PESQUIMAT, Revista de la Fac.CC.MM. de la
UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Vol.III,$N°$ 1,pag.40-50,LIMA-PERÚ. Julio 2000

MÉTODOS FACTORIALES EN MEDIDAS REPETIDAS

Emma Cambillo Moyano
e-mail: d230016@unmsm.edu.pe; ecambillo@yahoo.com

Universidad Nacional Mayor de San Marcos
Facultad de Ciencias Matemáticas
Laboratorio de Series de Tiempo

ABSTRACT Los métodos factoriales, concebidos para el estudio de una tabla de datos, pueden extenderse al caso de tablas múltiples. Se presenta la metodología propuesta por Scoffier [3] para el análisis de tablas mltiples en k ocasiones, una adaptación del Análisis de Componentes Principales para dos grupos propuesto por Krzanowski [2] y se propone una metodología para estudiar la relación de los cambios entre dos ocasiones. Para la aplicación, se utiliza los datos de una muestra panel de 50 hogares en la ciudad de Lima, observada durante cinco semanas.

Los métodos factoriales, específicamente el Análisis de Componentes Principales permite representaciones simplificadas de grandes tablas, facilitando la confrontación entre ocasiones diferentes.

1. INTRODUCCIÓN

vskip20pt

Desde hace una veintena de años, los métodos de análisis de datos han probado completamente su eficacia en el estudio de grandes masas complejas de información. El Análisis de Componentes Principales (ACP) ocupa un lugar primordial entre los métodos de análisis de datos, elección que se mantiene en parte debido a las representaciones geométricas de los datos que transforman en distancias euclídeas las proximidades estadísticas entre los elementos [2].

En los estudios longitudinales los elementos de una población se mantiene fijos y se miden repetidamente. La muestra permanece igual a lo largo del tiempo, de esta manera proporciona una serie de imágenes que, cuando se observan juntas, representan una ilustración muy real de la situación y de los cambios que tienen lugar con el paso del tiempo.

Los métodos factoriales, entre ellos el ACP, han sido concebidos para el estudio de una única tabla de datos. Ahora bien, en las encuestas longitudinales, los analistas de datos se encuentran frecuentemente confrontados al estudio simultáneo de varias tablas de datos, las cuales constituyen una

sucesión de tablas indexadas en el tiempo, llamadas tablas múltiples. Los datos están constituidos por un conjunto de individuos descritos por grupos de variables en cada ocasión. A cada ocasión le corresponderá una tabla.

En el transcurso de los años se han puesto a punto distintas metodologías [3]. Estas remiten generalmente al análisis de una tabla compleja formada por yuxtaposición de diferentes tablas. Estos métodos fundamentados sobre los métodos de análisis clásicos, concebidos en sí mismos para el estudio de una tabla simple, utilizan profusamente la técnica denominada de los elementos suplementarios. Pero estas técnicas tienen sus limitaciones y los objetivos específicos del análisis de tablas múltiples no son cubiertos en su totalidad.

El objetivo del presente trabajo fue encontrar una metodología que permita comparar encuestas realizadas en distintos momentos, en estudios longitudinales. Dicha metodología utiliza los principios fundamentales de los análisis clásicos, tomando en cuenta el carácter múltiple de las tablas y permitir el estudio simultáneo de varias tablas que cruzan los individuos y diferentes grupos de variables cuantitativas.

2. NOTACIÓN

Un conjunto de datos de medidas repetidas, es obtenida a partir de la medición u observación de p variables en K ocasiones para n individuos. Cuando $p = 1$, se genera un vector multivariado particular y el análisis de la estructura se resume al análisis de la matriz de covarianza de dimensión $K \times K$. Para $p > 1$, la matriz de datos de dimensión $n \times pK$,

$$X = \{X_1|X_2|\ldots|X_K\}$$

donde, la submatriz X_k es de dimensión $n \times p$, contiene las observaciones de las p variables en los n individuos para la k-ésima ocasión y las matrices de covarianza S_{kk} $(k = 1, 2, \ldots, K)$ de dimensión $p \times p$ de las p variables originales en la ocasión k son los elementos de la matriz de covarianza $S_{np\times np}$.

