TRABAJOS

Revista Investigación Educativa - Vol. 4 - n.8 - 1986 (P. 5-19)

UN ESTUDIO COMPARATIVO DEL ANÁLISIS DE CORRELACIÓN CANÓNICA

por
Maria Dolores Peris i Pascual

1. Breve panorámica histórica del Análisis Multivariado

El análisis multivariado, como generalización de las técnicas uni y bivariadas se ha desarrollado desde los comienzos de la investigación experimental siguiendo la siguiente pauta general: Pruebas empíricas; fundamentación matemática y pruebas de significación; modificación para casos especiales y comparación entre diferentes métodos.

Los modelos lineales y los análisis de mínimos cuadrados (1812) comienzan en la investigación diferencial (errores de medición en astronomía) con GAUSS y LE GENDRE, apareciendo la distribución normal (GAUS 1809) como un constructo para describir tales errores/diferencias de medición en las variables contínuas. En las variables discretas, el análisis de probabilidades con las definiciones de las distribuciones binomial y de Poisson (1837) llega a puntos comunes. Las series de FOURIER (1812) y las funciones racionales de LAGRANGE (1772) fundamentan la evaluación de la variabilidad temporal.

Antes de finalizar el siglo pasado, se habían descrito las principales distribuciones (HELMERT, 1875) en el campo de la antropología diferencial. Con GALTON y sus investigaciones diferenciales aparecen las nociones de correlación, regresión y análisis de varianza (1880) y con PEARSON los estadísticos elementales (así la Moda en 1894 o el coeficiente de variación en 1895). El desarrollo que a estos fundamentos se impartió a principios de nuestros siglo, es en gran parte debido a los estudios en genética de poblaciones, estableciéndose desde sus comienzos, una mútua interacción entre las ciencias humanas «diferenciales» y la ciencia experimental.

Haciendo una amplia generalización de la sucesiva aparición de las técnicas experimentales, podrían establecerse las siguientes etapas:

– Décadas de 1900, dedicada a los «coeficientes de asociación/dependencia»: χ^2 y r_t de PEARSON (1900); Q de YULE (1900); C de PEARSON (1904); ρ de SPEARMAN (1904); t de GOSSETT (Student de pseudóni-

mo, 1907); Asociación «latente» en ejes principales de PEARSON (1901) y del factor principal de SPEARMAN (1905); asociación «secuencial» en cadenas de MARKOV (1907); Coeficiente de fiabilidad de Spearman (1910); σ^2_r para muestras pequeñas de SOPER (1913) y F de FISHER (1915).

– Década de 1920 a 1929 y «parámetros de predicción»: Path Análisis de WRIGHT (1921); WISHART plantea MANOVA para p-variables (1920) y define la distribución multinormal (1928) que generaliza la de χ^2 en muestras p-variadas derivadas de las distribuciones normales, permitiendo los análisis multivariados. PEARSON (1921) genera la idea de A. Discriminante con su coeficiente de semejanza racial para la medida de la distancia entre dos muestras, que será reemplazado por la D^2 de MAHALANOBIS (1927), cuya primera aplicación en el estudio de razas indias se realiza en 1925. La razón de verosimilitud de NEYMAN y PEARSON, L. (1928), viene a considerarse como el más directo antecedente de la función Discriminante. FISHER define la suficiencia de un estadístico (1921), prueba rigurosamente la distribución t (1923), así como la distribución de la correlación parcial (1924) y la correlación intraclase en el análisis de varianza (1925). HALL, siguiendo a Fisher halla el valor aproximado de R^2 (1927).

– Década de 1930 a 1939, especialmente productiva y en la que se definen todas las «técnicas multivariadas» tradicionales. En 1931 WISHART obtiene el valor exacto de R^2 y HOTTELLING aplica T^2, como generalización de t^2 para p-variables. En 1932 WILKS, desde el principio de razón de verosimilitud genera el criterio de Lambda, análogo a la F de Fisher para probar la razón de varianza multivariada. HOTELLING presenta el análisis de Componentes Principales (1933) y HIRSGHFELD la regresión lineal simultánea (1935) como antecedente del análisis factorial de correspondencias. En 1936 aparece el Análisis de Correlación Canónica con HOTELLING, para resolver problemas diferenciales educativos y la Función Discriminante con FISHER, para maximizar las diferencias entre las variables de dos grupos y que WELCH (1939) valida al demostrar que es una aplicación del principio de razón de verosimilitud BARTLETT (1939) presenta su test de homogeneidad de varianzas y LAWLEY la distribución T^2 para varias muestras. El análisis de confluencia con FRISH intenta superar las dificultades de la multicolinearidad en el análisis de regresión. BOSE y ROY desarrollan las distribuciones T^2 y D^2 para p-muestras (1939) y tres autores a la vez, ROY, HSY y FISHER (1939), la distribución de las raíces latentes de las ecuaciones determinantes en las matrices de covarianza de distribuciones normales multivariadas. También los análisis clusters aparecen a finales de esta década con ZUBIN (1939) y TYRON (1939) como alternativa al análisis de componentes principales.

