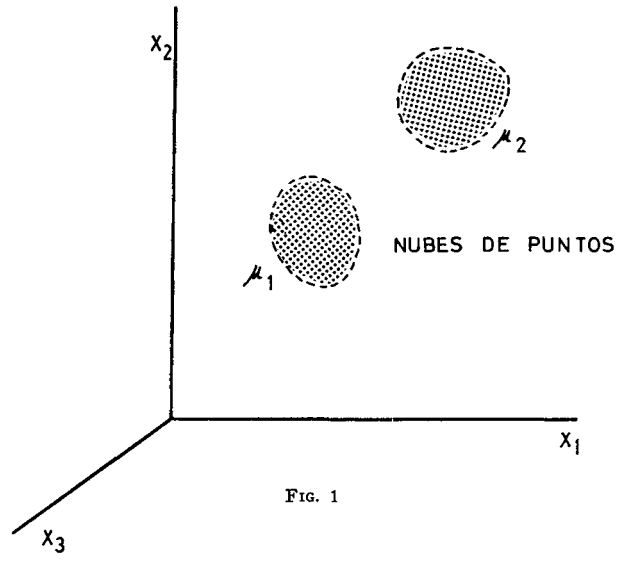

Fig. 1

Cada muestra se representa en un espacio de dimensión m como una nube de puntos, adoptando la forma de un elipsoide de centro el vector de medias de las variables.

Los ejes de estos elipsoides, que pueden ser calculados a través de un análisis factorial aplicado a

* Este trabajo forma parte de la Tesis Doctoral del segundo de los que suscriben y ha sido realizado en parte, con la Ayuda para el Fomento de la Investigación en la Universidad.

** Laboratorio de Cálculo. Universidad de Barcelona.

*** Departamento de Cristalografía y Mineralogía, Universidad de Barcelona. Sección de Mineralogía, Instituto "Jaime Almera", C. S. de I. C. Barcelona.

cada matriz, representan las direcciones de máxima variabilidad del conjunto de m variables dentro de cada grupo.

Pero nos interesa la variabilidad entre los grupos. Pues bien, se puede estudiar a partir de la matriz $\sum_E$ de varianzas-covarianzas entre grupos, hallando, a partir de ésta, los ejes de máxima diferenciación entre grupos.

El número de ejes llamados canónicos, no puede ser mayor a m o a $K-1$, si fuera $m > K-1$.

Indiquemos ahora:

X_{ijh} valor muestral del grupo j, variable i ($h = 1, 2, \ldots\ldots, n_j$).

$\bar{X}_i$ media común de la variable i.

$\bar{X}_{ij}$ media de la variable i en el grupo j.

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

Para el par de variables i, e, definamos a continuación

$$t_{ie} = \sum_{h=1}^{N} (X_{ijh} - \bar{X}_i)\,(X_{ejh} - \bar{X}_e)$$

suma total de los productos cruzados

$$W_{ie} = \sum_{j=1}^{k} \left(\sum_{h=1}^{nj} (X_{ijh} - \bar{X}_{ij})\,(X_{ejh} - \bar{X}_{ej}) \right)$$

suma de los productos cruzados dentro de los grupos

$$P_{ie} = \sum_{j=1}^{k} (\bar{X}_{ij} - \bar{X}_i)\,(\bar{X}_{ej} - \bar{X}_e)$$

suma de productos cruzados entre los grupos.

Es fácil demostrar que $t_{ie} = W_{ie} + P_{ie}$ o en notación matricial $T = W + P$.

Concluimos entonces, que la covariabilidad total, se descompone en dos partes:

a) la que existe entre grupos

b) la que existe dentro de los grupos.

De este modo queda justificado que el análisis debe concentrarse sobre la matriz (2)

$$\sum_E = \frac{1}{k-1} \cdot P$$

Condiciones previas para el análisis canónico

El análisis precisa la verificación de dos condiciones:

1. La homogeneidad entre las varianzas-covarianzas de las variables de los distintos grupos, es decir que se pueda admitir la igualdad

$$\sum_1 = \sum_2 = \ldots\ldots\ldots = \sum_m \text{ [1]}$$

2. Los vectores de media de los grupos, deben ser significativamente distintos.

Si la distribución conjunta de las variables es multinormal, existe una prueba para comprobar la primera condición.

