

Distr.  
RESTRINGIDA  
E/CEPAL/R.327  
25 de agosto de 1982  
ORIGINAL: ESPAÑOL

---

CEPAL  
Comisión Económica para América Latina



EL METODO DE LOS COMPONENTES PRINCIPALES: SU  
APLICACION EN EL ANALISIS SOCIOECONOMICO

Este documento fue preparado por la División de Desarrollo Social  
de CEPAL.

82-8-1903





## Índice

	<u>Página</u>
Introducción.....	1
1. Consideraciones previas.....	2
2. Planteamiento del problema.....	5
3. Calidad del índice.....	17
4. Visualización geométrica de la información e interpretación.....	19
Bibliografía.....	28



## Introducción

En los fenómenos socioeconómicos intervienen una infinidad de variables de naturaleza muy diversa estrechamente interrelacionados entre sí. Para poder interpretar esta gran masa de información se utilizan las técnicas de análisis de datos multidimensionales, destacándose entre ellas como la más importante el análisis de componentes principales. Mediante este método se reduce una gran cantidad de variables en un número pequeño de nuevas variables (componentes principales) que tienen la propiedad de estar descorrelacionadas entre sí. Estas nuevas variables resumen las características de las variables primitivas y cada valor de ellas representa una suma ponderada de las variables primitivas. El peso con que actúan las variables en cada componente es determinado por el método, el que además proporciona criterios para evaluar los resultados logrados.

El objetivo de este trabajo está orientado a divulgar en un lenguaje sencillo, aplicando sólo elementos básicos del álgebra lineal, este método, cuyo conocimiento es de gran utilidad como herramienta analítica para los investigadores que trabajan en la problemática del desarrollo.

Su aplicación es sumamente valiosa en la clasificación de regiones en un país, o en la clasificación de países, permitiendo considerar numerosos puntos de vista a la vez y estudiar o describir las relaciones entre conjuntos de variables.

El hecho de que las ponderaciones sean un elemento objetivo determinado por el método, favorece su uso en la construcción de índices y en la determinación del status socioeconómico de las familias.

La particularidad que las nuevas variables (componentes principales) estén descorrelacionadas entre sí hace especialmente útil su aplicación en tareas predictivas (construcción de modelos lineales).

Otra aplicación que puede mencionarse es su uso en problemas de filtrado de series de tiempo al eliminar las variaciones locales. El trabajo está presentado en cuatro secciones. En la primera, consideraciones previas, se introduce en el tema, planteando una clasificación de

países y señalando algunos pasos preliminares al desarrollo del problema; en la segunda, planteamiento del problema, se presenta el desarrollo matemático mediante un lenguaje sencillo con elementos básicos del álgebra lineal (vectores y matrices); en la tercera, calidad del índice, se explica como se puede medir la representatividad de las nuevas variables (componentes principales); y en la última parte, visualización geométrica de la información e interpretación, se expone la manera de interpretar los resultados mediante diagramas que facilitan el análisis y comprensión de ellos.

### 1. Consideraciones previas

Para ilustrar la aplicación del método supondremos que nuestra tarea es la de clasificar países. Por lo tanto consideraremos un conjunto de  $N$  países caracterizados por  $M$  variables, en base a esta información se pretende ordenar los países en función de las variables. El objetivo perseguido es reducir las  $M$  variables a una sola variable que contenga  $N$  valores. Cada uno de los valores de esta nueva variable corresponde a una suma ponderada de las variables originales.

Nuestros datos son los valores de las variables y nuestra incógnita, las ponderaciones. Estas últimas deben actuar sobre las variables asegurando una buena calidad del índice, es decir, dotándolo de las siguientes características:

- i) Una representación confiable de los valores del índice conseguidos minimizando la pérdida de información en el proceso de síntesis que supone todo índice, y
- ii) Una clara diferenciación de los valores que lo componen.

La información de los países esta representada por variables de diversa magnitud y expresadas en unidades diferentes. Por lo tanto, un paso previo al planteamiento del problema es considerar la forma de homogeneizar la información que se dispone. Ello se consigue estandarizando las variables.

Una variable  $x_j$  estandarizada de un país  $i$  se simbolizará por  $\hat{x}_j^i$  y será igual a

$$\hat{x}_j^i = \frac{x_j^i - \bar{x}_j}{\sqrt{\text{VAR } x_j}}$$

Esta variable ( $\hat{x}_j^i$ ) que mide la desviación de la media en unidades de desviación típica, se llama variable normalizada y sus cantidades son adimensionales (es decir, independientes de las unidades empleadas). El número de este cociente determina el centro de gravedad de la nube de puntos (observaciones). El nuevo origen representa la media aritmética de las variables.

Al normalizar las variables algunos momentos estadísticos asumen determinados valores y antes de plantear el método se hará una evaluación de ellos.

a) La media aritmética de una variable  $x_k$  estandarizada ( $\bar{\hat{x}}_k$ )

$$\bar{\hat{x}}_k = \frac{1}{N} \sum_{i=1}^N \hat{x}_k^i$$

$$\bar{\hat{x}}_k = \frac{1}{N} \sum_{i=1}^N \frac{(x_k^i - \bar{x}_k)}{\sqrt{\text{VAR } (x_k)}}$$

$$\bar{\hat{x}}_k = \frac{\frac{1}{N} \sum_{i=1}^N x_k^i - \bar{x}_k}{\sqrt{\text{VAR } (x_k)}}$$

$$\bar{\hat{x}}_k = \frac{\bar{x}_k - \bar{x}_k}{\sqrt{\text{VAR } x_k}} = 0$$

Luego la media aritmética de una variable estandarizada vale cero.

b) La varianza de una variable estandarizada

$$\text{VAR}[\hat{X}_k] = \frac{1}{N} \sum_{i=1}^N (\hat{X}_k^i - \bar{\hat{X}}_k)^2$$

$$\text{VAR}[\hat{X}_k] = \frac{1}{N} \sum_{i=1}^N (\hat{X}_k^i)^2$$

$$\text{VAR}[\hat{X}_k] = \frac{1}{N} \sum_{i=1}^N \left( \frac{X_k^i - \bar{X}_k}{\sqrt{\text{VAR}[X_k]}} \right)^2$$

$$\text{Como } \text{VAR}[X_k] = \frac{1}{N} \sum_{i=1}^N (X_k^i - \bar{X}_k)^2$$

$$\text{VAR}[\hat{X}_k] = \frac{\text{VAR}[X_k]}{\text{VAR}[X_k]} = 1$$

Luego la varianza de una variable estandarizada es igual a uno.

