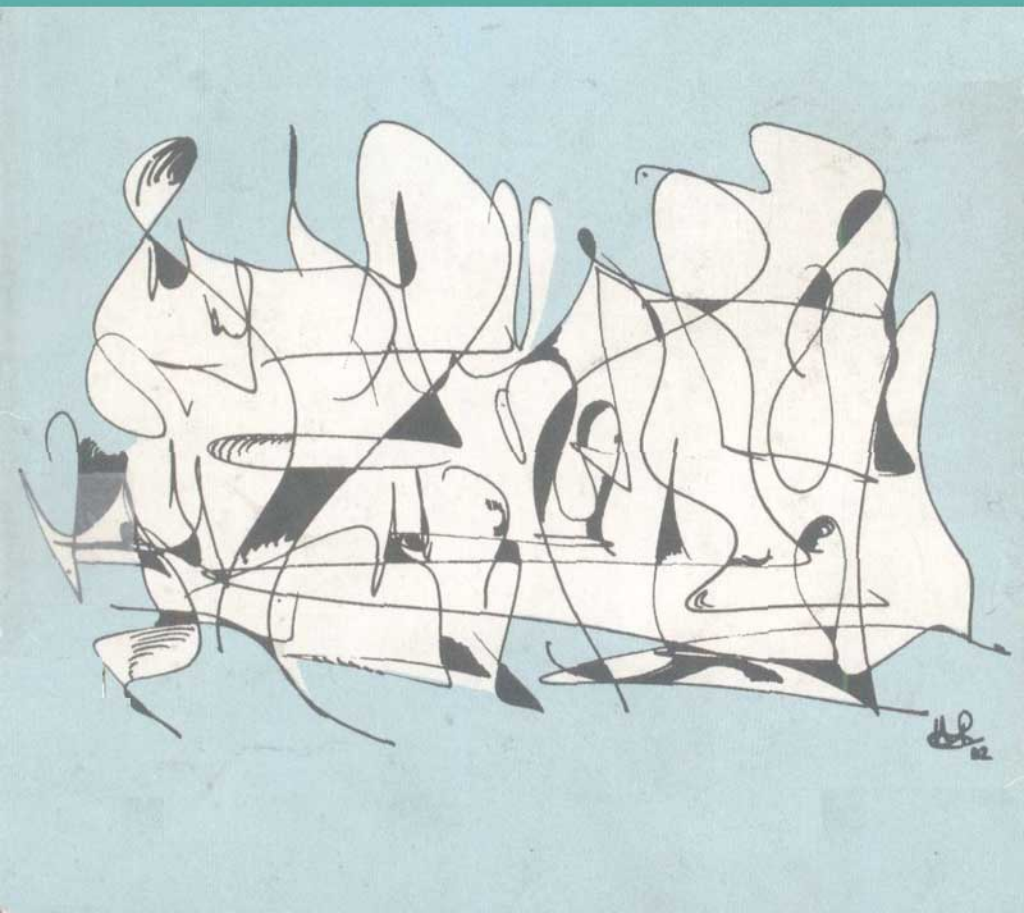
serie de matemática

monografía no. 27

ANALISIS MULTIVARIADO: METODO DE COMPONENTES PRINCIPALES

Programa Regional de Desarrollo Científico y Tecnológico
Departamento de Asuntos Científicos
Secretaría General de la
Organización de los Estados Americanos



ANALISIS MULTIVARIADO: METODO DE COMPONENTES PRINCIPALES

por

Laura E. Pla
Departamento de Producción Vegetal
Area de Ciencias del Agro y del Mar
Universidad Nacional Experimental
Francisco de Miranda
Coro, Falcón, VENEZUELA

Secretaría General de la
Organización de los Estados Americanos
Programa Regional de Desarrollo Científico y Tecnológico
Washington, D.C. - 1986

Esta monografía ha sido preparada para su publicación en el Departamento de Asuntos Científicos y Tecnológicos de la Secretaría General de la Organización de los Estados Americanos.

Editora: Eva V. Chesneau

Asesor Técnico: Prof. Evelio O. Fabbroni
Instituto Interamericano de Estadística
Organización de los Estados Americanos
Washington, D.C., EE.UU.

A los lectores

El programa de monografías científicas es una faceta de la vasta labor de la Organización de los Estados Americanos, a cargo del Departamento de Asuntos Científicos y Tecnológicos de la Secretaría General de dicha Organización, a cuyo financiamiento contribuye en forma importante el Programa Regional de Desarrollo Científico y Tecnológico.

Concebido por los Jefes de Estado Americanos en su Reunión celebrada en Punta del Este, Uruguay, en 1967, y cristalizado en las deliberaciones y mandatos de la Quinta Reunión del Consejo Interamericano Cultural, llevada a cabo en Maracay, Venezuela, en 1968, el Programa Regional de Desarrollo Científico y Tecnológico es la expresión de las aspiraciones preconizadas por los Jefes de Estado Americanos en el sentido de poner la ciencia y la tecnología al servicio de los pueblos latinoamericanos.

Demostrando gran visión, dichos dignatarios reconocieron que la ciencia y la tecnología están transformando la estructura económica y social de muchas naciones y que, en esta hora, por ser instrumento indispensable de progreso en América Latina, necesitan un impulso sin precedentes.

El Programa Regional de Desarrollo Científico y Tecnológico es un complemento de los esfuerzos nacionales de los países latinoamericanos y se orienta hacia la adopción de medidas que permitan el fomento de la investigación, la enseñanza y la difusión de la ciencia y la tecnología; la formación y perfeccionamiento de personal científico; el intercambio de informaciones, y la transferencia y adaptación a los países latinoamericanos del conocimiento y las tecnologías generadas en otras regiones.

En el cumplimiento de estas premisas fundamentales, el programa de monografías representa una contribución directa a la enseñanza de las ciencias en niveles educativos que abarcan importantísimos sectores de la población y, al mismo tiempo, propugna la difusión del saber científico.

La colección de monografías científicas consta de cuatro series, en español y portugués, sobre temas de física, química, biología y matemática. Desde sus comienzos, estas obras se destinaron a profesores y alumnos de ciencias de los primeros años de la universidad; de éstos se tiene testimonio de su buena acogida.

Este prefacio brinda al Programa Regional de Desarrollo Científico y Tecnológico de la Secretaría General de la Organización de los Estados Americanos la ocasión de agradecer a la Ing. Laura E. Pla, autora de esta monografía, y a quienes tengan el interés y buena voluntad de contribuir a su divulgación.

julio de 1986

INDICE

Página

CAPITULO 6. ANALISIS DE CALIFICACIONES POR ASIGNATURA

PARTE I
FUNDAMENTOS TEORICOS

1

EL UNIVERSO MULTIVARIADO Y SU ANALISIS

1. INTRODUCCION

Según Kendall (1980),[22] en el estudio propio del campo multivariado pueden utilizarse diferentes enfoques, tanto por los distintos tipos de situaciones que se presentan al obtener los datos, como por el objetivo específico del análisis. Los más importantes son:

a) **Simplificación de la estructura de los datos.** El objetivo es encontrar una manera simplificada de representar el universo de estudio. Esto puede lograrse mediante la transformación (combinación lineal o no lineal) de un conjunto de variables interdependientes en otro conjunto independiente o en un conjunto de menor dimensión.

b) **Clasificación.** Este tipo de análisis permite ubicar las observaciones dentro de grupos o bien concluir que los individuos están dispersos aleatoriamente en el multiespacio. También pueden agruparse variables.

c) **Análisis de la interdependencia.** El objetivo es examinar la interdependencia entre las variables, la cual abarca desde la independencia total hasta la colinealidad cuando una de ellas es combinación lineal de algunas de las otras o, en términos aun más generales, es una función f(x) cualquiera de las otras.

d) **Análisis de la dependencia.** Para ello se seleccionan del conjunto ciertas variables (una o más) y se estudia su dependencia de las restantes, como en el análisis de regresión múltiple o en el análisis de correlación canónica.

e) **Formulación y prueba de hipótesis.** A partir de un conjunto de datos es posible encontrar modelos que permitan formular hipótesis en función de parámetros estimables. La prueba de este nuevo modelo requiere una nueva recopilación de datos a fin de garantizar la necesaria independencia y validez de las conclusiones.

En los casos de poblaciones univariadas, casi siempre es posible caracterizar completamente la distribución de probabilidades a partir de dos parámetros: la media y la varianza. La inferencia estadística exige, entonces, tomar una muestra aleatoria y calcular los mejores estimadores de estos dos parámetros. El análisis termina con la interpretación de las dos estimaciones.

Sin embargo, para el caso multivariado en que se estudia una población p variada, es decir un conjunto de individuos donde se han observado o medido p características o propiedades, se dispondrá de p medias, p varianzas y $(1/2)p(p-1)$ covarianzas, que no sólo deben ser estimadas, lo cual no es difícil con las computadoras digitales, sino que *deben ser interpretadas*.

Si se logra una transformación que genere nuevas variables no correlacionadas se eliminan $(1/2)p(p-1)$ parámetros, y si se reducen las dimensiones de p a $(p-1)$, se pasa de $(1/2)p(p+3)$ parámetros poblacionales a ser estimados e interpretados a $(1/2)p(p+3) - (1/2)(p-1)(p+2) = (p+1)$ parámetros a ser estimados e interpretados.

Si bien puede no existir interés en todos los parámetros y, por lo tanto, no es necesario estimarlos, cuanto más sencillo sea el modelo poblacional, más cerca estará el investigador de encontrar una interpretación comprensible de la estructura original mediante la muestra efectivamente observada.

En la Tabla I se presentan el número de parámetros estimables en una población multivariada de diferentes dimensiones, el número de parámetros estimables si se efectúa una transformación que elimina una dimensión, y el número de parámetros estimables si se efectúa una transformación que genere nuevas variables no correlacionadas. Los datos de la Tabla I se representan gráficamente en la figura 1, donde puede apreciarse el crecimiento relativo del número de parámetros estimables.

Tabla I. Número de parámetros estimables según el número de variables por considerar

Número de Variables p	Número de Parámetros Estimables		
	Sin Transformar (1/2)(p(p + 3))	No Correlacionadas 2p	Reduciendo la Dimensión p + 1
1	2	-	-
2	5	4	2
4	14	8	5
6	27	12	7
8	44	16	9
10	65	20	11
20	230	40	21
30	495	60	31

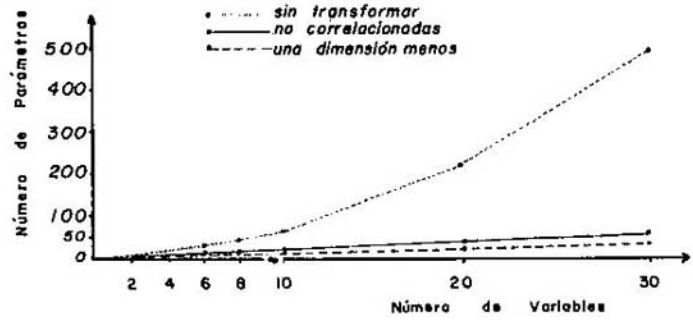


Fig. 1. Número de parámetros estimables en función del número de variables del sistema. El número de parámetros por estimar e interpretar disminuye rápidamente si se efectúa una transformación lineal que genere nuevas variables no correlacionadas; también si se logra disminuir una dimensión.

Cada situación requiere una evaluación particular para utilizar el método de análisis multivariado más adecuado, que permita extraer la máxima información posible del conjunto de datos, pero que a su vez garantice la validez de su aplicabilidad. Las técnicas multivariadas son muy potentes y pueden llevar al investigador a encontrar una justificación (¿por qué no la que "cree" más correcta?) que no se sustente necesariamente en el análisis objetivo de la información recopilada.

Para la mente humana, acostumbrada a pensar y a representar el espacio en dos dimensiones, o a lo sumo en tres, la noción de un multiespacio con cuatro, cinco o p dimensiones resulta difícil de comprender. Hay muchas maneras de acercarse a este concepto y quizás el enfoque matricial y matemático --base del análisis estadístico multivariado-- sea el más adecuado; aunque es, asimismo, el menos apropiado para el lector no familiarizado con el álgebra de matrices.

La medición de varias características de una misma unidad experimental, ya sea en forma simultánea o con ciertos intervalos de tiempo, genera una serie de datos que deben ser analizados con técnicas multivariadas. La unidad experimental puede ser un individuo, una parcela de experimentación, una finca, un animal, una planta, una porción de terreno, y las características serán una serie de atributos, mediciones, evaluaciones, estimaciones, tratamientos o propiedades correspondientes a esas unidades experimentales. No habrá independencia entre las diferentes propiedades utilizadas para caracterizar una unidad y no será posible asignar en forma aleatoria las características, como en un ensayo experimental típico. Habrá, sí, independencia entre las unidades experimentales que podrán constituir una muestra aleatoria de una población mayor.

Habiendo explicado qué se entiende por universo multivariado, se comprenderá por qué los métodos estadísticos multivariados pueden agruparse en dos conjuntos: los que permiten extraer información acerca de la interdependencia entre las variables que caracterizan a cada uno de los individuos y los que permiten extraer información acerca de la dependencia entre una (o varias) variable(s) con otra (u otras).

Entre los métodos de análisis multivariado para detectar la interdependencia entre variables y también entre individuos se incluyen el análisis de factores, el análisis por conglomerados o "clusters", el análisis de correlación canónica, el análisis por componentes principales, el análisis de ordenamiento multidimensional ("scaling"), y algunos métodos no paramétricos. Los métodos para detectar dependencia comprenden el análisis de regresión multivariado, el análisis de contingencia múltiple y el análisis discriminante.

En este trabajo se examina el método de análisis por componentes principales, uno de los más difundidos entre las técnicas multivariadas. Este método permite la estructuración de un conjunto de datos multivariados, obtenidos de una población cuya distribución de probabilidades no necesita ser conocida. Es, pues, una técnica matemática que no necesita un modelo estadístico para explicar la estructura probabilística de los errores. Aunque, si es posible suponer una distribución multinormal de la población, o el tamaño de la muestra es tal que puede asumirse multinormalidad ya sea por aumento en el número de variables consideradas o por el número de individuos que integran la muestra, podrá encontrarse significación estadística en los componentes, pues será posible asociar a cada uno de ellos una medida de confiabilidad.

A continuación se presenta un conjunto de conceptos necesarios para la interpretación del análisis por componentes principales y su aplicación a una matriz de datos.

2. ALGUNOS CONCEPTOS UNIFICADORES

En el caso multivariado, así como en el univariado, se dispone de ciertos estadísticos que permiten caracterizar una población y estudiar su comportamiento probabilístico. Algunos de ellos son una extensión directa de la noción univariada equivalente y otros son propios de las poblaciones y muestras multivariadas.

Definición 1. Se dice que un conjunto de datos constituye una muestra aleatoria multivariada si cada individuo ha sido extraído al azar de una población de individuos y en él se han medido u observado una serie de características. Sean x(ij) la observación de la j-ésima variable en el i-ésimo individuo, X(i) el vector fila que contiene las observaciones de todas las variables en el i-ésimo individuo y X(j) el vector columna que contiene todas las observaciones de la j-ésima variable. Se define una *matriz de datos multivariados* como el arreglo

$$X = (x(ij)) = \begin{bmatrix} x(1,1) & \cdots & x(1,p) \\ \vdots & & \vdots \\ x(i,1) & \cdots & x(i,p) \\ \vdots & & \vdots \\ x(n,1) & \cdots & x(n,p) \end{bmatrix}$$

de dimensión (n x p), que también puede expresarse como

$$X = \langle (X(1) \ \cdots \ X(p) \rangle = \begin{bmatrix} X(1) \\ \vdots \\ X(n) \end{bmatrix}.$$

A partir de esta matriz que contiene toda la información estadística de la muestra es posible calcular algunas funciones que, al igual que en el caso univariado, permitan extraer conclusiones de los datos. En el caso univariado se calcula la media para tener una estimación de la tendencia central de las observaciones y se calcula la varianza que informa acerca del grado de dispersión de los datos alrededor de esa media. La extensión al caso multivariado es posible; de ello resulta la media: un vector y la varianza: una matriz.

Definición 2. Dada una matriz de datos como la señalada en la definición 1, se define la media muestral de la j-ésima variable por:

$$\bar{x}(j) = 1/n \sum_{i=1}^{n} x(ij),$$

y el vector formado por los $\bar{x}(j)$ será el *vector promedio*

$$\bar{X} = \begin{bmatrix} \bar{x}(1) \\ \vdots \\ \bar{x}(p) \end{bmatrix}.$$

Definición 3. Dada una matriz de datos como la señalada en la definición 1, se define la varianza muestral de la j-ésima variable por:

$$s(jj) = 1/n \sum_{i=1}^{n} (x(ij) - \bar{x}(j))^2$$

y se define la covarianza entre la j-ésima y la k-ésima variable por:

$$s(jk) = 1/n \sum_{i=1}^{n} (x(ij) - \bar{x}(j))(x(ik) - \bar{x}(k))$$

$$j,k = 1,\ldots,p$$

La matriz formada por el arreglo de los s(jk) y los s(jj) será la *matriz de varianza-covarianza muestral* o, simplemente, matriz de covarianza muestral

$$S = \begin{bmatrix} s(11) & . & . & . & s(1p) \\ . & . & . & . & . \\ . & . & s(jk) & . & . \\ . & . & . & . & . \\ s(p1) & . & . & . & s(pp) \end{bmatrix}.$$

Definición 4. A partir de los elementos de la matriz S es posible calcular la matriz **R**, de igual dimensión que **S**, y cuyos elementos sean los coeficientes de correlación entre la j-ésima y la k-ésima variables:

$$r(jk) = \frac{s(jk)}{\sqrt{s(jj)\ s(kk)}} = \frac{s(jk)}{s(j)\ s(k)}.$$

También podrán ser arreglados en una matriz de *correlación muestral* cuya diagonal principal estará formada por números uno y será simétrica como la matriz de covarianza, por ser r(jk) = r(kj):

$$R = \begin{bmatrix} 1 & . & . & . & r(1p) \\ . & . & . & . & . \\ . & . & 1 & . & . \\ . & . & . & . & . \\ r(p1) & . & r(pk) & . & 1 \end{bmatrix}.$$

La matriz S de covarianza es una manera de expresar la dispersión de los datos alrededor de la media. Sin embargo, a veces es necesario disponer de un escalar que sintetice esta dispe· ión. Puede encontrarse un número que exprese la variabilidad multivariada a partir de la información contenida en la misma matriz S.

Definición 5. Dada una matriz S tal como la señalada en la definición 3, se denomina *varianza generalizada* al determinante de dicha matriz

$$V = |S|.$$

Definición 6. Dada una matriz S tal como la señalada en la definición 3, se denomina *variación total* a la traza de la matriz S.

$$\text{tr } S = \sum_{j=1}^{p} s(jj).$$

Tanto la varianza generalizada como la variación total serán mayores cuanto mayor sea la dispersión de los datos alrededor de la media. Sin embargo, como sostienen Mardia y colaboradores (1979),[23] cada medida refleja aspectos diferentes de la variabilidad de los datos. La primera desempeña un papel muy importante en la generación de los estimadores máximo verosímiles, mientras que la segunda se utiliza en el análisis de componentes principales. Ninguno de los dos conceptos tiene un equivalente en el análisis univariado.

Los datos multivariados ofrecen --y esto es novedoso-- la posibilidad de ser expresados en combinaciones lineales de las variables originales. Esta es quizás la herramienta más poderosa para realizar este tipo de análisis estadístico, el cual no es factible en el campo univariado. En un número reducido de combinaciones es posible sintetizar la mayor parte de la información contenida en los datos originales. Sin embargo, a veces resulta muy complicado deducir la distribución exacta de probabilidades, aunque respecto a las combinaciones más utilizadas se conocen resultados asintóticos.

Desde un punto de vista geométrico y espacial es posible conceptualizar la matriz de datos multivariados de dos maneras: como un conjunto de n individuos en un espacio definido por las p variables, o como un conjunto de p variables definidas en un espacio de n dimensiones. En el primer caso, las observaciones serán puntos que representan un individuo en el espacio definido por las variables (cada eje será una variable); en el segundo, cada punto representará una variable definida en el espacio cuyos ejes serán cada uno de los n individuos.

En el primer caso se comparan individuos considerados en función de sus características, es decir se comparan vectores fila X(i). Este procedimiento, denominado técnica Q, se utiliza en el análisis discriminante, en el análisis por conglomerados (aunque también es posible hacer conglomerados de variables) y en el análisis de ordenamiento multidimensional.

Si, por el contrario, se comparan columnas, se obtendrá información acerca de la relación entre características consideradas en función de los individuos que se estudian, es decir se comparan los vectores X(j) en un espacio de dimensión n. Esta técnica se llama técnica R, pues la matriz de correlación R debe ser calculada para poder iniciar los análisis; desempeña un papel importante en el análisis por componentes principales, en el análisis de factores y en el análisis de correlación canónica.

3. MODELO PROBABILISTICO MULTINORMAL

El conocimiento de la serie de definiciones formuladas a continuación ayudará a comprender las pruebas de hipótesis incluídas en este trabajo; sin embargo, su estudio no es imprescindible para el lector que se interesa sólo en los aspectos metodológicos de la aplicación de la técnica de componentes principales al análisis de datos.

Muchos de los fenómenos univariados que se estudian en ciencias agrícolas y biológicas se ajustan a un modelo cuyo componente aleatorio se distribuye normal e independientemente, o al menos puede ser objeto de una transformación adecuada que permita ajustar los valores observados a la distribución normal. Es por ello que muchas de las pruebas para los estadísticos calculados a partir de valores muestrales se basan en la distribución normal: chi cuadrado, F, t de Student, por mencionar sólo las más utilizadas. En esta serie de definiciones se sigue la nomenclatura de Mardia y colaboradores (1979).[23]

En los casos univariados se caracteriza una curva de distribución normal por los valores de su media poblacional μ, cuyo estimador insesgado es $\bar{x}$, la media muestral, y de su varianza poblacional σ^2 (sigma cuadrado), cuyo estimador de varianza mínima es S^2, la varianza muestral. Se utiliza su raíz cuadrada S para estimar el desvío de cada observación respecto a la media; este estimador se denomina desviación estándar. Cuando se desea obtener un valor que acote la desviación de la media muestral se divide la desviación estándar por la raíz del ta-

maño de la muestra, de forma tal que cuanto mayor sea el número de elementos a partir de los cuales se calcula la media y la varianza muestral, menor será la desviación de la media. Esta desviación estándar de la media se llama también "error estándar".

Definición 7. Se dice que un vector X de dimensión (p x 1) tiene la distribución normal p-variada (o distribución normal p-dimensional) con vector media μ y matriz de covarianza Σ si su función de densidad está dada por:

$$f(\mathbf{X}) = |2\pi \Sigma|^{-1/2} \exp < -1/2 (X - \mu)' \Sigma^{-1} (X - \mu) >,$$

donde $\Sigma > 0$ y μ finito.

Esto puede representarse por $X \sim Np(\mu, \Sigma)$, que debe leerse "**X** se distribuye normal p, con media mu y matriz de covarianza sigma".

La matriz S de la definición 3 es un estimador de la matriz Σ de la definición anterior y el vector $\bar{X}$ de la definición 2 es un estimador del vector μ de la función f(X). La matriz R de la definición 4 es un estimador de la matriz Σ cuando las variables han sido estandarizadas, es decir cuando se efectúa una transformación de las variables originales x(ij) para que tengan media cero y varianza uno. Generalmente una nueva variable resultante de esta transformación se simboliza por z(ij).

La matriz Σ en una distribución multinormal o normal multivariada puede adoptar otras formas particulares. Cuando las variables no están correlacionadas, es decir cuando la covarianza entre dos cualesquiera de ellas es cero, la matriz Σ será diagonal, o sea sólo tendrá elementos no nulos (distintos de cero) en la diagonal principal. Estos serán los valores de las varianzas de cada variable. Si esas variables no correlacionadas se estandarizan (media cero, varianza uno), la matriz Σ será la matriz identidad (que se simboliza por I), es decir una matriz con números uno en la diagonal principal y ceros fuera de ella. Si todas las variables estudiadas tuvieran una varianza común, como sucede en los modelos de diseño, la matriz de covarianza podría simbolizarse por $\sigma^2\mathbf{I}$. Se tendría una matriz diagonal en que todos los elementos de la diagonal principal serían iguales al valor de la varianza común.

$$\begin{bmatrix} s(11) & 0 & 0 & 0 & 0 \\ 0 & s(22) & 0 & 0 & 0 \\ 0 & 0 & s(33) & 0 & 0 \\ 0 & 0 & 0 & s(44) & 0 \\ 0 & 0 & 0 & 0 & s(55) \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} s(.) & 0 & 0 & 0 & 0 \\ 0 & s(.) & 0 & 0 & 0 \\ 0 & 0 & s(.) & 0 & 0 \\ 0 & 0 & 0 & s(.) & 0 \\ 0 & 0 & 0 & 0 & s(.) \end{bmatrix}$$

Variables no correlacionadas: *matriz diagonal*

Variables no correlacionadas estandarizadas: *matriz identidad*

Variables con varianza común: *matriz diagonal*

Conforme a la definición 7, para determinar la función de densidad normal multivariada es necesario calcular el determinante de la matriz de covarianza Σ, así como su inversa, por lo cual debe ser no singular, es decir poseer un determinante distinto de cero. Esto no ocurre cuando el rango de la matriz es menor que su dimensión. La dimensión está dada por el número de variables que se consideran, o sea la dimensión del vector X. El rango corresponde al número de vectores linealmente independientes de la matriz de datos y es por ello que si alguna de las variables consideradas es una combinación lineal de otras incluidas también en la matriz, el rango será menor que la dimensión y, por ende, la matriz de covarianza será singular. En este caso, la función

de densidad normal multivariada no puede determinarse y se dice que "no existe". Para definir la distribución es necesario recurrir a otras formas de expresión diferentes como la función generatriz de momentos.

4. VECTORES Y VALORES PROPIOS

El análisis por componentes principales se basa en una transformación lineal de las observaciones originales. Esta transformación lineal --que como se verá más adelante satisface las exigencias de dicho análisis-- es conocida en el campo del álgebra vectorial como generación de vectores y valores propios, o también vectores o valores característicos (el vocablo original de raíz anglosajona es "eigen" y, por eso, se lo ha traducido también como "eigen valores" o "eigen vectores").

Definición 8. Sea B una matriz cuadrada de dimensión (p x p); es posible encontrar un escalar λ (lambda) y un vector X de dimensión (p x 1), no nulo tal que

$$\mathbf{BX} = \lambda \mathbf{X}$$

lo que implica

$$\mathbf{BX} - \lambda \mathbf{X} = 0 = (\mathbf{B} - \lambda \mathbf{I})\ \mathbf{X} = 0$$

sacando el vector X como factor común a la derecha, de forma que la operación matricial sea conformable.

Por ejemplo, si B es una matriz de dimensión (2 x 2) deberá encontrarse la solución a la siguiente ecuación matricial:

10

$$\left[\begin{bmatrix} 6 & 3 \\ 3 & 4 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right] \begin{bmatrix} x(1) \\ x(2) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 6 - \lambda & 3 \\ 3 & 4 - \lambda \end{bmatrix} \begin{bmatrix} x(1) \\ x(2) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} (6 - \lambda)\ x(1) + 3\ x(2) \\ 3\ x(1) + (4 - \lambda)\ x(2) \end{bmatrix}$$

(2 x 2) (2 x 1) (2 x 1) (2 x 1)

Debe resolverse el sistema de ecuaciones siguiente:

$$(6 - \lambda)\ x(1) + 3\ x(2) = 0 \qquad [1]$$

$$3\ x(1) + (4 - \lambda)\ x(2) = 0 \qquad [2]$$

Despejando x(1) en la segunda ecuación y reemplazándolo en la primera se obtiene

$$x(1) = \frac{-(4 - \lambda)\ x(2)}{3}$$

$$\frac{-(6 - \lambda)(4 - \lambda)\ x(2)}{3} + 3\ x(2) = 0.$$

Sacando x(2) como factor común

$$\left[\frac{-(6 - \lambda)(4 - \lambda)}{3} + 3\right] x(2) = 0,$$

lo cual implica, si x(2) es distinto de cero, que el corchete debe ser necesariamente nulo:

$$\left[\frac{-(6-\lambda)(4-\lambda)}{3} + 3 \right] = 0. \qquad [3]$$

Sacando 1/3 factor común puede reordenarse como

$$(6-\lambda)(4-\lambda) - 3 \cdot 3 = 0.$$

Si se observa la estructura de esta ecuación se verá que es el determinante de la matriz original menos λI, es decir el producto de los dos valores de la diagonal principal, menos el producto de los dos valores de la diagonal secundaria. Esta es la solución encontrada cuando se sustituye x(2); la ecuación tiene dos soluciones por ser una ecuación de segundo grado. Si se la ordena como es costumbre, en el cálculo numérico se obtiene:

$$24 - 4\lambda - 6\lambda + \lambda^2 - 3 \cdot 3 = 0$$

que, reagrupando términos de igual potencia, será:

$$\lambda^2 - 10\lambda + 15 = 0.$$

Es decir que, aplicando la nomenclatura tradicional de una ecuación de segundo grado con valores

$$a = 1 \qquad b = -10 \qquad y \qquad c = 15$$

y resolviendo respecto a λ, se obtienen dos raíces que, en este caso, son $\lambda_1 = 1{,}8377$ y $\lambda_2 = 8{,}1623$.