Las técnicas no paramétricas y de bondad de ajuste, inician su desarrollo, ampliando las posibilidades de análisis cuantitativas: KOLMOGOROV y SMIRNOV (1933); corrección de χ^2 y prueba de Fisher para frecuencias pequeñas con YATES (1934), uso de Q de Yule en tablas de $2 \times 2 \times 2$ con BARTLETT (1935); dependencia de rangos con FRIEDMAN (1937); Tau de KENDALL

(1938); PSI^2 de KENDALL y SMITH (1938) y D_n de WALD y WOLFWITZ (1939), entre otras.

– Década de 1940 a 1949, con variaciones en las técnicas para permitir adaptarlas a casos especiales: Método de máxima verosimilitud en el Análisis Factorial de LAWLEY (1940); test de BARTLETT para la significación de la correlación canónica (1941); Coeficiente de correlación parcial por rangos de KENDALL (1942); Regresión múltiple (FISHER, 1946); ANOVA diadíca de TUKEY (1949) superando las limitaciones de los diferentes números de celdas y de sujetos por celda; pruebas de WALD (1944), de WILCOXON (1945) y de NEYMAN (1949). STEVENS (1946) define las cuatro escalas de medida básica, que despertarán el interés por los métodos de análisis de las de rango inferior.

– Década de 1950 a 1959. «Técnicas de agrupamiento», destacando las relaciones entre variables y las representaciones pictóricas. Las técnicas «taxonómicas» de «cluster» y «scaling» múltiple generan numerosas facetas. El primero en aplicar las segundas es considerado TORGERSON (1952) iniciándose en 1958 el escalamiento no-métrico (COOMBS y TORGERSON); FLOREK (1951) presenta el método de cluster dendrítico, SNEATH (1957) desarrolla el dendograma desde el algoritmo acumulativo, SOKAL y MICHENER (1958) el cluster aglomerativo politético y WILLIAMS (1959) el cluster por división monotética llamado análisis de asociación, entre otras variantes. ANDERSON (1954) desarrolla los «glyphs»; KRUSKAL (1956) y PRIM (1957) los algoritmos para la construcción de redes con grafos.

Versiones del análisis factorial que se desarrollan son la de Estructura latente de LAZARSFELD (1950), el método de AHMAVAARA (1954), el análisis de imágenes de GUTTMAN (1953) o el modelo modulador sobre los trabajos de SAUNDERS (1956). ROY y GNANADESIKEN (1959) aplican el cálculo matricial a ANOVA para diseños incompletos y números de subclases desproporcionados. A comienzos de la década se desarrollan formas no lineales del Análisis Discriminante (LUBIN, 1950) y se concluyen los intentos de ampliación a más de dos grupos (BRYAN, 1951). RAO (1952) al desarrollar las proyecciones y gráficas de D^2 en la comparación de perfiles y el modelo matemático del A. Discriminante, demuestra que son dos métodos del cálculo diferente del mismo análisis. Sigue la aplicación de nuevas pruebas noparamétricas como la de DUNCAN (1952), MOSES (1952), Q de COCHRAN (1950) o análisis RIDIT para datos subjetivamente categorizados (BROSS, 1958).

– Década de 1960 a 1969, decisiva por la incorporación de los modelos dinámicos de evaluación del cambio, básicos para la investigación educativa y por la divulgación de la informática que produce la «revolución» metodológica y teórica anunciada por YATES (1966). Con ella se descentraliza el desarrollo de la ciencia experimental de la escuela anglo-sajona surgiendo centros con perspectivas diferentes. Así la escuela francesa, como respuesta, entre otros, a los problemas planteados por la «Pedagogía Matemática», (LERMAN, 1966) yproductos como el Análisis factorial de Correspondencias (Benzecri, 1963), o

la escuela sueca con los modelos confirmatorios de Análisis Factorial (JORESKOG, 1967), y la incorporación de relaciones estructurales entre variables latentes como presenta el modelo LISREL (JORESKOG y M. van THILLO, 1972) o la danesa con el desarrollo de la «Teoría fuerte de la puntuación verdadera» (RASCH 1960; ANDERSEN 1973).

La variabilidad temporal se investiga con mayor detalle que las simples curvas de desarrollo (CONRAD 1931, MILES 1931, LORGE 1936) con propuestas como las de SCHAIE (1965) y BALTES (1968) para el control de validez y el desarrollo de series temporales multivariadas: Poliespectrum de BRILLINGER (1965), Biespectrum de GODFREY (1965), modelos uni y bidereccionales con tiempo incluido de COLEMAN (1964), aplicación de la Función Discriminante y D^2 a los modelos de desarrollo lineal en la evaluación del cambio (BURNABY, 1966) y de las cadenas de MARKOV a las investigaciones sobre diálogos en lenguaje (JAFFE, FELDSTEIN y CASSOTTA, 1967). La aplicación de la Correlación Canónica a estructuras factoriales, permite comprobar los cambios producidos en las mismas por el tiempo o por la intervención educativa siempre acaecida entre un tiempo anterior y otro posterior.

Los modelos causales anteriores se extienden y se generan, por diversas publicaciones, los modelos logit y log-lineal, como culminación del análisis de tablas de contingencia, que tiene en esta década una amplia profusión. Esta incidencia en los análisis de variables nominales (cualitativas) frente a las de intervalo, se interpreta como un desplazamiento de la investigación experimental desde la psicometría hacia la sociología, o de otra manera, desde las «dimensiones nomotéticas» (medidas) hacia las «relaciones observables», (frecuencias) no ajena a los cambios que en otros niveles se han definido como un paso de lo biológico a lo social. La obra de SOKAL y SNEATH «The principles of numerical taxonomy» (1963) se suele citar como un hito de este cambio de orientación en la investigación multivariada. Estas nuevas metodologías proporcionan a la educación, con abundantes variables de tipo cualitativo-categorial, nuevas perspectivas de investigación.