3. ANÁLISIS DE COMPONENTES PRINCIPALES

El análisis de componentes principales es uno de los métodos ms conocidos y utilizados del análisis multivariado. Tiene como objetivo la reducción de datos y la interpretación, motivo por el cual busca combinaciones lineales de las variables originales, tal que:

La primera componente principal es definida como la combinación lineal de las variables con máxima varianza. La varianza muestral de la combinación lineal $y = a'x$ es $S_y^2 = a'Sa$. Cuyo valor máximo λ es relativo a la longitud de a, obtenida de:

$$(S - \lambda I)a = 0$$

En consecuencia, existen p componentes principales , tales que $y_i = a_i'x, \quad i = 1, 2, \ldots, p$, los cuales son ortogonales $(a_i'a_j = 0; i <> j)$ y no correlacionados $(a_i'Sa_j = 0, i <> j)$. Si a es escalada tal que $a'a = 1$, las varianzas S_{yi}^2 son iguales a los autovalores λ_i.

Dado que los vectores de los coeficientes son ortogonales, las componentes principales son ortogonales, geométricamente equivalentes a una rotación de ejes para alinearse a la extensión natural de la nube de puntos.

Por otro lado uno de los objetivos principales del análisis de medidas repetidas es evaluar las diferencias que existen entre las ocasiones, por lo que es necesario comparar la estructura de datos en las ocasiones, luego de realizar un Análisis de Componentes Principales (ACP) en cada una de las submatrices de las ocasiones, y de superponer los espacios obtenidos en cada ocasión. Un procedimiento bastante intuitivo para comparar las ocasiones es describir cada ocasión en términos de un número pequeño de componentes principales y compararlas mediante ellas [1].

Desafortunadamente, una inspección visual no es muy confiable po- dría ocurrir que dos conjuntos de componentes principales que son completamente diferentes en apariencia pueden definir el mismo sub-espacio del espacio original p-dimensional. De esta manera se requiere un procedimiento analtico más confiable para la comparación. La interpretación geométrica de coeficientes congruentes entre dos vectores p dimensionales es mediante el coseno del ángulo entre estos vectores cuando ellos son representados como dos vectores en el espacio p dimensional. Por lo tanto, la naturaleza geométrica de componentes principales sugiere que una técnica análoga puede desarrollarse para la comparación de componentes principales.

4. MÉTODO

Si se realiza un ACP conjunto para las k ocasiones, la representación de los individuos en un mismo espacio ortogonal, permite observar si existe diferencias entre las K ocasiones; mientras que si se realizan K ACP, cada uno de ellos define un subespacio distinto de p dimensiones. Una comparación de la fuente de variación entre dos ocasiones requiere una comparación de dos sub espacios definidos por sus respectivos componentes principales, mientras que una comparación de las fuentes de variación de tres ocasiones requiere entonces de una comparación simultánea de los tres subespacios. Si el número de ocasiones se incrementa, dificulta la comparación simultánea entre los subespacios.

Es más apropiado, encontrar una medida que cuantifique el grado de concordancia de estos subespacios, dado que sub espacios coincidentes implican fuentes de variacin idénticas.

Para la comparación de dos subespacios, tomamos la matriz de datos de una ocasión, digamos A y la matriz de datos de otra ocasión, B; en cada una de las matrices se realiza un ACP y se retiene "m" componentes principales para representar cada una de las matrices. Todos los individuos son representados por puntos de un espacio p-dimensional con ejes de coordenadas $x_1, x_2, \ldots, x_p$. formando una nube de puntos en este espacio. Si cada una de las variables originales son relacionadas con un eje ortogonal en un espacio p-dimensional, las matrices A y B son representados como dos nubes de puntos en este espacio. El análisis de componentes principales es simplemente una rotación de ejes a una nueva posición $y_1, y_2, \ldots, y_p$ para la ocasión A y $z_1, z_2, \ldots, z_p$ para B. Estos nuevos ejes son tales que la proyección ortogonal de las varianzas muestrales son particionadas en forma decreciente en los ejes ortogonales.
Los coeficientes de las componentes principales l_{ij} y m_{ij} son los cosenos directores de la i-ésima componente de A y B respectivamente con el eje correspondiente a x_i.