Puede observarse que la suma algebraica de las dos soluciones de la ecuación, es decir la suma algebraica de los valores propios, es igual a la suma de los valores de la diagonal principal de la matriz original ($\lambda_1 + \lambda_2 = 8{,}1623 + 1{,}8377 = 10$).

Esta propiedad de los valores propios es importante, ya que cuando se calculan a partir de la matriz de covarianza, la suma de los valores propios es igual a la suma de las varianzas de las variables incluidas en la matriz, o sea es la variación total (definición 6).

Si el mismo procedimiento se aplica a continuación, pero respecto a x(1), se obtendrá la misma ecuación de segundo grado de la fórmula [3]. Conocidos los valores que puede tomar λ, es posible utilizar las ecuaciones [1] y [2] para determinar una función de los valores de x(1) y de x(2). Los posibles valores de λ se denominan *valores propios* y el vector formado por los valores de x(1) y x(2) generados por cada valor propio se llama *vector propio*. Así, habrá tantos vectores propios --con los valores de x(1) y x(2)-- como valores propios diferentes existan.

Para realizar esta sustitución se parte de las ecuaciones [1] y [2], las cuales se igualan y se obtiene:

$$(6-\lambda)\,x(1) + 3\,x(2) = 3\,x(1) + (4-\lambda)\,x(2), \qquad [4]$$

es decir que se dispone de una sola ecuación para resolver un sistema que posee dos incógnitas x(1) y x(2). Es posible reducir la ecuación [4] y expresarla como sigue:

$$x(1) = a\,x(2),$$

donde a será una función de los λ encontrados. Así, reagrupando y simplificando se obtiene:

$$x(1)((6 - \lambda) - 3) = x(2)((4 - \lambda) - 3)$$

$$x(1)(3 - \lambda) = x(2)(1 - \lambda)$$

$$x(1) = \left(\frac{1 - \lambda}{3 - \lambda}\right) x(2) \qquad [5]$$

sustituyendo en [5] por los valores de λ encontrados previamente se obtiene:

para $\lambda_1 = 8{,}1623$	para $\lambda_2 = 1{,}8377$
$x(1) = \frac{-7{,}1623}{-5{,}1623}\, x(2)$	$x(1) = \frac{-0{,}8377}{1{,}1623}\, x(2)$
$x(1) = 1{,}3847\, x(2)$	$x(1) = -0{,}7207\, x(2)$

Para encontrar otra ecuación que permita completar el sistema se establece la condición de que los vectores propios estén normalizados. Esto equivale en términos algebraicos a que la suma de los cuadrados de los elementos del vector debe ser 1. Así, deberá cumplirse que:

$$x(1)^2 + x(2)^2 = 1 \qquad [6]$$

y se completa así el sistema de dos ecuaciones que, para λ_1 será en este ejemplo:

$$x(1) = 1{,}3847\, x(2)$$

$$x(1)^2 + x(2)^2 = 1.$$

Despejando por el método de sustitución se encuentra que:

$$x(1) = 1{,}3847 \cdot \sqrt{1 - x(1)^2}$$

$$x(1)^2 = 1{,}3847^2\, (1 - x(1))^2 = 1{,}3847^2 - 1{,}3847^2\, x(1)^2$$

$$x(1)^2 + 1{,}3847^2\, x(1)^2 = 1{,}3847^2$$

$$x(1)^2\, (1 + 1{,}3847^2) = 1{,}3847^2$$

$$x(1) = \frac{1{,}3847}{\sqrt{1 + 1{,}3847^2}} \qquad [7]$$

y sustituyendo para encontrar x(2) se obtiene:

$$\frac{1{,}3847}{\sqrt{1 + 1{,}3847^2}} = 1{,}3847\, x(2)$$

$$x(2) = \frac{1}{\sqrt{1 + 1{,}3847^2}} \qquad [8]$$

En forma matricial esta solución puede ordenarse así:

$$\begin{bmatrix} x(1) \\ x(2) \end{bmatrix} = \frac{1}{1 + 1{,}3847^2} \begin{bmatrix} 1{,}3847 \\ 1 \end{bmatrix} = \frac{1}{1{,}7080} \begin{bmatrix} 1{,}3847 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} x(1) \\ x(2) \end{bmatrix} = \begin{bmatrix} 0{,}8107 \\ 0{,}5855 \end{bmatrix}. \qquad [9]$$

Puede confirmarse que la suma de cuadrados de los elementos del vector de la ecuación [9] cumple la condición establecida en [6]:

$$0{,}8107^2 + 0{,}5855^2 = 1{,}0000.$$

Para encontrar el vector propio generado por el segundo valor propio se utilizan las mismas fórmulas [7] y [8] y se obtiene:

$$\begin{bmatrix} x(1) \\ x(2) \end{bmatrix} = \frac{1}{1 + 0{,}7207^2} \begin{bmatrix} -0{,}7207 \\ 1 \end{bmatrix} = \frac{1}{1{,}2326} \begin{bmatrix} -0{,}7207 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} x(1) \\ x(2) \end{bmatrix} = \begin{bmatrix} -0{,}5847 \\ 0{,}8113 \end{bmatrix}. \qquad [10]$$

El lector puede confirmar que se cumple la condición establecida en [6].

El determinante de la ecuación [3], construido a partir de la matriz original, también se llama polinomio característico de la matriz B y es un polinomio de orden igual a la dimensión de la matriz. En todos los casos, el polinomio característico de una matriz será del mismo orden que la dimensión de dicha matriz y tendrá tantas soluciones como dimensiones --no todas necesariamente diferentes-- y, por lo tanto, existirán tantos valores propios como soluciones de la ecuación característica. Cada valor propio podrá generar una solución para el vector X y por esto existirán tantos vectores propios como valores propios. Cada vector propio tendrá una dimensión igual a la de la matriz original ya que deberán ser conformables.

2

EL METODO DE ANALISIS POR COMPONENTES PRINCIPALES

1. INTRODUCCION

Este es uno de los métodos de análisis más difundidos, que permite la estructuración de un conjunto de datos multivariados obtenidos de una población, cuya distribución de probabilidades no necesita ser conocida.

Se trata de una técnica matemática que no requiere un modelo estadístico para explicar la estructura probabilística de los errores. Sin embargo, si puede suponerse que la población muestreada tiene distribución multinormal, podrá estudiarse la significación estadística y será posible utilizar la muestra efectivamente observada para efectuar pruebas de hipótesis que contribuyan a conocer la estructura de la población original, con un cierto grado de confiabilidad, fijado *a priori* o *a posteriori*.

En este capítulo se explicará cómo se generan los componentes principales a fin de satisfacer el primero de los objetivos que se señalan a continuación. En el siguiente capítulo se expondrá la interpretación de los componentes encontrados y se indicará la forma de alcanzar los dos últimos objetivos.

2. OBJETIVOS

Los objetivos más importantes de todo análisis por componentes principales son:

- Generar nuevas variables que puedan expresar la información contenida en el conjunto original de datos.

- Reducir la dimensionalidad del problema que se está estudiando, como paso previo para futuros análisis.

- Eliminar, cuando sea posible, algunas de las variables originales si ellas aportan poca información.

Las nuevas variables generadas se denominan *componentes principales* y poseen algunas características estadísticas deseables, tales como independencia (cuando se asume multinormalidad) y en todos los casos no correlación. Esto significa que si las variables originales no están correlacionadas, el análisis por componentes principales no ofrece ventaja alguna.

La literatura acerca de la construcción de los componentes principales, su uso y sus propiedades es muy amplia. Casi en todos los libros de texto de análisis multivariado se dedica un capítulo al análisis por componentes principales. A manera de introducción pueden consultarse Chatfield y Collins, 1980;[8] Morrison, 1976;[28] Kendall, 1980;[22] Harris, 1975;[16] Cooley y Lohnes, 1971[7] y Seal, 1964.[32]

En Anderson, 1984;[2] Gnanadesikan, 1977;[13] Arnold, 1981;[3] Mardia y colaboradores, 1982;[23] Muirhead, 1982[29] y Srivastava y Carter,

1983,(33) se presenta un enfoque que acentúa los aspectos teóricos, aunque no por ello más complejo, que requiere cierto dominio inicial de álgebra matricial y de teoría estadística --aunque no de teoría de la medida. En los últimos cinco años ha aumentado considerablemente el número de textos dedicados al análisis multivariado aplicado a diferentes disciplinas del conocimiento.

3. ORIGENES

En 1901 Karl Pearson(30) publicó un trabajo sobre el ajuste de un sistema de puntos en un multiespacio a una línea o a un plano. Este enfoque fue retomado en 1933 por Hotelling,(20) quien fue el primero en formular el análisis por componentes principales tal como se ha difundido hasta nuestros días. El trabajo original de Pearson, 1901,(30) se centraba en aquellos componentes, o combinaciones lineales de variables originales, para los cuales la varianza no explicada fuera mínima. Estas combinaciones generan un plano, función de las variables originales, en el cual el ajuste del sistema de puntos es "el mejor", por ser mínima la suma de las distancias de cada punto al plano de ajuste.

El enfoque de Hotelling se centraba en el análisis de los componentes que sintetizan la mayor variabilidad del sistema de puntos; ello explica quizás el calificativo de "principal". Por inspección de estos componentes, que resumen la mayor proporción posible de la variabilidad total entre el conjunto de puntos, puede encontrarse un medio para clasificar o detectar relaciones entre los puntos.

Cada punto en el multiespacio p-dimensional es el extremo de un vector X tal que cada uno de sus elementos $x(j)$, para $j = 1,\ldots, p$, es una medida de la variable j-ésima en un individuo dado. Si se miden n individuos, se obtienen n vectores X y n puntos en el espacio de p dimensiones.

Desde sus orígenes, el análisis por componentes principales ha sido aplicado en situaciones muy variadas: en psicología, medicina, meteorología, geografía, ecología, agronomía.

Dicho análisis se aplica, pues, cuando se dispone de un conjunto de datos multivariados y no se puede postular, sobre la base de conocimientos previos del universo en estudio, una estructura particular de las variables. Cuando se conoce la existencia de una (o varias) variables independientes y, por lo tanto, otro conjunto de variables dependientes, pueden aplicarse las técnicas de regresión múltiple o las de regresión multivariada. Si se sabe que no existe ninguna relación entre las variables (hay independencia o, al menos, no hay correlación), habrá que abstenerse de buscar una explicación de la "relación" entre las variables, o entre los individuos a partir de dichas variables en forma conjunta. En este último caso, en estudios de tipo unidimensional se obtendrán los mismos resultados con técnicas más potentes y en forma menos tediosa, tanto desde el punto de vista computacional como por la facilidad de interpretación.

El análisis por componentes principales deberá ser aplicado cuando se desee conocer la relación entre los elementos de una población y se sospeche que en dicha relación influye de manera desconocida un conjunto de variables o propiedades de los elementos.

4. GENERACION DE LOS COMPONENTES PRINCIPALES

Se ha dicho que los componentes principales tienen ciertas características que son "deseables":

- Los componentes principales *no están correlacionados* y si, además, puede suponerse multinormalidad en los datos originales, son *independientes.*

- Cada componente principal sintetiza la máxima variabilidad residual contenida en los datos.

Al estudiar un conjunto de n individuos mediante p-variables es posible encontrar nuevas variables denominadas Y(k), k = 1,..., p que sean combinaciones lineales de las variables originales X(j), e imponer a este sistema ciertas condiciones que permitan satisfacer los objetivos del análisis por componentes principales.

Esto implica encontrar (p x p) constantes tales que:

$$Y(k) = \sum_{j=1}^{p} l(jk)\, X(j), \qquad k = 1,\ldots, p, \qquad [11]$$

donde l(jk) es cada una de esas constantes. Obsérvese que debido a la sumatoria, en cada nueva variable Y(k) intervienen todos los valores de las variables originales X(j). El valor numérico de l(jk) indicará el grado de contribución que cada variable original aporta a la nueva variable definida por la transformación lineal. Es posible que l(jk) tenga en algún caso particular el valor cero, o muy cercano a cero, lo cual indica que esa variable no influye en el valor de la nueva variable Y(k).

a. No Correlación

Sin pérdida de generalidad y para simplificar la presentación, supondremos que $E < X(j) > = 0$, j = 1,..., p.

Para satisfacer la condición de no correlación entre las nuevas variables definidas en la ecuación [11] se requiere que:

$$E < Y(k)\, Y(m) > = 0, \qquad k, m = 1,\ldots, p, \qquad k \neq m.$$

Reemplazando cada nueva variable por su definición en función de las variables originales se obtendrá:

$$E < \left(\sum_{j=1}^{p} l(jk)\, X(j) \right) \left(\sum_{h=1}^{p} l(hm)\, X(h) \right) > = 0. \qquad [12]$$

Dado que l(jk) y l(hm) son constantes, su esperanza matemática será la misma constante y, por lo tanto, pueden intercambiarse operadores con la sumatoria obteniendo una doble sumatoria:

$$E < Y(k)\, Y(m) > = \sum_{j=1}^{p} \sum_{h=1} l(jk)\, l(hm)\, E < X(j)\, X(h) >. \qquad [13]$$

Se ha identificado la constante que multiplica a cada valor de la variable original con dos subíndices diferentes (j y h) para destacar que al introducir el operador esperanza matemática en la ecuación [12] se generan todos los dobles productos posibles.

Al reemplazar en la ecuación [13] la expresión $E < X(j)\, X(h) >$ por su valor, éste será la covarianza entre las variables originales

X(j) y X(h); en otras palabras, serán los términos que quedan fuera de la diagonal principal de la matriz de varianza-covarianza.

Como se ha señalado que $k \neq m$, ya que la condición es que no exista correlación entre dos componentes principales, habrá $(1/2)p(p-1)$ restricciones sobre las constantes l(jk) que deben ser impuestas para que el sistema tenga una solución única.

Estas restricciones se establecen al aplicar a la transformación lineal definida en [11] las condiciones para que las nuevas variables originadas sean ortogonales. Si se concibe la transformación que produce los componentes principales como aquella que genera un nuevo conjunto de ejes o coordenadas que sean perpendiculares entre sí, el coseno del ángulo formado por dos cualesquiera de los ejes debe ser 0. Estas condiciones pueden expresarse así:

$$\sum_{j=1}^{p} l(jk)\, l(jm) = 0, \qquad k \neq m$$

$$\sum_{j=1}^{p} l(jk)\, l(jm) = 1, \qquad k = m$$

$$, \; k, m = 1, \ldots, p \qquad [14]$$

lo cual en álgebra vectorial se denomina "delta de Kronecker".

Es posible expresar la condición anterior en forma matricial si se define una matriz L como el arreglo de las (p x p) constantes l(jk)

$$L = \begin{bmatrix} l(11) & . & . & . & . & l(1p) \\ . & . & . & . & . & . \\ . & . & . & l(jk) & . & . \\ . & . & . & . & . & . \\ l(p1) & . & . & . & . & l(pp) \end{bmatrix}. \qquad [15]$$

Para safisfacer la condición de ortogonalidad es preciso que

$$L\,L' = L'\,L = L^{-1}\,L = I$$

y se dice, entonces, que L es una matriz ortogonal. Puede expresarse la transformación lineal de componentes principales en términos de esta matriz

$$\underset{(n \times p)}{Y} = \underset{(n \times p)}{X} \quad \underset{(p \times p)}{L} \qquad [16]$$

En la ecuación anterior se ha expresado la matriz original completa, es decir una matriz de datos como la de la definición 1, a la que se aplica la transformación ortogonal L y se obtiene una nueva matriz Y de dimensión igual a la matriz original (n x p). Para cada uno de los individuos (n en total), se calcularon nuevos valores correspondientes a las variables no correlacionadas. Esta nueva matriz Y tendrá también una matriz de varianza covarianza que será diagonal, ya que las variables --combinaciones lineales originadas por la transformación ortogonal L-- no estarán correlacionadas.

La condición de la ecuación [13] también puede expresarse en términos matriciales como:

$$E < Y'\,Y > = E < (X\,L)'\,(X\,L) >$$

Aplicando las propiedades de las operaciones con matrices

$$= E < L' X' X L >$$

e introduciendo el operador esperanza matemática, ya que **L** es una matriz de constantes

$$= L' E < X' X > L,$$

donde la esperanza de (**X' X**) es la matriz de covarianza de los datos originales. En las aplicaciones, ésta se reemplaza por **S**, su estimador muestral, con lo que la condición queda:

$$E < Y'Y > = L' S L = \Lambda, \qquad [17]$$

donde Λ es una matriz diagonal, ya que las covarianzas muestrales de las variables transformadas deben ser nulas. Esta matriz Λ tendrá en la diagonal principal los valores de las varianzas de las nuevas variables o componentes principales.

Si se premultiplican ambos miembros de la ecuación [17] por L y se recuerda que $L L' = I$, se obtiene:

$$L \Lambda = S L. \qquad [18]$$

La matriz Λ puede expresarse como:

$$\Lambda = \begin{bmatrix} \lambda(1) & 0 & 0 & 0 & 0 \\ 0 & \lambda(2) & 0 & 0 & 0 \\ . & . & \lambda(k) & . & . \\ 0 & 0 & 0 & \lambda(p-1) & 0 \\ 0 & 0 & 0 & 0 & \lambda(p) \end{bmatrix}.$$

La expresión matricial de la ecuación [18] puede desarrollarse como

$$\begin{bmatrix} l(11) & . & . & . & l(1p) \\ . & . & . & . & . \\ . & . & l(jk) & . & . \\ . & . & . & . & . \\ l(p1) & . & . & . & l(pp) \end{bmatrix} \begin{bmatrix} \lambda(1) & 0 & 0 & 0 & 0 \\ 0 & \lambda(2) & 0 & 0 & 0 \\ . & . & \lambda(k) & . & . \\ 0 & 0 & 0 & \lambda(p-1) & 0 \\ 0 & 0 & 0 & 0 & \lambda(p) \end{bmatrix} =$$

$$= \begin{bmatrix} s(11) & . & . & . & s(1p) \\ . & . & . & . & . \\ . & . & s(jj) & . & . \\ . & . & . & . & . \\ s(p1) & . & . & . & s(pp) \end{bmatrix} \begin{bmatrix} l(11) & . & . & . & . & l(1p) \\ . & . & . & . & . & . \\ . & . & . & l(jk) & . & . \\ . & . & . & . & . & . \\ l(p1) & . & . & . & . & l(pp) \end{bmatrix}.$$

Si se multiplica la primera fila de la primera matriz por la primera columna de la segunda matriz se obtiene:

$$l(11) \qquad \lambda(1)$$

debido a que se anula el resto de los términos. Si se efectúa la misma operación, pero con las dos matrices que están a la derecha del signo igual, se obtiene:

$$s(11)\ l(11) + s(12)\ l(21) + \ldots + s(1p)\ l(p1).$$

Para mantener la igualdad, los elementos homólogos de las dos matrices deberán ser iguales, es decir:

$$l(11)\ \lambda(1) = s(11)\ l(11) + s(12)\ l(21) + \ldots + s(1p)\ l(p1).$$

El segundo elemento de la primera columna de la matriz resultante se obtendrá al multiplicar la segunda fila de la primera matriz por la primera columna de la segunda matriz:

$$l(21)\ \lambda(1) = s(21)\ l(11) + s(22)\ l(21) + \ldots + s(2p)\ l(p1).$$

El p-ésimo elemento de la primera columna de la matriz resultante se obtendrá multiplicando la última fila de la primera matriz por la primera columna de la segunda matriz:

$$l(p1)\ \lambda(1) = s(p1)\ l(11) + s(p2)\ l(21) + \ldots + s(pp)\ l(p1).$$

Este sistema de p ecuaciones puede reordenarse ubicando todos los términos en un mismo miembro, igualando cada ecuación a cero y sacando factor común l(k1) entre los correspondientes términos (k = 1,..., p)

$$\begin{aligned}
0 &= \langle s(11) - \lambda(1) \rangle\ l(11) + s(12)\ l(21) + \ldots + s(1p)\ l(p1) \\
0 &= s(21)\ l(11) + \langle s(22) - \lambda(1) \rangle\ l(21) + \ldots + s(2p)\ l(p1) \\
0 &= \quad . \qquad\qquad . \qquad\qquad . \\
. &= \quad . \qquad\qquad . \qquad\qquad . \\
0 &= s(p1)\ l(11) + \ldots . \qquad + \langle s(pp) - \lambda(1) \rangle\ l(p1).
\end{aligned}$$

Este mismo sistema de ecuaciones puede expresarse en forma matricial:

$$\begin{bmatrix} s(11) & . & . & . & s(1p) \\ . & . & . & . & . \\ . & . & s(jj) & . & . \\ . & . & . & . & . \\ s(p1) & . & . & . & s(pp) \end{bmatrix} - \begin{bmatrix} \lambda(1) & 0 & 0 & 0 & 0 \\ 0 & \lambda(1) & 0 & 0 & 0 \\ 0 & 0 & \lambda(1) & 0 & 0 \\ 0 & 0 & 0 & \lambda(1) & 0 \\ 0 & 0 & 0 & 0 & \lambda(1) \end{bmatrix} \begin{bmatrix} l(11) \\ l(21) \\ . \\ . \\ l(p1) \end{bmatrix} = 0. \qquad [19]$$

Nótese que en esta ecuación sólo interviene el primer elemento de la diagonal principal de la matriz L, o sea $\lambda(1)$. Si se compara esta ecuación con la transformación de la definición 8, se verá que es igual y que la matriz B de la definición es ahora la matriz de covarianza S. Como ya se ha señalado, esta ecuación tiene tantas soluciones como dimensiones tenga la matriz S, en este caso la ecuación tendrá p soluciones.

Si este mismo procedimiento que se detalló para la primera columna de la matriz resultante de la ecuación [18] se aplica a las columnas restantes se obtiene una ecuación idéntica, lo que implica que las p soluciones se encuentran al igualar a cero el determinante de la matriz de covarianza menos λI.

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

que se conoce como *polinomio característico de la matriz* **S**. En la ecuación [19] existen p + 1 incógnitas (p valores de l(j1) y $\lambda(1)$) y sólo p ecuaciones, por lo cual el sistema no tiene solución única. Para lograrla debe agregarse la condición ya señalada en la ecuación [6] respecto a la normalización del vector propio. Así, deberá cumplirse la condición

$$\sum_{j=1}^{p} l(j1)^2 = 1$$

si se consideran los p posibles sistemas generados por la ecuación [18], se tendrán (p x p) incógnitas correspondientes a los p elementos de cada

uno de los p vectores propios (los l(jk)) y p incógnitas de los $\lambda(j)$, pero sólo (p x p) ecuaciones. El resto de las ecuaciones se encuentran a partir de la condición de normalización, ya que deberá cumplirse que:

$$\sum_{j=1}^{p} l(jk)^2 = 1, \text{ para todo } k.$$

Este sistema de ecuaciones dará lugar a p vectores que satisfacen la ecuación [19]; cada uno de éstos se denomina *vector propio*. Es importante señalar que la matriz L, formada por los p vectores propios, *no es simétrica*, ya que cada una de las columnas identifica la nueva variable o componente principal, el cual es una combinación lineal de todas las variables originales.

Conocida la matriz L es posible posmultiplicar la matriz original de observaciones X por L y obtener una nueva matriz de datos Y, tal como se definió en la ecuación [16]

$$\underset{(n \times p)}{Y} = \underset{(n \times p)}{X} \quad \underset{(p \times p)}{L}$$

b. Máxima Variabilidad

Por la forma en que son generados los componentes principales, también satisfacen la condición de sintetizar en forma decreciente la varianza del conjunto original de datos.

Si se desea hallar la combinación lineal que sintetice la máxima varianza, deberá encontrarse el máximo de la expresión de la varianza de la ecuación [11] que define un componente principal

$$Y = \sum_{j=1}^{p} l(j)\, X(j),$$

de la cual sólo se ha eliminado el subíndice que identifica a la nueva variable:

$$\text{Var}(Y) = \text{Var}\left[\sum_{j=1}^{p} l(j)\, X(j)\right]. \qquad [20]$$

Para calcular la varianza de la expresión entre corchetes debe recordarse que las variables X(j) están correlacionadas, por lo cual a la varianza de las X(j) deberán sumarse las covarianzas. Si se recuerda que la covarianza s(jh) es igual a la covarianza s(hj), la expresión queda como sigue:

$$\text{Var}(Y) = \sum_{j=1}^{p} l(j)^2\, s(jj) + 2 \sum_{\substack{j=1 \\ h=2 \\ j<h}}^{p} l(j)\, l(h)\, s(jh)$$

si se recuerda que l(j) es una constante y que la varianza del producto de una variable aleatoria por una constante es igual al producto de la varianza de dicha variable aleatoria por el cuadrado de la constante. La sumatoria del segundo término se efectúa para todos los valores de j y h diferentes, siendo j menor que h. Esta misma ecuación puede escribirse:

$$\text{Var}\ (Y) \;=\; \sum_{j=1}^{p} l(j)^2 s(jj) + \sum_{\substack{j=1 \\ h=1 \\ j \neq h}}^{p} l(j)\ l(h)\ s(jh);$$

de ella sólo se ha eliminado el 2 y se ha efectuado la suma para todas las combinaciones de j y h diferentes. Es posible sintetizar aún más la ecuación anterior si se agrupan los dos términos en uno:

$$\text{Var}\ (Y) \;=\; \sum_{\substack{j=1 \\ h=1}}^{p} l(j)\ l(h)\ s(jh), \qquad [21]$$

donde j y h asumen todas las combinaciones de valores posibles. En la ecuación anterior se observa que en la determinación de la varianza de un componente intervienen las varianzas y las covarianzas (o sea los valores de s(jh)) entre todas las variables originales. Las constantes l(j) y l(h) que intervienen en la ecuación son los valores de las p constantes que forman el vector asociado con esa nueva variable o componente principal. Es decir que son los elementos del vector de la ecuación [19], o los elementos de *una columna* de la matriz L definida en [15].

Es necesario encontrar el máximo de la ecuación [21], con sujeción a las restricciones expresadas en [14]. En este caso, se trata de los elementos de un mismo vector y, por lo tanto, siguiendo la nomenclatura de la ecuación [21] será:

22

$$\sum_{j=h=1}^{p} l(j)\ l(h) \;=\; 1, \qquad [22]$$

o sea, la suma de los cuadrados de los valores debe ser 1.