Los análisis de escalamiento múltiple siguen ampliándose con aportaciones como el análisis de proximidades de SHEPARD (1962) cuyo algoritmo desarrolla KRUSKAL (1964), la de COOMBS (1964), las Coordenadas Principales para las parejas de distancias de GOWER (1966), el «K-mean» de McQUEEN (1967) o el algoritmo aglomerativo para los clusters de LANCE y WILLIAMS (1967), aplicando este tipo de técnicas CARROLL (1969) a las taxonomías en diferencias individuales.

Por su parte, las técnicas tradicionales siguen completándose: HORST (1961) desarrolla la correlación Canónica Múltiple y STEWART y LOVE (1968) el Índice de Redundancia para la misma; McDONALD (1962) propone el análisis Factorial no-lineal, KAISER y DICKMAN (1965) el análisis Factorial Alpha, KENDALL (1966) en análisis Discriminante no-paramétrico y COOLEY y LOHNES (1971) la Correlación Parcial Múltiple a partir de la Correlación Canónica; FISHER (1966) desarrolla el Diseño Factorial y GABRIEL (1968) las inferencias simultáneas de p-variables en ANOVA.

– Década de 1970 a 1979: Desarrollo de los «modelos matemáticos», desplazando la investigación desde la prueba de hipótesis hacia la prueba de modelos de estructura más acordes con las teorías «interaccionistas» de las ciencias sociales. La diferenciación entre análisis «exploratorio» y «confirmatorio» viene a coincidir con los dos momentos de relación entre los datos empíricos y la teoría en el método hipotético-deductivo, integrándose generalmente el segundo en el primero. La ciencia experimental, con esta ampliación de su alcance a sectores más complejos y globales adquiere, (o refuerza), su status de «necesaria» para la validación de toda teoría, ley, modelo o constructo en las ciencias empíricas, como es el caso de la Pedagogía.

Clasificaciones de modelos matemáticos en procesos sociales los proponen BARTOLOMEU (1973), BUNGE (1975) y BUGUEDA (1975) entre otros. OVERTON y REESE (1973) presentan los metamodelos del desarrollo; LEIK y MEEKER (1975) los modelos matemáticos de grafos para las estructuras organizacionales. Las cadenas de Markov se utilizan ampliamente en modelos dinámicos de relaciones de cambio (WAISERMAN, 1979) y en representaciones de la estructura secuencial de la conducta social (RAUSH, 1972). BOX y JENKINS (1970) permite con su modelo el control y la prospectiva en las series temporales. GOODMAN (1973) incluye en los modelos Path de estudios longitudinales el modelo log-lineal, sobre el que se publican numeroso textos en esta década, siendo el más representativo el de BISHOP, FIENBERG y HOLLAND (1975) con los de GOODMAN. El proceso causal adquiere múltiples derivaciones, tales como el «d-systems» de DAVIS (1975) y el ya citado modelo LISREL. La obra de BARNET (1981), resultado de la 2.ª Conferencia sobre Datos Multivariados de Sheffield (1980), con su centenar de diferentes representaciones multivariadas, es muestra del desarrollo y posibilidades que ha alcanzado la «modelización de interacciones». Los diferentes enfoques de aproximación a la estructura real (distancias y proximidades) ha producido una amplia gama de índices de relación, desde la primera dicotomía correlación vs. distancia».

Diversos autores apuntan el futuro de las técnicas multivariadas hacia el análisis exploratorio a través de métodos gráficos, dando prioridad a teorías como la de los grafos frente a los precedentes cálculos matriciales, y su traducción en modelos estocásticos, y hacia la optimización de la replicación y validación de estructuras en los análisis confirmatorios incluyendo técnicas de Investigación Operativa. En el A. Discriminante es fácil observar estas fases sucesivas (HAND, 1981): Intuitiva de FISHER, probabilística de WELCH, RAO et al. y basada en los principios de la Teoría de la Decisión de WILDIAN.

2. Comparaciones entre las técnicas multivariadas.

Diferentes autores han relacionado los objetivos perseguidos por las técnicas multivariadas (BALL, 1971; GNANADESIKAN, 1977; JOHNSON y WICHERN, 1982) siendo los más comunes los de reducción de la dimensionali-

dad, simplificción de estructuras, agrupamiento y clasificación, exploración de dependencias, predicción, construcción y prueba de hipótesis, evaluación de modelos y organización expositiva.

La clasificación de estas técnicas en función de dichos objetivos es inadecuada porque muchos de ellos, al no ser mútuamente exclusivos, son cubiertos a la vez por varias técnicas. El problema que se plantea es cuál resulta la idónea para cada problema concreto.

Por otra parte, los análisis multivariados han sido comparados con un microscopio «que da una visión parcial y deformada de la estructura, necesitando múltiples tratamientos para llegar a una interpretación válida» (LEFEBVRE, 1976). De donde se sigue la necesidad de multiplicar las aplicaciones. En otros casos se han criticado estudios parciales, como la aplicación de predicción usando una sola variable criterio, por ser una sobresimplificación del problema (DUNNETTE, 1963; THORNDIKE, 1978).