Si m componentes principales son retenidas con el propósito de representar las dos muestras a través de las matrices $L = \{lij\}$ y $M = \{\text{mij }\}$ de dimensión $m \times p$. La nube de puntos para cada muestra puede considerarse encajada en un sub espacio m dimensional del espacio p dimensional y estos dos subespacios son definidos por los ejes ortogonales y_1, y_2, y_p y $z_1, z_2, \ldots, z_p$ respectivamente.

Además para comparar los dos conjuntos de componentes principales, es necesario comparar los dos subespacios m dimensionales que ellos generan. Es decir lo que interesa es encontrar una medida de proximidad o de similaridad entre estos dos subespacios. Si la similaridad se define como la proximidad de dos conjuntos de dimensiones en el espacio de las componentes principales antes que la proximidad de las nubes de puntos, es conveniente una medida de proximidad basada en los ángulos entre segmentos antes que la distancia entre puntos.

Una posibilidad, puede ser medir el ángulo entre pares correspondientes de CP para las dos ocasiones, presentándose el siguiente problema: que los ángulos entre pares correspondientes de CP pueden ser lo suficientemente grandes a pesar que los dos conjuntos de componentes definen el mismo subespacio m dimensional. Por lo que antes de calcular los ngulos debemos encontrar primero el mejor conjunto de comparacin de los ejes ortogonales que estn siendo comparados.

Para la comparación de subespacios, es preciso conocer los siguientes teoremas:

Teorema 1. El mínimo ángulo entre un vector arbitrario en el espacio de las m componentes principales de A y un vector paralelo lo mas cercano a él, en el espacio de las componentes principales de B esta dado por,$\cos^{-1}(\sqrt{\lambda_1})$; donde λ_1 es el mayor autovalor de $N = LM'ML'$ (Demostración en Krzanowski [2]).

Teorema 2. Sea λ_1 el mayor autovalor de N, e_1 su autovector asociado, $b_i = L'a_i\,(i = 1, 2, \ldots, m)$. Entonces $b_1, b_2, \ldots, b_m$ forman un conjunto de vectores mutuamente ortogonales incrustado en el subespacio de A y $M'Mb_1, M'Mb_2, \ldots, M'Mb_m$; el conjunto de vectores mutuamente ortogonales correspondientes en el subespacio B dentro de la cual las diferencias entre los subespacios pueden descomponerse en forma decreciente a través de los ejes ortogonales. El ángulo entre el p-ésimo par b_i y $M'Mb_i$, esta dado por (Demostración en Krzanowski [2]):

$$\cos^{-1}(\sqrt{\lambda_i}), \quad i = 1, 2, \ldots, m$$

Los teoremas 1 y 2 permiten evaluar la similaridad entre las ocasiones A y B a través de los pares b_i y $M'Mb_i$, donde λ_i es una medida de la contribución del i-ésimo par a la similaridad global (total). Además el vector en el espacio p-dimensional original que es cercano a b_i y $M'Mb_i$ es la bisectriz del ángulo entre ellos, en el plano en el cual ellos se encuentran. La bisectriz está dada por,

$$c_i = (2(I + \sqrt{\lambda_i}))^{-1}(I + \frac{1}{\sqrt{\lambda_i}}M'M)b_i$$

El conjunto $c_1, c_2, \ldots, c_m$ define el espacio m-dimensional que es el promedio de los subespacios A y B. .

Por otro lado, dado que uno de los objetivos del análisis de medidas repetidas es evaluar los cambios que tienen lugar con el paso del tiempo. Consideremos dos tiempos u ocasiones 1 y 2, es posible calcular las diferencias entre los valores observados de la misma variable en las dos ocasiones y así,

$$X_{D(n\times p)} = X_{1(n\times p)} - X_{2(n\times p)}$$

donde $X_{1(n\times p)} - X_{2(n\times p)}$ son la matrices de datos de las p-variables observadas en los n individuos en la ocasión 1 y 2 respectivamente, $X_{D(n\times p)}$ es la matriz de la diferencia entre las dos ocasiones .

El análisis de componentes principales a partir de la matriz $X_{D(n\times p)}$, permitirá estudiar la relación de los cambios entre las dos ocasiones, mediante una representación en menor dimensión.

5. APLICACIÓN

Se realizó una encuesta piloto a 50 familias en Lima Metropolitana, con el objetivo de evaluar el comportamiento de los consumidores durante cinco semanas respecto a los gastos semanales de productos de la canasta familiar: tales como carne roja, carne de aves, pescado, frutas, verduras, papa, yuca, camote y mostrar aplicaciones de las metodologías.