Derivando la ecuación [21] respecto a los valores de l(j), teniendo en cuenta la restricción de la ecuación [22] mediante el uso de multiplicadores de Lagrange, se obtiene:

$$\partial \Big< \sum_{\substack{j=1 \\ h=1}}^{p} l(j)\ l(h)\ s(jh) - g\Big(\sum_{j=1}^{p} l(j)^2 - 1\Big) \Big> / \partial\, l(j) = 0. \qquad [23]$$

Esta ecuación puede derivarse respecto a p posibles valores de l(j). Supóngase que j=1, se obtendrá:

$$\sum_{h=1}^{p} l(h)\ s(1h) + \sum_{h=1}^{p} l(h)\ s(h1) - g\ (2\ l(1)) \;=\; 0,$$

ya que, como se ha visto al deducir la ecuación [21], la doble sumatoria incluye todas las combinaciones posibles. Al examinar la ecuación anterior y teniendo en cuenta las propiedades de simetría de la matriz de covarianza puede observarse que:

$$2 \sum_{h=1}^{p} l(h)\ s(1h) - 2\ g\ l(1) \;=\; 0, \qquad [24]$$

de donde puede eliminarse, sin alterar, el número 2 en ambos miembros. A partir de la ecuación [23], y derivando respecto a l(2),..., l(p)

podrán encontrarse ecuaciones similares a la [24] que, en conjunto, formarán un sistema donde la incógnita es el valor de g. Este sistema será:

$$\sum_{h=1}^{p} l(h)\, s(1h) - g\, l(1) = 0$$

$$\sum_{h=1}^{p} l(h)\, s(2h) - g\, l(2) = 0$$

$$\cdot \quad .. \quad .. \quad = \quad .$$
$$\cdot \quad .. \quad .. \quad = \quad .$$

$$\sum_{h=1}^{p} l(h)\, s(ph) - g\, l(p) = 0$$

el cual puede expresarse sin alterarlo como:

$$\sum_{h=1}^{p} l(h)\, s(1h) = g\, l(1)$$

$$\sum_{h=1}^{p} l(h)\, s(2h) = g\, l(2)$$

$$.. \quad = \quad ..$$
$$.. \quad = \quad ..$$

$$\sum_{h=1}^{p} l(h)\, s(ph) = g\, l(p)$$

Cada una de las ecuaciones de este sistema puede expresarse como el producto de dos vectores, los cuales podrán arreglarse luego en forma matricial. Para la primera ecuación será:

$$< l(1)\ l(2)\ \ldots\ l(p) > \begin{bmatrix} s(11) \\ s(21) \\ . \\ . \\ s(p1) \end{bmatrix} = l(1)\ g$$

que podrá repetirse para cada una de las restantes. Estos vectores pueden disponerse en una matriz

$$< l(1)\ l(2)\ \ldots\ l(p) > \begin{bmatrix} s(11) & s(12) & \ldots & s(1p) \\ s(21) & s(22) & \ldots & s(2p) \\ . & . & \ldots & . \\ . & . & \ldots & . \\ s(p1) & . & \ldots & s(pp) \end{bmatrix} = < l(1)\ l(2)\ \ldots\ l(p) > g.$$

Sin restarle generalidad, es posible expresar el segundo término de la ecuación anterior en forma matricial y obtener:

$$< l(1)\ l(2)\ \ldots\ l(p) > \begin{bmatrix} s(11) & s(12) & \ldots & s(1p) \\ s(21) & s(22) & \ldots & s(2p) \\ . & . & \ldots & . \\ . & . & \ldots & . \\ s(p1) & . & \ldots & s(pp) \end{bmatrix} =$$

$$< l(1)\ l(2)\ \ldots\ l(p) > \begin{bmatrix} g & 0 & \ldots & 0 \\ 0 & g & \ldots & 0 \\ . & . & .g. & . \\ . & . & \ldots & . \\ 0 & . & \ldots & g \end{bmatrix} . \qquad [25]$$

Reordenando la ecuación [25], igualando a cero y sacando el vector de las constantes como factor común a la izquierda, se obtiene:

$$< l(1)\ l(2)\ \ldots\ l(p) > \left(\begin{bmatrix} s(11) & s(12) & \ldots & s(1p) \\ s(21) & s(22) & \ldots & s(2p) \\ . & . & \ldots & . \\ . & . & \ldots & . \\ s(p1) & . & \ldots & s(pp) \end{bmatrix} - \begin{bmatrix} g & \ldots & 0 \\ 0 & \ldots & 0 \\ . & .g. & . \\ . & \ldots & . \\ 0 & \ldots & g \end{bmatrix} \right) = 0.$$

$(1 \times p) \qquad (p \times p) \qquad (p \times p) = (1 \times p)$

A la izquierda del signo igual (=) se encuentra un vector para el cual existen p incógnitas, una matriz cuyos valores son conocidos ya que representan las varianzas y covarianzas de las variables originales, y una segunda matriz con una incógnita, g. Existen p ecuaciones y (p + 1) incógnitas. En consecuencia, para que el sistema tenga solución única, debe encontrarse otra ecuación. Esta es la condición que se ha impuesto a las l(j), cuya suma elevada al cuadrado debe ser 1 (ecuación [14]). Es decir, el sistema tiene solución única.

Es posible reordenar este sistema y obtener la transpuesta de ambos miembros, es decir transponer filas por columnas, y se obtiene, entonces, un vector (p x 1), o sea un vector columna, en vez del vector fila:

$$\left(\begin{bmatrix} s(11) & s(12) & \ldots & s(1p) \\ s(21) & s(22) & \ldots & s(2p) \\ . & . & \ldots & . \\ . & . & \ldots & . \\ s(p1) & . & \ldots & s(pp) \end{bmatrix} - \begin{bmatrix} g & 0 & \ldots & 0 \\ 0 & g & \ldots & . \\ . & . & .g. & . \\ . & . & \ldots & . \\ 0 & . & \ldots & g \end{bmatrix} \right) \begin{bmatrix} l(1) \\ l(2) \\ . \\ . \\ l(p) \end{bmatrix} = 0. \qquad [26]$$

$(p \times p) \qquad (p \times p) \qquad (p \times 1)\ (p \times 1)$

La diferencia entre las dos matrices no se altera ya que es simétrica; la matriz y el vector cambian posiciones a fin de hacer conformable la operación (ésta es una propiedad de las operaciones matriciales).

Si se compara la ecuación [26], que refleja la condición que debe cumplir una transformación lineal para sintetizar la máxima variabilidad, con la ecuación [19], que expresa la condición que debe satisfacer una transformación lineal para que las variables resultantes no estén correlacionadas, se observará que son idénticas, excepto que en un caso los valores de una de las incógnitas se han llamado $\lambda(1)$, y en el otro, g. En ambos casos, y como ya se demostró para la ecuación [19], tiene p soluciones que son los valores propios generados por **S**, cada uno de los cuales origina un conjunto de valores para el vector de constantes l(j).

La transformación lineal que sintetiza la máxima variabilidad corresponderá, pues, a la generada por el valor de $\lambda(j)$ que sea mayor. Convencionalmente esta solución máxima se ha denominado $\lambda(1)$ y la notación se utiliza de manera que se cumpla:

$$\lambda(1) \geq \lambda(2) \geq ,\ldots, \geq \lambda(p). \qquad [27]$$

Así, la primera transformación lineal, o primera variable generada, o primer componente principal, sintetiza la máxima variabilidad posible en el conjunto de datos originales. La segunda transformación lineal, o segundo componente principal, sintetiza la máxima variabilidad residual, sujeta a la condición de no correlación con el primer componente principal, y así hasta el p-ésimo componente.

5. NUEVA EXPRESION DE LOS DATOS

Si se conocen los valores propios generados por la matriz de covarianza de un conjunto de datos, es posible calcular todas las constantes que forman la matriz L de transformación, definida en la ecuación [16]. Una vez encontrada esta matriz, es posible posmultiplicar la matriz original X y obtener una nueva matriz de datos Y.

Esta matriz de datos transformada tendrá las características deseables que se mencionaron antes:

- Para cada observación o individuo tendrá p valores que corresponden a cada uno de los componentes principales o nuevas variables.

- La matriz de covarianza de este conjunto de datos será diagonal (ya que las nuevas variables no están correlacionadas) y los valores de las varianzas de cada variable serán los valores propios encontrados al resolver el polinomio característico de la matriz de covarianza de los datos originales.

- La varianza del primer componente principal será la mayor, y cada uno de los siguientes componentes tendrá una varianza menor, hasta que el último componente será el que posea la menor varianza.

- El vector promedio de la nueva matriz también experimentará la misma transformación lineal:

$$\underset{(1 \times p)}{\bar{Y}'} = \underset{(1 \times p)}{\bar{X}'} \underset{(p \times p)}{L} \qquad [28]$$

Se utiliza la transpuesta de los vectores promedio, ya que en la definición 2 $\bar{X}$ es un vector columna.

En esta nueva expresión los datos cumplen una serie de *propiedades* que pueden sintetizarse en:

a) $E < Y(k) > = E < X' > l(k)$, donde $l(k)$ es el k-ésimo vector propio.

b) $Var < Y(k) > = \lambda(k)$, donde $\lambda(k)$ es el k-ésimo valor propio.

c) $Cov < Y(k), Y(m) > = 0$, para $k \neq m$.

d) $Var [Y(i)] \geq Var [Y(2)] \geq \ldots \geq Var [Y(p) \geq 0$.

e) $\sum_{k=1}^{p} Var < Y(k) > = tr\ S$.

$$f)\ \prod_{k=1}^{p} \mathrm{Var} < Y(k) > = |S|.$$

Si se desea que la media de las nuevas variables generadas sea cero, deberá efectuarse una traslación, es decir restarle a la matriz Y de los nuevos datos el valor de los promedios de las nuevas variables como se definieron en a), calculados a partir de la ecuación [28]. Así, una matriz de datos transformados y con media cero será

$$Y^* = Y - \bar{X}' L = X L - \bar{X}' L = Y^* = [X - 1\bar{X}'] L \qquad [29]$$

al reemplazar Y por su expresión en función de los datos originales, donde 1 representa un vector columna de números 1 que, al multiplicarse por el vector fila $\bar{X}'$, conforma una matriz que permite centrar las variables originales; luego, se saca a la derecha la matriz de transformación como factor común L y se obtiene la ecuación [29].

6. USO DE LA MATRIZ DE CORRELACION

Hasta ahora se ha utilizado siempre la matriz de covarianza S de los datos originales presentada en la definición 3. Es posible calcular los valores y los vectores propios y, por lo tanto, la matriz de transformación L si se emplean los datos estandarizados, en cuyo caso la matriz de covarianza será la matriz de correlación de la definición 4.

Los valores de la diagonal principal de R --la matriz de correlación-- son números 1, ya que las nuevas variables estandarizadas poseen varianza unitaria. Esto significa que en el conjunto de datos a partir del cual se generarán los componentes principales se otorga la misma importancia a todas las variables observadas. Esta situación puede o no ser deseable, pero importa destacar que el uso de la matriz de correlación implica una ponderación de las variables originales, otorgándole a cada una la misma importancia, independientemente de los valores relativos de sus varianzas. En el capítulo 4 se describe el efecto de la estandarización sobre los resultados del análisis por componentes principales.

Cuando se utiliza la matriz R, cambian algunas de las propiedades de los valores y vectores propios generados. De las propiedades antes mencionadas cabe destacar que las correspondientes a e) y f) deben expresarse en términos de R:

$$e)\ \sum_{k=1}^{p} \mathrm{Var} < Y(k) > = \mathrm{tr}\ R = p.$$

ya que la matriz R tiene números 1 en la diagonal principal y es de dimensión p;

$$f)\ \prod_{k=1}^{p} \mathrm{Var} < Y(k) > = |R|.$$

La matriz de transformación L así generada será diferente de la obtenida a partir de S. Esta característica de los valores y vectores propios determinan que el análisis por componentes principales sea sensible a los cambios de escala. Por este motivo, es muy importante examinar cuidadosamente los datos originales en sus promedios y sus varianzas y covarianzas a fin de poder decidir qué tipo de matriz conviene utili-

zar, qué cambio de escala puede introducirse antes de encontrar S o R y qué interpretación deberá darse a los componentes encontrados.

Si se utiliza la matriz de correlación para generar los componentes principales deberá usarse la matriz X estandarizada (variables con media cero y varianza unitaria) para aplicar la ecuación [16], ya que el método de componentes principales es sensible a los cambios de escala y será imposible obtener nuevas variables no correlacionadas (componentes principales) si se emplean expresiones diferentes para el cálculo de la matriz L y de la matriz Y. Si se utilizan los datos estandarizados no es necesario corregir según la ecuación [29], ya que la media de cualquiera de las variables originales será cero.

3

INTERPRETACION DE LOS COMPONENTES PRINCIPALES

1. INTRODUCCION

No importa cuán simple o complicado sea un método de análisis de datos; una vez concluida la manipulación algebraica, es necesario interpretar correctamente los resultados obtenidos. En los análisis más difundidos, como el análisis de varianza o la regresión lineal simple, luego de lo que para muchos son cálculos tediosos que hoy día pueden hacerse rápida y eficientemente con computadoras de mesa y hasta con calculadoras de bolsillo, hay que interpretar un par de estimadores, por ejemplo: la ordenada en el origen y la pendiente en una ecuación de regresión. Para verificar la hipótesis acerca de los valores encontrados basta generalmente comparar un valor calculado con otro tabulado.

En el análisis por componentes principales es necesario calcular e *interpretar* tanto los valores propios generados como los vectores propios. Deberá decidirse cuántos valores propios serán considerados si se desea reducir la dimensión original de p variables a m (siendo m < < p). Habrá que ser muy cuidadoso al interpretar los vectores propios, ya que el método no es independiente de la escala de medición de las variables originales.

En este capítulo se presentarán los criterios para seleccionar el número de componentes a considerar cuando se reduce la dimensión original, la correlación entre las variables originales y los componentes principales, como asimismo algunos casos de aplicación directa del método a la utilización de componentes no lineales, a la solución de los problemas de multicolinealidad en regresión y a la detección de marginales.

2. SELECCION DEL NUMERO DE COMPONENTES

Se ha dicho que la suma de las varianzas de las variables originales, es decir la traza de S, es igual a la suma de los valores propios de la matriz S. A su vez, la varianza de cada componente principal es el valor propio que le dió origen; así se cumplirá que

$$\sum_{j=1}^{p} s(jj) = \sum_{k=1}^{p} \lambda(k) \qquad j,k = 1,\ldots, p,$$

lo que se deduce de las propiedades b) y e) mencionadas en el capítulo 2.

Cada componente principal explica una proporción de la variabilidad total y esa proporción puede calcularse mediante el cociente entre el valor propio y la traza de S. Este cociente se denomina proporción de la variabilidad total explicada por el componente k-ésimo y se calcula así:

$$\frac{\lambda(k)}{tr\ S} = \text{variación explicada.} \qquad [30]$$

Como los valores propios se ordenan en forma creciente, es posible seleccionar los primeros m valores propios (siendo $m < p$) y la *eficiencia* del ajuste de los datos originales por los nuevos m componentes principales estará dada por la proporción de la variación total explicada por la suma de los m primeros valores propios.

$$\frac{\sum_{k=1}^{m} \lambda(k)}{\text{tr } S} 100 = \text{porcentaje de la variación total} \qquad [31]$$

Tanto la ecuación [30] como la [31] pueden expresarse en forma de proporción o de porcentaje; en el primer caso, la variación total del conjunto original de datos será la unidad y en el segundo, será 100.

Aplicando la propiedad e) (punto 5, capítulo 2) a las ecuaciones anteriores es posible expresar la variación explicada por cada componente o por los m primeros componentes en función de la suma total de los valores propios. Así se obtiene:

$$\frac{\lambda(k)}{\sum_{k=1}^{p} \lambda(k)} 100 = \text{porcentaje de variación explicada por el k-ésimo componente}$$

$$\frac{\sum_{k=1}^{m} \lambda(k)}{\sum_{k=1}^{p} \lambda(k)} 100 = \text{porcentaje de variación explicada por los m primeros componentes}$$

Cuando se consideran todos los componentes, es decir si $m = p$, la proporción de la variación explicada es 1 y el porcentaje es 100%.

Al decidir cuántos componentes se mantienen en una situación particular, deberá examinarse cuántos componentes son necesarios incluir para que el porcentaje de variación explicada sea satisfactorio. No es posible aplicar una prueba de hipótesis que tenga validez para toda situación y que permita decidir cuándo se ha alcanzado el "nivel satisfactorio". En principio, se recomienda preparar un gráfico donde se represente el porcentaje de variación explicada por cada componente en las ordenadas y los componentes en orden decreciente en las abscisas (véase la Fig. 2).

Pueden presentarse varias situaciones respecto de las cuales se dará un resultado para cinco variables originales y, por lo tanto, cinco valores propios y cinco vectores propios. La Tabla II muestra los porcentajes de la varianza total explicada por cada componente en cuatro casos designados a), b), c) y d).

Tabla II. Variación explicada por cada componente

Caso	Porcentaje de la Variación Total Explicada por el Componente					Suma Total
	1°	2°	3°	4°	5°	
a)	35	30	28	4	3	100
b)	45	30	9	8	8	100
c)	75	7	7	6	5	100
d)	22	21	20	19	18	100

En el caso a) los primeros tres componentes sintetizan porcentajes similares de la varianza total y los últimos dos contribuyen con valores mucho más pequeños; podrían considerarse sólo los tres primeros y se habría explicado el 93% de la variación total, lo cual indica que sin perder casi información la dimensión del problema puede reducirse de cinco a tres.

En el caso b), hay una disminución del primero al segundo componente y los últimos tres aportan muy poco; si se consideran los dos primeros componentes se explicaría el 75% de la variación original, resultado no muy bajo y que para ciertos casos puede ser utilizado. Procede señalar que los criterios de selección no son fijos y que dependerán de la posible interpretación que se haga de los componentes generados.

La misma proporción es explicada en el caso c) por el primer componente; los cuatro componentes restantes contribuyen con valores bajos y muy similares entre sí. En este caso, sólo se tomará en cuenta el primer componente, aunque debe prestarse suma atención a la representación de las diferentes variables originales en este componente, lo cual puede medirse mediante su correlación con las variables, como se ilustrará más adelante.

En el último caso todos los componentes sintetizan proporciones similares de la variación total y, en consecuencia, es difícil decidir cuántos componentes deberán seleccionarse o si debe aceptarse que la verdadera dimensión del problema analizado es la original. Puede facilitar esta decisión la inspección de la correlación con las variables originales y la inspección de la matriz S o R utilizada.

En la figura 2 se han graficado los valores de la Tabla II. De la inspección visual puede deducirse que deberán considerarse los componentes anteriores al punto de inflexión de la curva.

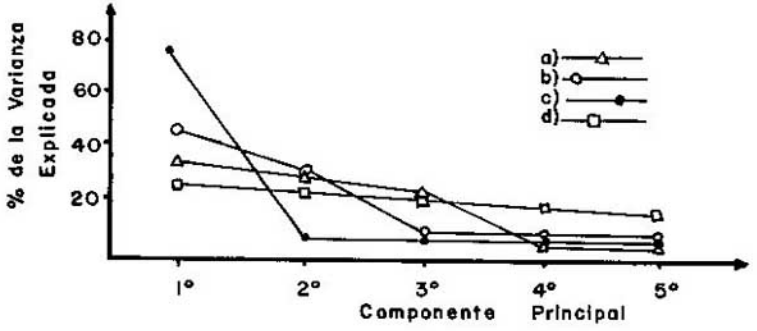


Fig. 2. Proporción de la variación explicada por cada componente. En todos los casos, los cinco componentes sintetizan el 100% de la varianza de las variables originales.

Otro criterio para seleccionar el número de componentes consiste en incluir sólo aquellos cuyos valores propios sean superiores al promedio. Si se utiliza la matriz R, se incluirán los componentes cuyos valores propios sean mayores que 1. Sin embargo, este criterio, debido a Kaiser

(citado por Mardia y colaboradores[23]), tiende a incluir muy pocos componentes cuando el número original de variables es inferior a 20.

El criterio gráfico, presentado en la figura 2, fue sugerido por Cattell (1966)[6] y tiene el efecto contrario, ya que tiende a incluir un número alto de componentes. Por esto, se requiere una solución de compromiso y una inspección detallada de las correlaciones.

3. CORRELACION ENTRE VARIABLES ORIGINALES Y COMPONENTES PRINCIPALES

En el caso bivariado, por ejemplo al estudiar un modelo lineal de relación entre dos variables X e Y, como en la regresión lineal simple, la correlación se calcula así

$$r = \frac{cov(x,y)}{var(x)\ var(y)} \quad . \qquad [32]$$

El cuadrado de la expresión [32], conocido como coeficiente de determinación, constituye una medida de la asociación entre las dos variables. Esta medida podrá ser contrastada por una prueba de hipótesis para valores cercanos a cero cuando se supone binormalidad; de efectuarse una prueba para valores absolutos elevados del coeficiente de correlación (varía entre -1 y +1), deberá disponerse de una muestra grande y aplicarse la conocida distribución asintótica del estadístico z de Fisher calculado a partir de r.

Para estudiar la correlación entre las variables originales y los componentes principales habrá que calcular todas las correlaciones de cada variable original con cada nueva variable.

Sean **x** el vector de dimensión (p x 1) de las p variables originales e **y** el vector, también (p x 1) de las nuevas variables --combinaciones lineales de las originales--, debe encontrarse una expresión de la esperanza del producto de los vectores, y reemplazando el vector y por su expresión en función de la transformación lineal según la ecuación [16]

$$E < \mathbf{x},\mathbf{y}' > = E < \mathbf{x},\mathbf{x}'\mathbf{L} >$$

si se aplica la ecuación [18] y se recuerda que el estimador de Σ es S, ya que Σ es la matriz de covarianza de las variables originales:

$$E < \mathbf{x},\mathbf{y}' > = E < \mathbf{x},\mathbf{x}' > \mathbf{L} = \Sigma\,\mathbf{L} = \mathbf{L}\mathbf{\Lambda}\mathbf{L}'\,\Sigma\,\mathbf{L} = \mathbf{L}\mathbf{\Lambda}.$$

Así la covarianza entre X(j) e Y(k) será el elemento de la matriz $\mathbf{L}.\mathbf{\Lambda}$ ubicado en la posición (jk). La matriz **L** posee en la posición (jk) el elemento j-ésimo del k-ésimo vector propio, es decir l(jk). La matriz $\mathbf{\Lambda}$ es diagonal y los elementos no nulos son los valores propios de la matriz S, por lo cual

$$Cov < X(j),\ Y(k) > = l(jk)\ \lambda(k) \quad j,k = 1,\ldots,p \qquad [33]$$

Para calcular la correlación deberá dividirse por la raíz de las varianzas de las variables, es decir por s(jj) y por $\lambda(k)$, y se obtendrá:

$$r(jk) = \frac{l(jk)\ \lambda(k)}{\sqrt{s(jj)\ \lambda(k)}}$$

que puede expresarse por:

$$r(jk) = l(jk) < \lambda(k) / s(jj) >^{1/2} \qquad [34]$$

Si se utiliza la matriz de correlación de los datos originales R, las varianzas s(jj) serán unitarias y la fórmula [27] se reduce a

$$r(jk) = l(jk) < \lambda(k) >^{1/2} \qquad [35]$$

Las ecuaciones [34] y [35] representan la correlación entre la variable original X(j) y el k-ésimo componente principal. El cuadrado del coeficiente de correlación es una medida de la asociación entre ellos y una manera de cuantificar la proporción de la variación total de una variable original explicada por el componente k. Obsérvese que el denominador de la ecuación [34] es la raíz de la varianza de la variable X(j); así, al elevar al cuadrado, se obtiene una ponderación de la variación explicada por la combinación lineal k-ésima.

Si se efectúa una sumatoria en k, es decir se suman las proporciones de la varianza explicada por los p componentes principales para la variable original X(j), se obtendrá el valor 1. Puede efectuarse la sumatoria para los m primeros componentes incluidos luego de la selección destinada a reducir la dimensión de un conjunto de datos y determinar cuál es la proporción de la varianza de cada variable original considerada en el nuevo subconjunto. Esto es posible ya que los componentes principales no están correlacionados entre sí; de lo contrario, no se aplicaría lo anterior por no tomar en cuenta las covarianzas. Así, en términos de la matriz de covarianza será:

$$r^2(jk) = \frac{\lambda(k)\ l^2(jk)}{s(jj)} \qquad j,k = 1,\ldots, p \qquad [36]$$

para el k-ésimo componente, y sumando para los primeros m:

$$\sum_{k=1}^{m} r^2(jk) = \frac{1}{s(jj)} \sum_{k=1}^{m} \lambda(k)\ l^2(jk).$$

Del análisis detallado de estas proporciones y de los elementos de cada vector propio podrán inferirse las conclusiones necesarias para explicar la estructuración de un conjunto de datos multivariados.

4. PRUEBA DE HIPOTESIS PARA LOS VALORES PROPIOS

Al seleccionar los componentes principales a considerar en una descripción puede ser útil disponer de una prueba para afirmar que los últimos (p - m) componentes son cero. Esto implicaría que A es, en realidad, de rango m (existen p - m variables originales que son combinaciones lineales de las restantes m). Sin embargo, en ese caso también S, la matriz de covarianza estimada a partir de la muestra de n observaciones, sería de rango m con probabilidad 1 y una prueba de que l(p) --el menor valor propio estimado-- es igual a cero sería trivial (Mardia y colaboradores, 1979).[23]

Cuando puede suponerse *multinormalidad* en la población de la cual se extrae la muestra será posible verificar dos tipos de hipótesis si se calculan los valores propios de la matriz S.

a. La proporción de la variación explicada por los últimos p - m componentes es menor que un cierto valor. Esto sería:

$$Ho: \frac{\sum_{k=1}^{m} \lambda(k)}{\sum_{k=1}^{p} \lambda(k)} = W \qquad m < p.$$

$$Ho: \frac{\sum_{k=1}^{m} \lambda(k)}{\sum_{k=1}^{p} \lambda(k)} < W$$

El estimador muestral de W, llámese w, seguirá una distribución aproximadamente normal con media W y varianza

$$Var(w) = T^2 = \frac{2 \operatorname{tr} \Sigma}{(n-1)(\operatorname{tr} \Sigma)^2} (W^2 - 2aW + a^2),$$

donde

$$a = \frac{\sum_{k=1}^{m} \lambda^2(k)}{\sum_{k=1}^{p} \lambda^2(k)} \qquad m < p.$$

Es posible estimar T^2 utilizando la matriz de covarianza S y los valores propios de esa matriz. La raíz cuadrada del valor estimado, llámese t, puede ser usada para construir un estadístico de prueba

$$z = \frac{w - W}{t}$$

que tiene distribución normal tipificada, es decir con media cero y varianza 1, para lo cual existen tablas de probabilidad que permiten tomar decisiones acerca de la hipótesis planteada y construir intervalos de confianza si se desea.

b. Probar la hipótesis de que los últimos p - m valores propios son iguales

$$Ho: \lambda(p) = \lambda(p-1) = \ldots = \lambda(m+1).$$

Esta prueba se conoce también como la prueba de isotropía ya que implica que en las últimas p - m dimensiones los datos están dispersos en una hiperesfera y, por lo tanto, el incluir uno de los componentes en el análisis debería implicar la inclusión de todos los restantes.

El estadístico utilizado para probar esta hipótesis se obtiene por el método de la razón de verosimilitud y la función $-2 \log L^* = np(a - 1 - \log g)$ se distribuye aproximadamente como chi-cuadrado, siendo L^* la razón de verosimilitud cuando se supone multinormalidad, n el tamaño de la muestra, p el número total de variables observadas, a la media aritmética de los valores propios y g la media geométrica de los valores propios.

Después de un manejo algebraico es posible utilizar

$$a'' = \frac{\sum_{k=m+1}^{p} l(k)}{(p - m)}$$ donde $l(k)$ es el estimador de $\lambda(k)$,

a'' es la media aritmética de los últimos $p - m$ valores propios estimados que corresponden a los incluidos en la hipótesis nula, y

$$g'' = < l(m + 1) \cdot l(m + 2) \ldots l(p) >^{(1/(p - m))}$$

es la media geométrica de los últimos $p - m$ valores propios estimados; y el estadístico puede escribirse como

$$-2 \log L^* = np(a'' - 1 - \log g'').$$

Barlett propone una aproximación (citado por Mardia y colaboradores, 1979[23]), de forma tal que la ecuación será

$$\chi^2 = \left(n - \frac{2p + 11}{6}\right) (p - m) \log \frac{a''}{g''}$$

que se distribuye asintóticamente como chi-cuadrado con $(1/2)(p - m + 2)(p - m - 1)$ grados de libertad.

5. ALGUNAS APLICACIONES DEL ANALISIS POR COMPONENTES PRINCIPALES

Cuando se aplican los métodos estadísticos a casos concretos en diferentes disciplinas del conocimiento, el estadístico está obligado a explicar cuidadosamente el significado de los resultados alcanzados. Asimismo, éste deberá sondear al especialista hasta que responda a todos los interrogantes acerca de las características de la información disponible y del uso que se dará a los resultados obtenidos.

En el campo multivariado esto es aún más necesario ya que sus métodos y técnicas no están todavía muy difundidas y cualquier interpretación debe ser en extremo cuidadosa porque las técnicas utilizadas son muy poderosas.

El problema se asemeja al de aquel padre que concurre al colegio de su hijo e interroga al maestro: ¿Cómo se desempeña mi hijo? La pregunta no podrá ser contestada con propiedad hasta que el padre aclare para qué necesita la información: ¿Para predecir su triunfo en los estudios? ¿en su vida social? ¿en los deportes?

Es preciso comprender que cuando el estadístico enfrenta la solución de un problema construye un *modelo artificial* que debe reflejar los supuestos y los objetivos del estudio. En el caso multivariado ello equivale a formular una *combinación matemática artificial* de variables aleatorias observadas.

Nunca se destacará en demasía la naturaleza artificial de estos modelos. Serán útiles para elucidar fenómenos y mecanismos de funcionamiento, pero muy peligrosos si se les otorga existencia física real (Bargmann, 1969[4]). Y como el ser humano interviene en la definición del universo y de las variables-respuesta, probablemente nunca se encuentre un modelo matemático único. Por ello, antes de aplicar las técnicas habrá que definir claramente el propósito del modelo que se haya de establecer.

El análisis por componentes principales es esencialmente descriptivo y tiene una interpretación geométrica, a partir de Pearson (1901[30]),

en planos de mejor ajuste y vectores de máxima concentración, en función de distancias euclidianas. A continuación se resumen algunas de las aplicaciones derivadas del análisis de componentes principales.]