c) Evaluemos la correlación entre dos variables estandarizadas

$$\text{CORR}[\hat{X}_j, \hat{X}_k] = \frac{\text{COV}[\hat{X}_j, \hat{X}_k]}{\sqrt{\text{VAR}[\hat{X}_j] \text{VAR}[\hat{X}_k]}}$$

$$\text{COV}[\hat{X}_j, \hat{X}_k] = \frac{1}{N} \sum_{i=1}^N (\hat{X}_j^i - \bar{\hat{X}}_j)(\hat{X}_k^i - \bar{\hat{X}}_k)$$

$$\text{COV}[\hat{X}_j, \hat{X}_k] = \frac{1}{N} \sum_{i=1}^N \hat{X}_j^i \hat{X}_k^i$$



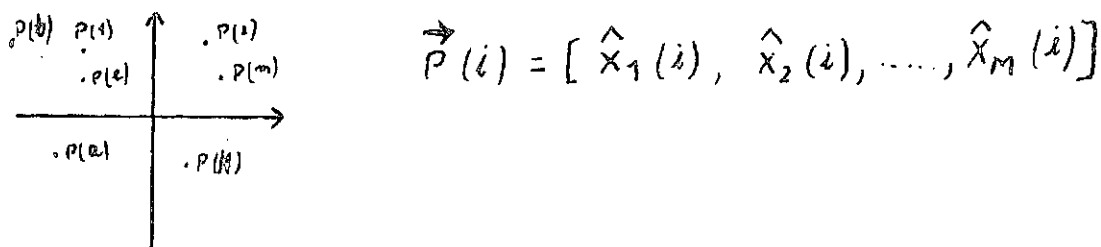
$$\text{CORR} [\hat{X}_j, \hat{X}_k] = \frac{1}{N} \sum_{i=1}^N \hat{X}_j^i \hat{X}_k^i$$

$$r_{j.k} = \frac{1}{N} \sum_{i=1}^N \hat{X}_j^i \hat{X}_k^i$$

## 2. Planteamiento del problema

Una manera de presentar el problema es ubicar a los países en un espacio de M dimensiones, en el que las variables constituyen los ejes de un sistema ortogonal. Por lo tanto, cada uno de estos ejes contiene los valores de los N países. Y la posición de cada país en el espacio está determinada por M coordenadas. En otras palabras, se tienen N vectores cada uno de ellos con M componentes, que fijan la posición de los países en el espacio  $R^M$ .

Así, por ejemplo, la posición de un país cualquiera (i) sería:



$$\vec{P}(i) = [\hat{X}_1(i), \hat{X}_2(i), \dots, \hat{X}_M(i)]$$

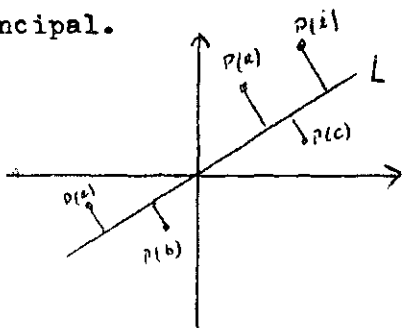
Este planteamiento a pesar de contener la totalidad de la información, no permite realizar ninguna ordenación de los países; a lo sumo sería posible calcular las distancias entre países. Si bien es cierto que ello proporcionaría una manera de compararlos, sería de difícil interpretación debido a la multidimensionalidad en que están expresadas las observaciones. Dado que la dimensión del espacio es lo que hace engorrosa la interpretación, se busca reducir el número de variables a una única dimensión, logrando que cada país quede determinado por una sola coordenada en lugar de las M.

Este proceso de reducción sólo es posible suponiendo un cierto grado de asociación entre las variables. Lo que se busca es una nueva

variable, que contenga tantos valores como observaciones existan. En nuestro caso, esta variable tendrá N valores, cada uno de ellos perteneciente a un país. El conjunto de valores de esta variable se denomina componente o factor, y cada uno de los valores que asume corresponde a una combinación lineal de las M variables iniciales. Así por ejemplo, un país cualquiera i tendrá un solo valor y este será igual a:

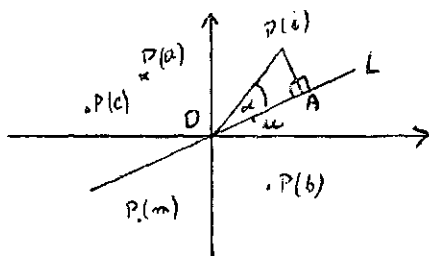
$$\text{País (i)} = \alpha_1 \hat{X}_1(i) + \alpha_2 \hat{X}_2(i) + \dots + \alpha_M \hat{X}_M(i)$$

Esta nueva variable esta contenida en una recta L, que pasa por el origen, los valores de ella son determinados por las proyecciones de las observaciones sobre la recta L, denominada recta solución o primer eje principal.



Cada uno de los puntos encontrados en la recta L corresponde a una combinación lineal de las variables primitivas y representa un valor del índice. El conjunto de valores es el índice, que como se había mencionado se denomina también componente o factor.

No se debe confundir el componente o factor con la recta solución ya que esta última contiene infinitos puntos, en cambio, la componente sólo tiene tantos como sean las observaciones proyectadas sobre la recta solución, en nuestro caso serán N puntos. En otras palabras, la componente es un subconjunto de la recta solución.



Expresemos matemáticamente lo que hemos formulado conceptualmente. Consideremos hipotéticamente que hemos determinado la recta solución L. Sobre esta recta se proyecta la observación del país (i)  $\angle P(i)$  del espacio  $R^M$ ).  $\sphericalangle$  La proyección es el trazo OA  $OA = OP(i) \cos \alpha$ .

Por otra parte, definamos los vectores  $\vec{OP}$  y  $\vec{OU}$

$$\vec{OP} = [\hat{X}_1(i), \hat{X}_2(i), \dots, \hat{X}_M(i)]$$

$$\vec{OU} = [\alpha_1, \alpha_2, \dots, \alpha_M]$$

imponiendo la restricción que el vector  $\vec{OU}$  sea unitario,

tenemos que  $|\vec{OU}| = \sum_{j=1}^M \alpha_j^2 = 1$

luego  $\vec{OP} \cdot \vec{OU} = |\vec{OP}| |\vec{OU}| \cos \alpha$

$$|\vec{OU}| = 1$$

$$\vec{OP} \cdot \vec{OU} = OP \cos \alpha$$

luego  $\vec{OP} \cdot \vec{OU} = OA$

como  $\vec{OP} \cdot \vec{OU} = \alpha_1 \hat{X}_1(i) + \alpha_2 \hat{X}_2(i) + \dots + \alpha_M \hat{X}_M(i) = \sum_{j=1}^M \alpha_j \hat{X}_j(i)$

con la restricción de que  $\sum_{j=1}^M \alpha_j^2 = 1$

O sea, el producto punto entre 2 vectores, uno de los cuales es unitario, es igual a la proyección del vector no unitario sobre la recta definida por el vector unitario y el origen.