Desde otra perspectiva, se critica el problema de que cada criterio de análisis, «predispone» a encontrar tipos particulares de resultados y puede distorsionar de diferente modo los datos, que no se escapan a la dependencia del modelo explicativo subyacente.

Especialmente desde mediados de los 60, con la accesibilidad a los medios informáticos, aparece una nueva línea de investigación como respuesta a estas cuestiones, que es la comparación entre las técnicas multivariadas. Todas ellas presentan aspectos comunes tales como el de resaltar las relaciones interdependientes y la importancia *relativa* de las características incluidas en la información. «Un aspecto de generalización de todas los modelos es que proporcionan un predictor lineal, basado en una combinación lineal de las variables» (McCULLAGH y NELDER, 1983) y toman la máxima verosimilitud como algoritmo común para la estimación de parámetros.

La comparación entre los numerosos métodos taxonómicos, especialmente clusters, ha sido la más desarrollada. ROHLF y SOKAL (1965) analizan las ventajas de las correlaciones y de las distancias. WILLIAMS, LAMBERT y LANCE (1966), recomiendan tras su estudio comparativo el agrupamiento controide más que las correlaciones y las distancias para los datos estadísticos. GOWER (1967) compara tres métodos de cluster; RUBIN (1967) describe como su criterio de agrupamiento está relacionado con el Análisis Multivariado de Varianza. Tentativas de aproximación entre el Análisis Factorial y las técnicas de clasificación han sido frecuentes (HOWARD, 1969; BENZECRI, 1971; GONDRAND 1975; JAMBU, 1976; LERMAN, 1981). Las investigaciones de semejanzas entre las clasificaciones y otras metodologías más formales se propone como tarea futura necesaria, con el fin de facilitar información acerca de sus propiedades.

Métodos formales de comparación de clusters han sido propuestos por ADAMS (1972) y GORDON (1980). ROHLF (1970) compara los esquemas jerárquicos de clusters, WILLIAMS (1971) diversos procedimientos también de clusters y en 1972 ROHLF compara empíricamente 3 técnicas de ordenamiento

en taxonomía numérica: Escalamiento multivariado nomotético de Kruskal, Análisis de Componentes Principales y Análisis de Coordenadas Principales de Gower.

En 1969 BISHOP compara modelos de análisis de tablas de contingencia. GOODMAN, en 1977, compara el análisis de variables cualitativas según el análisis de Estructura Latente y el modelo Log-Lineal, pretendiendo terminar con la separación entre «correlacionistas» (economistas, biometras y psicólogos) y «crostabuladores» (sociólogos y pedagogos).

KRZANOWSKI (1977) indica que hay que estudiar más el análisis conjunto de variables cualitativas y cuantitativas, insistiendo JOHNSON y WICHERN (1982) en la escasez de literatura sobre este punto.

KUIPER y FISHER (1975) examinan 6 algoritmos de análisis bi y multivariados de muestras normales, llegando a conclusiones como la de que los métodos de cuadrados de sumas son buenos para muestras de igual tamaño, pero no cuando el n es diferente.

Muchos de estos estudios comparativos se apoyan en técnicas de simulación. La simulación de muestras grandes desde n pequeños a través de los números aleatorios, es especialmente útil cuando es muy costosa su obtención. Sobre una muestra así simulada, THONGUTAI (1980) compara la planificación de la selectividad y orientación universitaria aplicando el A. Discriminante, la Programación Lineal y un índice obtenido por la aplicación de Reyes a dos índices de probabilidad obtenidos previamente por regresión múltiple. PRESS y WILSON (1978) al compara la función discriminante con la regresión logística y definir sus diferencias, halla superioridad en la primera para fines clasificatorios. EVANS (1978) define el A. Discriminante como «el mejor método de comparación de grupos». BARNET (1981) propone que el A. de Correspondencias se llame A. Discriminante doble, puesto que investiga la dependencia de dos divisiones de las mismas unidades de muestreo.

De entre todas las técnicas, es la Correlación Canónica la que va adquiriendo desde estas comparaciones el status Central por ser la de mayor generalización, presentándose las demás como casos límites de ella. Para LEFEBVRE (1976) el A. Canónico, al relacionar dos grupos de variables, establece un puente entre los bi y los multivariados.

En 1957 KENDALL se lamenta de que no se use más en las investigaciones. Los programas de COOLEY (1962) y LOHNES (1971) facilitaron su divulgación, aunque en general se suele atribuir a sus dificultades de interpretación la falta de mayores aplicaciones. En 1963 BARNETT y LEWIS recopilan la aplicación de este análisis en educación. Recuérdese que Hotelling la presenta para resolver un problema de Pedagogía Diferencial.

Las relaciones más patentes con los demás modelos lineales múltiples son las siguientes:

- Es, según BARTLETT, el caso más extremo del Análisis Factorial externo (entre grupos de variables diferentes).
- Tanto la correlación bivariada (dos grupos con un solo componente cada

uno), como la Múltiple, (uno de los grupos con un solo componente) han sido definidos como casos especiales de la Correlación Canónica.