5.1. Análisis de Componentes Principales no Lineales

Si el conjunto de datos que se analiza tiene una expresión lineal (o casi lineal) singular, es posible determinar un sistema de coordenadas lineales que se ajuste mejor a la configuración de los datos. Así, su expresión en el nuevo sistema es más simple ya que la descripción podrá efectuarse con menor número de coordenadas. Este es el objetivo del análisis por componentes principales.

Sin embargo, si el conjunto de datos tiene una configuración no lineal (o cercana a la no linealidad) singular, el objetivo será encontrar un nuevo sistema de coordenadas no lineales que se ajuste mejor a la distribución de los datos en el multiespacio original.

Este procedimiento de selección de combinaciones no lineales será similar al del análisis por componentes principales, ya que deberá escogerse la combinación no lineal que sintetice la máxima variación de los datos y luego obtener otra, no correlacionada con la anterior, cuya varianza sea la siguiente máxima, y así sucesivamente (Gnanadesikan y Wilk, 1969[12]). Estos autores han elaborado dos ejemplos, con datos obtenidos a partir de configuraciones no lineales.

5.2. Uso de Componentes Principales en Regresión

El uso de los componentes principales para ajustar una ecuación de regresión (múltiple o multivariada) fue propuesto por Kendall,[21] en 1957, y posteriormente por Marquardt (1970)[24] para aquellas situaciones en que las variables independientes presentan colinealidad. Aunque estos autores no lo mencionan, los estimadores generados son sesgados, pero la estimación es mejor que la obtenida por el método de los mínimos cuadrados.

Greenberg (1975)[14] examina las propiedades de los estimadores así generados y llega a la conclusión de que al incluir en el modelo los últimos componentes principales (pequeños valores propios) aumenta la varianza de los estimadores, aunque, si hay una elevada correlación de éstos con la variable dependiente (o con las variables dependientes), disminuye el sesgo. Gunst y colaboradores (1976)[15] compararon el método de mínimos cuadrados con el de componentes principales y concluyeron que en presencia de multicolinealidad es preferible el segundo método, tanto para estimar los parámetros como para seleccionar variables.

Entre los criterios de formación de estimadores sesgados, la regresión Ridge es cada vez más utilizada. Hocking (1976)[19] demuestra cómo puede relacionarse la regresión Ridge con la regresión en componentes principales utilizando la caracterización de la primera sugerida por Allen (1974).[1]

Según Weisberg (1980)[36] si se parte de un modelo de regresión en función de los componentes principales a los cuales identifica con z, se obtiene un modelo de la forma:

$$Y = \beta z + \varepsilon$$

siendo Y la variable dependiente, β el vector de los coeficientes de regresión, z la matriz de los componentes principales y ε el vector del componente aleatorio del modelo. Si $Var(e) = \sigma^2$, es po-

sible demostrar que $\hat{\beta}_1$, el valor estimado de β_1, tiene menor varianza que cualquier otro coeficiente estimado con cualquier combinación lineal de las variables originales X. La varianza de $\hat{\beta}_1$ será σ^2/λ_1. Del mismo modo entre todas las posibles combinaciones lineales de las X, ortogonales a z(1) (el primer componente principal), la especificada por z(2) tiene un $\hat{\beta}_2$ con varianza σ^2/λ_2 menor que cualquier otra combinación lineal. Esto se repite para todos los demás componentes principales.

Como se cumple que

$$\sigma^2/\lambda_1 \leq \sigma^2/\lambda_2 \leq \ldots \leq \sigma^2/\lambda_p$$

la cantidad de información sintetizada en el primer coeficiente β_1 será mayor que en β_2, y ésta que la de β_3, y así sucesivamente.

Sin embargo, el uso de componentes principales en el modelo de regresión puede ser inconveniente cuando entre las variables originales (las X's) exista un alto grado de heterogeneidad, ya que, como se ha señalado, el método es sensible a los cambios de escala. Una solución puede encontrarse, como lo sugiere Weisberg (1980),[36] empleando más de un grupo de variables independientes y realizando el análisis por componentes principales dentro de cada grupo y, luego, ajustando el modelo de regresión en función de los grupos de componentes principales. Así no se eliminará completamente la multicolinealidad, pero se garantizará la no correlación dentro de cada grupo.

Chatterjee y Price (1977)[9] también han elaborado un ejemplo y mostrado las correlaciones existentes entre los coeficientes de regresión calculados a partir de los datos originales y los calculados a partir de los componentes principales. El ejemplo ilustra cómo la eliminación de algún componente principal del modelo de regresión equivale a imponer una restricción a los coeficientes del modelo generado con las variables originales. De tal forma que al examinar los contrastes que menos contribuyan a explicar la variación total, no sólo se simplifica el modelo, sino que se conocen nuevas relaciones entre las variables originales.

5.3 Detección de Marginales por Componentes Principales

Un dato marginal puede definirse como una observación (o subconjunto de observaciones) que parece ser inconsistente con el resto de los datos (Barnett y Lewis, 1978[5]). En esta definición, como lo señalan los autores, la frase "parece ser inconsistente" es fundamental. La idea es poder detectar cuándo una observación constituye un elemento genuino de una población principal. Por estos motivos, los métodos para procesar los marginales tienen características condicionadas al modelo que se suponga.

Así, un marginal sorprende por su valor extremo y debido a esa discordancia es estadísticamente poco probable que pertenezca al modelo. La aparición de marginales no está restringida al campo univariado; al contrario, se presenta también en el campo multivariado, aunque su análisis y detección son más complejos. Al tomar en cuenta un conjunto de variables para cada observación, los marginales no se manifiestan de manera sencilla como una observación "extrema", ya que no existe una forma única de ordenar un conjunto de datos multivariados. Además, la presencia de marginales en el caso multivariado puede distorsionar no sólo las medidas de posición (media aritmética, por ejemplo) y de dispersión (desviación estándar), sino también las de orientación, es decir las correlaciones entre las variables. En tercer lugar, existe una

variedad de marginales, ya que un vector-respuesta (una observación) puede ser marginal porque tiene un gran error en uno de sus componentes o por existir un error sistemático en todos sus componentes.

Por tanto, con un método para detectar marginales debería ser posible distinguir entre estas posibilidades. Pero no existe un método único y, como sostienen Gnanadesikan y Kettenring (1972),[13] en el campo multivariado una observación puede ser marginal para un propósito y no serlo necesariamente para otro.

Se ha sugerido que los componentes principales se usen para detectar marginales, para lo cual han de examinarse los valores de los componentes de varios órdenes. Gnanadesikan y Kettenring (1972)[13] demuestran que los valores extremos de los primeros componentes sirven para detectar aquellas observaciones que contribuyen a aumentar en alto grado la varianza y la covarianza (o la correlación si el análisis por componentes principales se realiza a partir de la matriz R), y que los últimos componentes son sensibles para detectar observaciones que agregan dimensiones espúreas a los datos. Sugieren que para detectar visualmente los marginales se elabore un diagrama de dispersión entre pares de los primeros y de los últimos componentes principales. Si puede asumirse multinormalidad es posible efectuar gráficos en papel probabilístico y los marginales serán aquellos que se alejen de la línea recta o que tengan valores extremos sobre esta línea. McCabe (1984)[27] propone un método basado en componentes principales para seleccionar variables orginales.

Respecto al caso multinormal, Hawkins (1974)[17] propone otros tres criterios basados en los componentes principales y los compara con la prueba de un componente por vez (un chi-cuadrado) en dos hipótesis alternas diferentes. En los tres casos, los primeros componentes tienen menor capacidad para detectar marginalidad y, por ello, las pruebas se basan en los últimos componentes. Para seleccionar cuántos componentes han de utilizarse se aplican los mismos criterios ya descritos, sólo que se tendrá en cuenta el subconjunto de $\lambda\ (m+1),\ldots,\lambda\ (p)$, es decir los últimos $p - m$ componentes en vez de los m primeros. El autor incluye una tabla que permite conocer la verdadera probabilidad a la cual se efectúa una comparación postulada al 0,05 de significación cuando se usa el criterio de un componente por vez recurriendo a un chi-cuadrado. Así, por ejemplo, si se trabaja con 20 variables y 50 observaciones, el verdadero nivel de significación es 0,54 en vez de 0,05. A medida que p aumenta es necesario aumentar n para acercarse al verdadero nivel del 5%.

PARTE II
ESTUDIO DE CASOS

4

CARACTERIZACION DE LA PRODUCCION LECHERA DE UN DISTRITO

1. INTRODUCCION

Los datos que se utilizan en este capítulo derivan del Proyecto de Diagnóstico Lechero realizado por la Universidad Francisco de Miranda y su empresa de servicios INUFALCA conjuntamente con el Fondo de Crédito Agropecuario (UNEFM-FCA, 1984)[35] y corresponden al Distrito Federación del Estado Falcón, Venezuela.

Con el objeto de conocer la situación del sector lechero en ese distrito, se realizó una encuesta durante la cual se visitó a productores en su finca. Se reunió información acerca de una serie de variables que influyen en la producción total por finca y la productividad por finca y por vaca, y se efectuó un análisis por componentes principales.

Los resultados de esta caracterización fueron utilizados posteriormente para organizar la asistencia técnica que debía proporcionarse y para formular un plan de evaluación técnico-económica que forma parte del proyecto global.

La Tabla III sintetiza los valores promedio y las desviaciones para cada variable. Los índices de maquinaria y los de instalaciones se calcularon teniendo en cuenta el costo de reposición de maquinaria y equipo y su estado al momento de la visita. El índice sanitario se calculó ponderando el costo de las vacunaciones preventivas y del tratamiento curativo brindado por cada productor.

Tabla III. Variables utilizadas en el análisis por componentes principales

Nombre	Código	Promedio	Desviación Estándar	Coeficiente de Variación (%)
Superficie total de la finca (ha)	SUP	383,51	328,25	86
Número total de vacas	VACA	100,13	73,23	73
Indice sanitario	SANI	18,17	4,00	21
Indice de instalaciones	INST	71,16	16,43	23
Indice de maquinaria	MAQ	12,38	21,57	174
Promedio de leche/vaca (l)	PROM	2,20	1,50	68

Los altos valores de los coeficientes de variación indican una gran heterogeneidad entre los productores del distrito. Es necesario, entonces, estudiar cuál o cuáles variables influyen más sobre la productividad y qué fincas son las más afectadas. Este es el primer paso para elaborar un plan de asistencia técnica y localizar aquellos productores que no hicieran buen uso de sus recursos.

2. CALCULO DE LOS VALORES Y VECTORES PROPIOS DE LA MATRIZ DE COVARIANZA

Se calculó la matriz de covarianza con las seis variables utilizando los datos originales, es decir sin estandarizar los valores. La matriz resultante fue

		SUP	VACA	SANI	INST	MAQ	PROM
	SUP	107746,30					
	VACA	11875,63	5362,88				
S =	SANI	268,34	24,44	15,97			
	INST	389,19	156,36	10,64	269,93		
	MAQ	1704,49	293,25	14,02	34,59	465,30	
	PROM	-25,30	-31,44	0,76	3,64	2,68	2,25

No se transcriben los valores por encima de la diagonal principal porque, como se desprende de la definición [3], S es una matriz simétrica. Tampoco se indicarán para la matriz R, también numérica.

Los valores de las varianzas de SUP y de VACA son mucho mayores que para las otras cuatro variables, lo cual se debe fundamentalmente a que sus magnitudes absolutas son también mayores (véanse los promedios en la Tabla III). Las covarianzas asumen valores del mismo orden de magnitud que las varianzas, lo que indica una dependencia entre las variables. Se observa una covarianza negativa entre PROM y SUP y entre PROM y VACA debido a la disminución del rendimiento promedio por día y por vaca al aumentar la superficie y el número de animales de la finca. Este puede ser un indicador del mal manejo a que están sometidas las fincas grandes.

En la Tabla IV se presentan los valores propios y la proporción de la variación total explicada por cada uno de los componentes al usar la matriz S. Se observa que el primer componente resume el 95,85% de la variabilidad total. Esto significa que la combinación lineal de las variables originales representada por el primer componente principal sintetiza casi el 96% de la variación total del conjunto de datos. Esta situación, que a primera vista puede ser muy ventajosa, pues podrían reemplazarse las seis variables originales sólo por el primer componente, puede llevar a interpretaciones equivocadas.

Tabla IV. Valores propios y proporción de la variación explicada (Cálculos a partir de la matriz S)

Componentes	Valor Propio	Proporción de la Varianza Total Explicada Absoluta (%)	Acumulada (%)
1°	109135,00	95,85	95,85
2°	4009,67	3,52	99,37
3°	439,17	0,39	99,76
4°	261,79	0,23	99,99
5°	14,74	<0,01	100,00
6°	1,91	<<0,01	100,00

Como se ha señalado en los capítulos anteriores, es necesario examinar no sólo los valores propios, sino también los vectores y los coeficientes de correlación de las nuevas variables con las originales. Las Tablas V y VI muestran los valores que se obtienen utilizando la matriz de covarianza S.

Para encontrar los elementos de cada vector propio debe resolverse el sistema de ecuaciones de la fórmula [19] o de la [26], sustituyendo $\lambda(k)$ o g por cada uno de los valores propios de la Tabla IV. Los valores propios, como asimismo los vectores propios, se obtienen con programas de computación muy sencillos, escritos en casi todos los lenguajes. Los grandes "paquetes estadísticos" como el SAS, el BMDP y el SPSS facilitan estos resultados como parte de las transformaciones disponibles. En este ejemplo se utilizó el programa BMDP en una computadora IBM 370/105.

Al calcular los coeficientes de correlación entre las variables y los componentes principales se recurre a la fórmula [36]. A menudo es necesario proceder *a posteriori*, ya que los "paquetes estadísticos" no los incluyen.

Tabla V. Vectores propios de la matriz de covarianza

Variable	Vectores Propios 1°	2°	3°	4°	5°	6°
SUP	0,9934	-0,1142	-0,0126	-0,0015	-0,0023	-0,0003
VACA	0,1137	0,9926	-0,0331	0,0254	0,0032	0,0076
SANI	0,0025	-0,0014	0,0265	-0,0335	0,9982	-0,0424
INST	0,0037	0,0299	0,1443	-0,9882	-0,0376	-0,0149
MAQ	0,0159	0,0275	0,9885	0,1462	-0,0216	-0,0070
PROM	-0,0003	-0,0070	0,0104	-0,0154	0,0417	0,9989

La Tabla V muestra que el primer componente tiene un coeficiente de 0,9934 para la variable SUP y coeficientes mucho más pequeños para las cinco restantes. Esto indica que la primera combinación lineal refleja la variación en superficie de las fincas. En este ejemplo cada componente puede asociarse con una variable, con aquella que posee el valor mayor en el vector, el cual ha sido recuadrado en la Tabla V. Se observa una correspondencia entre el 1° componente principal y la variable SUP, entre el 2° componente principal y VACA, entre el 3° y MAQ, entre el 4° e INST, entre el 5° y SANI, y entre el 6° y PROM.

Esta misma situación se refleja cuando se examinan los valores de la correlación entre las variables y los componentes principales en la Tabla VI. El 99,56% de la variación total de SUP queda explicada por el primer componente; el 0,05% restante por el segundo. Esto significa que si se consideran los dos primeros componentes principales se habrá explicado el 100% de la variación de SUP.

Tabla VI. Proporción de la variación original explicada por cada componente principal utilizando la matriz de covarianza

Variable	Componentes 1°	2°	3°	4°	5°	6°
SUP	0,9956	0,0005	0,0000	0,0000	0,0000	0,0000
VACA	0,2631	0,7366	0,0001	0,0000	0,0000	0,0000
SANI	0,0427	0,0005	0,0193	0,0184	0,9199	0,0002
INST	0,0055	0,0133	0,0339	0,9471	0,0001	0,0000
MAQ	0,0593	0,0652	0,9222	0,0120	0,0000	0,0000
PROM	0,0044	0,0874	0,0200	0,0276	0,0114	0,8485

El 26,31% de la variable VACA está sintetizado en el primer componente y el 73,66% en el segundo. Al considerar sólo los dos primeros, se conocerá el 99,99% de la variación total de VACA.

Sin embargo, el aporte de las otras cuatro variables será muy pequeño y, por ello, la variación explicada por estos dos primeros componentes principales será muy baja: 4,32% para SANI, 1,88% para INST, 12,45% para MAQ y 9,18% para PROM.

Se comprende la aparente contradicción entre la alta proporción de la variación total explicada por los dos primeros componentes (99,37%, Tabla IV) y el hecho que cuatro de las seis variables consideradas aporten muy poco a estos dos componentes, si se observan las magnitudes de las varianzas y de las medias para SUP y para VACA en comparación con cualquiera de las otras (Tabla III). Esta situación, frecuente cuando se trabaja con variables que se expresan en distintas unidades y cuyos recorridos de valores diferen en uno o más órdenes de magnitud, puede influir de manera apreciable en la interpretación de los resultados.

Se recomienda examinar el efecto que tiene sobre el resultado del análisis la estandarización de los datos originales o la transformación de unidades de alguna(s) de las variables a fin de homogeneizar las magnitudes. No hacerlo implica suponer que la variable que asume valores absolutos mayores es la que más influye, o sea la determinante en el análisis. En ciertos casos puede ser conveniente esta ponderación y en otros casos, cuando el orden de magnitud no refleje la importancia relativa de cada uno de los factores estudiados, ocurre todo lo contrario.

Más adelante en el análisis se utilizaron datos estandarizados (media cero y varianza uno) y el resultado obtenido difiere del anterior.

3. USO DE LA MATRIZ DE CORRELACION

A partir de la matriz S es posible efectuar los cálculos necesarios para obtener la matriz de correlación R, tal como se expresó en la definición [4], la cual se transcribe a continuación:

$$R = \begin{array}{c} \\ \text{SUP} \\ \text{VACA} \\ \text{SANI} \\ \text{INST} \\ \text{MAQ} \\ \text{PROM} \end{array} \begin{bmatrix} \text{SUP} & \text{VACA} & \text{SANI} & \text{INST} & \text{MAQ} & \text{PROM} \\ 1{,}0000 & & & & & \\ 0{,}4940 & 1{,}0000 & & & & \\ 0{,}2046 & 0{,}0835 & 1{,}0000 & & & \\ 0{,}0722 & 0{,}1300 & 0{,}1620 & 1{,}0000 & & \\ 0{,}2407 & 0{,}1856 & 0{,}1627 & 0{,}0976 & 1{,}0000 & \\ -0{,}0514 & -0{,}2865 & 0{,}1265 & 0{,}1477 & 0{,}0829 & 1{,}0000 \end{bmatrix}$$

En la Tabla VII se presentan los valores propios de esta matriz y la proporción de la variación total explicada por cada uno de los componentes. Obsérvese que en este caso la suma de los valores propios es cinco, el número total de variables consideradas.

En este caso, el primer componente sólo sintetiza el 29,90% y el primero más el segundo alcanzan el 52,25%. Es necesario considerar hasta cuatro componentes para obtener el 81% de la variación y hasta el quinto para conseguir el 93%. Aplicando el criterio de selección de componentes principales de Kaiser, ya reseñado, se utilizarían los dos primeros, cuyos valores son mayores que el promedio, es decir, superiores a 1. El método gráfico de Cattell (1966),[6] representado para este ejemplo en la figura 3, conduce a la misma conclusión, aunque podría considerarse también la inclusión del tercer componente a fin de alcanzar el 67% de la variación total explicada.

Tabla VII. Valores propios y proporción de la variación explicada (Cálculos a partir de la matriz R)

Componentes	Valor Propio	Proporción de la Varianza Total Explicada Absoluta (%)	Acumulada (%)
1°	1,79412	29,90	29,90
2°	1,34091	22,35	52,25
3°	0,90316	15,05	67,30
4°	0,82129	13,69	80,99
5°	0,71244	11,88	92,87
6°	0,42805	7,13	100,00

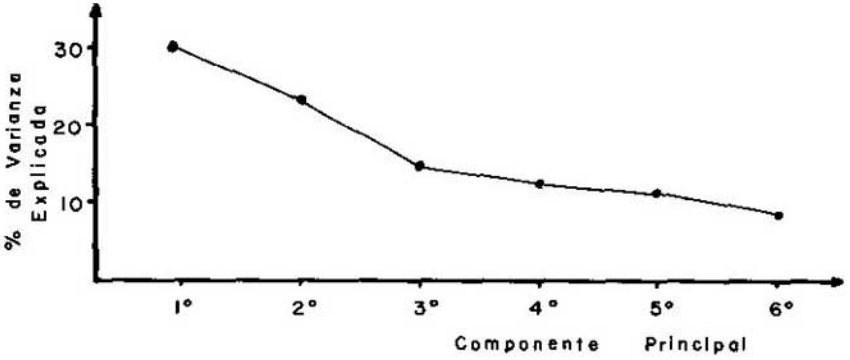


Fig. 3. Variación explicada por cada componente: caso producción lechera. En el tercer componente se observa un punto de inflexión.

Tabla VIII. Vectores propios de la matriz de correlación

Variable	Vectores Propios 1°	2°	3°	4°	5°	6°
SUP	0,5866	0,0999	-0,1344	-0,0883	-0,5341	-0,5787
VACA	0,5692	0,3415	0,1840	0,0625	-0,1361	0,7093
SANI	0,3247	-0,4176	-0,0818	-0,7734	0,3267	0,0925
INST	0,2418	-0,4118	0,7935	0,2820	0,1554	-0,1927
MAQ	0,3961	-0,2286	-0,5297	0,5397	0,4671	-0,0283
PROM	-0,1053	-0,6908	0,1767	0,1389	0,5892	0,3376

Para adoptar una decisión definitiva al respecto, es necesario examinar los vectores propios y la correlación entre las variables originales y los componentes principales, los cuales se presentan en las Tablas VIII y IX.

Los coeficientes del primer vector propio señalado en la Tabla VIII indican que esta combinación lineal es aproximadamente un promedio entre

Tabla IX. Proporción de la variación original explicada por cada componente principal de la matriz de correlación

Variable	Componentes 1°	2°	3°	4°	5°	6°
SUP	0,6174	0,0134	0,0163	0,0064	0,2032	0,1434
VACA	0,5813	0,1564	0,0306	0,0032	0,0132	0,2154
SANI	0,1892	0,2338	0,0060	0,4913	0,0760	0,0037
INST	0,1049	0,2274	0,5687	0,0653	0,0172	0,0166
MAQ	0,2815	0,0701	0,2534	0,2392	0,1554	0,0003
PROM	0,0111	0,6399	0,0282	0,0158	0,2474	0,0489

todas las variables con ponderación relativa mayor para SUP y para VACA (0,6174 y 0,5813), y algo menor para SANI, INST y MAQ. El coeficiente para PROM es negativo, lo que indica que disminuirá el valor del primer componente si aumenta la productividad por vaca. Sin embargo, su contribución es pequeña y las fincas con valores elevados del primer componente estarán asociadas con valores elevados de las variables SUP, VACA, SANI, INST y MAQ.

En el segundo vector propio se observa muy exigua contribución de SUP (0,0999) y coeficientes negativos para las variables SANI, INST, MAQ y PROM. La variable VACA tiene un coeficiente positivo de 0,3415, constituyendo un contraste con las otras cuatro variables. De tal forma que las fincas con los valores más elevados del 2° componente principal serán aquellas que posean los mayores rebaños lecheros, la menor productividad por vaca, las peores condiciones sanitarias y deficiencias en las instalaciones y equipo.

Si se consideran los dos primeros componentes principales puede concluirse que en las mejores fincas se obtendrán valores elevados del primer componente y valores negativos del segundo. Es decir que estarán ubicadas en el segundo cuadrante si se grafican los dos primeros componentes. Recuérdese que se trabaja con valores estandarizados de los datos, donde cada variable tiene media cero y varianza 1.

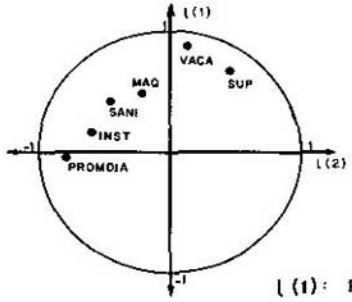
4. INTERPRETACION DE LOS RESULTADOS

Puede graficarse la correlación r(jk) de cada variable original con los dos primeros componentes principales. Para ello se extrae la raíz cuadrada de los valores de la Tabla IX, respetando el signo del coeficiente correspondiente de la Tabla VIII.

En la Tabla X se presentan estos resultados y en la figura 4 se han graficado en un par de ejes ortogonales, que representan el primer componente principal l(1) y el segundo componente principal l(2). El círculo de radio unitario, trazado con centro en el origen del par de coordenadas, permite identificar las variables cuya correlación con los componentes sea mayor (se acercan al círculo). Aquellas variables que se agrupan cerca del centro serán las que menos se correlacionan con los componentes principales. Como la proporción de la variación de cada variable explicada en los dos primeros componentes principales es la suma de sus correlaciones al cuadrado, las variables mejor explicadas se ubicarán cerca del círculo y las menos explicadas, cerca del origen de coordenadas.

Tabla X. Correlación de las variables originales con los dos primeros componentes

Variable	Componentes 1°	Componentes 2°	Variación Total Explicada por los Dos Primeros Componentes (%)
SUP	0,79	0,10	63
VACA	0,76	0,40	74
SANI	0,44	-0,48	42
INST	0,32	-0,48	33
MAQ	0,53	-0,26	35
PROM	-0,01	-0,80	65



L(1): Primer Componente

L(2): Segundo Componente

Fig. 4. Correlación de las variables originales con los dos primeros componentes, caso producción lechera. Los dos primeros componentes explican mejor las variables que se ubican cerca del círculo unitario.

Además de graficar las variables originales en función de los dos primeros componentes, es posible graficar los valores de los dos primeros componentes principales para cada observación. Si el gráfico se superpone al anterior se denomina "birrepresentación" ("biplot" en inglés). En este ejemplo se han elaborado los gráficos por separado; en la figura 5 se representan con un punto cada una de las 127 fincas que forman la muestra original. Es posible identificarlas con un número, que se prefirió omitir en este caso para facilitar la interpretación general. Cada zona del plano definido por estos componentes principales sintetiza una problemática diferente, de manera que una vez conocida la ubicación de una finca en el plano es posible sacar conclusiones acerca de su situación respecto a la producción lechera.

Las fincas situadas en el sector A estarán en las mejores condiciones: tendrán buena dotación de maquinaria y equipo, sus condiciones sanitarias serán apropiadas y el rebaño lechero será grande y productivo.

Las condiciones de las fincas del sector B también serán satisfactorias en cuanto a equipo y maquinaria, como asimismo desde el punto de vista sanitario, pero la superficie total de la explotación no será tan grande ni el rebaño lechero tan numeroso, aunque los índices de productividad por vaca serán buenos.

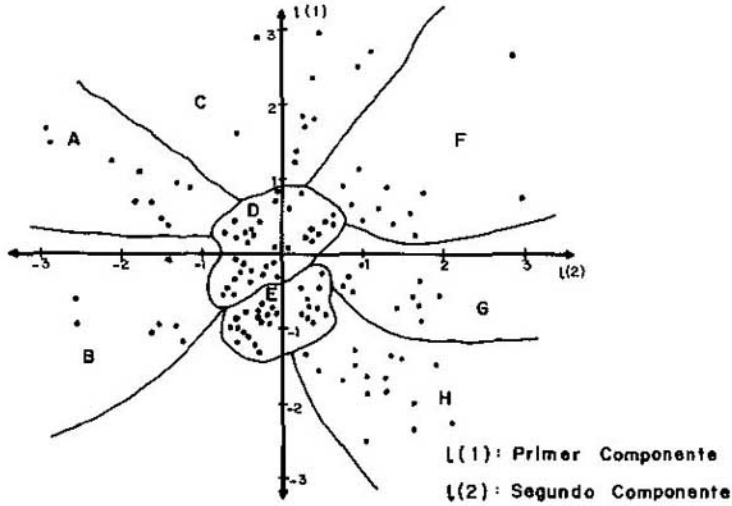


Fig. 5. Rrepresentación bivariada de resultados. Cada sector, identificado con una letra mayúscula, constituye un conjunto de fincas con características comunes. Véase explicación en el texto.

El sector C comprenderá las fincas grandes con dotación heterogénea de recursos, pero posiblemente no muy productivas (segundo componente cercano a cero). Los sectores F y G estarán constituidos por fincas casi iguales, con superficie cercana al promedio, con cierto equipo, pero con baja productividad. Los sectores D y E, que se sitúan en el promedio de la zona, afrontarán diversos problemas, pero en todos los casos la productividad será también cercana al promedio de la zona. Entre ellos, el sector D contará con más recursos que el E para mejorar su productividad. Las fincas pertenecientes al sector H serán las más pobres, ya que tendrán baja productividad debido a la falta de recursos para explotarlas.