Destaquemos las siguientes relaciones:

1) La proyección del país (i) sobre la recta Solución L, es igual a la combinación lineal de las variables primitivas  $[\sum_{j=1}^M \alpha_j \hat{X}_j(i)]$

Por lo tanto, la proyección del conjunto de observaciones determinan el índice.

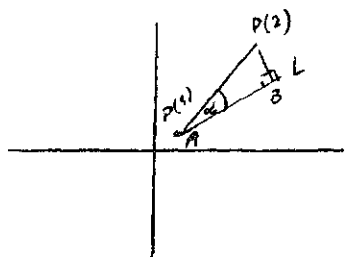
2) La recta solución está determinada por el origen y el vector unitario siendo las componentes de este último las ponderaciones de las combinaciones lineales de las variables.

El paso siguiente es encontrar algún criterio que permita determinar la recta solución. Este se puede deducir de los requisitos exigidos a las ponderaciones para asegurar una buena calidad del índice. Recordemos que estos eran 1) lograr la menor pérdida de información en el proceso de síntesis que supone todo índice, y 2) conseguir una clara diferenciación entre los valores que componen el índice.

En relación a la primera condición, habíamos mencionado que las distancias que mantienen las observaciones en el espacio conservan la totalidad de la información. Por lo tanto, uno de los requisitos que se debe cumplir es que el conjunto de proyecciones sobre la recta buscada permita reconstruir en forma óptima las distancias que mantienen entre sí las observaciones en el espacio  $R^M$ .

Como la operación de proyección reduce, en general, la distancia entre dos puntos, el criterio natural es usar aquel que maximiza la dispersión de las proyecciones sobre la recta.

Gráficamente se puede ilustrar con el ejemplo siguiente:



Dados dos puntos  $P(1)$  y  $P(2)$ , la proyección de ambos sobre  $L$  es  $AB = P(1) P(2) \cos \alpha$   
Si  $\alpha = 0$ ,  $\cos \alpha = 1$   
 $AB = P(1) P(2)$

O sea, en la medida que el  $\alpha$  tienda a cero la proyección se hace máxima reproduciendo con mayor fidelidad la distancia de las observaciones.

Recapitulando, la primera condición (evitar pérdida de información) se satisface logrando que la dispersión de las observaciones sobre la recta sea máxima, ello también implica cumplir la segunda condición (una clara diferenciación de los valores que componen el índice).

Por lo tanto, el medio para determinar las ponderaciones (componentes del vector unitario) es hacer máxima la varianza de las proyecciones.

Si designamos por  $F(i)$  la proyección de una observación  $i$ , recordemos que esta era igual a  $F(i) = \sum_{j=1}^M \alpha_j \hat{x}_j(i)$ , con la restricción de que  $\sum_{j=1}^M \alpha_j^2 = 1$

Luego, la varianza del conjunto de proyecciones será igual a

$$\text{VAR} [F] = \frac{1}{N} \sum_{i=1}^N [F(i) - \bar{F}]^2$$

$$\text{VAR} [F] = \frac{1}{N} \sum_{i=1}^N [F(i)]^2$$

Por lo tanto, la varianza del conjunto de las proyecciones es igual al promedio de la suma cuadrática de la combinación lineal de las variables.

O sea, la expresión que debe maximizarse es la varianza de las proyecciones, con la restricción que el vector de las ponderaciones sea unitario, es decir,  $\sum_{j=1}^M \alpha_j^2 = 1$

o lo que es igual  $\sum_{j=1}^M \alpha_j^2 - 1 = 0$

Por lo tanto, nos encontramos frente a un problema de máximo condicionado, que se resuelve mediante el método de los multiplicadores de Lagrange:

$$\text{VAR}[F] - \lambda \left[ \sum_{j=1}^M \alpha_j^2 - 1 \right] = V$$

$$\text{VAR}[F] = \frac{1}{N} \sum_{i=1}^N F_i^2$$

$$\text{VAR}[F] = \frac{1}{N} \sum_{i=1}^N \left( \sum_{j=1}^M \alpha_j \hat{X}_{j,i} \right)^2$$

luego, 
$$\frac{1}{N} \sum_{i=1}^N \left( \sum_{j=1}^M \alpha_j \hat{X}_{j,i} \right)^2 - \lambda \left( \sum_{j=1}^M \alpha_j^2 - 1 \right) = V$$

Derivando respecto a un  $\alpha_K$  se tiene  $\frac{\partial V}{\partial \alpha_K} = 0$

$$\frac{\partial V}{\partial \alpha_K} = \frac{2}{N} \sum_{i=1}^N \left( \sum_{j=1}^M \alpha_j \hat{X}_{j,i} \right) \hat{X}_{K,i} - 2\lambda \alpha_K = 0$$

luego 
$$\frac{1}{N} \sum_{i=1}^N \left( \sum_{j=1}^M \alpha_j \hat{X}_{j,i} \right) \hat{X}_{K,i} = \lambda \alpha_K$$

Reagrupando el primer término de la ecuación

$$\sum_{j=1}^M \alpha_j \frac{1}{N} \sum_{i=1}^N \hat{X}_{j,i} \hat{X}_{K,i} = \lambda \alpha_K$$

Recordemos que en las consideraciones previas evaluamos la correlación entre dos variables estandarizadas, la cual era:

$$\text{CORR}[\hat{X}_j, \hat{X}_k] = \frac{1}{N} \sum_{i=1}^N \hat{X}_{j,i} \hat{X}_{k,i}$$

Expresión que puede expresarse como

$$r_{j,k} = \frac{1}{N} \sum_{i=1}^N \hat{X}_{j,i} \hat{X}_{k,i}$$

Reemplazando esa expresión en la ecuación anterior se tiene

$$\sum_{j=1}^M \alpha_j \pi_{j,k} = \lambda \alpha_k$$

Esta ecuación corresponde a una derivada cualquiera de V (función a maximizar).

Como V tiene M variables, el máximo se encontrará haciendo nulas las derivantes parciales, resultando ese un sistema de M ecuaciones.