- La correlación parcial se puede entender como la correlación de un grupo parcializando el otro, en términos de correlación Canónica. Así, cuando cada grupo tiene dos variables, tenemos una correlación parcial de segundo orden. COOLEY y LOHNES (1971) han descrito la correlación parcial múltiple a partir de la Canónica.
- Es una generalización de la regresión de una variable sobre otra, con un vector respecto a otro (KSHIRSAGAR).
- La diferencia con D^2 estriba en que las variables canónicas son soluciones ortogonales (hecho que permite su adición) mientras que D^2 proporciona soluciones oblículas. No obstante, existen métodos de transformación de las segundas en las primeras.
- En 1968 GLAHN demuestra como el A. Discriminante es un caso especial de la Correlación Canónica: aquel en que uno de los grupos es una matriz lógica que indica las categorías a discriminar. Las correlaciones canónicas medirán la capacidad de las variables para discriminar linealmente entre estas categorías.
- Para KLECKA (1980) la correlación canónica proporciona un valor de la función Discriminante y de ella se botiene cúales serán las funciones útiles.
- Como en el LISREL se halla una estructura latente, combinación lineal de dos conjuntos de variables observadas. Pero las restricciones del Análisis canónico son mayores, puesto que las variables latentes son ortogonales (equivale a la restricción de $\gamma/\beta = 0$ en el LISREL) y cada variable latente maximiza en orden sucesivo las relaciones lineales entre los dos conjuntos de variables. Esta maximización de relación está ausente en el LISREL, así como la secuencialización de varianza explicada entre las variables latentes. Por ello la correlación Canónica puede considerarse un caso límite de optimización de tales criterios en el LISREL. VAN DE GEER (1971) en su enfoue del análisis multivariado como paht analisis, presenta así la correlación canónica con múltiples facetas estructurales, una de ellas correspondiendo al «análisis factorial doble». Por otro lado, las ventajas que la correlación múltiple permite al aumentar el número de grupos de variables a más de dos le confiere una indudable ventaja como análisis estructural latente. BAGOZZI, FORNELL y LARCKER (1981) desarrollan este punto.

Las aplicaciones más comunes son las de tipo diferencial, con las variables diferenciales en uno de los grupos y los rendimientos en el otro y en los diseños antes-después (pre-post), para relacionar los cambios producidos por la intervención en la segunda evaluación (segundo grupo de variables) respecto a la primera (primer grupo), o bien una serie de intervenciones educativas (variables de un grupo) y sus efectos múltiples (segundo grupo). A su vez se puede aplicar a matrices de tipo Q relacio-

nando grupos distintos de sujetos. También compara estructuras factoriales, o bien dichas estructuras con sus variables originales.

3. Conceptos en el Análisis de Correlación Canónica

El análisis de Correlación Canónica es la maximización de las correlaciones entre dos grupos de variables (y/z) cuando se tiene más de un criterio y más de un predictor, hallando una relación lineal de ambas. El punto de partida es la división de la matriz de datos por una línea en p × q columnas, produciendo una partición de la matriz de correlaciones R tal que

$$R = \left| \begin{array}{c|c} R_{11} & R_{12} \\ \hline R_{21} & R_{22} \end{array} \right|$$

donde
R_{11} es la matriz de intercorrelaciones entre los p-predictores.
R_{22} idem entre los q-criterios.
$R_{21} = R'_{12}$ es la intercorrelación entre criterios y predictores.

El problema consiste en hallar las q raíces, λ_i (i – 1,2,...q) siendo q p (es indiferente que el grupo menor sea el primero o el segundo), que son el cuadrado de las correlaciones canónicas ($\lambda_i = \rho_i^2$) en la ecuación determinante $\left| R_{12} R_{22}^{-1} R_{21} - \lambda R_{11} \right| = 0$ así como en $\left| R_{21} R_{21}^{-1} R_{12} - \lambda R_{22} \right| = 0$

Para cada raíz canónica λ_i el correspondiente vector canónico de pesos para las variables predictoras, que tiene como elementos los coeficientes de la combinación lineal de las variables x, se determina por la ecuación

$$a_i = (R_{11}^{-1} R_{12} b_i) / \sqrt{\lambda_i}$$

Los elementos en los vectores a_i y el correspondiente b_i para la correlación canónica ρ_i se puede usar como directo o derivarlo en puntuación de desviación, dividiendo cada elemento por la de su variable correspondiente.

De este modo se maximiza la correlación entre el vector a_i, combinación lineal de las p-variables predictoras en z

$$a_i = \sum_{i=k}^{P} a_i x_i \; (i = 1,2,...p)$$

Una descripción detallada puede encontrarse en los textos de análisis multivariado como los de THORNDIKE (págs. 175 a 202), KSHIRSAGAR (págs. 247 a 288), LEFEBVRE (págs. 122 a 136), CATTELL (págs. 403 a 416), LEBART et als. (págs. 91 a 96) a HOPE (págs. 151 a 177 y 219 a 240) y en monografías como las de «SAGE», n.º 6 de LEVINE y n.º 47 de THOMPSON.

y el b_i, combinación lineal de las q-variables criterio en y

$$b_i = \sum_{j=1}^{q} b_j y_j \ (j = 1,2,...q).$$

Hay tantas soluciones posibles como el menor p o q, siendo las sucesivas correlaciones tales que $\rho_1 \rangle \rho_2 \rangle ... \rangle \rho_q$

El tamaño de la raíz latente está relacionado con el poder discriminante de esta función. Para compararlos, se suman los valores de todos y se divide por cada uno, convirtiéndolos en proporciones.