La matriz de datos de dimensión 50×40, dado que fueron observadas 8 variables durante cinco semanas en 50 familias, fue ordenada de la siguiente forma: $X = \{X_1|X_2|\ldots|X_K\}$, donde X_k, es la submatriz de dimensin 50×8, que agrupa al conjunto de las ocho variables observadas durante la k-ésima semana en los 50 individuos, $k = 1, 2, \ldots, 5$.

Siguiendo la metodología propuesta por Scofier [3], se realizó un ACP en cada una de las submatrices X_k, luego se pondero cada una de las submatrices por el inverso del autovalor correspondiente a la primera componente principal, obteniéndose $X^* = \{X_1^*|X_2^*|\ldots|X_K^*\}$, donde X_k^*, es la submatriz ponderada. Finalmente, se realizó un ACP con la matriz ponderada global X^*.

Los resultados obtenidos se presentan en el gráfico $N°1$. En la primera componente principal, observamos que no existe diferencias entre los gastos semanales en relación al gasto de fruta, verdura, ave y carne; además de existir relación en el gasto semanal de dichos productos, pues las variables relacionadas al gasto en un determinado producto se encuentran próximas durante las cinco semanas. A diferencia de la segunda componente principal, en la que podemos observar que no existen diferencias entre los gastos semanales en relación a la papa, pero si existen diferencias entre los gastos semanales en relación al gasto en camote y yuca, además de existir diferencias muy marcadas en el gasto semanal en pescado. La primera componente principal representa los gastos en alimentos con bajo contenido de carbohidratos y la segunda componente principal a los gastos en alimentos con alto contenido de carbohidratos.

Para utilizar la metodología propuesta por Krzanowski [2] comparemos las componentes principales de las dos primeras semanas. En la tabla $N°1$ se presentan las componentes principales para las dos semanas. La primera componente principal es un promedio ponderado del gasto semanal de todos los artículos durante la primera semana, predominando el gasto semanal en frutas, verduras, carne, ave y pescado tanto en la primera como en la segunda semana.

La segunda componente principal está relacionada al gasto semanal en papa, yuca y camote en la primera semana, a diferencia de la segunda semana, en la que la segunda componente principal está relacionada al gasto semanal en yuca, camote y pescado. Existe una ligera diferencia cuando comparamos las dos primeras componente principales obtenidas en las dos primeras semanas y, que se va incrementando a medida que comparamos las componentes principales restantes.

Una comparación entre las dos semanas, pede realizarse mediante la comparación de los subespacios generados por las componentes principales, $r = 1, 2, 3, 4$. Procederemos a comparar las dos primeras componentes principales, es decir, comparar los planos definidos y los resultados se presentan en la tabla $N°2$.

En la parte a) los ángulos de separación de los espacios generados por las dos componentes principales y sus autovalores correspondientes.

En la parte b) se presentan las bisectrices de estos dos planos, los vectores c_i, los cuales son los vectores más cercanos a los dos espacios para la dimensión en comparación.

En la tabla $N°1$, se observa, que las primeras componentes principales en las dos semanas son casi similares, con una ligera diferencia, la cual se incrementa a medida que observamos las componentes principales restantes.

Las primeras fuentes de variación entre las dos semanas son similares, puede concluirse por el ángulo 4.07 cuando se comparan los subespacios de dimensión 1 y por el ángulo 9.97 cuando se comparan los subespacios de dimensión 2 (tabla $N°2$); más ellas se vuelven diferentes a medida que se incrementan las dimensiones. Los ángulos cuando son comparados los subespacios de dimensión 2 y 3 son 44.42 y 67.39 respectivamente.

La primera componente principal expresa la similitud de los gastos semanales de papa, yuca, camote y pescado en las dos semanas y la segunda componente principal los gastos semanales en fruta, verdura y ave en las dos semanas.

Como uno de los objetivos del análisis de medidas repetidas es el estudio de los cambios entre ocasiones, es decir en este caso se debe estudiar los cambios en los gastos semanales de los artículos en las dos semanas. Al utilizar el análisis de componentes principales permite estudiar las relaciones entre los cambios en los gastos semanales de los artículos mencionados. Para lo cual utilizaremos la matriz de las diferencias entre los gastos por consumo entre las dos primeras semanas. La tabla $N°3$, muestra los resultados de la aplicación del análisis de componente principales en la matriz de las diferencias.