A esta interpretación de los componentes principales puede agregarse una tabla donde se presenten los valores promedios para cada una de las variables estudiadas en cada uno de los sectores y donde se comparen esos valores con los promedios generales a fin de corroborar la explicación derivada del análisis de la figura 5.

Este procedimiento de formación de los sectores, efectuado en este caso por simple inspección visual, puede realizarse con una técnica de formación de conglomerados utilizando los componentes principales seleccionados como variables del análisis. En este ejemplo sólo se procedió a la interpretación visual, ya que lo que se desea es hacer hincapié en la interpretación de los resultados de un análisis por componentes principales.

5

ANALISIS FLORISTICO DE VEGETACION SEMINATURAL

INTRODUCCION

Los métodos estadísticos multivariados pueden utilizarse para estudia. la cubierta vegetal de los continentes y de los mares. La vegetación, entendiendo por ella el conjunto de especies vegetales interactuantes que existen en una zona como resultado de la acción de los factores ambientales, constituye un universo multivariado que para su interpretación requiere la aplicación de técnicas y procedimientos holísticos.

Dos grandes enfoques podrían usarse en los estudios sobre vegetación. Los métodos no formales, que se excluyen de esta monografía, son los más antiguos y se basan en una característica cualitativa determinada según la experiencia, el sentido común y la intuición de los investigadores. Sus resultados a menudo no pueden ser comparados con los de otros estudios y, por el alto grado de subjetividad implícito, su repetibilidad es muy reducida o nula.

Los métodos formales son de origen relativamente reciente; se han difundido en forma rápida a partir de los años 1960 del mismo modo que las computadoras digitales. Se basan en la aplicación de técnicas estadísticas que hacen posible la planificación del muestreo, la uniformidad y consistencia en la recolección de datos, la eliminación del sesgo entre investigadores, el procesamiento computarizado de datos y la repetibilidad de los resultados.

Sin embargo, la estadística no podrá en ningún caso suplantar la experiencia y formación del ecólogo, ya que si bien los procedimientos y técnicas de manejo de datos pueden objetivizarse, es imposible sin la ayuda de éste tomar las decisiones que implican los análisis. Existe una serie de alternativas y ciertas decisiones que se han de tomar con base en las características de la vegetación (conocida por el ecólogo) y no de acuerdo con los resultados numéricos.

En el análisis de tipo formal se utilizan en general datos cuantitativos acerca de las especies presentes en una zona: densidad, cobertura, área basal, frecuencia. Pueden también considerarse datos cualitativos del tipo presencia-ausencia de alguna característica, o de las especies. Estos datos se presentan en tablas de doble entrada o matrices primarias de datos (véase la Tabla XII) en las que constan los valores de los atributos para todas las muestras del estudio.

Es imposible extraer conclusiones de la inspección visual de la matriz primaria dado el alto número de atributos y de muestras o elementos que contiene. En la monografía no. 22 (Matteucci y Colma, 1983)[26] se describen con detalle los métodos de clasificación y de ordenación de la vegetación. Dichos métodos permiten simplificar su presentación, agrupar individuos o atributos en clases, reducir su dimensionalidad o ubicarlos a lo largo de un gradiente.

En este capítulo se presenta un ejemplo de aplicación del método de componentes principales al estudio de la vegetación del Estado Falcón en el nordeste de Venezuela. Se utilizó una parte de la información

florística generada en el Proyecto "Análisis Regional de la Vegetación y el Ambiente del Estado Falcón" (Matteucci y colaboradores, 1979)[25].

Tabla XI. Especies utilizadas en el análisis de la vegetación en la zona semiarida del Estado Falcón

Especie Número	Nombre Científico	Familia	Censos Presente No.	Frecuencia
1	*Acacia tortuosa* (L.) Willd.	Mimosoideae	61	0,59
2	*Bourreria cumanensis* (Loefl.) O.E. Shultz	Boraginaceae	53	0,51
3	*Bulnesia arborea* (Jacq.) Eng.	Zigofilaceae	74	0,71
4	*Caesalpinia coriaria* (Jacq.) Willd.	Caesalpinoideae	82	0,79
5	*Capparis odoratissima* Jacq.	Capariaceae	86	0,83
6	*Cercidium praecox* (R. & P.) Harms	Caesalpinoideae	69	0,66
7	*Mimosa arenosa* (Willd.) Poir	Mimosoideae	53	0,51
8	*Pereskia guamacho* Weber	Cactaceae	56	0,54
9	*Pithecellobium unguis-cati* Benth.	Mimosoideae	75	0,72
10	*Prosopis juliflora* (Swartz) DC.	Mimosoideae	101	0,97
11	*Ritterocereus* spp.	Cactaceae	96	0,92
12	*Castela erecta* Turp.	Simarubaceae	70	0,67
13	*Ipomoea carnea* Jacq.	Convolvulaceae	61	0,59
14	*Jatropha gossipifolia* L.	Euphorbiaceae	71	0,68
15	*Cnidosculus urens* (L.) Arthur	Euphorbiaceae	52	0,50
16	*Melocactus caesius* (Wendl.) Britt & Rose	Cactaceae	87	0,84
17	*Opuntia wentiana* Britt & Rose	Cactaceae	103	0,99
18	*Sporobolus pyramidatus* (Lam.) Hitche	Poaceae	69	0,66
19	*Loranthaceas* (varias especies)	Loranthaceae	63	0,61

En la Tabla XI se indica el nombre científico de cada especie, el número que la identificará en este análisis, el total de censos en que estaba presente cada especie y la frecuencia de cada una. La frecuencia por especie fue calculada como el total de censos en que la misma está presente, dividido entre el número total de censos. La Tabla XII contiene los datos originales de cobertura, expresada en porcentaje, para las 19 especies cuya frecuencia es igual o superior al 50% en un total de 104 censos o sitios de muestra de la zona semiárida del Estado Falcón.

En los cálculos posteriores se transformaron los valores originales utilizando lo que se denomina "transformación angular", la cual permite expresar un conjunto de observaciones consignadas en porcentajes, que generalmente no se distribuyen en forma normal, en una nueva escala en la cual se alcance la normalidad. En los gráficos de la figura 6 se aprecia la distribución de los datos antes de ser transformados y luego de la transformación angular para la especie *Prosopis juliflora*. La transformación utilizada en este análisis es:

$$X' = \text{sen}^{-1} \sqrt{X + 0{,}01} \qquad [37]$$

donde X' indica el valor transformado, X el valor original expresado en porcentaje en una escala de 0 a 1 y sen^{-1} indica la inversa del seno

Tabla XII. Datos originales de cobertura por especie expresados en porcentajes de 0 a 100

	ESPECIE																		
CENSO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	0,5	0,0	0,0	0,5	0,5	3,0	0,0	0,0	10,5	20,5	3,0	20,5	10,5	0,0	0,5	0,5	3,0	0,5	0,0
2	3,0	0,5	3,0	0,5	3,0	3,0	10,5	0,0	0,5	20,5	10,5	0,5	10,5	0,5	0,5	0,5	3,0	0,5	0,5
3	0,5	0,5	0,0	0,5	0,5	10,5	0,0	0,0	0,5	10,5	10,5	0,5	0,5	0,0	0,5	0,5	0,5	0,5	0,5
4	20,5	0,5	0,5	0,5	0,5	10,5	3,0	0,0	0,5	0,5	0,5	37,5	0,0	0,5	0,0	0,5	0,5	0,0	0,5
5	10,5	0,0	0,5	0,5	0,5	3,0	3,0	0,0	0,5	0,5	0,5	20,5	0,0	0,0	0,0	0,5	0,5	0,0	0,0
6	3,0	0,0	10,5	10,5	0,0	10,5	0,5	0,0	10,5	3,0	10,5	10,5	0,0	0,0	0,0	0,5	0,5	0,0	0,5
7	20,5	0,5	0,5	0,5	0,5	0,5	0,5	0,0	3,0	0,5	0,5	37,5	0,5	0,0	0,5	0,5	0,5	0,0	0,5
8	3,0	0,0	0,0	0,0	0,0	3,0	0,0	0,0	0,0	3,0	0,0	20,5	0,0	0,0	0,0	0,5	0,5	0,5	0,5
9	10,5	0,0	0,5	0,5	0,0	0,5	0,0	0,0	0,5	0,0	0,0	10,5	0,0	0,0	0,0	0,5	0,5	0,5	0,5
10	3,0	0,5	0,5	20,5	0,5	10,5	0,5	0,5	10,5	3,0	0,5	20,5	0,0	0,5	0,0	0,5	0,5	0,5	0,5
11	0,0	0,0	0,0	20,5	0,0	0,5	3,0	3,0	0,5	20,5	62,5	0,0	0,0	0,0	0,0	0,0	0,5	0,0	0,0
12	0,0	0,0	0,0	20,5	0,5	0,5	0,0	0,0	0,0	20,5	20,5	37,5	0,0	0,0	0,0	0,5	0,5	0,5	0,0
13	0,0	0,0	0,5	0,0	0,5	3,0	0,0	0,0	0,0	37,5	20,5	10,5	0,0	0,0	0,0	0,5	3,0	0,5	0,0
14	0,0	0,0	0,0	0,0	3,0	20,5	0,0	20,5	0,5	3,0	20,5	0,5	82,5	0,5	0,0	0,5	20,5	20,5	0,0
15	10,5	3,0	3,0	10,5	10,5	10,5	0,5	10,5	10,5	20,5	20,5	10,5	3,0	0,0	0,0	0,5	20,5	0,0	0,0
16	0,0	3,0	10,5	0,5	10,5	20,5	3,0	10,5	10,5	20,5	20,5	20,5	20,5	0,0	3,0	0,5	10,5	0,0	0,0
17	0,0	0,0	0,0	10,5	3,0	10,5	0,0	3,0	0,0	3,0	10,5	10,5	0,5	0,0	0,0	0,5	10,5	0,5	0,0
18	0,0	0,0	0,0	3,0	3,0	20,5	0,0	0,0	0,0	10,5	20,5	10,5	0,0	0,0	0,0	0,0	20,5	0,5	0,0
19	0,0	0,0	0,0	0,0	0,5	20,5	0,0	0,0	0,0	67,5	0,5	20,5	0,0	0,0	0,0	0,5	0,5	0,5	0,0
20	3,0	0,0	0,0	0,5	0,0	0,5	0,0	0,0	0,0	3,0	0,5	0,5	0,0	0,5	0,0	0,5	0,5	0,5	0,0
21	0,0	0,0	0,0	0,0	0,0	0,5	0,0	0,0	0,0	10,5	0,5	3,0	0,0	0,5	0,0	0,5	0,5	0,5	0,0
22	0,0	0,0	0,0	3,0	3,0	3,0	0,0	0,0	0,5	10,5	10,5	20,5	0,0	0,0	0,0	3,0	3,0	0,5	0,0
23	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	82,5	0,0	3,0	0,0	0,0	0,0	0,0	3,0	0,5	0,5
24	0,0	10,5	10,5	0,5	10,5	10,5	3,0	0,0	3,0	37,5	10,5	20,5	3,0	0,5	0,5	0,5	10,5	3,0	0,0
25	0,0	0,0	0,0	0,0	0,0	3,0	0,5	0,5	0,0	10,5	0,5	10,5	0,5	0,5	0,0	0,5	0,5	0,0	0,0
26	0,5	0,5	0,5	10,5	0,5	3,0	0,5	0,0	0,5	10,5	3,0	37,5	0,0	0,5	0,5	0,5	3,0	0,5	0,0
27	0,0	0,5	20,5	10,5	3,0	10,5	10,5	10,5	10,5	37,5	0,0	3,0	67,5	0,5	0,5	0,5	3,0	0,0	0,0
28	0,0	0,0	0,0	0,0	37,5	0,0	0,0	0,0	0,0	82,5	10,5	0,5	0,0	0,0	0,0	0,5	10,5	0,5	0,0
29	0,0	0,0	0,0	3,0	0,5	0,5	0,0	0,5	3,0	0,5	10,5	3,0	0,5	0,5	0,5	0,5	0,5	0,5	0,5
30	0,0	0,0	0,5	0,5	0,5	0,5	0,0	0,0	0,5	67,5	20,5	10,5	3,0	3,0	0,5	0,5	10,5	0,5	0,0

Continuación de la Tabla XII

	ESPECIE																		
CENSO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
31	0,5	0,5	10,5	37,5	0,5	0,5	0,5	0,5	3,0	10,5	10,5	0,0	0,0	0,5	0,5	0,5	20,5	0,5	0,5
32	20,5	0,5	20,5	20,5	10,5	20,5	0,0	20,5	10,5	37,5	37,5	20,5	20,5	0,5	0,5	0,5	10,5	0,0	0,0
33	0,0	0,5	0,0	20,5	3,0	0,0	0,0	0,0	3,0	3,0	10,5	3,0	0,0	0,0	0,0	0,5	3,0	0,0	0,0
34	0,0	0,5	0,5	10,5	0,5	0,0	0,5	37,5	0,5	3,0	3,0	0,0	37,5	0,5	0,5	0,0	3,0	0,0	0,5
35	0,5	0,5	3,0	0,0	0,0	0,0	0,0	0,0	0,5	20,5	0,5	0,0	0,5	0,0	0,5	0,5	0,5	0,5	0,5
36	0,5	0,5	0,5	3,0	0,0	0,0	0,0	0,0	0,0	20,5	20,5	0,0	0,5	0,5	0,5	0,5	20,5	0,5	0,5
37	3,0	0,0	0,0	0,0	0,5	3,0	3,0	10,5	0,5	20,5	0,5	0,0	10,5	0,5	0,0	0,5	62,5	10,5	0,5
38	10,5	0,5	0,5	3,0	3,0	20,5	10,5	10,5	3,0	20,5	20,5	37,5	10,5	0,5	0,5	0,5	37,5	0,5	0,5
39	3,0	0,0	3,0	10,5	0,5	10,5	10,5	0,0	10,5	3,0	37,5	3,0	0,0	0,5	0,0	0,5	20,5	0,0	0,5
40	3,0	0,0	0,5	10,5	0,5	3,0	0,0	0,0	0,0	3,0	3,0	0,5	0,0	0,0	0,0	0,0	3,0	0,0	0,5
41	0,0	0,0	0,0	3,0	0,0	0,5	0,0	0,0	0,0	87,5	3,0	0,5	0,0	0,5	0,0	0,0	10,5	0,0	0,0
42	3,0	0,0	3,0	0,5	20,5	10,5	0,5	0,5	0,5	37,5	10,5	37,5	10,5	0,5	0,5	0,5	10,5	0,5	0,0
43	0,5	0,0	0,5	37,5	0,5	0,0	0,0	0,0	0,0	10,5	37,5	37,5	0,5	0,5	0,5	0,5	10,5	0,5	0,0
44	3,0	0,0	0,5	37,5	3,0	0,0	0,0	0,5	0,0	37,5	20,5	10,5	10,5	0,5	0,5	0,0	20,5	0,5	0,5
45	0,0	0,5	0,5	0,0	10,5	3,0	0,5	0,5	3,0	62,5	3,0	10,5	3,0	0,5	0,0	0,0	20,5	0,5	0,5
46	0,0	0,0	0,0	0,0	3,0	37,5	0,0	3,0	0,0	20,5	20,5	3,0	0,5	0,5	0,0	0,0	62,5	0,5	0,0
47	0,5	0,0	3,0	10,5	10,5	20,5	10,5	3,0	0,5	20,5	10,5	20,5	0,0	0,0	3,0	0,5	10,5	0,5	0,0
48	0,0	3,0	3,0	3,0	10,5	20,5	10,5	0,5	3,0	20,5	10,5	10,5	37,5	0,5	0,5	0,5	20,5	0,0	0,5
49	0,5	0,0	0,0	0,5	0,5	3,0	0,0	0,0	0,5	3,0	3,0	37,5	0,0	0,0	0,0	3,0	3,0	0,0	0,5
50	0,0	0,0	0,0	0,0	0,0	0,5	0,0	0,0	0,0	3,0	0,5	10,5	0,0	0,0	0,0	0,5	0,5	0,5	0,0
51	3,0	0,0	0,5	3,0	3,0	10,5	3,0	3,0	3,0	10,5	10,5	0,5	0,5	0,5	0,0	0,5	3,0	0,5	0,0
52	3,0	0,5	3,0	3,0	3,0	10,5	3,0	0,0	3,0	10,5	10,5	20,5	0,0	0,0	0,0	0,5	10,5	0,5	0,5
53	0,0	0,0	3,0	0,0	0,5	0,0	0,0	0,5	10,5	37,5	0,5	0,0	0,5	0,5	0,5	0,0	3,0	0,0	0,5
54	0,5	0,0	10,5	0,5	0,5	37,5	0,5	0,0	3,0	10,5	20,5	0,5	0,5	0,5	0,0	0,5	0,5	0,5	0,5
55	10,5	0,5	3,0	20,5	3,0	0,5	3,0	0,5	3,0	10,5	3,0	0,5	0,0	0,5	0,0	3,0	20,5	0,5	0,5
56	3,0	0,0	0,5	3,0	0,5	3,0	0,5	0,0	3,0	3,0	0,5	20,5	0,0	0,5	0,0	0,5	0,5	0,5	0,0
57	3,0	0,5	0,5	10,5	0,5	0,5	3,0	0,0	3,0	10,5	10,5	0,0	0,5	0,5	0,0	0,5	0,5	0,5	0,5
58	0,5	3,0	3,0	0,5	0,5	0,5	3,0	0,0	3,0	37,5	0,5	0,0	10,5	0,5	0,5	0,0	10,5	0,5	0,5
59	0,5	10,5	0,5	0,5	0,0	0,0	0,0	0,5	0,5	0,5	10,5	0,0	0,5	0,5	0,5	0,5	3,0	0,5	0,5
60	3,0	0,0	3,0	20,5	0,5	0,5	0,5	0,5	3,0	20,5	3,0	0,5	0,5	0,5	0,5	0,5	3,0	0,5	0,5

Continuación de la Tabla XXII

61	0,0	0,0	0,5	0,0	0,0	0,5	0,5	0,0	0,0	37,5	0,0	0,5	0,5	0,5	0,5	0,0	0,5	0,0	0,0
62	3,0	0,0	0,5	10,5	0,5	0,5	3,0	0,0	3,0	10,5	10,5	20,5	0,5	0,5	0,5	0,5	10,5	0,5	0,5
63	0,5	0,0	0,5	10,5	3,0	0,0	0,5	0,0	3,0	10,5	3,0	20,5	0,0	0,5	0,0	0,5	0,5	0,5	0,0
64	0,5	0,5	0,5	10,5	10,5	0,5	0,5	3,0	10,5	20,5	37,5	10,5	0,0	0,5	0,5	0,0	62,5	0,5	0,5
65	0,5	3,0	0,5	37,5	0,5	0,5	0,5	3,0	0,0	3,0	10,5	10,5	20,5	0,5	0,0	0,5	37,5	0,5	0,5
66	0,5	0,5	0,5	10,5	3,0	0,0	0,5	20,5	0,0	10,5	10,5	0,0	20,5	3,0	0,5	0,5	3,0	0,5	0,5
67	0,5	0,5	0,0	10,5	0,5	0,0	0,0	10,5	3,0	3,0	0,5	0,0	3,0	0,5	0,5	0,5	0,0	0,5	0,5
68	0,0	0,5	0,5	10,5	0,5	0,0	0,5	0,5	0,5	0,0	0,5	0,0	0,5	0,5	0,5	0,5	0,5	0,0	0,5
69	3,0	3,0	3,0	0,0	0,5	0,0	10,5	0,5	0,5	0,5	0,5	0,0	0,5	0,0	0,5	0,5	0,5	0,5	3,0
70	0,5	20,5	0,5	20,5	3,0	0,0	0,0	3,0	0,0	20,5	62,5	0,0	0,5	0,5	0,5	0,5	20,5	0,5	0,5
71	0,5	0,0	0,5	10,5	0,5	0,0	0,0	0,0	0,0	20,5	20,5	0,0	0,5	0,5	0,5	0,5	20,5	0,5	0,5
72	0,5	0,5	0,0	3,0	0,0	0,5	0,0	0,0	0,0	87,5	0,5	0,0	0,5	0,5	0,5	0,0	20,5	0,5	0,5
73	3,0	3,0	0,5	10,5	0,5	10,5	3,0	0,5	3,0	3,0	3,0	3,0	0,5	0,5	0,5	0,5	10,5	0,5	0,5
74	0,5	0,5	20,5	0,5	0,0	0,0	0,0	0,5	3,0	3,0	20,5	0,0	0,5	0,5	0,5	0,5	37,5	0,0	0,5
75	20,5	10,5	3,0	10,5	0,5	20,5	3,0	0,0	10,5	10,5	3,0	3,0	0,5	0,0	0,0	0,5	37,5	0,0	0,5
76	20,5	0,0	3,0	3,0	3,0	10,5	10,5	0,0	3,0	3,0	20,5	37,5	0,0	0,5	0,5	0,5	3,0	0,0	0,5
77	10,5	0,5	0,5	20,5	0,5	3,0	10,5	0,0	3,0	0,5	0,5	20,5	0,0	0,0	0,0	0,5	3,0	0,5	0,5
78	3,0	0,0	0,5	20,5	0,5	0,5	0,5	0,5	3,0	0,0	0,5	10,5	0,0	0,0	0,5	0,5	0,5	0,5	0,5
79	0,5	0,5	0,5	0,5	0,5	0,0	0,5	0,5	0,5	0,5	0,0	0,0	0,0	0,0	0,5	0,5	0,5	0,0	0,5
80	0,5	0,0	0,5	0,5	0,5	0,0	0,0	0,5	3,0	10,5	0,5	0,0	0,5	0,5	0,5	0,0	0,5	0,0	0,5
81	0,0	0,0	0,0	0,0	10,5	0,5	0,0	0,0	0,0	87,5	3,0	0,5	0,0	0,0	0,0	0,5	0,5	0,5	0,0
82	0,0	0,0	0,0	0,0	0,5	0,0	0,0	0,0	0,0	87,5	3,0	0,5	0,0	0,0	0,0	0,5	0,5	0,5	0,0
83	0,0	0,0	0,0	0,0	0,5	0,0	0,0	0,0	0,0	20,5	20,5	3,0	0,0	10,5	0,0	0,5	0,5	10,5	0,0
84	0,0	0,5	0,5	10,5	0,5	0,0	0,0	10,5	0,5	0,5	0,5	0,0	0,5	0,5	0,0	0,5	0,5	0,0	0,5
85	0,0	0,0	0,0	0,0	0,5	0,0	0,0	0,0	0,0	0,5	0,5	0,0	0,0	0,5	0,0	0,5	0,5	0,5	0,0
86	0,0	0,0	20,5	20,5	0,0	0,5	0,5	3,0	20,5	10,5	37,5	0,0	0,0	0,0	0,0	0,5	62,5	0,5	0,5
87	0,0	0,0	10,5	3,0	10,5	0,0	0,5	0,5	3,0	62,5	0,5	0,0	0,0	0,5	0,0	0,0	0,5	0,0	0,0
88	0,5	0,5	20,5	3,0	0,5	0,5	0,0	0,5	0,5	20,5	20,5	0,0	3,0	0,5	0,5	0,5	37,5	0,5	0,5
89	0,0	3,0	3,0	20,5	3,0	0,0	10,5	3,0	3,0	3,0	3,0	0,0	3,0	20,5	0,0	0,5	0,5	0,5	0,0
90	0,0	10,5	10,5	10,5	3,0	0,0	3,0	0,5	3,0	20,5	37,5	0,0	0,5	0,5	0,0	0,5	62,5	0,0	0,5
91	0,0	20,5	20,5	0,5	0,5	0,0	0,0	0,5	10,5	20,5	0,0	0,0	0,5	0,5	0,0	0,5	10,5	0,5	0,5
92	0,0	20,5	10,5	10,5	0,5	0,0	0,5	0,5	3,0	10,5	10,5	0,0	0,5	0,5	0,5	0,5	3,0	0,5	0,5

Continuación de la Tabla XII

	ESPECIE																		
CENSO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
93	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	3,0	0,0	0,5	0,5	37,5	0,0	0,5	0,5	20,5	0,5
94	0,0	0,0	0,0	0,5	0,5	0,5	0,0	0,5	10,5	20,5	62,5	3,0	0,5	0,0	0,5	0,5	62,5	0,5	0,0
95	0,0	0,0	0,0	0,0	3,0	0,0	0,0	0,0	0,5	87,5	0,5	0,5	0,5	0,5	0,5	0,0	3,0	0,0	0,5
96	0,5	0,0	3,0	20,5	3,0	0,0	0,5	3,0	10,5	20,5	20,5	0,0	0,5	0,5	0,5	0,5	37,5	0,0	0,5
97	0,5	0,5	0,5	20,5	0,5	0,5	0,0	0,5	3,0	10,5	3,0	0,5	0,5	0,5	0,0	0,5	0,5	0,5	0,5
98	0,5	0,5	0,5	0,5	0,5	0,5	0,0	3,0	0,5	10,5	3,0	0,5	10,5	0,5	0,5	0,5	3,0	0,5	0,5
99	0,5	0,5	0,5	0,5	0,5	0,0	0,0	0,5	0,5	3,0	0,5	0,0	0,0	0,5	0,0	0,5	3,0	0,0	0,5
100	0,5	0,5	0,5	20,5	0,5	0,0	0,5	0,5	3,0	3,0	0,5	0,0	0,5	0,5	0,5	0,5	3,0	0,0	0,5
101	0,5	0,5	0,0	3,0	0,0	0,0	0,5	3,0	3,0	10,5	0,5	0,0	3,0	0,5	0,5	0,5	20,5	0,0	0,5
102	0,5	0,5	3,0	0,5	3,0	10,5	0,5	0,5	0,5	20,5	0,5	0,5	0,5	0,5	0,5	0,5	20,5	0,0	0,0
103	0,0	0,5	0,0	0,5	0,5	10,5	0,5	0,5	0,5	37,5	3,0	20,5	0,5	0,5	0,5	0,5	10,5	0,5	0,5
104	0,0	0,0	0,5	20,5	10,5	3,0	0,0	3,0	3,0	10,5	20,5	10,5	0,0	0,5	0,5	0,5	10,5	0,0	0,0

trigonométrico. Así, el valor de X' es el ángulo cuyo seno es la raíz cuadrada de (X + 0,01). La constante 0,01 se ha utilizado para eliminar los problemas del exceso de ceros en la matriz original que dificultan el cálculo de los valores y vectores propios. La distribución de los datos originales (Fig. 6a) exhibe una marcada asimetría, lo que puede apreciarse visualmente por la posición relativa de la línea punteada vertical que indica el valor del promedio. En la figura 6b, los datos transformados están mejor centrados, se agrupan los valores alrededor del promedio (línea vertical punteada) y se reduce *la cola* de los valores positivos.

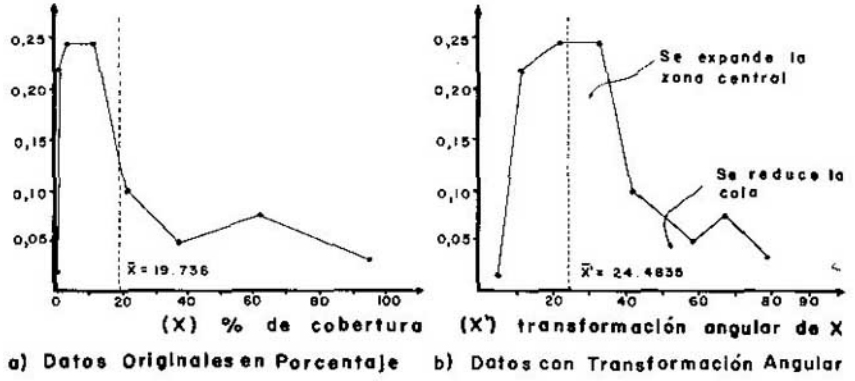


Fig. 6. Representación frecuencial del porcentaje de cobertura de *Prosopis juliflora* L. Al aplicar la transformación angular, la distribución de las observaciones --expresadas como porcentaje de cobertura-- es aproximadamente normal.

Todos los cálculos de este ejemplo fueron hechos con los servicios que brinda el paquete estadístico STAT MOD (1982)[34] en una microcomputadora Apple IIe de 128 kbites de memoria (el programa y la memoria son suficientes si se utiliza la versión de 64 kbites). Los valores propios y los coeficientes de los vectores propios pueden variar a nivel del tercer o cuarto dígito significativo si se utilizan los mismos datos con otro programa, esto se debe a que existen distintos algoritmos para los cálculos y porque la memoria del computador mantiene diferentes grados de precisión al realizar los pasos intermedios.