Desarrollando la sumatoria y tomando K diferentes valores de la M se llega al siguiente sistema de ecuaciones

Para K = 1  $\alpha_1 \pi_{1.1} + \alpha_2 \pi_{2.1} + \dots + \alpha_M \pi_{M.1} = \lambda \alpha_1$

Para K = 2  $\alpha_1 \pi_{1.2} + \alpha_2 \pi_{2.2} + \dots + \alpha_M \pi_{M.2} = \lambda \alpha_2$

Para K = M  $\alpha_1 \pi_{1.M} + \alpha_2 \pi_{2.M} + \dots + \alpha_M \pi_{M.M} = \lambda \alpha_M$

O lo que es lo mismo

$$(\pi_{1.1} - \lambda) \alpha_1 + \pi_{1.2} \alpha_2 + \dots + \pi_{1.M} \alpha_M = 0$$

$$\pi_{2.1} \alpha_1 + (\pi_{2.2} - \lambda) \alpha_2 + \dots + \pi_{2.M} \alpha_M = 0$$

$$\pi_{M.1} \alpha_1 + \pi_{M.2} \alpha_2 + \dots + (\pi_{M.M} - \lambda) \alpha_M = 0$$

O sea, se tiene un caso especial de un sistema de ecuaciones homogéneas, el problema de vectores y valores propios.

Designando por R a la matriz de correlaciones y por  $\vec{\alpha}$  al vector de las ponderaciones se puede escribir en forma matricial

$$[R] \vec{u} = \lambda \vec{u}$$

o bien  $[R - \lambda I] \vec{u} = 0$

En esta clase de problemas el valor propio ( $\lambda$ ) debe cumplir la función de hacer que la matriz sea singular. Este es el único medio de obtener soluciones diferentes a la trivial. Ya que de ocurrir que la matriz no fuera singular, sería posible:

$$\begin{aligned} [R - \lambda I]^{-1} [R - \lambda I] \vec{u} &= [R - \lambda I]^{-1} \cdot 0 \\ I \vec{u} &= 0 \\ \vec{u} &= 0 \end{aligned}$$

En este caso la única solución posible es la trivial. Luego para encontrar soluciones diferentes a la trivial, es necesario que la matriz sea singular.

Por lo tanto, el primer paso es resolver el determinante

$$|R - \lambda I| = 0$$

La resolución de este determinante origina una ecuación característica de orden M en  $\lambda$

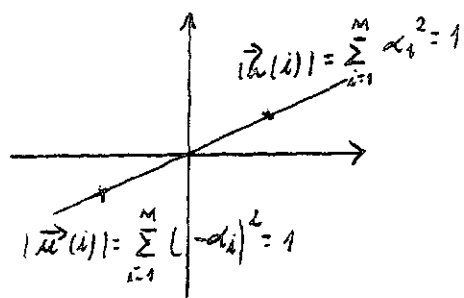
$$\lambda^M + C_1 \lambda^{M-1} + C_2 \lambda^{M-2} + \dots + C_{M-1} \lambda + C_M = 0$$

Cada uno de los valores propios  $\lambda(i)$  da origen a un vector propio  $\vec{u}(i)$ , o sea, en nuestro caso se obtienen M vectores propios determinados por los M valores propios. El procedimiento para encontrar un vector propio cualquiera  $\vec{u}(i)$ , es reemplazar el valor de  $\lambda(i)$  en la matriz de coeficientes. Sin embargo, ello da origen a un sistema de ecuaciones lineales homogéneas, cuya resolución tendría infinitas soluciones. Por lo tanto, para obtener un número determinado de soluciones es necesario imponer la restricción establecida en el planteamiento inicial,



que cada vector propio  $\vec{u}(i)$  sea unitario  $|\vec{u}(i)| = \sum_{i=1}^M \alpha_i^2 = 1$

De ese modo, por cada valor propio  $\lambda(i)$  se obtienen dos soluciones, cuyos valores absolutos son idénticos difiriendo únicamente en el signo de sus componentes



$$|\vec{u}(i)| = \alpha_1^2 + \alpha_2^2 + \dots + \alpha_M^2 = 1$$

$$|\vec{u}(i)| = (-\alpha_1)^2 + (-\alpha_2)^2 + \dots + (-\alpha_M)^2 = 1$$

Resumamos lo que se ha logrado:

- 1) Se han determinado los vectores propios unitarios, y, por lo tanto, las ponderaciones de las combinaciones lineales de la variable
- 2) Se han encontrado tantos vectores propios unitarios como variables primitivas que intervienen en el planteamiento del problema, es decir, M vectores propios unitarios que originan M rectas solución. En cada una de ellas es posible proyectar las N observaciones; además hay que tener presente que cada recta solución está determinada por un valor propio  $\lambda(i)$  diferente, de modo que las ponderaciones de las combinaciones lineales de las variables son distintas en cada componente.
- 3) Por cada valor propio  $\lambda(i)$  se obtienen dos soluciones, cuyos valores absolutos son idénticos difiriendo únicamente en el signo de sus componentes, pudiendo ocurrir que el algoritmo usado elija aquel vector que asigne valores con signo negativo a los países mejor clasificados en el índice.

El paso siguiente es determinar cual de los M vectores propios unitarios cumple mejor los requerimientos de confiabilidad que debe contar el índice.

Para comenzar, examinemos las propiedades que poseen los vectores propios unitarios de una matriz simétrica.

Recordemos que  $[R] \vec{u} = \lambda \vec{u}$

en que R es la matriz de correlaciones, y por lo tanto, es una matriz simétrica, y  $\vec{u}$  y  $\lambda$  son vector y valor propio de R.

Ahora bien, supongamos que se tienen dos valores propios distintos

$\lambda_1$  y  $\lambda_2$ , queremos indagar los atributos de los vectores propios  $\vec{u}_1$  y  $\vec{u}_2$  que originan esos valores propios.