La primera componente principal representa los cambios entre los gastos semanales por consumo de fruta, verdura y ave. La segunda componente principal representa los cambios entre los gastos semanales por consumo de papa, camote, (obsérvese el gráfico $N°2$); similar a las interpretaciones obtenidas a partir de la tabla $N°2$, cuando comparamos los sub espacios.

6. CONCLUSIONES

Los métodos factoriales, y dentro de ellos el ACP, permite representaciones simplificadas de grandes tablas y se manifiestan como un instrumento de síntesis notable, permitiendo la confrontación entre diferentes ocasiones, lo que es infinitamente más rico que su examen por separado. Permiten abordar estudios nuevos más ricos y más complejos, en todos los dominios: marketing, seguros, banca, ecología, economía, etc.; representando una ilustración real de la situación y de los cambios que tienen lugar con el paso del tiempo.

En el ejemplo de aplicación ha sido posible evaluar el comportamiento de los consumidores durante cinco semanas respecto a los gastos semanales de ciertos productos de la canasta familiar, así como fue posible comparar los

gastos semanales por consumo de los diferentes artículos en las dos primeras semanas y los cambios que se producen en los gastos semanales por consumo de dichos productos.

BIBLIOGRAFIA

[1] G. ARNOLD, *Interpretation of Transformed Axes in Multivariate Analysis* Royal Statistical Society. Serie C. Vol. 42 $N°2$, 1993.

[2] KRZANOWSKI. *Principles of Multivariate Analysis.* A user's Perspective, 1988.

[3] B. SCOFIER, PAGES. *Análisis Factoriales Simples y Multiples.* DUNOD. Bordas, París. 1992

	C. P. 1		C. P. 2		C. P.3		C. P.4	
Variable	Sem. 1	Sem. 2	Sem.1	Sem. 2	Sem.1	Sem.2	Sem.1	Sem.2
Frutas	-0.494	0.560	0.200	-0.167	0.286	-0.205	-0.257	0.248
Verduras	-0.443	0.445	0.072	-0.376	0.109	0.024	0.503	-0.248
Papa	-0.173	0.233	-0.553	-0.070	0.135	0.349	0.246	-0.787
Yuca	-0.186	0.187	-0.508	0.599	-0.485	-0.135	0.212	-0.041
Camote	-0.068	0.283	-0.538	0.465	0.432	0.246	-0.171	0.164
Carne	-0.358	0.364	0.056	0.170	-0.665	-0.664	-0.220	-0.154
Ave	-0.405	0.415	0.297	-0.190	0.135	0.424	0.412	0.478
Pescado	-0.444	0.119	-0.100	0.431	0.070	0.369	-0.573	0.012

Tabla *N*°1
Resultado del Análisis de Componentes Principales en las Dos Primeras Semanas

a) ángulo	$\lambda_1 = 0.994701$ $\alpha_1 = 4.17$ $\lambda_2 = 0.970042$ $\alpha_2 = 9.97$ $\lambda_3 = 0.510151$ $\alpha_3 = 44.42$ $\lambda_4 = 0.147836$ $\alpha_4 = 67.39$
b) bisectriz	c1 c2 c3 c4
	-.0085
Fruta	.0014
Verdura	-.3727
Papa	-.4788
Yuca	-.3502
Camote	.-2704
Carne	.1679
Ave	-.3136
Pescado	.4123
	.5027
	.2323
	-.1067
	.1421
	-.1169
	.5521
	.1155
	-.2774
	-.0697
	.3276
	-.1242
	.2707
	-.5870
	-.0789
	-.0308
	-.2237
	.2447
	.2166
	.1829
	-.2072
	0775
	.0311
	-.3487

Tabla N° 2
Comparación de los Espacios de las Componentes Principales en las dos Semanas

VARIABLES	C.P.1	C.P.2
Fruta	0,716	0,239
Verdura	0,838	-0,002
Papa	-0,190	0,768
Yuca	-0,311	0,356
Camote	-0,005	0,751
Carne	-0,390	0,008
Ave	0,769	0,033
Pescado	0,199	0,499