2. ANALISIS DE LOS DATOS ORIGINALES

2.1. Matriz de Covarianza

A partir de los datos de la Tabla XII se calculó el vector promedio y la matriz de varianza-covarianza que se presentan en la Tabla XIII. La especie que tiene la mayor cobertura promedio es *Prosopis juliflora* (19,5%); le sigue en orden decreciente *Opuntia caracasana* (11,5%), *Ritterocereus* spp (10,6%), *Castela erecta* 8,4% y *Caesalpinia coriaria* (7,4%). Los porcentajes de cobertura promedio de todas las demás especies son inferiores al 5%, siendo la menos abundante *Cnidoscolus urens* y las plantas de la familia Loranthacea. El promedio de cobertura total por sitio fue 84,68%.

Tabla XIII: Vector promedio y matriz de covarianza datos sin transformar

Especie	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Cobertura (%)	2,29	1,39	2,68	7,37	2,55	4,75	1,48	2,21	2,69	19,54	10,59	8,42	4,17	1,03	0,30	0,49	11,47	0,93	0,33

Especie	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	23,270																		
2	0,082	15,343																	
3	1,417	6,337	27,393																
4	2,006	2,003	4,276	89,966															
5	0,221	0,015	2,224	-2,559	24,998														
6	8,664	-1,013	5,861	-11,880	7,126	60,777													
7	4,083	0,386	1,637	1,738	1,185	6,228	8,801												
8	1,948	-0,723	3,545	5,205	2,524	7,105	0,817	29,746											
9	3,342	2,047	10,116	6,838	1,199	4,130	1,749	2,333	14,696										
10	-21,314	-4,498	2,890	-47,572	42,879	-5,042	-9,358	-11,941	-7,809	509,632									
11	-1,676	9,157	12,720	38,744	8,006	13,881	1,308	7,962	14,010	-5,742	196,793								
12	24,001	-6,489	-6,639	5,104	7,227	21,575	6,506	-4,712	-0,444	-28,400	3,342	134,654							
13	-2,233	-1,363	12,477	1,563	6,953	25,529	7,785	42,255	4,528	-1,952	3,888	-5,844	145,751						
14	-1,715	-0,261	-1,131	-0,656	-1,145	-3,412	1,020	-0,271	-1,345	-8,220	-4,177	-5,692	-1,075	18,067					
15	-0,110	0,078	0,344	0,028	0,398	0,580	0,297	0,392	0,172	0,038	0,534	0,634	0,703	-0,188	0,206				
16	0,226	-0,003	0,001	0,147	-0,064	-0,138	0,049	-0,228	-0,067	-2,621	-0,407	1,253	-0,269	-0,006	-0,018	0,224			
17	-0,882	7,800	21,697	20,139	7,487	21,056	2,254	8,518	22,370	13,365	118,456	-24,412	12,706	-6,919	0,099	-0,461	283,339		
18	-1,374	-0,534	-1,556	-4,412	-0,625	1,616	-0,606	3,461	-1,603	-5,449	-0,081	-4,049	14,434	7,951	-0,178	0,017	4,173	9,665	
19	0,198	0,210	0,092	0,094	-0,424	-0,447	0,266	-0,067	0,038	-1,810	-0,686	-0,768	-0,366	-0,034	0,007	0,002	0,284	-0,051	0,130

Si se efectúa una prueba de asociación entre la frecuencia por especie y la cobertura promedio se obtiene un valor de r, el coeficiente de correlación lineal de 0,72, el cual al ser sometido a una prueba de t resulta significativamente diferente de cero. Esta asociación positiva entre la cobertura y la frecuencia no puede considerarse como un modelo de predicción, pues con los datos disponibles es imposible determinar su bondad de ajuste.

Si se estudian las varianzas de las especies puede detectarse una asociación con el promedio. Así, las especies más abundantes son también las que tienen una mayor varianza (especies 10, 17, 11, 13, 12, 4, 6) y las especies con menor cobertura promedio tienen las menores varianzas (especies 15, 19, 16, 7, 18).

Al estudiar la estructura de las covarianzas, se advierte que *Prosopis juliflora* tiene covarianzas negativas (lo que indica una relación inversa) con la mayoría de las especies, salvo con *Capparis odoratissima* y *Opuntia caracasana*. Sin embargo, es necesario estudiar la composición de la matriz de correlación para conocer la magnitud de estas relaciones, ya que --como se ha señalado-- las varianzas son muy diferentes entre sí, lo que dificulta la interpretación mediante estudio directo de la matriz de covarianzas.

Si se examinan los coeficientes de cada variable en el primer vector propio de la Tabla XIV se verá --como era de esperar-- que éste representa casi exclusivamente a la especie 10 *(Prosopis juliflora)*, ya que le corresponde el máximo valor de la varianza determinado por su alta cobertura en los sitios estudiados. El segundo componente sintetiza a las cactáceas, pues posee elevados coeficientes para las especies 17 y 11 *(Ritterocereus* spp. y *Opuntia caracasana)*. El tercer componente sintetiza el aporte de la especie 13 *(Ipomoea carnea)* y en mucho menor grado de las especies 6 y 8 *(Cercidium praecox* y *Pereskia guamacho)*. El cuarto componente corresponde casi exclusivamente a la especie 12 *(Castela erecta)* y en menor grado a la especie 8. Si se observan los valores de las covarianzas para los pares de especies que aparecen en un mismo componente se comprueba que son siempre los máximos valores si se consideran las covarianzas entre cualquiera de las dos especies y cada una de las restantes. Así, por ejemplo, la covarianza entre las especies 17 y 11 es 118,456 (Tabla XIII), la cual es mayor que cualquiera de las covarianzas de la especie 17 con las restantes y mayor también que cualquiera de las covarianzas de la especie 11 con las demás. Igual situación se presenta si se compara la covarianza de las especies 13 y 6 y de las especies 13 y 8. Nunca se insistirá demasiado en que estos altos valores absolutos de covarianza no corresponden necesariamente a los más altos valores de correlación. Si en la Tabla XV se observa el valor de la correlación entre la especie 13 y la 6, se verá que es inferior a la correlación de la especie 13 con la especie 18.

Desde el sexto vector propio hasta el último, se observa que existe una correspondencia con alguna especie en particular. Esto indica que cada componente principal resumirá la información aportada por una sola especie; su interpretación no será muy diferente de la que se hace a partir de los datos originales y este análisis no sirve para sintetizar la información contenida en los datos originales. Los coeficientes de las variables que más se asocian con cada componente han sido subrayados en la Tabla XIV para facilitar su lectura. Asimismo, en la Tabla XIII se han subrayado aquellos valores de covarianza a que se han hecho referencia hasta ahora.

Dado que los componentes principales a partir de la matriz de covarianza no constituyen una buena síntesis de las variables en estudio,

Tabla XIV. Valores y vectores propios para los datos sin transformar a partir de la matriz de covarianza

Valor Propio	523,445	381,256	168,000	151,568	117,118	75,962	42,601	30,076	21,255	19,676	18,526	13,669	12,559	7,859	7,028	2,371	0,205	0,160	0,089
Varianza Acumulada (%)	32,85	56,78	67,32	76,83	84,18	88,95	91,62	93,51	94,84	96,08	97,24	98,10	98,89	99,38	99,82	99,97	99,98	99,99	100

VECTORES PROPIOS

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0,0465	0,0040	0,0071	0,1599	0,1161	0,0420	0,1415	0,1868	-0,0707	-0,3108	-0,6852	0,4733	0,1264	-0,2382	0,1654	-0,0787	-0,0100	0,0194	-0,0109
0,0076	-0,0343	0,0156	-0,0352	-0,0414	-0,0307	0,0854	0,3128	0,1007	0,0525	0,0795	0,3441	-0,8269	0,2513	0,0769	-0,0187	0,0024	0,0194	-0,0109
-0,0070	-0,0792	-0,0806	-0,0196	-0,0130	0,0235	0,2888	0,7268	0,1953	0,1572	0,0824	-0,2979	0,0788	-0,4362	-0,1185	0,0730	-0,0032	-0,0127	-0,0027
0,1088	-0,1329	0,0835	0,0958	-0,4283	0,7914	0,3127	-0,1926	0,0505	0,0467	0,0114	0,0156	-0,0361	-0,0343	0,0324	0,0345	-0,0008	0,0022	-0,0002
-0,0842	-0,0290	-0,0464	0,1057	0,0106	-0,0052	0,0956	0,0317	0,1949	-0,7751	0,5386	0,1504	0,0848	-0,0690	0,0690	0,0066	-0,0023	-0,0116	0,0104
0,0103	-0,0784	-0,2141	0,2366	0,2203	-0,3044	0,7664	-0,3616	0,0255	0,1243	0,0111	-0,0822	-0,0831	0,0082	0,0481	-0,0177	0,0070	0,0009	0,0091
0,0195	-0,0106	-0,0493	0,0531	0,0272	0,0097	0,0855	0,0356	0,0980	-0,0126	-0,0548	0,2974	0,1429	0,2574	-0,8220	0,3436	0,0215	-0,0340	-0,0407
0,0240	-0,0482	-0,2821	0,0134	-0,0843	0,0388	-0,0554	-0,0129	-0,1893	-0,4658	-0,3604	-0,5991	-0,3326	0,1007	-0,1823	0,0975	0,0171	-0,0108	-0,0023
0,0150	-0,0777	-0,0208	0,0101	0,0086	0,0458	0,1640	0,3462	0,0333	-0,0220	-0,0668	-0,1176	0,3578	0,7646	0,3289	0,0599	0,0086	0,0027	0,0044
-0,9842	0,0201	0,0060	0,0940	-0,0493	0,0866	0,0217	-0,0105	0,0040	0,0546	-0,0876	-0,0057	-0,0173	0,0187	-0,0065	0,0100	0,0046	0,0036	0,0022
0,0132	-0,5624	0,1639	0,3268	-0,5912	-0,4286	-0,1023	0,0072	-0,0183	0,0163	-0,0403	0,0277	0,0496	-0,0233	-0,0083	-0,0085	0,0022	0,0006	0,0050
0,0779	0,0641	0,0805	0,8506	0,3316	0,1804	-0,2640	0,0757	0,0563	0,1032	0,0729	-0,1223	-0,0806	0,0327	-0,0074	0,0011	-0,0055	-0,0092	0,0078
0,0051	-0,0795	-0,9018	0,0774	-0,1350	0,0780	-0,2022	0,0252	-0,0169	0,1545	0,0879	0,2177	0,0857	-0,0135	0,0590	-0,1350	-0,0020	0,0030	0,0062
0,0156	0,0218	-0,0016	-0,0528	-0,0301	-0,0254	-0,0822	-0,1287	0,8283	-0,0222	-0,2322	-0,1200	0,0020	0,0848	-0,1076	-0,4446	0,0021	0,0096	0,0166
0,0000	-0,0014	-0,0046	0,0063	-0,0006	-0,0020	0,0066	0,0077	-0,0037	-0,0077	0,0170	-0,0125	-0,0043	0,0159	-0,0388	0,0074	-0,4999	0,8543	-0,1335
0,0052	0,0017	0,0022	0,0051	0,0046	0,0028	-0,0064	0,0060	0,0045	0,0008	0,0111	0,0044	0,0033	-0,0140	0,0033	0,0117	0,8623	0,5058	0,0069
-0,0469	-0,7948	0,0181	-0,2105	0,5134	0,2024	-0,1055	-0,0551	0,0010	0,0008	0,0104	0,0065	-0,0269	-0,0206	-0,0217	-0,0307	-0,0012	0,0008	-0,0025
0,0090	-0,0109	-0,0936	-0,0313	0,0068	-0,0513	-0,1352	-0,1505	0,4112	0,0405	-0,1187	0,0022	-0,0471	-0,1125	0,3306	0,7980	-0,0162	0,0040	-0,0191
0,0033	0,0003	0,0017	-0,0082	0,0027	0,0033	0,0009	0,0089	-0,0049	0,0011	-0,0121	0,0176	-0,0037	0,0049	-0,0321	0,0379	-0,0730	0,1105	0,9896

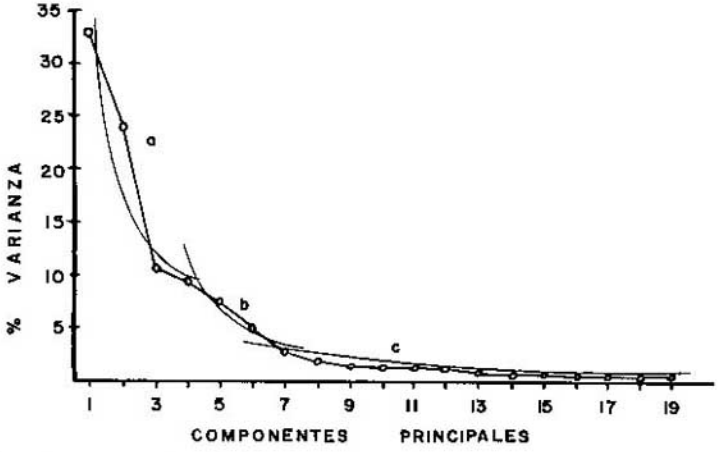


Fig. 7. Porcentaje de la varianza total explicada por cada componente principal (datos originales-covarianza). Cada onda, identificada con una letra minúscula, representa grupos de componentes. Véase el texto.

se efectuará un análisis a partir de la matriz de correlación. Esto equivale a suponer que todas las especies tienen la misma importancia respecto a la información que suministran para la interpretación de los ambientes estudiados. En el caso anterior, al trabajar con los valores de covarianza, se está asignando una ponderación relativa diferente a cada una; esa relatividad está determinada por el valor de la varianza, que --como se ha visto-- puede atribuirse a las diferencias en las magnitudes de la cobertura. Efectuar un análisis a partir de la matriz de correlación implica aceptar el criterio de que todas las especies, independientemente de su abundancia, tienen la misma importancia para caracterizar un sitio.

2.2. Matriz de Correlación

En la Tabla XV se consignan los valores de correlación para las variables sin transformar. A partir de esta matriz se generaron los valores y vectores propios presentados en la Tabla XVI. El porcentaje de la varianza explicada por cada componente se ha graficado en la figura 8.

De la inspección de las correlaciones de la Tabla XV se observará que la mayor correlación se encuentra entre *Ipomoea carnea* y *Pereskia guamacho*, seguida por la de *Sporobolus pyramidatus* con *Jatropha gossipifolia*; en ambos casos los valores son superiores a 0,60. La correlación es también alta entre *Pithecelobium unguis-cati* y *Bulnesia arborea* y entre *Ritterocereus* spp. y *Opuntia caracasana*. De menor orden, pero superior a 0,30 es también la correlación entre *Acacia tortuosa* y *Castela erecta*, entre *Sporobolus pyramidatus* e *Ipomoea carnea*, entre *Prosopis juliflora* y *Capparis odoratissima*, y entre *Phitecelobium unguis-cati* y *Opuntia caracasana*. La mayor correlación negativa se encuentra entre *Prosopis juliflora* y *Melocactus caesius*. Le siguen en orden de magnitud las correlaciones negativas entre las Loranthacea y las especies *Capparis odoratissima*, *Prosopis juliflora*, *Castela erecta* y *Cercidium praecox*, lo cual indica que es poco frecuente la aparición conjunta de Loranthacea en ambientes con elevados porcentajes de cobertura de

Tabla XV. Matriz de correlación de los datos sin transformar

Especie	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1,0000																		
2	0,0043	1,0000																	
3	0,0561	0,3091	1,0000																
4	0,0438	0,0539	0,0861	1,0000															
5	0,0092	0,0008	0,0850	-0,0540	1,0000														
6	0,2304	-0,0332	0,1437	-0,1607	0,1828	1,0000													
7	0,2853	0,0332	0,1054	0,0618	0,0799	0,2693	1,0000												
8	0,0740	-0,0338	0,1242	0,1006	0,0925	0,1671	0,0505	1,0000											
9	0,1807	0,1363	0,5042	0,1881	0,0626	0,1382	0,1538	0,1116	1,0000										
10	-0,1957	-0,0509	0,0245	-0,2222	0,3799	-0,0287	-0,1397	-0,0970	-0,0902	1,0000									
11	-0,0248	0,1667	0,1732	0,2912	0,1141	0,1269	0,0314	0,1041	0,2605	-0,0181	1,0000								
12	0,4288	-0,1428	-0,1093	0,0464	0,1246	0,2385	0,1890	-0,0744	-0,0100	-0,1084	0,0205	1,0000							
13	-0,0384	-0,0288	0,1975	0,0137	0,1152	0,2712	0,2174	0,6417	0,0978	-0,0072	0,0230	-0,0417	1,0000						
14	-0,0836	-0,0157	-0,0509	-0,0163	-0,0539	-0,1030	0,0809	-0,0117	-0,0826	-0,0857	-0,0701	-0,1154	-0,0210	1,0000					
15	-0,0503	0,0441	0,1447	0,0064	0,1754	0,1641	0,2207	0,1582	0,0987	0,0037	0,0839	0,1215	0,1284	-0,0977	1,0000				
16	0,0992	-0,0018	0,0003	0,0328	-0,0270	-0,0374	0,0347	-0,0885	-0,0369	-0,2453	-0,0612	0,2281	-0,0471	-0,0032	-0,0839	1,0000			
17	-0,0109	0,1183	0,2463	0,1261	0,0890	0,1605	0,0451	0,0928	0,3467	0,0352	0,5016	-0,1250	0,0625	-0,0967	0,0129	-0,0579	1,0000		
18	-0,0916	-0,0438	-0,0956	-0,1496	-0,0402	0,0667	-0,0657	0,2041	-0,1345	-0,0776	-0,0019	-0,1122	0,3846	0,6017	-0,1265	0,0113	0,0797	1,0000	
19	0,1141	0,1485	0,0486	0,0274	-0,2350	-0,1592	0,2488	-0,0343	0,0278	-0,2225	-0,1357	-0,1837	-0,0842	-0,0221	0,0417	0,0100	0,0468	-0,0451	1,0000

estas especies. La correlación es negativa entre *Prosopis juliflora* y *Caesalpinia coriaria;* ello indica que en ambientes con alto porcentaje de cobertura de una de ellas es baja la cobertura de la otra. Es también negativa la correlación entre *Prosopis juliflora* y *Acacia tortuosa,* aunque no existe correlación entre *Caesalpinia coriaria* y *Acacia tortuosa.* Esto refleja la existencia de ambientes donde predomina *P. juliflora* con baja cobertura de *C. coriaria,* o a la inversa; estas dos situaciones se darán en algunos casos con altas coberturas de *A. tortuosa* y en otros con baja cobertura. En otros ambientes predomina *P. juliflora* con baja cobertura de *A. tortuosa,* o a la inversa, y en estos ambientes puede encontrarse *C. coriaria* sin un patrón de cobertura definido.

En la figura 8 se han graficado los porcentajes de la varianza total explicados por cada componente. La gráfica puede dividirse en cuatro partes o cuatro ondas, la primera, a, que incluye del primero al cuarto componente sintetiza el 44,29% de la variación total; la segunda onda, b, contiene del quinto al décimoprimer componente y sintetiza un 38,26% de la varianza; la tercera, c, abarca del décimosegundo al décimooctavo componente con un total de 16,75%; en la última onda, d, figura sólo el décimonoveno componente con un 0,76% de la variación total. Este último componente es el único que explica menos del 1% de la variación. Al seleccionar el grupo de componentes que se utilizarán en cualquier análisis posterior para sintetizar todo el conjunto de datos debe tenerse en cuenta que al pasar de una onda a otra existe un punto de inflexión en la curva, es decir un cambio de magnitud de las varianzas explicadas por cada componente. Por tal motivo, se recomienda mantener, para el análisis posterior, la totalidad (o ninguno) de los componentes de una onda. En este caso, habría que decidir si mantener cuatro componentes --sólo la primera onda-- u once componentes --las dos primeras ondas.

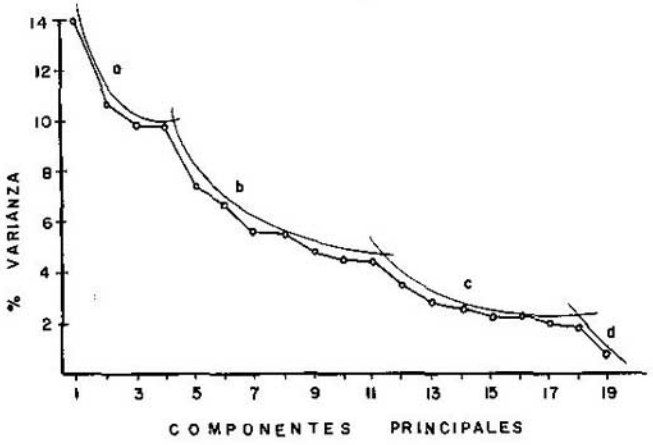


Fig. 8. Porcentaje de la varianza total explicada por cada componente principal (datos originales - correlación). El número de ondas es mayor que el de la figura 7. Se observa que los porcentajes explicados por los primeros componentes son menores en este caso en que se utiliza la matriz de correlación.

En el análisis que se hará posteriormente se utilizará la información sintetizada por los cuatro primeros componentes por tratarse de la búsqueda de relaciones gráficas y asociaciones con factores ambientales. Si la información se ha de usar en un análisis por conglomerados (formación de grupos o clases con vegetación similar), o para un análisis

Tabla XVI. Valores y vectores propios para los datos sin transformar a partir de la matriz de correlación

Valor Propio	2,6653	2,0350	1,8631	1,8664	1,3866	1,2658	1,0458	1,0340	0,9050	0,8339	0,8112	0,6562	0,5302	0,4701	0,4257	0,4046	0,3670	0,3252	0,1417
Varianza Acumulada (%)	14,00	24,70	34,48	44,29	51,58	58,23	63,72	69,15	73,91	78,29	82,55	86,00	88,79	91,26	93,49	95,62	97,55	99,26	100

V E C T O R E S P R O P I O S

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0,1766	-0,1784	0,4068	-0,1906	-0,0004	-0,0722	-0,3058	-0,0208	0,3369	0,1664	0,2807	0,3311	-0,0260	0,0113	0,3485	0,0506	-0,2534	-0,3245	0,1387
0,1419	-0,1644	-0,0322	0,2723	-0,0455	0,3573	0,0587	0,3024	-0,2226	0,0802	0,6922	-0,1068	0,0120	-0,0915	-0,0734	0,2822	0,0752	0,0740	0,0014
0,3575	-0,1151	-0,1169	0,1660	-0,0745	0,2422	-0,0976	0,4189	0,1611	-0,2372	-0,1313	-0,0155	0,1977	0,2902	0,0167	-0,4851	-0,3210	0,0846	-0,1308
0,1519	-0,1990	0,0304	0,2200	0,1340	-0,3656	0,5186	-0,0986	0,3359	0,0631	0,0670	-0,2415	-0,0300	0,4886	-0,0194	0,2115	-0,0131	-0,0755	-0,0778
0,1688	0,0823	-0,1686	-0,3562	0,1862	0,3338	0,2822	0,0871	0,1427	0,3734	-0,0412	0,0886	-0,5414	-0,0177	-0,2731	-0,1522	-0,0707	-0,0602	0,0137
0,3125	0,1171	0,0948	-0,3084	0,0430	0,0367	-0,3628	-0,1092	-0,2727	-0,2080	0,0719	-0,3416	-0,3084	0,4629	0,1273	0,0504	0,2320	0,0512	0,0118
0,2658	-0,0364	0,3225	-0,0881	-0,2276	0,2462	0,1086	-0,3155	0,0702	0,1323	-0,1628	-0,5038	0,1781	-0,3474	0,0798	0,0302	-0,1403	0,0612	-0,2764
0,3039	0,3506	-0,0635	0,0187	-0,2397	-0,3593	0,0895	0,1057	0,0754	0,1745	0,1424	0,2317	-0,1328	-0,1428	0,3070	-0,0550	0,0951	0,4812	-0,2956
0,3814	-0,1991	-0,0564	0,1438	0,0560	0,0418	-0,1680	0,1371	0,3307	-0,2610	-0,3012	0,1352	-0,1278	-0,3020	-0,1615	0,3183	0,4745	-0,1137	-0,0342
-0,0683	0,0847	-0,4233	-0,2623	0,0874	0,3632	-0,0178	-0,0077	0,2318	0,2676	-0,0487	0,0111	0,4781	0,2031	0,3232	0,1852	0,2726	0,0073	-0,0517
0,3068	-0,1143	-0,1688	0,1374	0,4299	-0,1029	0,1340	-0,2698	-0,2287	0,0570	0,1494	-0,0034	0,1260	-0,2294	0,2257	-0,5024	0,2726	-0,2358	0,0782
0,0839	-0,1382	0,3566	-0,4313	0,2022	-0,1316	0,0919	0,0193	0,0036	-0,0895	0,1460	0,1946	0,3983	0,0549	-0,4430	-0,0763	0,1798	0,3588	-0,0209
0,3175	0,4429	-0,0239	-0,0263	-0,2323	-0,1954	0,0347	0,1732	-0,0034	0,1519	-0,0103	-0,1837	0,2719	-0,0410	-0,2414	0,0507	-0,0022	-0,2619	0,5654
-0,0878	0,3545	0,2814	0,2672	0,2567	0,3183	0,1564	-0,1043	0,2462	-0,2378	-0,0272	0,0454	-0,0823	0,0186	0,2245	-0,0371	0,0488	0,3276	0,4164
0,2184	-0,0158	-0,0448	-0,1780	-0,2915	0,1723	0,4744	-0,1787	-0,3205	-0,4028	-0,0716	0,4168	0,0373	0,0689	0,1479	0,1824	-0,1024	-0,1751	0,0340
-0,0305	-0,1018	0,3478	-0,0152	0,1783	-0,0877	0,1576	0,5479	-0,3829	0,2577	-0,3994	-0,0168	0,0126	0,0153	0,2991	0,0965	0,1029	-0,0409	0,0132
0,3163	-0,0767	-0,1863	0,2068	0,3085	0,0147	-0,2242	-0,2630	-0,2401	0,2610	-0,2188	0,1808	0,0766	0,0662	-0,1128	0,3334	-0,4164	0,3109	0,1067
0,0098	0,5471	0,1885	0,2204	0,2271	0,0924	-0,0773	-0,0271	-0,0636	0,0098	0,0324	0,1833	0,1134	0,1391	-0,2157	0,0736	0,0299	-0,3636	-0,5185
-0,0014	-0,1643	0,2055	0,3062	-0,4571	0,1588	-0,0823	-0,2403	-0,0753	0,3841	-0,1247	0,2366	-0,0356	0,3066	-0,1553	-0,2261	0,3682	0,0208	0,0922

discriminante (asignación de censos a clases predeterminadas), se recomienda incluir hasta el décimoprimer componente, con lo cual se utilizará casi el 85% de la variabilidad total contenida en los datos.

Si se recuerda que los coeficientes de cada variable en un vector propio indican el grado de influencia de la variable en el componente principal, es posible interpretar estos resultados inspeccionando los coeficientes de los cuatro primeros vectores. Se observa que:

a) El primer componente tendrá valores elevados (superiores al promedio) cuando la cobertura de las especies 3, 6, 8, 11, 13 y 17 sea alta y no tendrán influencia las especies 10, 12, 14, 18 y 19. Así, en un ambiente con valores similares de cobertura de *Bulnesia arborea* (3), *Cercidium praecox* (6), *Pereskia guamacho* (8) y *Ritterocereus* (11), todas del estrato superior, y en que el piso tenga coberturas de *Ipomoea carnea* (13) y de *Opuntia caracasana* (17) también similares entre sí, pero que en un caso tenga altos porcentajes de cobertura de *Prosopis juliflora* (10) y en otro esta especie no esté ni siquiera presente, el primer componente tendrá valores similares. Para diferenciarlos habrá que observar el valor de otro componente donde *P. juliflora* tenga una incidencia detectable; por ejemplo, el tercero, donde *P. juliflora* contribuye con signo negativo. En el caso anterior, en un censo con una alta cobertura de *P. juliflora* el valor del tercer componente quedará por debajo del promedio y en otro con bajo porcentaje de cobertura el valor del tercer componente estará por encima del promedio.

b) El segundo componente tendrá valores superiores al promedio cuando la cobertura de las especies 8, 13, 14 y 18 sea alta; y en él no influyen las especies 5, 10, 15 y 17. Se puede decir que este componente sintetiza el aporte de *Pereskia guamacho* (8) y de las especies del estrato inferior *Ipomoea carnea* (13), *Jatropha gossipifolia* (14) y *Sporobolus pyramidatus* (18). Estas últimas son las tres especies que tienen la máxima varianza del conjunto total de especies del estrato inferior (12 a 18), con excepción de *Opuntia caracasana* (17), ya presente en el primer componente.

c) El tercer componente constituye un contraste entre la especie 10 y las especies 1, 7, 12 y 16, ya que el signo del coeficiente de la especie 10 *(Prosopis juliflora)* es negativo y el de las especies *Acacia tortuosa* (1), *Mimosa arenosa* (7), *Castela erecta* (12) y *Melocactus caesius* (16) es positivo.

d) El cuarto componente constituye también un contraste entre cuatro especies que están presentes con coeficientes negativos: *Capparis odoratissima* (5), *Cercidium praecox* (6), *Prosopis juliflora* (10) y *Castela erecta* (12), y cinco especies que tienen coeficientes positivos: *Bourreria cumanensis* (2), *Jatropha gossipifolia* (14) y *Loranthaceas* (19) y, en menor proporción, *Sporobolus pyramidatus* (18) y *Opuntia caracasana* (19).