Por ser  $\vec{u}_1$  y  $\lambda_1$  vector y valor propio de R, se tiene

a)  $R \vec{u}_1 = \lambda_1 \vec{u}_1$

y por la misma razón

b)  $R \vec{u}_2 = \lambda_2 \vec{u}_2$

Operando en a):  $(R \vec{u}_1)^t = \lambda_1 \vec{u}_1^t$

$\vec{u}_1^t R = \lambda_1 \vec{u}_1^t$  multiplicando por la derecha por  $\vec{u}_2$   
 $\vec{u}_1^t R \vec{u}_2 = \lambda_1 \vec{u}_1^t \vec{u}_2$

Operando en b):  $R \vec{u}_2 = \lambda_2 \vec{u}_2$  multiplicando por la izquierda  $\vec{u}_1^t$

$\vec{u}_1^t R \vec{u}_2 = \lambda_2 \vec{u}_1^t \vec{u}_2$

Luego a)  $\vec{u}_1^t R \vec{u}_2 = \lambda_1 \vec{u}_1^t \vec{u}_2$

b)  $\vec{u}_1^t R \vec{u}_2 = \lambda_2 \vec{u}_1^t \vec{u}_2$

de donde a) - b): 
$$\lambda_1 \begin{matrix} \rightarrow t \\ \mu_1 \end{matrix} \begin{matrix} \rightarrow \\ \mu_2 \end{matrix} - \lambda_2 \begin{matrix} \rightarrow t \\ \mu_1 \end{matrix} \begin{matrix} \rightarrow \\ \mu_2 \end{matrix} = 0$$
$$\begin{matrix} \rightarrow t \\ \mu_1 \end{matrix} \begin{matrix} \rightarrow \\ \mu_2 \end{matrix} (\lambda_1 - \lambda_2) = 0$$

como por hipótesis  $\lambda_1 \neq \lambda_2$

luego 
$$\begin{matrix} \rightarrow t \\ \mu_1 \end{matrix} \begin{matrix} \rightarrow \\ \mu_2 \end{matrix} = 0$$

Es decir, los vectores propios unitarios de una matriz simétrica son perpendiculares entre sí.

Por lo tanto, las M rectas solución, determinadas por los M vectores propios unitarios y el origen, constituyen un sistema de ejes ortogonales, o sea, los componentes originados en cada recta solución están totalmente descorrelacionados entre sí.

Otra derivación importante surge al asociar las propiedades que presentan los vectores propios unitarios de la matriz simétrica R.

Es decir: 
$$|\mu(i)| = \sum_{j=1}^M \alpha_j^2 = 1$$

y 
$$\begin{matrix} \rightarrow t \\ \mu_1 \end{matrix} \begin{matrix} \rightarrow \\ \mu_2 \end{matrix} = 0$$

Estas propiedades determinan que el conjunto de vectores propios de R forman una matriz ortogonal U, en que todos sus vectores son perpendiculares entre sí, y cada uno de ellos es unitario.

La propiedad de esta matriz ortogonal es que  $U U^t = I$

o sea que el inverso de la matriz ortogonal U es igual a su traspuesta.

Procedamos ahora a evaluar la traza de la matriz simétrica R aplicando las propiedades de U.

$$\text{traza } R = \text{traza } R U U^t$$

Designemos por B = RU

$$\text{traza } R = \text{traza } B U^t$$

$$\text{traza } B U^t = \text{traza } U^t B$$

$$\text{Reemplazando B traza } R U U^t = \text{traza } U^t R U$$

$$\text{por lo tanto: traza } R = \text{traza } U^t R U$$

Evaluemos, por lo tanto, el producto matricial  $U^t R U$

$$U^t R U = U^t [R \vec{u}_1 \quad R \vec{u}_2 \quad \dots \quad R \vec{u}_M]$$

en que los  $\vec{u}(i)$  representan vectores columnas de la matriz ortogonal U

como  $R \vec{u}(i) = \lambda \vec{u}(i)$

luego  $U^t R U = U^t [\lambda \vec{u}_1 \quad \lambda \vec{u}_2 \quad \dots \quad \lambda \vec{u}_M]$

$$U^t R U = \lambda [I]$$

$$\lambda [I] = [D]$$

matriz diagonal en que todo  $u_{ij} = 0$

para todo  $i \neq j$ , y  $u_{ij} = \lambda$  para todo  $i = j$

$$\text{luego: traza } U^t R U = \text{traza } R = \text{traza } D = \sum_{i=1}^M \lambda_i$$

Considerando, además, que cada elemento de la diagonal de la matriz de correlaciones  $R, (r_{ij})$ , representa la varianza normalizada de una variable. Por lo tanto la traza de R será igual a la varianza total y será igual a M.

$$\text{Luego: traza } R = \sum_{i=1}^M \lambda_i = M = \text{VARIANZA TOTAL}$$

### 3. Calidad del índice

La varianza total es igual al promedio de la suma cuadrática de las distancias que mantienen las observaciones (países) con el origen. Habíamos mencionado la importancia de reproducir esas distancias al hacer las proyecciones ya que ellas contienen toda la información original.

Sin embargo, al proyectar las observaciones sobre una recta solución, se altera, en alguna medida, esas distancias originales. La mayor o menor fidelidad con que se cumple esta condición básica es expresada por la varianza de la componente.

Calculemos la varianza de una componente cualquiera

$$\text{VAR} [F_k] = \frac{1}{N} \sum_{i=1}^N F_k^i{}^2$$

$$\text{VAR} [F_k] = \frac{1}{N} \sum_{i=1}^N \left( \sum_{j=1}^M \alpha_j \hat{X}_j^i \right)^2$$

$$\text{VAR} [F_k] = \sum_{k=1}^M \alpha_k \sum_{j=1}^M \alpha_j \frac{1}{N} \sum_{i=1}^N \hat{X}_j^i \hat{X}_k^i$$

recordemos que  $r_{j,k} = \frac{1}{N} \sum_{i=1}^N \hat{X}_j^i \hat{X}_k^i$

$$\text{luego } \text{VAR} [F_k] = \sum_{k=1}^M \alpha_k \sum_{j=1}^M \alpha_j r_{j,k}$$

Del desarrollo de página 9 se dedujo que  $\sum_{j=1}^M \alpha_j r_{j.k} = \lambda \alpha_k$

Como  $F_K$  está contenida en una recta  $L_K$  originada por un vector propio unitario  $\vec{u}_K$  y éste, a su vez, es determinado por un valor propio  $\lambda_K$

por lo tanto  $\sum_{j=1}^M \alpha_j r_{j.k} = \lambda_K \alpha_k$

luego  $\text{VAR} [F_K] = \sum_{k=1}^M \alpha_k \lambda_K \alpha_k$

$$\text{VAR} [F_K] = \lambda_K \sum_{k=1}^M \alpha_k^2$$

como la restricción impuesta es  $\sum_{k=1}^M \alpha_k^2 = 1$

luego  $\text{VAR} [F_K] = \lambda_K$

O sea, la varianza de una componente  $F_K$  es igual al valor propio  $\lambda_K$

Como se había demostrado que la varianza total era igual a la sumatoria de todos los valores propios y era igual a M.