Tabla $N°3$
Matriz de Coeficientes de las Componentes Principales en la Matriz de las Diferencias

GRAFICO N° 01

ACP DE LA MATRIZ PONDERADA POR LAS OCASIONES

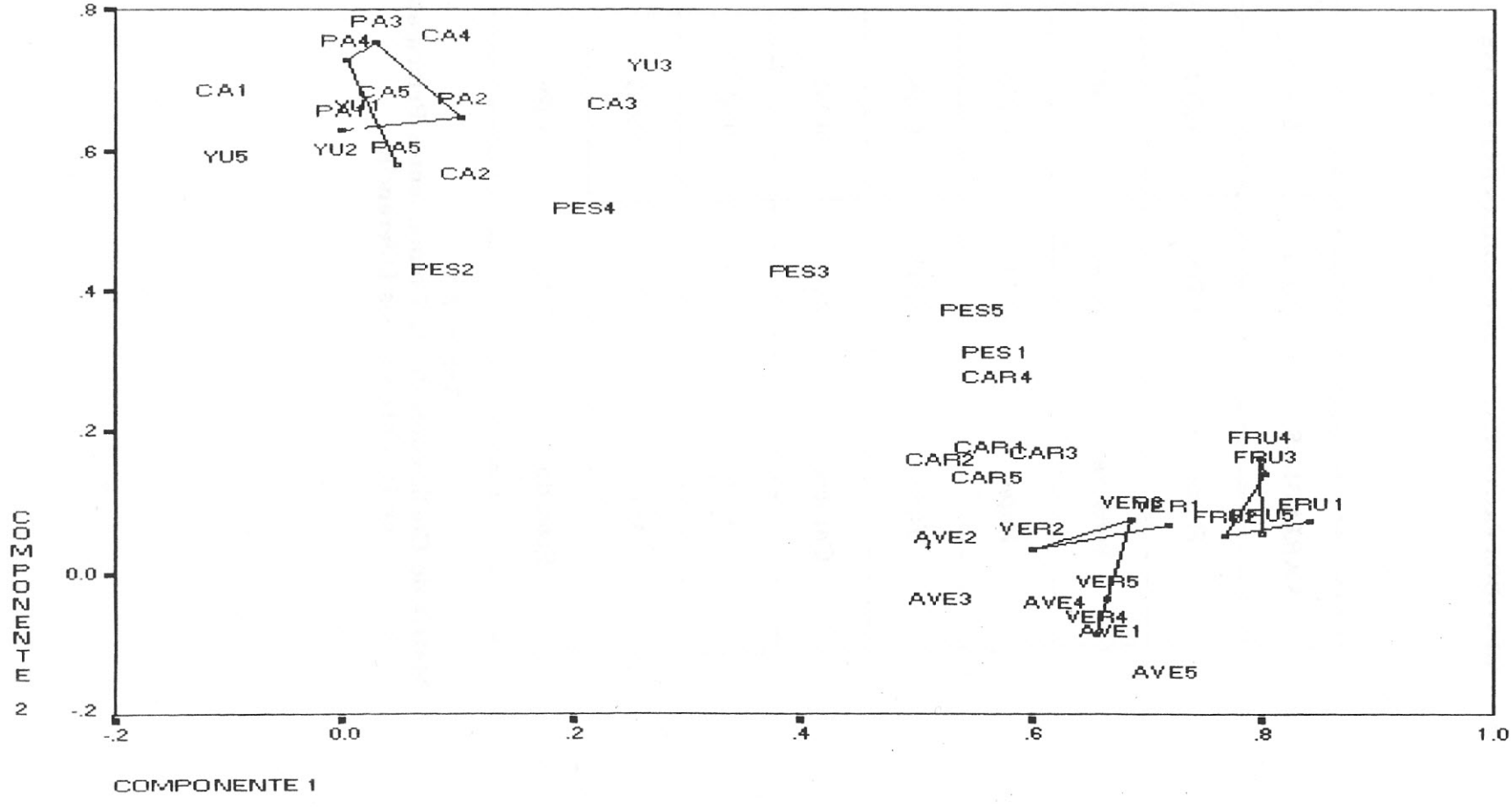

GRAFICO N° 02 ACP DE LA MATRIZ DE DIFERENCIAS ENTRE LAS DOS SEMANAS