3. ANALISIS DE LOS DATOS TRANSFORMADOS

Si a cada observación de la matriz de datos originales presentada en la Tabla XI se aplica la transformación angular señalada en la ecuación [37] se obtiene una nueva matriz de datos de dimensión (104 x 19) con la que puede calcularse una nueva matriz de covarianza y una nueva matriz de correlación. En la Tabla XVII se consigna el promedio de las variables transformadas calculado a partir de la matriz completa, ya que no podrán obtenerse los promedios transformados aplicando la transformación angular a los promedios de los datos originales. En la Tabla XVIII puede verse la matriz de covarianza y en la Tabla XIX la matriz de

correlación. A partir de cada una de estas matrices se calcularon los valores propios que se presentan en la Tabla XX; se indica el porcentaje de varianza explicado en forma acumulativa. En las figuras 9 y 10 se grafican los porcentajes de varianza explicados por cada componente.

Tabla XVII. Vector promedio - datos transformados

Especie	Cobertura Promedio (%)
1	9,05
2	7,79
3	9,54
4	14,49
5	9,62
6	11,68
7	8,16
8	8,80
9	10,02
10	24,48
11	17,19
12	14,95
13	10,26
14	7,32
15	6,49
16	6,95
17	17,64
18	7,28
19	6,57

Si se comparan los valores promedios de la Tabla XI con los de la Tabla XVII se observará que la dispersión de estos últimos es menor y que su distribución es más simétrica que la de los datos originales. Así, la transformación angular permite no sólo acercarse a la distribución normal en cada variable, como ya se ha señalado, sino también normalizar el promedio de las variables. Conviene recordar que para poder efectuar pruebas de hipótesis se requiere multinormalidad de los datos y que --como se ha expresado-- ésta puede lograrse mediante una transformación adecuada o aumentando el tamaño de la muestra o aumentando el número de variables incluidas en el análisis.

Si se comparan las figuras 7 y 9 se advertirá que es menos brusco el descenso de la variación explicada del primer al segundo valor propio. Sin embargo, como sucede con los valores generados a partir de los datos sin transformar, los cuatro primeros valores propios constituyen un primer grupo, ya que a partir del quinto la proporción de varianza explicada por cada uno es menor. Considerar los cuatro primeros componentes equivale a incluir en cualquier análisis posterior casi el 70% de la variación total, como se observa en la Tabla XX; en cambio, en el caso de los datos sin transformar este porcentaje alcanzaba el 76,83% (Tabla XIV). Si bien el porcentaje total ha disminuido, no es uniforme la proporción de la varianza de cada variable incluida si se utilizan los datos sin transformar. En la Tabla XXI se muestran los vectores propios para los valores generados a partir de la matriz de covarianza y puede observarse que la contribución de cada especie en los primeros cuatro componentes está mejor distribuida que cuando se utilizan los datos sin transformar.

Tabla XVIII. Matriz de covarianza de los datos transformados

Especie	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	30,449																		
2	0,110	20,571																	
3	1,996	9,373	34,947																
4	4,790	3,630	6,690	86,103															
5	0,388	1,058	4,856	-0,074	29,005														
6	11,967	-0,879	7,612	-10,682	12,092	64,624													
7	7,679	2,194	4,252	4,626	3,110	10,396	16,805												
8	1,282	0,362	4,617	9,733	5,826	8,538	2,239	33,380											
9	5,791	3,949	15,165	11,446	3,167	6,702	4,687	4,473	24,304										
10	-23,116	-3,590	4,360	-34,278	32,172	-1,493	-9,839	-8,623	-6,986	255,428									
11	-3,018	7,083	12,803	33,780	13,822	15,716	1,778	10,807	12,040	8,273	129,391								
12	24,259	-7,765	-7,450	4,306	10,356	31,081	8,386	-6,321	1,720	-16,950	7,874	116,478							
13	-2,037	1,287	11,317	1,756	10,815	22,258	8,296	38,339	5,065	4,914	6,940	-1,726	96,920						
14	-2,691	0,114	-1,020	-0,551	-1,246	-4,646	1,183	0,524	-1,883	-4,882	-2,803	-6,419	0,822	16,091					
15	-0,268	0,364	1,078	0,236	0,993	0,898	0,768	1,100	0,539	0,490	1,249	1,005	1,837	-0,394	0,921				
16	0,657	0,061	0,116	0,411	-0,095	0,067	0,234	-0,481	-0,045	-4,211	-0,224	2,245	-0,570	-0,013	-0,079	0,865			
17	-0,756	10,203	19,565	18,766	13,011	17,385	4,551	13,687	19,497	28,017	78,116	-12,779	18,175	-5,695	0,922	-0,694	170,065		
18	-2,209	-0,703	-2,628	-6,095	-0,713	1,445	-0,993	3,472	-2,980	-2,949	0,390	-3,847	10,801	7,956	-0,498	0,122	3,767	12,112	
19	0,774	0,733	0,421	0,529	-1,155	-1,252	0,698	-0,152	0,299	-3,428	-1,312	-2,021	-0,565	-0,108	0,073	0,009	0,636	-0,160	0,644

Tabla XIX. Matrix de correlación de los datos transformados

Especie	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1,0000																		
2	0,0044	1,0000																	
3	0,0612	0,3496	1,0000																
4	0,0935	0,0862	0,1220	1,0000															
5	0,0131	0,0433	0,1525	-0,0015	1,0000														
6	0,2698	-0,0241	0,1602	-0,1432	0,2793	1,0000													
7	0,3403	0,1180	0,1755	0,1216	0,1409	0,3155	1,0000												
8	0,0402	0,0138	0,1352	0,1816	0,1872	0,1838	0,0946	1,0000											
9	0,2129	0,1766	0,5204	0,2502	0,1193	0,1691	0,2319	0,1571	1,0000										
10	-0,2621	-0,0495	0,0462	-0,2311	0,3738	-0,0116	-0,1502	-0,0934	-0,0887	1,0000									
11	-0,0481	0,1373	0,1904	0,3200	0,2256	0,1719	0,0381	0,1644	0,2147	0,0445	1,0000								
12	0,4073	-0,1586	-0,1168	0,0430	0,1782	0,3582	0,1896	-0,1014	0,0323	-0,0983	0,0641	1,0000							
13	-0,0375	0,0288	0,1945	0,0192	0,2040	0,2812	0,2056	0,6740	0,1044	0,0312	0,0620	-0,0162	1,0000						
14	-0,1216	0,0063	-0,0430	-0,0148	-0,0577	-0,1441	0,0719	0,0226	-0,0952	-0,0761	-0,0614	-0,1483	0,0208	1,0000					
15	-0,0507	0,0837	0,1900	0,0265	0,1922	0,1125	0,1952	0,1985	0,1140	0,0320	0,1144	0,0971	0,1944	-0,1024	1,0000				
16	0,1280	0,0146	0,0212	0,0476	-0,0189	0,0089	0,0614	-0,0894	-0,0098	-0,2833	-0,0212	0,2237	-0,0623	-0,0035	-0,0891	1,0000			
17	-0,0105	0,1725	0,2538	0,1551	0,1853	0,1658	0,0851	0,1817	0,3033	0,1344	0,5266	-0,0908	0,1416	-0,1089	0,0737	-0,0572	1,0000		
18	-0,1150	-0,0445	-0,1277	-0,1887	-0,0380	0,0516	-0,0696	0,1727	-0,1737	-0,0530	0,0099	-0,1024	0,3153	0,5699	-0,1490	0,0377	0,0830	1,0000	
19	0,1748	0,2014	0,0887	0,0710	-0,2672	-0,1941	0,2120	-0,0328	0,0755	-0,2672	-0,1437	-0,2333	-0,0714	-0,0336	0,0942	0,0124	0,0608	-0,0574	1,0000

Tabla XX. Valores propios para los datos transformados

A partir de la matriz de covarianza

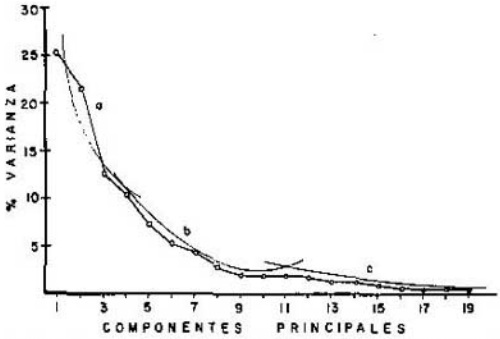
Valor Propio	Varianza Acumulada (%)
287,175	25,21
246,124	46,82
142,293	59,31
119,902	69,84
83,797	77,19
60,066	82,46
48,068	86,68
31,359	89,44
21,462	91,32
20,891	93,16
19,723	94,89
17,205	96,40
13,657	97,60
11,624	98,62
10,080	99,50
3,766	99,83
0,839	99,91
0,666	99,96
0,402	100

A partir de la matriz de correlación

Valor Propio	Varianza Acumulada (%)
3,0261	15,92
2,1559	27,26
1,9234	37,38
1,8767	47,26
1,3748	54,49
1,2235	60,93
0,9817	66,09
0,9570	71,13
0,9155	75,94
0,7616	79,95
0,7175	83,72
0,6069	86,92
0,4880	89,48
0,4373	91,79
0,3978	93,83
0,3544	95,69
0,3345	97,45
0,2931	98,99
0,1916	100

En la Tabla XXII se muestra el porcentaje de la variación de cada especie que sintetiza cada uno de los cuatro primeros componentes principales cuando se utilizan los datos sin transformar y los transformados. Los cálculos se basaron en la fórmula [36] y en los valores de las Tablas XIII y XIV para los datos sin transformar y de las Tablas XX y XXI para los datos transformados.

Igual comparación puede efectuarse entre las figuras 8 y 10; en ambos casos se distinguen cuatro ondas que corresponden a cuatro grupos de valores propios (indicados por a, b, c y d en las figuras 8 y 10).



Fig, 9. Porcentaje de la varianza total explicada por cada componente principal (datos transformados - covarianza). El número de ondas es igual al de la figura 7, pero los porcentajes han cambiando por haberse utilizado datos con transformación angular y porque el método de componentes principales es sensible a los cambios de escala.

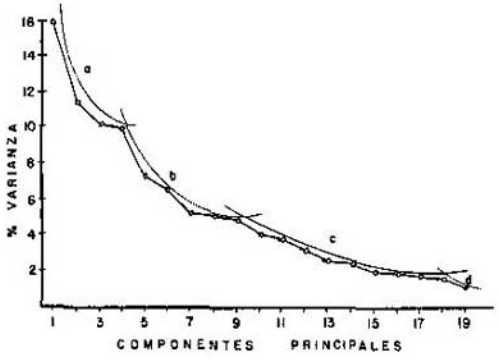


Fig. 10. Porcentaje de la varianza total explicada por cada componente principal (datos transformados - correlación). Se observan diferencias con la figura 8, ya que en el método influye la transformación angular aplicada a los datos.

En las dos figuras el primer grupo está constituido por los cuatro primeros valores, con la diferencia de que al utilizar los datos transformados el porcentaje de la varianza total sintetizada es ligeramente mayor (pasa de 44,29 a 47,26%). En la segunda onda o grupo de variables se encuentran desde el quinto al noveno valor propio, sintetizando un 28,26% de la varianza total. Tanto el número de componentes que lo integran como la variación explicada por ellos es menor que en el caso de los datos sin transformar.

De la Tabla XXII puede concluirse que se presentan diferencias significativas entre la varianza explicada para cada variable cuando se utiliza la matriz de covarianza calculada a partir de los datos originales y la varianza explicada cuando se usa la matriz calculada a partir de los datos transformados. Así, de las 14 variables en que menos del 50% de la varianza total ha sido explicada por los cuatro primeros componentes, en 13 aumenta el porcentaje explicado si se utiliza la transformación angular y en sólo una disminuye. En cambio, de las cinco especies con más del 50% de la varianza total explicada, en dos dis-

Tabla XXI. Vectores propios de los datos transformados a partir de la matriz de covarianza

VECTORES PROPIOS																		
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0,0845	0,0760	-0,1962	-0,0430	0,0667	-0,2099	-0,1653	-0,0287	-0,7802	0,1151	-0,3127	-0,0883	0,0821	0,2631	0,2461	0,0705	0,0200	0,0173	-0,0351
-0,0203	0,0581	0,0710	0,0156	0,0096	-0,0392	-0,2342	0,2640	0,1213	0,2248	-0,0223	-0,7423	0,3396	-0,3435	0,1187	0,0080	-0,0093	-0,0069	-0,0176
-0,0768	0,1072	0,0256	0,1014	-0,0225	-0,1495	-0,5725	0,4573	0,2213	-0,0755	-0,0424	0,1722	0,1311	0,5003	-0,2278	-0,0760	-0,0047	-0,0315	-0,0003
0,0471	0,2983	0,1126	-0,1759	-0,6958	-0,4639	0,0015	-0,3404	0,1742	0,0276	-0,0706	-0,0407	0,0383	0,0660	0,0674	-0,0412	0,0095	0,0060	-0,0020
-0,1440	0,0186	-0,1582	0,0403	-0,0794	0,0036	-0,1104	-0,0225	-0,0834	-0,5533	0,7321	0,0369	-0,1507	0,1559	0,1862	-0,0141	0,0043	-0,0211	0,0232
-0,0566	0,1415	-0,4247	0,1787	0,2287	0,1540	-0,4275	-0,6411	0,2465	-0,0635	-0,0788	-0,0196	0,1331	-0,0237	0,0494	0,0432	0,0031	0,0176	0,0181
0,0171	0,0666	-0,0951	0,0606	0,0043	-0,1001	-0,1571	-0,0419	-0,1053	0,2858	-0,2002	-0,1557	-0,5929	-0,2092	-0,5803	-0,2284	-0,0429	-0,0373	-0,0387
-0,0322	0,1314	-0,0202	0,3470	-0,1676	0,0120	0,0967	-0,0610	-0,3402	-0,0369	0,2293	0,1650	0,5585	-0,2216	-0,5002	-0,1009	-0,0337	-0,0043	0,0104
-0,0311	0,1295	-0,0132	0,0201	-0,0149	-0,1960	-0,3495	0,2137	-0,0762	-0,1133	0,0215	0,4553	-0,1245	-0,6491	0,3326	-0,0328	-0,0029	0,0161	0,0042
-0,8050	-0,5032	-0,1257	-0,0702	-0,1855	-0,1031	-0,0235	-0,0383	-0,0459	-0,0313	-0,1488	-0,0077	0,0481	-0,0454	-0,0235	-0,0175	-0,0072	0,0181	0,0080
-0,2880	0,4951	-0,0123	-0,2842	-0,2773	0,6810	-0,0538	0,1149	-0,1362	-0,0461	-0,0917	-0,0079	-0,0690	0,0039	0,0078	0,0100	-0,0032	0,0012	0,0117
0,1009	0,0897	-0,8063	-0,2861	-0,0282	-0,1547	0,2885	0,3052	0,1612	-0,0289	0,0016	0,0192	0,1002	-0,0517	-0,0883	-0,0079	0,0036	-0,0211	0,0216
-0,1023	0,1486	-0,1756	0,7845	-0,2329	0,0217	0,2215	0,1750	0,0557	-0,1096	-0,1194	-0,1292	-0,2491	0,0762	0,2300	0,1411	-0,0067	0,0008	0,0059
0,0254	-0,0160	0,0585	0,0401	-0,0232	0,0688	0,0631	0,0389	0,1678	0,6209	-0,3633	0,3170	0,1631	-0,0646	-0,0839	0,5459	-0,0071	0,0243	0,0267
-0,0056	0,0075	-0,0112	0,0116	-0,0087	0,0002	-0,0123	0,0217	0,0076	-0,0060	0,0363	-0,0120	-0,0206	-0,0147	-0,0684	0,0283	0,7800	0,5569	-0,2697
0,0145	0,0077	-0,0098	-0,0078	0,0052	-0,0030	0,0040	0,0203	0,0141	0,0144	0,0103	-0,0041	-0,0058	0,0296	0,0076	-0,0486	-0,5797	0,8119	-0,0056
-0,4533	0,5500	0,1642	-0,0374	0,5040	-0,3584	0,2688	-0,0286	0,0504	0,0161	0,0256	-0,0069	0,0036	0,0385	-0,0294	0,0466	-0,0018	-0,0019	-0,0091
-0,0037	0,0118	0,0142	0,1117	0,0522	0,1096	0,1633	0,0051	0,1232	0,3491	-0,2833	0,1741	0,1815	0,0517	0,2450	-0,7728	0,0683	-0,0041	-0,0251
0,0104	0,0054	0,0204	0,0030	0,0122	-0,0222	-0,0111	0,0098	-0,0295	0,0012	-0,0226	-0,0347	-0,0248	-0,0024	-0,0237	-0,0338	0,2169	0,1599	0,9597

minuye y en tres aumenta el porcentaje explicado, aunque estos aumentos son pequeños.

Tabla XXII. Proporción de la variación total de cada variable explicada por los cuatro primeros componentes principales (matriz de covarianza)

	Porcentaje de la Variación Explicada									
	Datos sin Transformar					Datos Transformados				
Vector Propio	1°	2°	3°	4°	Suma	1°	2°	3°	4°	Suma
VARIABLE										
1	4,86	0,03	0,04	16,65	21,58	6,26	4,97	19,02	0,71	30,95
2	0,02	2,92	0,27	1,22	4,61	0,53	4,30	3,69	0,14	8,66
3	0,09	8,73	3,98	0,21	13,02	4,50	8,61	0,28	3,43	16,83
4	6,89	7,48	1,30	1,55	17,22	0,69	27,07	2,22	4,19	34,16
5	14,65	1,28	1,45	6,77	24,35	19,07	0,31	12,98	0,65	33,02
6	0,09	3,86	12,67	13,96	30,58	1,32	8,12	41,99	5,76	57,18
7	2,26	0,49	4,64	4,86	12,24	0,46	6,91	8,10	2,55	18,02
8	1,01	0,74	44,95	0,09	46,79	0,83	13,55	0,18	42,02	56,59
9	0,80	15,66	0,49	0,11	17,06	1,06	18,07	0,11	0,19	19,44
10	99,49	0,03	0,00	0,26	99,78	67,69	25,96	0,93	0,22	94,81
11	0,05	61,28	0,02	8,23	69,57	17,10	49,62	0,02	7,27	74,01
12	2,36	1,16	0,81	81,44	85,77	2,33	1,81	83,96	8,19	96,29
13	0,01	1,65	93,74	0,62	96,02	2,88	5,97	4,79	73,98	87,61
14	0,71	1,00	0,00	2,34	4,05	1,07	0,42	3,20	1,16	5,85
15	0,00	0,36	1,73	2,92	5,01	0,91	1,60	2,05	1,70	6,26
16	6,32	0,49	0,36	1,76	8,93	6,49	1,80	1,67	0,82	10,77
17	0,41	85,00	0,02	2,37	87,80	32,24	46,59	2,38	0,10	81,31
18	0,44	0,47	15,23	1,54	17,67	0,03	0,30	0,25	12,00	12,58
19	4,38	0,03	0,37	7,84	12,62	4,48	1,19	9,71	0,16	15,54

Este comportamiento diferente también puede apreciarse si en un gráfico se relaciona el porcentaje de varianza explicada por cada uno de los valores propios generados a partir de la matriz de covarianza con el de los generados a partir de la matriz de correlación. Las figuras 11 y 12 presentan esos valores para los datos sin transformar y con transformación angular, respectivamente. Se utilizan los datos de las Tablas XIV y XVI para la figura 11 y los de la Tabla XX para la figura 12. La relación queda muy bien ajustada por una función de la raíz cuadrada de los porcentajes de varianza de la matriz de covarianza. Para calcular los valores de Fc indicados en las figuras se utilizó el método de los mínimos cuadrados; r-cuadrado es el coeficiente de determinación. En ambos casos, el ajuste es muy bueno. Se ha graficado la función ajustada y también una línea que corresponde a la relación y = x a fin de poder comparar la distancia entre ambas curvas. Cuanto más se alejen las curvas ajustadas de la línea y = x, mayor será la diferencia entre los resultados obtenidos utilizando la matriz de covarianza y la de correlación, ya que si con ambas se obtuviera el mismo resultado, los porcentajes de varianza sintetizados por cada uno de ellos sería igual. Esta información puede extraerse a partir de una figura; por ejemplo, en el caso de la figura 11 puede observarse que la línea de ajuste se aleja de la diagonal. Lo mismo ocurre en la figura 12.

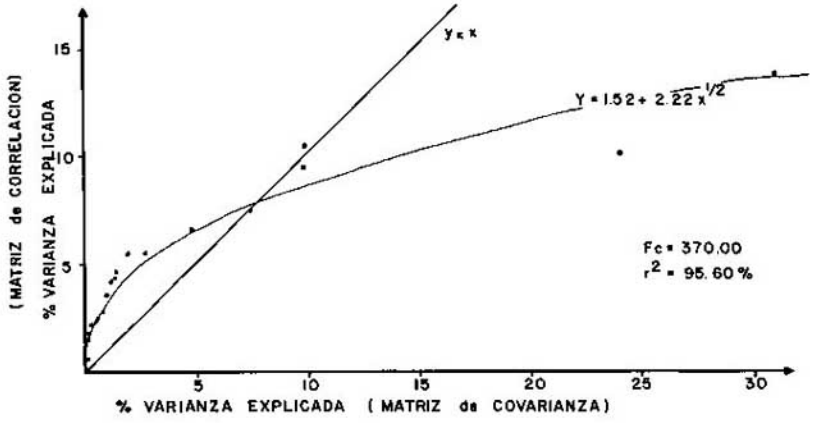


Fig. 11. Relación entre la varianza explicada por cada componente a partir de la matriz de covarianza y de la matriz de correlación (datos sin transformar). Véase explicación en el texto.

Estas figuras permiten tomar decisiones acerca de la importancia de los cambios que ha introducido la transformación angular. El valor del coeficiente de regresión que acompaña a $\sqrt{x}$ indica el grado de inclinación que tiene la función; a menores valores de β, más se alejará de la diagonal de referencia $x = y$. La ordenada al origen crecerá a medida que sean mayores los porcentajes de varianza explicados por los últimos componentes de la matriz de correlación y sean menores las varianzas de los últimos componentes de la matriz de covarianza. Es decir, que un aumento en la estimación de α y una disminución en la estimación de β serán indicadores de diferencias entre los resultados obtenidos con los datos transformados y sin transformar.

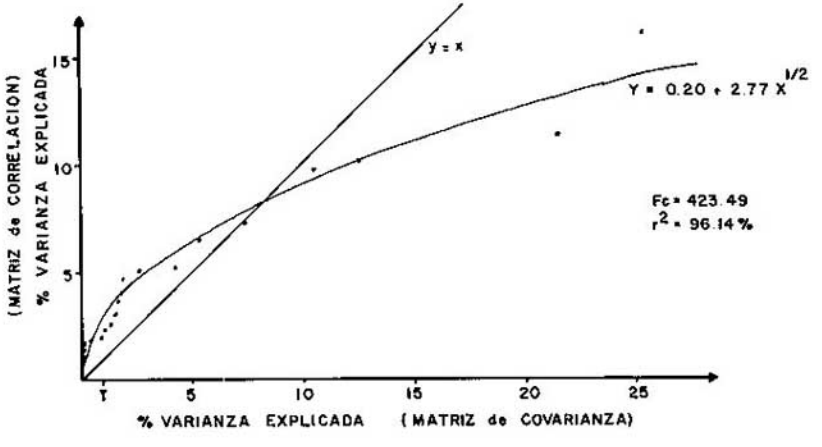


Fig. 12. Relación entre la varianza explicada por cada componente a partir de la matriz de covarianza y de la matriz de correlación (datos transformados). Véase explicación en el texto.

Es posible comparar las dos ecuaciones de regresión siguiendo el método propuesto por Rao (1973)[31] y confrontar las hipótesis:

$$\text{Hipótesis nula} : \alpha_t = \alpha_o \text{ y } \beta_t = \beta_o$$
$$\text{Hipótesis alterna:} \alpha_t \neq \alpha_o \text{ o } \beta_t \neq \beta_o$$

donde α_o y α_t son las ordenadas en el origen para los datos originales y transformados y β_o y β_t las pendientes para los datos originales y transformados, respectivamente. La dócima se realiza mediante el cálculo de un valor de F que permite comparar la diferencia entre el residual encontrado al combinar las dos muestras en una (Rc) y el de las dos regresiones por separado (Ro) con el de las regresiones por separado (Ro):

$$F = \frac{Rc - Ro}{Ro} \times \frac{19 + 19 - 4}{2}.$$

Los grados de libertad del denominador corresponden a la suma del tamaño de las dos muestras (19 para cada una en este ejemplo), menos cuatro que es el número de parámetros que se desea comparar: ordenada en el origen y pendiente más el número de constantes calculadas (un valor promedio por cada serie). Este valor se compara con un valor de F tabulado y si el valor calculado es mayor que el tabulado, para una determinada probabilidad (5 y 1%, son los más utilizados), la diferencia entre las ecuaciones se declara significativa. En este ejemplo el valor de F calculado fue 5,8081, el cual resulta significativo al 4%, por lo cual debe concluirse que no existe evidencia para aceptar la hipótesis nula y se declaran significativamente diferentes las dos ecuaciones de regresión ajustadas.

En vista de lo que antecede se decidió trabajar con los datos transformados y con cuatro componentes calculados a partir de la matriz de covarianza. Esto obedece a tres causas fundamentales:

a) La transformación angular acerca los datos a un comportamiento normal y permite obtener resultados significativamente diferentes de los alcanzados con los datos originales.

b) La matriz de covarianza mantiene la diferencia relativa de las coberturas de cada especie dentro de los ambientes estudiados. Si el análisis tiene fines descriptivos, como en este caso, las especies más abundantes dominan el paisaje y este aspecto se manifiesta si se utiliza la matriz de covarianza. Si los fines del estudio consistieran, por ejemplo, en detectar especies indicadoras asociadas con algún factor ambiental, sería necesario tomar una decisión acerca de si la importancia de ciertas especies con poca cobertura justifica el uso de la matriz de correlación. En ese caso, todas las especies tendrían la misma importancia relativa, independientemente de su cobertura absoluta.

c) Se utilizan los cuatro primeros componentes que pertenecen a la primera onda o primer grupo y que sintetiza el 69,84% de la variación total contenida en los datos originales.

4. INTERPRETACION DE LOS RESULTADOS

4.1. Relaciones entre Variables

Del examen de la Tabla XXII pueden derivarse las relaciones entre las especies que caracterizarán a cada componente principal. Si se analiza la Tabla tomando especie por especie pueden identificarse aquellas con altos coeficientes en un mismo vector, lo cual implica que

están relacionadas para el conjunto de datos analizado. A los efectos de detectar similitudes, estas relaciones pueden compararse con la estructura de la matriz de covarianza. Cuando se utilice la matriz de covarianza para generar los valores y vectores propios, será la estructura de la covarianza la que se refleje en los resultados. La comparación puede hacerse en forma gráfica.

Es posible construir un polígono, donde en cada vértice se indique una de las 19 especies consideradas en el análisis y que estos vértices sean unidos por líneas cuyo grosor indique la correlación existente entre las variables. La interpretación de este polígono ayudará al investigador a conocer las relaciones entre las especies, pero hasta que no se idee algún método para verificarlas, las conclusiones serán sólo de carácter indicativo.

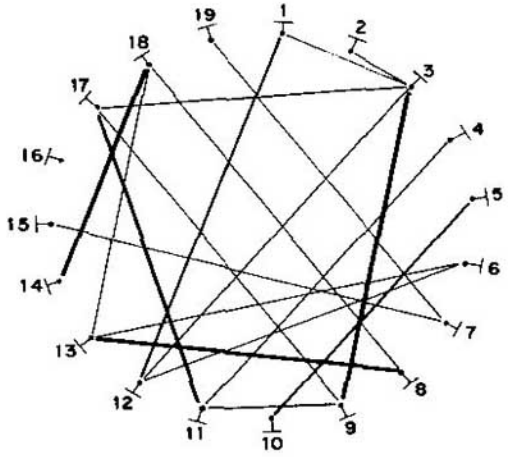


Fig. 13. Correlaciones entre especies. El número indica la especie conforme a la Tabla XI. Las líneas unen especies con altas correlaciones. Las líneas más gruesas indican correlaciones más altas. Para efectuar el gráfico y seleccionar las líneas que debían trazarse se tomaron en cuenta las máximas correlaciones de una especie y se las unió si también la correlación resultaba alta para la otra. Véase el texto.