Cada componente estaría representando una parte de la varianza total; y la calidad de cada una de ellas estaría determinada por el cuociente entre el valor propio respectivo y M. Esta última operación representaría el porcentaje de distorsión de la nube de puntos (observaciones) al ser proyectadas sobre la recta solución.

Resumiendo, la interpretación será que en la medida que el resultado del cuociente sea cercano a uno, el resultado obtenido es bastante bueno ya que indicaría que las distancias iniciales son reproducidas casi

íntegramente. Si, por el contrario, el resultado se aleja de uno hacia cero, mayor es la pérdida de información y por lo tanto, menos confiable la ordenación. Luego, los componentes pueden ordenarse en relación al porcentaje de varianza total que representan, denominándose primera componente a aquella que representa el porcentaje más elevado de varianza, y así sucesivamente.

#### 4. Visualización geométrica de la información e interpretación

Los resultados pueden presentarse:

- a) en forma de índice
- b) mediante el plano principal.

Y es posible interpretarlos mediante:

1. La correlación entre variables y factores (cuya representación gráfica es el círculo de correlaciones) y
2. La matriz de correlaciones de las variables.

##### a) El índice

Aquella componente que retiene el mayor porcentaje de la varianza es el que se presenta como índice. Si en el índice aparecen con signo negativo aquellos países mejor representados en la ordenación, este hecho no altera en absoluto la validez del resultado. Recordemos que ello se debe a que cada valor propio tiene dos soluciones, ambas con valores absolutos idénticos, difiriendo únicamente en el signo de sus componentes. Por lo tanto, puede ocurrir que el algoritmo usado elija aquel vector que origina valores negativos.

Por el hecho de estar normalizadas las variables la media de la componente es igual a cero. Por lo tanto, los valores del índice próximos a cero representarían los valores medios del índice. El índice puede considerarse como una nueva variable cuyo nombre se deriva del carácter de las variables primitivas, así por ejemplo, si estas describen aspectos del desarrollo socioeconómico, el índice tomará ese nombre.

b) Plano principal (visualización geométrica de la información)

Si el porcentaje de la varianza total que retiene la primera componente es muy bajo es posible incorporar la segunda componente para describir la información.

Como ambas componentes,  $F_1$  y  $F_2$  están contenidas en las rectas solución  $L(1)$  y  $L(2)$ , que representan los ejes de un sistema ortogonal. Ellas originan por lo tanto un plano denominado plano principal, en el cual la posición de los países estaría determinado por las nuevas coordenadas  $F_1(i)$  y  $F_2(i)$ .

El hecho de que las componentes esten descorrelacionadas entre sí, implica que la segunda componente incorporada al análisis aporta aspectos adicionales de la información no descritos por la primera componente.

La calidad de esta representación plana se mide por el cociente entre la suma de las varianzas de ambos componentes y la varianza total, o sea,

$$\frac{\text{VAR } [F(1)] + \text{VAR } [F(2)]}{\text{VAR total}}$$

por el hecho ya mencionado que la correlación entre  $F(1)$  y  $F(2)$  es nula.

El plano principal permite juzgar la homogeneidad del conjunto de observaciones ya que mediante él se puede visualizar aquellas observaciones que estan distanciadas en el plano y las que estan muy próximas entre sí. Las primeras corresponderían a países con características muy diferentes entre sí, y las segundas a países semejantes. Ello puede constituir la base para construir eventualmente una tipología de países.

#### 4.1. Correlación entre variables y factores

Tanto el índice como el plano principal constituyen los resultados del análisis de componentes principales; el primero expresa una ordenación de las observaciones, mientras el segundo muestra la dispersión de los países en el plano principal.

Si se pretende hacer un análisis más fino, es necesario explicar las causas que determinan la diferente posición de los países en el



índice, o bien poder interpretar la dispersión de estos en el plano principal.

Al comparar dos países de la nube de puntos, éstos pueden diferir por un solo parámetro o varios, o todos a la vez. Por lo tanto, identificar estas diferencias posibilita la interpretación de los resultados.

Para facilitar la explicación supongamos que se tiene un conjunto de países caracterizados por dos variables  $\hat{X}_1$  y  $\hat{X}_2$ . De ese conjunto de observaciones consideraremos dos países P(1) y P(2), estableciendo las causas que originan la diferente ordenación de estos países en las dos primeras componentes.

Para desarrollar el análisis se presenta un diagrama en el que aparecen: la nube de puntos, los ejes de las variables primitivas ( $\hat{X}_1$  y  $\hat{X}_2$ ) y los nuevos ejes, o sea las rectas solución L(1) y L(2) que contienen las componentes F(1) y F(2).

Los datos con que se cuentan para la interpretación de este diagrama son:

- 1) Los vectores propios unitarios que originan las rectas solución

$$\vec{u}(1) = [\alpha_1 \quad \alpha_2] \quad \text{que determina L(1)}$$

$$\vec{u}(2) = [\beta_1 \quad \beta_2] \quad \text{que determina L(2)}$$

- 2) La posición de los países en los ejes primitivos ( $\hat{X}_1$  y  $\hat{X}_2$ ).