La mejor interpretación de los resulados se hace a partir de las correlaciones de las variables originales con las variables canónicas, tomando tan sólo las que resulten significativas tras la aplicación del test de Bartlett.

El test de Bartlett (1985) prueba la significación de la correlación entre los dos grupos. Halla $\chi^2 = [N - 0.5 (p + q + 1) \log_e \Lambda$
con $v = (p - m) (q - m)$ siendo

$$\Lambda = \prod_{i(m+1)}^{q} (1 - \lambda_i)$$

Cuando χ^2 es significativo, ambos grupos son dependientes. Las raíces latentes significativas serán aquellas cuyos χ^2 lo sean.

STEWART y LOVE (1968) aplican el Índice de Redundancia (R_d) como el porcentaje de varianza del grupo 1 referido al grupo 2. No es simétrico, como en el caso del coeficiente de determinación. Los índices de redundancia se dan para todo el grupo (total) y para cada componente del grupo.

IVERSEN y NORPOTH (1976) plantean una interpretación alternativa del coeficiente de correlación canónica desde el Análisis de Varianza, bajo el nombre de «eta» y «razón de correlación». Los grupos se entienden como variables independientes y la función como la dependiente. Eta mide el grado de diferencia entre las medias de grupos sobre la función. El cuadrado de eta (el cuadrado de la correlación canónica) es la proporción de variación en la función discriminante explicada por los grupos. Es así que la correlación canónica proporciona un valor de la función discriminante.

4. Comparación empírica del Análisis de Correlación Canónica con el A. Discriminante y con el Path Análisis.

Presentamos dos ejemplos de investigación educativa con la utilización de un par de técnicas multivariadas cada uno, con el doble objetivo de evaluar sus posibilidades y de determinar el peso relativo de la educación en el rendimiento escolar, junto con otras variables de las que, tradicionalmente, se hace depender

éste. Con ello pretendemos llegar al resultado, ya obtenido en otros trabajos precedentes, de que el «producto educativo», dependerá, en primer término, del «trabajo educativo», siendo muy inferior el peso de las variables no dependientes del mismo y que, comunmente, son alegadas como causas directas de los bajos rendimientos. Estas variables actúan cuando no alcanzan ciertos umbrales mínimos, pero su incidencia no es en absoluto decisiva cuando sus valores oscilan dentro de los umbrales de «normalidad» (especialmente este hecho se constata en las aptitudes intelectuales del primer ejemplo).

Insistimos en la «relatividad» de los resultados multivariados. Los valores obtenidos están limitados por el muestreo de sujetos y de variables, junto con otros aspectos como el tipo de escalas de medida. El efecto que produce el incluir una variable muy homogeneamente (hetereogeneamente) distribuida, es el aumento (disminución) del peso, en el conjunto de la «variabilidad» analizada, de otras variables con mayor (menor) dispersión. Nuestras muestras son representativas del sector general de la población escolar (Centros Subvencionados), con distribuciones dentro de los límites «normales» en todas las variables analizadas, y en ello fundamentamos la validez externa de los resultados, frente a resultados distintos obtenidos en muestras con abundantes valores extremos en algunas de sus variables.

En el primer ejemplo, se analizan un conjunto de los tests psicológicos más usuales aplicados al principio y al final del I ciclo, sobre la misma muestra, en un diseño longitudinal: En base al tiempo transcurrido se aplica el Path Análisis analizando como variables endógenas las calificaciones finales y como exógenas las variables más correlacionadas con ellas de la evaluación inicial (dos años antes). A su vez se plantean tres agrupamientos distintos de las 25 variables y los tres análisis canónicos correspondientes.

En todos los casos se obtiene una mayor asociación de los resultados finales con las evaluaciones escolares que con las pruebas cognitivas y sensoriomotoras. Puede concluirse que estos tests pierden su utilidad para la predicción del rendimiento en el caso de los sujetos cuyas capacidades caen dentro de los umbrales normales. En este caso, la mejor predicción, de entre los datos analizados, es el dominio en las actividades escolares básicas (lectura, escritura, cálculo). O de otro modo, superado un determinado nivel, las variaciones en las aptitudes psicológicas, apenas influyen en los rendimientos escolares.

En la aplicación de la correlación canónica, se obtiene una visión de la estructura completa, mejor interpretable cuando más simple sea, tal como ocurre con el A. Factorial. Cuando la «multicolinearidad» es elevada, como es el caso de este ejemplo, para conocer de forma independiente el efecto sobre cada variable, es preferible el Path Analisis. Se puede argumentar que esta independiencia no es reflejo exacto de la realidad, porque la interconexión, tal como muestra la correlación canónica es muy elevada. Pero siempre se puede contraargumentar que, habida cuenta de la interconexión de otras variables no incluidas en el análisis, que siempre es cuestionable, cualquier estudio, como parcialización de la población de variables estudiada, es una sectorización que debe integrarse y coordinarse con otros estudios del mismo campo.