La estimación conjunta de la matriz de varianza-covarianza (suponiendo que todas son iguales) es:

$$\sum = \frac{1}{N-k} \cdot W$$

calculando entonces los determinantes

$$|\sum_1|,\ |\sum_2|\ \ldots,\ |\sum_m| \text{ y } \sum$$

se demuestra que el estadístico

$$\chi^2 = -2\left[1 - \left\{ \sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{N-k} \right\} \frac{2m^2 + 3m - 1}{6(m+1)\,(k-1)} \right.$$
$$\left. L_n \left\{ \frac{\prod_{i=1}^{k} |\sum_i|^{\frac{n_i - 1}{e}}}{|\sum|^{\frac{N-k}{2}}} \right\} \right]$$

sigue aproximadamente una distribución ji — cuadrado con $(k-1) \cdot m \cdot (m+1)/2$ grados de libertad, si es cierta la hipótesis [1].

Si el valor χ^2 obtenido se desvía mucho, puede admitirse que hay diferencias significativas entre las $\sum_i$ (3) (4).

Suponiendo ahora comprobadas las igualdades [1], existe también un estadístico que permite verificar si las poblaciones son significativamente distintas, es decir, si los vectores de medias de los grupos son desiguales. Se basa en la generalización de la distancia (al cuadrado) entre dos poblaciones, o distancia de Mahalanobis. En el caso de K poblaciones, siendo $K \geqslant 2$, esta distancia al cuadrado es:

$$D^2 = \sum_{i=1}^{m} \sum_{e=1}^{m} d_{ij} \sum_{h=1}^{k} n_k\,(\bar{X}_{ih} - \bar{X}_i)\ \ (\bar{X}_{eh} - \bar{X}_e);$$
$$(d_{ij}) = \sum^{-1}$$

y se demuestra, en el caso de multinormalidad y de igualdad de vectores de medias, que sigue una distribución ji — cuadrado con $m \cdot (k-1)$ grados de libertad.

En consecuencia, si el valor D^2 resulta muy desviado puede admitirse la hipótesis de desigualdad entre los vectores de media, teniendo el análisis canónico sentido (5).

Obtención de los ejes canónicos

Si se desea que los ejes representen las direcciones de máxima variabilidad entre grupos, es evidente que se obtendrá un análisis de las componentes principales de la matriz $\sum_E$, es decir, hallando sus vectores y valores propios. Ahora bien, las posicio-

nes angulares de las variables quedan determinadas por la matriz

$$\Sigma = \frac{1}{N-k} W$$

luego en lugar de resolver la ecuación

$$|\Sigma_E - \lambda I| = O$$

debemos sustituir la matriz identidad I, por Σ, es decir, resolver

$$|\Sigma_E - \lambda \Sigma| = O$$

Utilizando las matrices

$$P = (k-1)\ \Sigma_E,\ W = (N-k)\ \Sigma$$

el problema equivale a resolver

$$|P - \phi W| = O$$

siendo

$$\phi = \frac{k-1}{N-k}\ \lambda$$

o lo que es lo mismo, a obtener los vectores y valores propios de $W - P$, siendo W y P matrices simétricas y W definida positiva.

Cada vector propio C_i deberá verificar la condición

$$C_i^1 \cdot W \cdot C_i = N - k$$

lo cual determina su módulo.

Las direcciones de los ejes canónicos vienen dadas por estos vectores (3).

Coordenadas canónicas

Si C es la matriz cuyos vectores fila son los vectores propios en número igual al menor de los números m y $(k-1)$, sea este nv, las nuevas coordenadas en los ejes canónicos se obtienen mediante la transformación lineal:

$$Y = C \cdot X$$

siendo Y de dimensión nv

X de dimensión m.

Aplicando esta transformación a los vectores de medias, se obtienen unos puntos en el espacio de dimensión nv, que pueden representarse con referencia al transformado del vector de medias comunes que se toman como origen.