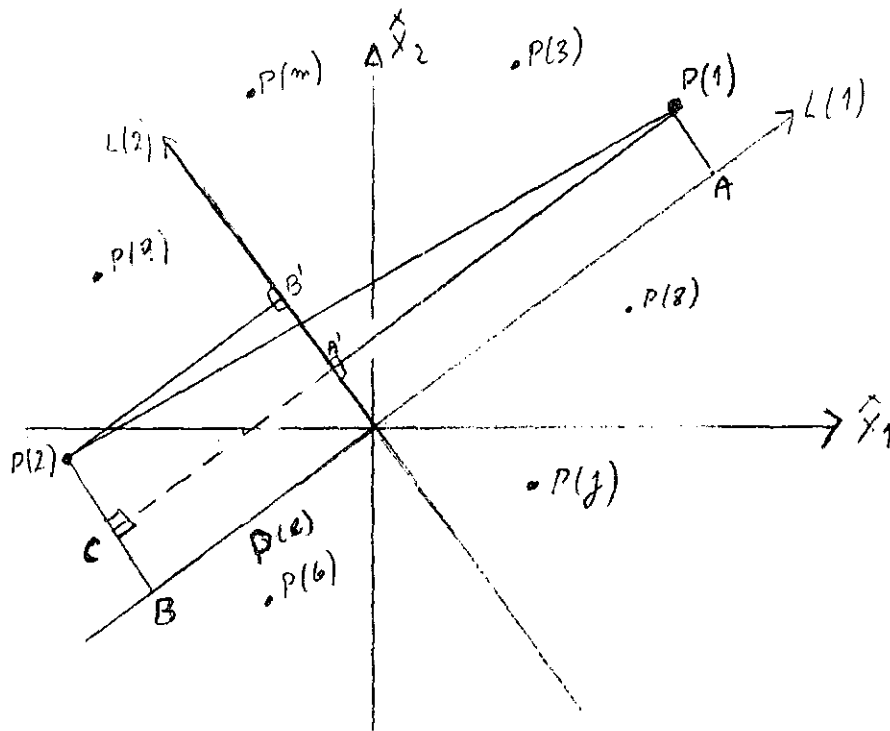
$$P(1) = [\hat{X}_1(1), \hat{X}_2(1)]$$

$$P(2) = [\hat{X}_1(2), \hat{X}_2(2)]$$

- 3) Los valores de las componentes

$$F(1) = [A \quad B]$$

$$F(2) = [A' \quad B']$$



La primera consideración que se puede deducir del diagrama es que las proyecciones de los países  $P(1)$  y  $P(2)$  sobre la recta  $L(1)$  reproducen con mayor fidelidad la distancia entre  $P(1)$  y  $P(2)$  que las proyecciones que recaen sobre  $L(2)$ .

La consideración geométrica de ello es:

La distancia  $P(1) P(2)$  = hipotenusa del triángulo rectángulo  $P_2, C, P(1)$

$$AB = CP(1)$$

$CP(1)$  = cateto mayor del triángulo rectángulo  $P_2, C, P(1)$

$$A'B' = CP(2)$$

$CP(2)$  = cateto menor del triángulo rectángulo  $P_2, C, P(1)$

Una segunda consideración surge al analizar la proyección del país uno (P(1)) sobre las rectas L(1) y L(2). En el diagrama se puede apreciar que la proyección del país uno sobre la recta L(1), (el valor A) representa un valor elevado de ese eje, y por lo tanto indicaría una clasificación buena del país uno en el índice (F(1)). En cambio, al examinar la proyección del mismo país (P(1)) en la recta L(2) (el valor A'), representa un valor próximo a la media de la clasificación en la componente F(2).

Comparando ambas proyecciones:

$$\begin{aligned} \text{Proyección de P(1) en la recta L(1)} \quad A &= \alpha_1 \hat{X}_1(1) + \alpha_2 \hat{X}_2(1) \\ \text{Proyección de P(1) en la recta L(2)} \quad A' &= \beta_1 \hat{X}_1(1) + \beta_2 \hat{X}_2(1) \end{aligned}$$

Se constata que siendo iguales los valores de las variables lo que establece la diferencia en la clasificación de los países es el hecho de que las variables no intervienen con los mismos pesos en la determinación de A y A'.

Una tercera consideración se desprende al analizar la proyección de ambos países sobre una misma recta solución.

Comparemos la proyección de dos países (P(1) y P(2)) en la recta L(1)

$$\begin{aligned} \text{Proyección de P(1) en la recta L(1)} \quad A &= \alpha_1 \hat{X}_1(1) + \alpha_2 \hat{X}_2(1) \\ \text{Proyección de P(2) en la recta L(1)} \quad B &= \alpha_1 \hat{X}_1(2) + \alpha_2 \hat{X}_2(2) \end{aligned}$$

En este caso se puede verificar que las ponderaciones son iguales y lo que determina la diferencia en la clasificación de los países son los valores de la variable. Por lo tanto, el paso siguiente es identificar aquellas variables que inciden con más fuerza en la ordenación de las observaciones. Mientras más acentuadas son las diferencias del valor de una variable en los países mayor es su poder de discriminación sobre las observaciones. En consecuencia, las variables que presentan una correlación más elevada con la componente son las que mejor explican la clasificación de los países.

Calculemos la correlación entre una componente cualquiera  $F_K$  y una variable  $\hat{X}_t$

$$\text{CORR}[F_K, \hat{X}_t] = \frac{\text{COV}[F_K, \hat{X}_t]}{\sqrt{\text{VAR}F_K} \cdot \sqrt{\text{VAR}\hat{X}_t}}$$

$$\text{COV}[F_K, \hat{X}_t] = \frac{1}{N} \sum_{i=1}^N F_K^i \hat{X}_t^i$$

$$\text{COV}[F_K, \hat{X}_t] = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M \alpha_j \hat{X}_j^i \hat{X}_t^i$$

$$\text{COV}[F_K, \hat{X}_t] = \sum_{j=1}^M \alpha_j \frac{1}{N} \sum_{i=1}^N \hat{X}_j^i \hat{X}_t^i$$

$$\text{COV}[F_K, \hat{X}_t] = \sum_{j=1}^M \alpha_j r_{j \cdot t}$$

Recordemos que del desarrollo hecho en página 9 se dedujo que:

$$\sum_{j=1}^M \alpha_j r_{j \cdot t} = \lambda \alpha_t$$

Como  $F_K$  está contenida en una recta  $L_K$ , originada por un vector propio unitario  $\vec{u}_K$  y este a su vez por un valor propio  $\lambda_K$

luego  $\text{COV}[F_K, \hat{X}_t] = \lambda_K \alpha_t$

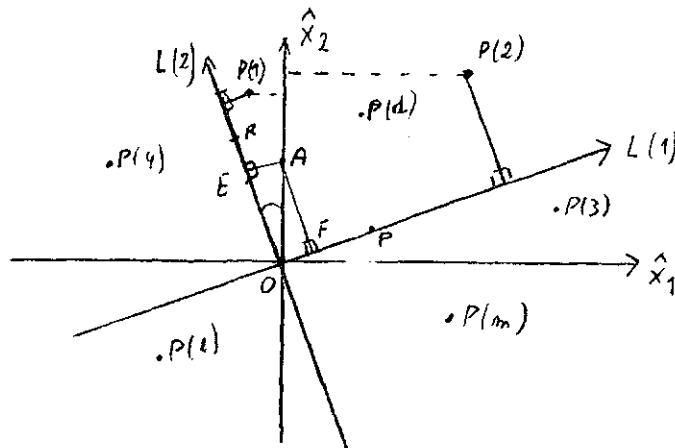
Por lo tanto,  $\text{CORR}[F_K, \hat{X}_t] = \frac{\lambda_K \alpha_t}{\sqrt{\text{VAR}[F_K]}}$

como  $\text{VAR}[F_K] = \lambda_K$

luego:  $CORR [F_k, \hat{X}_k] = \frac{\hat{T}_k \alpha t}{\sqrt{\hat{T}_k}}$

O bien:  $WRR [F_k, \hat{X}_k] = \alpha t \sqrt{\hat{T}_k}$

Representación simultánea de variables y observaciones en el plano principal



En el plano principal además de proyectar las observaciones se pueden proyectar también las direcciones de los antiguos ejes, eso permite - aún si estas direcciones son sólo aproximaciones - interpretar las diferencias que existen entre las observaciones.