De las tres divisiones entre las variables analizadas, obtenemos los tres resultados siguientes:

a) Cuando en un grupo sólo aparecen los resultados finales, estos se asocian en la primera correlación, con las demás evaluaciones referidas a contenidos escolares en el segundo grupo.

b) Cuando se sitúan todas las evaluaciones del rendimiento escolar (tanto finales como parciales) del 2.º curso en uno de los grupos, la multicolinearidad vuelve a asociarlas en la primera variable canónica, con las evaluaciones escolares del primer curso entre las variables del segundo grupo asociadas con esta primera variable canónica.

c) Dividiendo todas las evaluaciones en función del momento de su obtención (antes o después del I Ciclo), se observa que, manteniéndose de forma equivalente la primera variable canónica, la heterogenidad de aspectos en los dos grupos a la vez, permite la aparición de una segunda variable canónica correlacionada con los tests de Inteligencia (Raven, Rey y ¡cálculo!). Estos son precisamente los tres factores que se obtienen con la rotación Varimax en el A. Factorial (GASCO). La ortogonalidad de las variables canónicas permite llegar al mismo resultado, siempre que, como se ha posido ver, variables de cada factor esten presentes en uno y otro grupo del análisis.

Frente al A. Factorial, el A. Canónico, al separar las estructuras de uno y otro tiempo, nos confirma que éstas no han cambiado, no pudiéndose justificar en base a estos resultados, ningún modelo de cambio de tipo divergente en el sector de variables y edades analizado. Obsérvese a la vez, como la significación de las variables proporciona una información equivalente a la comunalidad en el A. Factorial.

En el segundo ejemplo se compara el rendimiento en Matemáticas entre 8.º curso de E.G.B. y 1.º de F.P. a partir de la resolución de 5 problemas, aplicando los análisis Discriminantes y de Correlación Canónica. En este caso, hemos obtenido una correlación de 1.0 entre la primera variable canónica y una de las originales, proporcionando una estructura muy simple. (PERIS, 1985).

Los problemas asociados en el segundo grupo con dicha variable del primer grupo (Nivel de Estudios), obtienen un 91,3% de aciertos, al clasificar a los alumnos en dichos niveles con el A. Discriminante. Las variables que no obtienen correlaciones canónicas significativas, tampoco serán criterios útiles de discriminación. Y de nuevo la variable educativa es la única que produce diferencias en los rendimientos.

Hay que destacar que el sexo no tiene ningún poder diferenciador en la capacidad de resolver problemas matemáticos. Por ser éste uno de los aspectos en que más diferencia se encontraba con alumnos que no habían recibido coeducación, se puede afirmar que estas diferencias se debían a los efectos diferenciadores de la educación recibida por cada sexo, y no a aptitudes diferenciales entre los mismos. Como en otros estudios (PERIS, 1984) al controlar la variabilidad educativa desaparecen las diferencias en otras variables de tipo

sociopsicológico que covarían con la educación, invirtiéndose la relación y apareciendo en última instancia, como variable moduladora del rendimiento educativo, las diferencias debidas a sus propios procesos.

Comparando la estructura obtenida entre las variables canónicas en éste y en el anterior ejemplo, podemos llegar a la conclusión de que este análisis es especialmente idóneo cuando el grupo q está incorrelacionado, es decir, las variables son mutuamente independientes. Tratándose de soluciones ortogonales, de no existir independencia entre las variables, difícilmente podrá aparecer más de una variable canónica significativa, sin necesidad de referirse a componentes subyacentes a los datos examinados, con la consiguiente dificultad de interpretación tantas veces aludida (LEFEBRE, 1976). Cuando existen grupos de variables mutuamente indepndientes, las superioridad de este análisis frente a otros es evidente, siendo entonces muy sencilla su interpretación. Hay que resaltar como el A. Canónico define la significación de cada variable como posible criterio discriminante, pudiendo utilizarse como análisis exploratorio previo cuando se cuednta con un gran número de posibles criterios.

Sobre este punto, en un trabajo previo (PERIS, 1985b) se obtuvieron idénticos resultados con un solo A. de Correlación canónica y con Cinco análisis Discriminantes (uno para cada variable de las incluidas en el grupo q). En aquel caso, y con una clara ventaja de la tecnología educativa sobre otras variables diferenciales (100% de aciertos en la predicción del rendimiento en una prueba objetiva), cada variable canónica se asociaba a una variable diferencial (grupo q) y a los ítems de la prueba objetiva con mayor poder discriminante sobre la misma (grupo p), en una estructura relativamente simple, que mostraba a su vez las variables diferenciales con cierta interacción. El orden en que las variables canónicas referidas a cada variable diferencial original aparecía, correspondía con el orden decreciente de porcentajes de acierdos obtenidos al aplicar la función discriminante en cada una de ellas. Es así que, cuando el número de variables a analizar sean muchas, la Correlación Canónica puede suponer una gran economía frente al A. Discriminante.

Hay que señalar que en los dos últimos casos comentados, la escala de medida era dicotómica, correspondiendo a «aciertos y errores». Como diferentes autores han interpretado, esta escala puede considerarse un caso límite de la escala de intervalo (un solo intervalo) y operar con ella tal como se hace con las demás escalas de intervalo.

Destaquemos que el signo de la correlación de cada variable original con la canónica, muestra una información superior que F en el análisis Discriminante, al especificar la dirección de las mútuas relaciones, sin necesidad de recurrir a las medias de cada variable en cada grupo. Y es esta información la que proporciona mayores posibilidades de interpretación «cualitativa» de los resultados.

REFERENCIAS BIBLIOGRÁFICAS

ADAMS, E.N. (1972). «Consensus techniques and the comparison of taxonomic trees». *Syst. Zool.*, 21, 390-397.