En la práctica se consideran solamente los primeros ejes significativos, utilizándose a tal fin el Test de Bartlett, basado en el hecho que el estadístico

$$\chi_i^2 = \left\{(N-1) - (m+i)/2\right\} L_n \left\{\prod_{j=i+1}^{m} (1+\phi_j)\right\}$$

sigue una distribución ji — cuadrado con $(m-i)\cdot(h-i-1)$ grados de libertad si

$$\phi_i + 1 = \phi_{i+2} = \ldots\ldots = \phi_{nv} = O \quad (3)\ (4).$$

El test es el siguiente: si χ_i^2 no resulta desviado, puede admitirse que los ejes $i-1$, $i-2$, ..., nv, no son significativos. La representación es entonces suficiente en un espacio de i dimensiones.

Por último, una vez determinada la dimensión y representados los vectores de medias, cada grupo o población queda ubicado en una región confidencial con un 90 % de probabilidad, igual a una circunferencia de radio $1.645 / \sqrt{nj}$ (3).

La yuxtaposición de dos poblaciones se interpreta como una coincidencia significativa.

La alineación de dos poblaciones respecto al origen, se interpreta como influidas por un mismo factor y la separación angular, se interpreta como una medida de independencia entre ambas poblaciones (fig. 2).

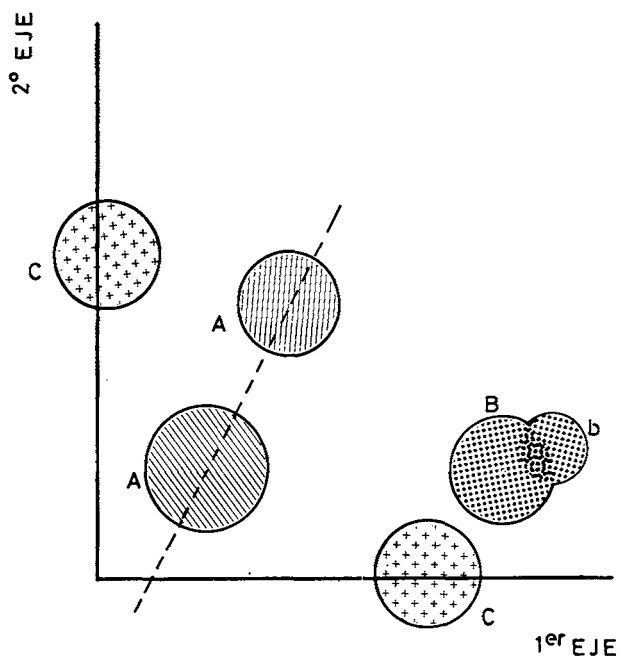

Fig. 2. — A: grupos influidos por un factor; B: grupos yuxtapuestos; C: grupos independientes.

Campos de aplicación geológica del análisis canónico

El análisis canónico presenta múltiples aplicaciones en geología, porque responde a dos preguntas fundamentales:

a) ¿Son iguales dos poblaciones?

b) ¿De qué forma se relacionan dos o más poblaciones con unos determinados factores en relación con unos ejes canónicos?

Aún cuando estas dos preguntas tienen una exposición simplista, abarcan un amplio espectro geológico, ya que las formaciones pueden ser identificadas con faunas fósiles, estructuras, minerales, etc., no teniendo más limitaciones que la selección de variables mesurables y homogéneas.

Es obvio que no vamos a detallar a qué poblaciones puede ser aplicado dicho análisis, ya que cada una de ellas posee variables específicas y características diferenciales. Sin embargo, vamos a insistir en el hecho que el análisis canónico puede aportar a la geología:

1) la comprobación experimental de hipótesis acerca de la relación entre entidades;
2) la detección de relaciones desconocidas;
3) suministrar modelos matemáticos para la explicación de relaciones difícilmente demostrables por simple extrapolación.

BIBLIOGRAFÍA

1. CUADRAS-AVELLANA, C.: Métodos de análisis Factorial. *Pub. Laboratorio de Cálculo,* Universidad de Barcelona 1970,
2. COOLEY, W. W. and LOHNES, P. R.: *Multivariate Procedures for the Behavioral Sciences.* John Wiley & Sons, Inc., London 1962.
3. SEAL, H. L.: *Multivariate Statistical Analysis for Biologist.* Nethuen and Co. Ltd. 1964.
4. ANDERSON, T. W.: *Introduction to multivariate Statistical Analysis.* John Wiley & Sons, Inc., London 1958.
5. RAO, R.: *Advanced Statistical Methods in Biometric Research.* John Wiley & Sons., Inc., London 1952.