Para ilustrar el procedimiento supondremos un conjunto de  $N$  países caracterizados por dos variables  $\hat{X}_1$  y  $\hat{X}_2$ . En este ejemplo, primero se calculará la dirección en que se proyecta la variable  $\hat{X}_2$  en el plano principal y luego se hará una interpretación de la clasificación de dos países ( $P(1)$  y  $P(2)$ ) de la nube de puntos en la componente  $F(2)$ .

En el diagrama se presentan la nube de puntos, los antiguos ejes  $\hat{x}_1$   $\hat{x}_2$ , los nuevos ejes L(1) y L(2), los puntos P y R que representan los vectores propios unitarios  $|\vec{u}_1|=1$  y  $|\vec{u}_2|=1$  que determinan las rectas L(1) y L(2), y la magnitud OA = 1 de la variable  $\hat{x}_2$ ; es decir, el punto A tiene coordenadas (0,1) en el antiguo eje.

La proyección del punto A sobre las rectas L(1) y L(2) determinan los puntos F y E respectivamente.

$$OF = \vec{OA} \cdot \vec{u}_1 = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} u_{1.1} \\ u_{1.2} \end{bmatrix} = u_{1.2}$$
$$OE = \vec{OA} \cdot \vec{u}_2 = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} u_{2.1} \\ u_{2.2} \end{bmatrix} = u_{2.2}$$

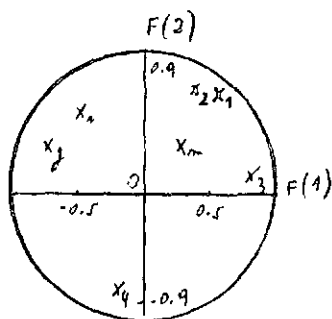
Luego el eje que corresponde a la variable  $\hat{x}_2$  se proyecta en el plano principal en la dirección del vector  $\begin{bmatrix} u_{1.2} \\ u_{2.2} \end{bmatrix}$ . Mientras más próxima sea la dirección de una variable a una de las rectas solución, mayor es la asociación de esa variable con la componente contenida en esa recta.

En el diagrama es posible constatar que la variable  $\hat{x}_2$  está más próxima a la recta L(2) que a L(1); indicando con ello una asociación más estrecha con la componente F(2). Ello significaría que la variable  $\hat{x}_2$  tiene valores muy diferentes en el conjunto de países, y por lo tanto, ejerce una influencia decisiva en la clasificación de los países en esa componente. Sin embargo al establecer la comparación de dos observaciones del conjunto, los países P(1) y P(2) se verifica que la variable  $\hat{x}_2$  tiene valores muy semejantes en ambos países; y en este caso específico la ordenación de estos dos países en la componente F(2) estaría determinada por la variable  $\hat{x}_1$ .

O sea que una representación de esta naturaleza permite visualizar rápidamente las causas que explican la diferencia en la ordenación en algunos países del conjunto, aspecto que no es tan fácil de descubrir mediante el análisis matemático cuya preocupación principal es

interpretar la ordenación del conjunto de países.

Círculo de correlaciones



Este gráfico permite visualizar correlaciones entre variables. Cada punto representa a una variable; se la caracteriza por sus correlaciones respectivas con la primera y segunda componentes principales. Siendo el módulo de las correlaciones inferior a uno, todos los puntos se encuentran al interior del círculo de centro cero y radio uno. Mientras más alejado del centro del círculo este la variable menos deformada es su representación original y más importante es su peso en la construcción de las componentes principales. Si dos variables bien representadas (alejadas del centro de la circunsferencia) son vecinos ( $x_2, x_1$ ), tienen una correlación alta positiva, mientras que si son ortogonales ( $x_3, x_4$ ) tienen entre sí una correlación baja y finalmente si son opuestos ( $x_2, x_4$ ), tienen una correlación alta negativa.

Bibliografía

- L. Lebart et Fenelon, Statistique et informatique Appliquées, Dunond, Paris, 1973.
- F. Caillez y J.P. Pazes, Introduction a l'analyse des donnees, SMASH, Paris, 1976.
- T.W. Anderson, Introduction to multivariate statistical analysis, Wiley, New York, 1958.
- J.P. Benzecri, L'analyse des donnees, Dunond, Paris, 1973.
- M. Tatsouka, Multivariate analysis: techniques for educational and psychological research, Wiley, New York, 1971.
- N.Norman, C.Hull, J.G.Jenkins, K.Steinbrenner, D.Bent: Statistical package for social sciences, McGraw Hill, New York, 1975.
- N.Lacourly, "Panorama de las técnicas de análisis de datos multidimensionales";
- J.Muñoz y H. Aliaga, "Determinación objetiva de indicadores socioeconómicos"  
En Sigma, Revista de matemáticas aplicadas.  
Primeras jornadas de matemáticas aplicadas,  
26 al 30 de julio de 1976, Santiago de Chile.
- N. Lacourly, R. Cruz Coke, C. Valenzuela y M.Benavente, Estudio en la distribución de la morbilidad en fratrias hospitalizadas, mediante el análisis de componentes principales. Sigma, Revista del Departamento de Matemáticas, vol.2, Nº3, junio 1976. U. de Chile, Santiago de Chile.
- J.Johnston, Métodos de econometría. Editorial Vicens-Vives, 3a. edición, 1980. Impreso en España.
- Arnoldo de Hoyos y Pablo Mandler, Agrupamiento de países de América Latina por indicadores socio-económicos. Análisis de componentes principales. Seminario sobre métodos y problemas en tipificación de empresas agropecuarias. Instituto Interamericano de Ciencias Agrícolas, Zona Sur-OEA. Dirección de Investigaciones Agropecuarias, MAP, vol.Z. Editor Hugo Cohen, Montevideo, diciembre 1975.
- Claudio Arenas, Nancy Lacourly y Servet Martínez, Técnicas para el análisis de información clínica. Curso dictado en 1978. U.de Chile, Depto. Matemáticas. Santiago de Chile, agosto 1978.