BALL, G.H. (1971) *Classification Analysis.* Stanford Research Inst. SRI. Proyect. 5533.

BARNET, V. (1981). *Interpreting Multivariate Data.* Wiley. Chichester.

BENZECRI, J.B. *L'analyse des données.* Dunod. Paris. (1973).

BAGOZZI, R.P.; FORNELL, C.; LARCKER, D.F. (1981). «Canonical correlation analysis as a special case of a structural relations model». *Multivariate Behavioral Research, 16* 437-454.

BISHOP, Y.M.M. y FIENBERG, E. (1969) «Incomplete two-dimensional contingency tables». *Biometrica* 22, 119-128.

CATTELL, R.B. *Handbook of Multivariate Experimental Psychology.* Rand McNally. Chicago. 1966.

DUNNETTE, M.D. «A note on the criterion». *J. of Applied Psychology,* 1963, 47, 251-254.

EADES, D.C. (1965) «The inappropriateness of the correlation coefficient as a measure of taxonomic resemblance». *Syst. Zool.*, 14.

GASCO, J. «Predicción multivariada del rendimiento escolar en el Ciclo Inicial de EGB» *Tesis de Licenciatura.* Fac. de F.ª y CCEE. Valencia 1984.

GNANADESIKAN, R. (1977). *Methods for Statistical data Analysis of Multivariate Observation.* WILEY. N. York.

GONRAD, M. *Valeurs propes et vecteurs propes en classification hiérarchique.* RAIRO, Paris. (1975).

GORDON, A.D. (1980). «On the assessment and comparison of classifications» en R. TOMASSONE *Analyse des Données et Informatique.* I.R.F.A. Le Chenay.

HOWARD, N. (1969) «Least squares classification and principal component analysis: a comparison» en M. DOGAN y S. ROKKAN *Quantitative ecological analysis in the social sciences.* Cambridge. MIT Press.

JAMBU, M. «Programme de calcul des contributions mutuelles entre classes d'une hiérachie et facteuns d'une correspondance». *Les cahiers de l'analyse des données.* Vol. 1, 1976, n.º 1. Paris.

JOHNSON, R.A. y WICHERN, D.W. *Applied Multivariate Statistical Analysis.* Prentice-Hall. N. Jersey. (1982).

KENDALL, M.G.A. *Cours in multivariate analysis.* N. York. Hafner (1957).

KRZANOWSKI, W.J. «The Performance of Fisher's Linear Discriminant Function Under Nonoptional Conditions» *Technometrics,* 19, 2.º (1977), 191, 200.

KSHIRSAGAR, A.M. (1978). *Multivariate Analysis.* N. York. Dekker.

HOPE, K. *Métodos de análisis multivariante.* Inst. Est. Polit. Madrid. 1982.

KUIPER, F.K. y FISHER, L. (1975). «A Monte Carlo comparison of six clustering procedures». *Biometrics,* 31, 777-783.

LEBART, L.; MORINEAU, A. Y TABARD, N. *Techniques de la description statistique.* Bordas, Paris. (1977).

LEFEBVRE, J. *Introduction aux Analyses statistiques Multidimensionnelles.* Masson. Paris. (1976).

LERMAN, I.C. (1981) *Classification et analyse ordinales des données.* Paris. Dunod.

LEVINE, M.S. (1977). *Canonical Analysis & Factor Comparison.* Sage, Bev. Hills.

McCULLAGH, P. y NELDER, J.A. *Generalized Linear Models.* London. Chapman and Hall, (1983).

PERIS, M.D. (1984). «Relación educación-inteligencia como alternativa a la relación clase social-inteligencia». *VIII C. Nal. de Pedagogía.* Santiago.

—, (1985). «Cours d'animateurs culturels: L'éducation comme la variable differentielle principal». *9.º C. Int. WAER-AMSE.* Madrid.

PRESS, S.J. y WILSON, S. (1978). «Choosing between logistic regression and discriminant analysis» *J. Amer. Statis. Assoc.* 73, 699-705.

ROHLF, F.J. (1970). «Adaptative hierarchical clustering schemes». *Syst. Zool.*, 19.

—, (1972). «An empirical comparison of three ordination techiques in numerical taxonomy». *Syst. Zool.*, 21.

RUBIN, J. (1967). «Optimal classification into groups: an approach for solving the taxonomy problem». *J. Theor. Biol.*, 15.

THOMPSON (1985). *Canonical Correlation Analysis: uses and Interpretation.* Sage. Bev. Hi.
THONGUTAI, U. *Prediction of college success and optimal assignment of studens using linear programming.* Univ. Calif. (1980).
THORNDIKE, R.M. (1978). *Correlational Procedures for Research.* N. York. Gardner Press.
VAN DE GEER, J.P. (1971). *Introduction to Multivariate Analysis for the Social Sciencies.* S. Francisco: Freeman & Comp.
YATES, F. (1966). «The first Fisher Memorial Lecture. Computers, the second revolution in statistics». *Biometrics,* 22 (2), 233-251.

Dolores Peris es Profesora titular del Área de Métodos de Investigación y Diagnóstico en Educación. Imparte la materia de Pedagogia Experimental en la Universidad de Valencia. Domicilio: Avda. Menéndez Pidal, 7 - 46009 Valencia.