En la figura 13 se muestra este polígono construido a partir de la matriz de correlaciones de los datos transformados presentada en la Tabla XIX. Las líneas unen los vértices correspondientes a dos especies que tienen alta correlación, lo cual indica que aparecerán juntas en los sitios estudiados. Hay seis pares de especies muy relacionadas, a saber: *Acacia tortuosa* (1) y *Castela erecta* (12); *Bulnesia arborea* (3) y *Pithecelobium unguis-cati* (9); *Capparis odoratissima* (5) y *Prosopis juliflora* (10); *Pereskia guamacho* (8) e *Ipomoea carnea* (13); *Ritherocereus* spp (11) y *Opuntia caracasana* (17); y *Jatropha gossipifolia* (14) y *Sporobolus pyramidatus* (18). Estas especies aparecen unidas por un trazo más grueso que las restantes. Las líneas más delgadas unen especies con correlación moderada y las especies que no tienen líneas que las unen no se encuentran correlacionadas. A partir de este polígono es posible agrupar las variables en núcleos entre los cuales hay una muy estrecha relación y a su vez relacionar los grupos entre sí.

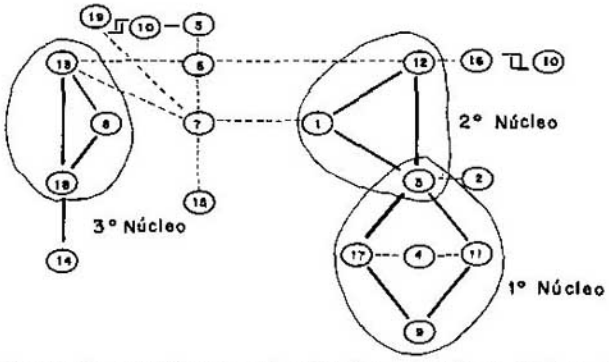


Fig. 14. Grupos de especies correlacionadas. El número indica la especie conforme a la Tabla XI. Las líneas unen especies correlacionadas. Las líneas de puntos indican la correlación de algunas especies que conectan los núcleos principales. El símbolo ⌐¬ denota correlación negativa de alto valor absoluto.

La figura 14 muestra este ordenamiento, donde se distinguen claramente tres núcleos: uno formado por las especies 3, 11, 9 y 17 que tienen relación con la especie 4 y se halla unido por un vértice (una especie en común) con el segundo núcleo formado por las especies 3, 1 y 12; el tercer núcleo está constituido por las especies 8, 13 y 18. Entre el segundo y el tercero existe una cadena de unión que, a su vez, forma un grupo que se ha indicado con línea punteada porque las correlaciones que presentan las especies no son tan altas como en el caso de los núcleos principales. A los lados de la figura se han indicado con un símbolo aquellas especies con una alta correlación negativa con alguna de las especies: las Loranthaceas (19) con *Prosopis juliflora* (10) en la parte superior izquierda de la figura y *Prosopis juliflora* (10) con *Melocactus caesius* (16) al lado derecho.

Si se estudia la composición de los cuatro primeros vectores propios, que son los seleccionados para sintetizar la información contenida en el conjunto de datos, puede observarse que los coeficientes con valores absolutos más altos en el primer componente corresponden a especies del primer núcleo (11 y 17), aunque posee el mayor coeficiente para *Prosopis juliflora* (10) que no se halla directamente en ningún núcleo. El segundo componente también tiene los máximos coeficientes en especies del primer núcleo (4, 11 y 17) y también para la especie 10. Así, las relaciones entre las especies de este primer núcleo y *Prosopis juliflora* son las que más influyen en la diferenciación de sitios a nivel de la vegetación de la zona semiárida del Estado Falcón.

El tercer componente exhibe el mayor coeficiente para la especie *Castela erecta* (12) y está expresando la variación debida a las especies que forman el segundo núcleo. Asimismo, posee un coeficiente relativamente alto para *Cercidium praecox*, el cual pertenece al núcleo secundario que enlaza el segundo y el tercero de los núcleos principales. El cuarto componente tiene el máximo coeficiente para *Pereskia guamacho*, que pertenece al tercer núcleo.

De este análisis pueden extraerse conclusiones complementarias acerca de la relación entre las especies, pues no sólo se conocen aquellas que están relacionadas entre sí, sino la importancia relativa de

estos grupos. Por ejemplo, las especies que pertenecen al primer núcleo, por estar asociadas al primer componente principal, tendrán una ponderación relativa mayor en el conjunto total de censos analizados.

4.2. Información en los Componentes Principales

En la Tabla XXIII se consignan, a manera de ejemplo, los valores de los cuatro primeros componentes principales obtenidos en los diez primeros censos. Cada valor de esta Tabla se obtuvo multiplicando el coeficiente del primer vector propio por la cobertura original de cada especie en el censo correspondiente. Así, para el primer censo con los datos de la Tabla XII transformados por la ecuación [37] y los coeficientes de la Tabla XXI, el primer componente será:

$$
\begin{aligned}
-31,7897 = {} & (0,0845 \times \operatorname{sen}^{-1}\sqrt{0,005 + 0,01}) + (-0,0203 \times \operatorname{sen}^{-1}\sqrt{0,00 + 0,01}) + \\
& + (-0,0768 \times \operatorname{sen}^{-1}\sqrt{0,00 + 0,01}) + (0,0471 \times \operatorname{sen}^{-1}\sqrt{0,005 + 0,01}) + \\
& + (-0,1440 \times \operatorname{sen}^{-1}\sqrt{0,005 + 0,01}) + (-0,0566 \times \operatorname{sen}^{-1}\sqrt{0,03 + 0,01}) + \\
& + (0,0171 \times \operatorname{sen}^{-1}\sqrt{0,00 + 0,01}) + (-0,0322 \times \operatorname{sen}^{-1}\sqrt{0,00 + 0,01}) + \\
& + (-0,0311 \times \operatorname{sen}^{-1}\sqrt{0,105 + 0,01}) + (-0,8050 \times \operatorname{sen}^{-1}\sqrt{0,205 + 0,01}) + \\
& + (-0,2880 \times \operatorname{sen}^{-1}\sqrt{0,03 + 0,01}) + (0,1009 \times \operatorname{sen}^{-1}\sqrt{0,205 + 0,01}) + \\
& + (-0,1023 \times \operatorname{sen}^{-1}\sqrt{0,105 + 0,01}) + (0,0254 \times \operatorname{sen}^{-1}\sqrt{0,00 + 0,01}) + \\
& + (-0,0056 \times \operatorname{sen}^{-1}\sqrt{0,005 + 0,01}) + (0,0145 \times \operatorname{sen}^{-1}\sqrt{0,005 + 0,01}) + \\
& + (-0,4533 \times \operatorname{sen}^{-1}\sqrt{0,03 + 0,01}) + (-0,0037 \times \operatorname{sen}^{-1}\sqrt{0,005 + 0,01}) + \\
& + (0,0104 \times \operatorname{sen}^{-1}\sqrt{0,00 + 0,01}).
\end{aligned}
$$

De igual forma se obtienen los valores para los censos restantes y para los otros componentes. El promedio de todos los valores del primer

Tabla XXIII. Valores de los cuatro primeros componentes principales para los diez primeros censos

Censo	Componentes Principales 1°	2°	3°	4°
1	-31,7897	12,7578	-34,0399	7,2725
2	-37,0596	15,8090	-22,1693	13,1335
3	-26,7082	14,1615	-18,2848	2,7399
4	- 7,6539	18,9258	-46,0290	-3,2146
5	- 8,9466	16,1366	-32,5248	-1,3671
6	-19,0511	26,5943	-26,7975	-2,2130
7	- 7,3198	17,6224	-40,5471	-4,7026
8	-12,7228	11,5187	-30,8096	-1,1430
9	- 7,9738	14,4320	-23,3565	0,2798
10	-13,2712	22,1592	-32,2197	-2,8132

componente es -33,8517 y su varianza es 287,178 que, salvo por errores de redondeo, es el primer valor propio. El valor promedio también puede obtenerse aplicando la transformación lineal (es decir utilizando los coeficientes del primer vector propio) a los promedios de las variables originales. El valor mínimo para el primer componente fue -73,3946 correspondiente al censo número 28. Para no trabajar con valores negativos, se sumó 74 a cada dato (el censo 28 tomó un valor de 0,6054) y los valores positivos obtenidos fueron graficados en la figura 15. De igual

manera se procedió respecto al segundo, tercer y cuarto componentes que se presentan en las figuras 16, 17 y 18, respectivamente. Las ordenadas representan el valor del componente principal y se ha levantado una delgada barra para representar los censos por separado. Cada una de estas figuras representa un perfil que indica el valor de cada censo en el componente considerado. Así, es fácil detectar valores comparativamente pequeños y comparativamente grandes y encontrar una explicación que relacione esos censos.

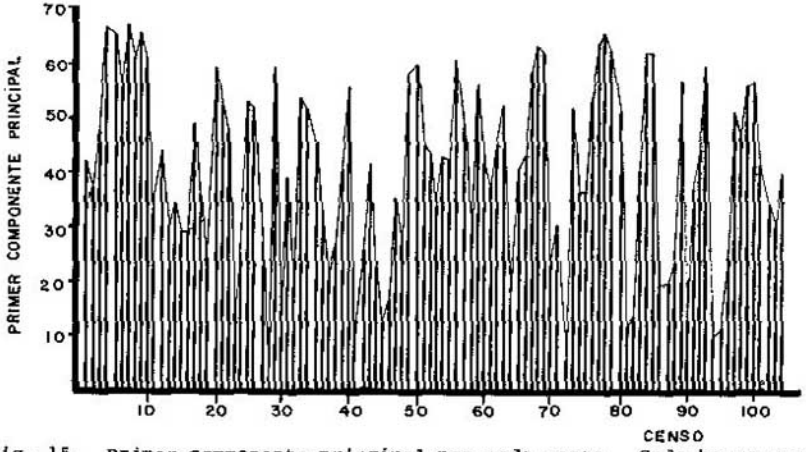


Fig. 15. Primer componente principal por cada censo. Cada barra representa el valor del primer componente principal en cada uno de los 104 censos que comprende la muestra. Véase explicación en el texto.

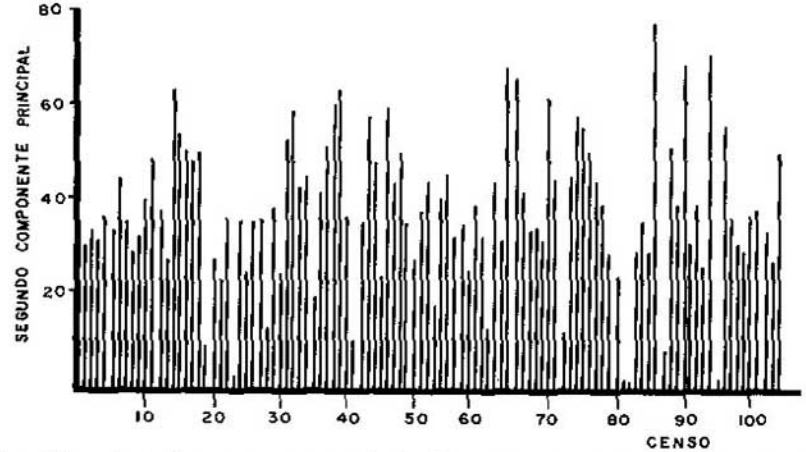


Fig. 16. Segundo componente principal por cada censo. Cada barra representa el valor del segundo componente principal en cada uno de los 104 censos que comprende la muestra. Véase explicación en el texto.

En la figura 15 se ha sombreado todo el perfil a fin de destacar la existencia de depresiones en algunos censos, resaltan el 28, 41, 65, 72

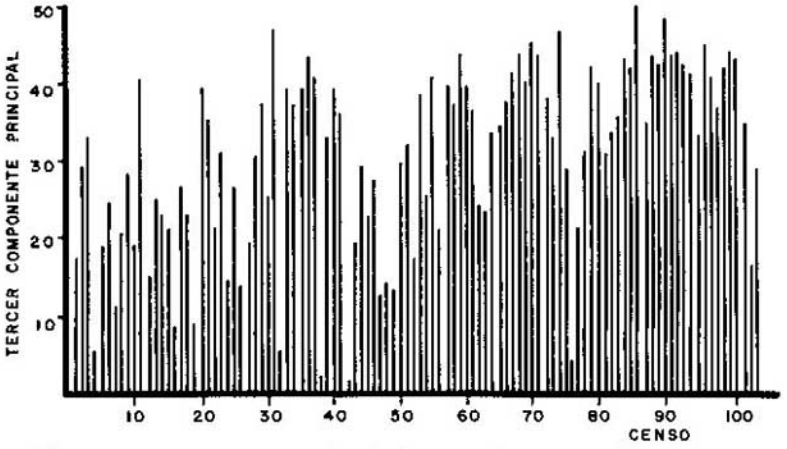


Fig. 17. Tercer componente principal por cada censo. Cada barra representa el valor del tercer componente principal en cada uno de los 104 censos que comprende la muestra. Véase explicación en el texto.

y 95. Todos ellos corresponden a vegas de los ríos o quebradas que corren por la planicie costera de Falcón con dirección predominante sur a norte. Aunque también en otros censos se obtuvieron valores bajos que no corresponden a vegas de ríos ni a quebradas, como el 94, 90 y 64. Si se examina la figura 16, donde se ha representado el segundo componente, puede verse que en algunos censos los valores son muy bajos; en el perfil aparecen como depresiones que corresponden a vegas de ríos y quebradas. Así, es posible identificar los censos en que la vegetación característica es la de vegas por la existencia simultánea de estas depresiones en los dos primeros componentes principales. El perfil del tercer componente principal no presenta diferencias tan marcadas entre

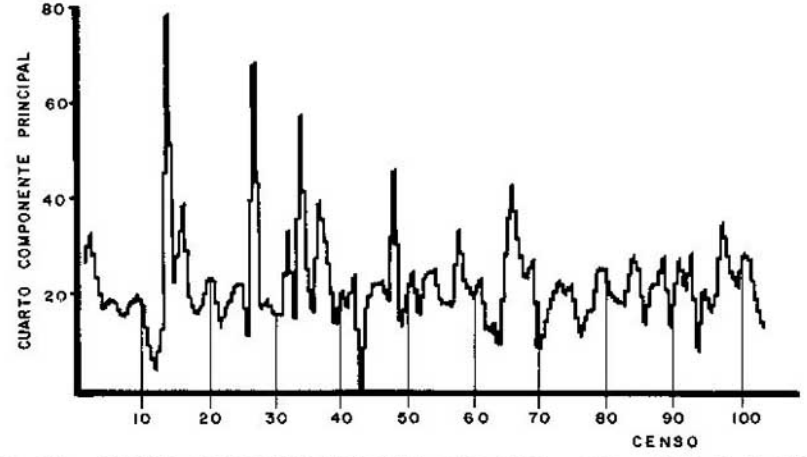


Fig. 18. Cuarto componente principal por cada censo. Representación alternativa de los componentes. La línea une los valores del cuarto componente en cada uno de los censos.

los censos y en el del cuarto destacan algunos censos con valores altos, como el 14, 27 y 34. Los elevados valores de estos censos se deben a la presencia de *Castela erecta* y es posible detectar visualmente que en muy pocos sitios abunda esta especie. En la figura 18 se observa el perfil para el cuarto componente, el cual resulta muy uniforme con excepción de elevados valores en los censos 14, 27, 34, 48 y 66. En estos censos domina el estrato arbustivo *Ipomoea carnea*, acompañada en el piso por *Sporobolus pyramidatus*.

6

ANALISIS DE CALIFICACIONES POR ASIGNATURA

1. INTRODUCCION

En las universidades y liceos de América Latina y de la mayoría de los países del mundo, el progreso en las asignaturas realizado por un estudiante se evalúa con una nota que según los países varía de 1 a 5, de 0 a 9, de 0 a 10, o de 0 a 20. En estos sistemas de calificación a partir de un cierto valor el alumno se considera aprobado y por debajo de ese valor debe repetir el examen o la asignatura completa, conforme al régimen particular de promoción.

Al término de su carrera el estudiante es juzgado por el promedio de sus calificaciones. Se considera a menudo que será mejor profesional aquel que haya obtenido las "mejores" calificaciones, expresadas en el promedio simple de calificaciones logradas durante su carrera (suma de todas las notas dividida entre el número total de asignaturas); en el promedio ponderado correspondiente a la carga o importancia relativa de cada asignatura, que a menudo se valoriza como "unidades de crédito" en cada una (suma de cada calificación multiplicada por las "unidades de crédito" correspondientes, dividida por el total de esas "unidades de crédito") y que se denomina en algunos países "índice académico"; o por cualquiera de los anteriores, pero considerando sólo la última calificación lograda en cada asignatura. Este último procedimiento enmascara los intentos que el estudiante debe realizar antes de aprobar una materia.

En este capítulo se presenta un ejemplo de análisis por componentes principales de las notas aprobatorias obtenidas por tres cohortes de estudiantes de la Facultad de Agronomía de la Universidad Central de Venezuela en las llamadas "materias básicas" del pensum que comprenden 17 asignaturas de diferentes departamentos y esferas del conocimiento. Los datos originales forman parte de la Tesis de Maestría del profesor W. Henriquez, quien gentilmente nos ha permitido utilizarla (Henriquez, 1985[18]).

2. REDUCCION DE LA DIMENSIONALIDAD POR COMPONENTES PRINCIPALES

El objeto de este análisis es detectar las materias que contribuyen con la mayor variabilidad en el conjunto total de las asignaturas básicas. Aquel estudiante que obtenga una nota promedio mayor en su carrera será comparado favorablemente con otro que tenga un promedio menor. Ningún empleador y ninguna institución pública o privada tendrá manera alguna de comparar el rendimiento académico relativo de diferentes alumnos en las distintas esferas del conocimiento. Dispondrá solamente del conjunto de notas por asignatura y en la mayoría de los casos ni siquiera dispondrá de personal que pueda hacer un análisis superficial de ese rendimiento del alumno.

A continuación se presenta un ejemplo, donde se han agrupado las 17 asignaturas básicas del pensum de Agronomía y se busca la forma de sintetizar la información contenida en estas asignaturas a fin de obtener tanto relaciones entre materias, como una medida sintética del aprovechamiento estudiantil que pueda ser utilizada con fines descriptivos. El análisis puede continuarse tratando de relacionar el comportamiento

Tabla XXIV: Matriz de correlación para asignaturas básicas del pensum de Agronomía

1. Propagación de Plantas	1,0000															
2. Agricultura	0,2223	1,0000														
3. Oleoricultura	0,2515	-0,0521	1,0000													
4. Química General	0,6468	0,2086	0,1739	1,0000												
5. Botánica I	0,3299	0,2223	-0,1720	0,2040	1,0000											
6. Botánica II	0,3767	0,4291	-0,1153	0,2120	0,4479	1,0000										
7. Entomología I	0,0481	0,1195	0,2048	0,1411	0,1679	0,0659	1,0000									
8. Producción	0,0343	0,1232	0,1165	0,0754	0,3139	0,1226	-0,3292	1,0000								
9. Inv. Bibliograf.	0,3504	0,4457	0,4967	0,1909	0,3746	0,3907	0,1440	0,3372	1,0000							
10. Extensión Agríc.	-0,1425	-0,1225	0,1450	0,0196	-0,1810	0,0951	-0,4056	0,4532	-0,0526	1,0000						
11. Economía Agríc.	0,0061	0,4267	0,2301	0,0197	-0,1309	0,2603	-0,0209	0,0769	0,2967	0,3471	1,0000					
12. Dibujo	0,3133	0,2971	0,1801	0,1142	0,1350	0,2378	0,3960	-0,0769	0,4066	-0,2536	0,0209	1,0000				
13. Algebra y Geom.	-0,2126	0,1514	0,0000	-0,1364	-0,3622	-0,3523	0,2893	-0,3040	-0,0000	-0,1353	0,0795	0,3300	1,0000			
14. Física Aplicada	0,1241	0,0557	-0,0897	0,4772	0,1858	-0,0722	0,0046	0,2559	-0,0351	0,2295	0,0439	-0,2051	0,0868	1,0000		
15. Anat. y Fisiol.I	0,3285	0,0856	0,4130	0,5379	0,2086	0,0520	0,2905	0,1429	0,2603	0,0868	0,2078	0,3508	0,1513	0,2348	1,0000	
16. Anat. y Fisiol.II	0,4288	0,3422	0,4113	0,2358	0,4876	0,3970	0,3001	0,4273	0,7173	-0,0116	0,1948	0,3573	-0,1254	0,2077	0,3506	1,0000
17. Zoología	0,4142	0,3135	0,0651	0,0319	0,3586	0,3200	-0,1230	0,4367	0,3740	0,1611	-0,0186	0,3941	-0,1542	-0,0382	0,0766	0,4275	1,0000

profesional de una muestra de estudiantes con su rendimiento durante el paso por la universidad. Este no es el objetivo del ejemplo y sólo podrá emprenderse una investigación de este tipo si se dispone de un equipo de personal multidisciplinario y del tiempo de seguimiento adecuado.

En la Tabla XXIV se aprecia la matriz de correlación para las 17 materias que se indican en la primera columna. En la Tabla XXV se consignan los valores propios y el porcentaje de la varianza total explicado por cada uno y en la Tabla XXVI se presentan los vectores propios.

Tabla XXV. Valores propios y varianza explicada para las asignaturas básicas

Valor propio	4,34924	2,38302	1,95874	1,77997	1,38369	1,13231	
	1,02221	0,75120	0,51153	0,42180	0,37085	0,29435	
	0,21401	0,18284	0,14515	0,05254	0,04626		
Varianza acumulada (%)	25,58	39,60	51,13	61,60	69,34	76,40	82,41
	86,83	89,84	92,32	94,50	96,23	97,49	98,57
	99,42	99,73	100,00				

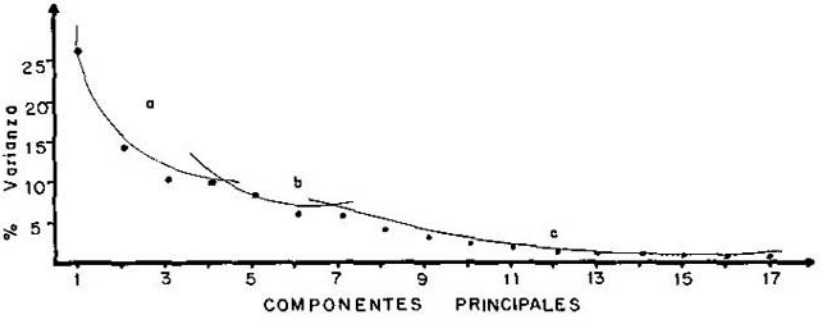


Fig. 19. Variación explicada por cada componente, caso pensum de Agronomía. La primera onda incluye los cuatro primeros componentes que se mantendrán en los análisis posteriores.

En la figura 19 se han graficado los porcentajes de varianza explicados por cada componente utilizando los datos de la Tabla XXV. Por haberse calculado los valores propios a partir de la matriz de correlación, la suma de todos ellos será 17 (el número total de variables originales) y si se utiliza el criterio de mantener en el análisis aquellos componentes generados por valores propios superiores al promedio, deberá incluirse hasta el séptimo sintetizando un 82,41% de la variación total. Si se utiliza el criterio de las ondas, señaladas en la figura 19, habrá que decidir si se deben mantener los cuatro primeros componentes que corresponden a la primera onda (a) y que sintetizan un 61,60% de la variabilidad total, o si se deben incluir las dos primeras ondas (a y b) que totalizan los siete primeros componentes con 82,41%. Esta última decisión coincide con el criterio de mantener los componentes que correspondan a valores propios que están por encima del promedio.

En la Tabla XXVI se observará que el primer vector propio, que contiene los coeficientes de la combinación lineal que dará origen al primer componente principal, exhibe los mayores valores para las asignaturas Anatomía y Fisiología II, Investigación Bibliográfica, Propagación de Plantas, Botánica II, Botánica I y Zoología y que es prácticamente nula la contribución de las asignaturas Extensión, Física Aplicada y Algebra y Geometría. De esta manera, en el caso de un alumno que ha recibido una nota alta en las asignaturas que poseen los mayores coeficientes, el primer componente principal tendrá un valor alto. Este primer componente se asemeja mucho a un promedio ponderado, donde la ponderación está dada por los coeficientes del vector. Por lo antes señalado, en este promedio no influyen las notas de las asignaturas Extensión, Física Aplicada y Algebra y Geometría.

Por consiguiente, si se desea un profesional con buena preparación y las aptitudes necesarias para efectuar trabajos relacionados con los aspectos biológicos de los cultivos, deberá preferirse el egresado para quien el primer componente posea el valor máximo. Sin embargo, si los trabajos que deberá desempeñar se relacionan con la ingeniería de maquinaria, sistemas de riego u otros aspectos donde sea esencial contar con una sólida base matemática, el valor del primer componente principal no entregará ninguna información acerca de la preparación adecuada en ese campo.

La interpretación del segundo componente no es tan sencilla, ya que algunos coeficientes tienen signo positivo y otros signo negativo. Si se consideran los mayores valores absolutos se observa que las asignaturas Entomología I, Producción, Extensión, Algebra y Geometría y Dibujo determinan el valor de este componente. Pero de estas asignaturas Entomología y Dibujo tienen signo negativo, en tanto que las restantes llevan signo positivo. Así, para un alumno con altas calificaciones en Entomología y Dibujo y bajas en las tres restantes, el valor del segundo componente será bajo y para un alumno con altas calificaciones en Producción, Extensión y Algebra y Geometría el valor del segundo componente será elevado.

El tercer componente tiene altos coeficientes positivos para Física Aplicada, Anatomía y Fisiología I, Química General, Extensión y Oleoricultura y coeficientes con valores absolutos altos, pero negativos, para las asignaturas Zoología, Botánica II y Botánica I. El resto de las materias contribuyen con coeficientes muy cercanos a cero. El alumno que obtenga un alto valor para el tercer componente principal tendrá altas calificaciones en las asignaturas Física Aplicada, Anatomía y Fisiología I, Química General, Extensión y Oleoricultura y bajas calificaciones en las asignaturas Botánica I, Botánica II y Zoología (ya que si fueran altas harían disminuir el valor de este componente pues el signo del coeficiente es negativo.

Cuando se dice haber trabajado, como en este ejemplo, con la matriz de correlación, las expresiones "bajas calificaciones" o "altas calificaciones" o, en forma más general, altos y bajos valores de una variable para un determinado alumno, indican términos relativos. Por tener las variables estudiadas, en este caso las calificaciones en cada una de las 17 materias, media cero y varianza unitaria, una "calificación alta" será aquella que esté por encima del promedio y una "calificación baja" la que esté por debajo del promedio.

De modo que si se busca un profesional con buena preparación en los aspectos matemáticos y de ingeniería deberá prestarse atención al valor del tercer componente.

Tabla XXVI. Vectores propios para las asignaturas básicas

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0,3204	-0,0271	0,0076	0,2735	0,1224	0,4282	-0,1847	-0,2461	0,1646	0,0038	-0,4018	-0,0616	-0,0769	-0,0980	0,4049	-0,3937	-0,0438
0,2588	-0,0629	-0,1273	-0,1554	-0,5303	0,0050	-0,1244	-0,3120	-0,2384	-0,3563	0,2551	0,0448	0,0185	0,4424	0,0857	-0,1903	-0,0642
0,1854	-0,1164	0,3053	-0,2903	0,4569	0,1558	0,2911	-0,1715	0,0377	-0,1011	0,0768	0,1235	0,1179	0,3429	0,3239	0,3939	-0,0417
0,2466	-0,0259	0,3317	0,3908	-0,0610	0,3031	-0,1657	-0,0903	-0,0378	-0,1414	0,3431	0,1190	0,1398	-0,3105	0,3355	0,2768	-0,2935
0,2769	0,1214	-0,2301	0,3424	-0,0438	-0,3206	0,1506	0,1924	-0,2824	0,1467	-0,3048	0,2791	0,3710	0,1210	0,1583	0,1440	-0,3223
0,2801	0,1353	-0,2595	0,0413	-0,3143	0,2641	0,1791	0,3402	0,3193	0,2304	0,2444	0,1525	-0,3696	0,0178	0,2212	0,2785	0,1279
0,1215	-0,4687	0,0163	0,0950	-0,0202	-0,1978	0,3009	0,3396	0,1960	-0,5517	0,0570	0,0551	0,1114	-0,2698	0,1391	-0,1947	0,1393
0,1940	0,4348	0,0826	-0,1083	0,1333	0,3972	-0,1307	-0,0457	-0,1391	-0,1030	0,3178	-0,2494	-0,1496	-0,3840	0,4308	-0,0111	-0,1064
0,3747	-0,0362	-0,0433	-0,2299	0,0574	0,1171	0,2075	-0,3207	0,0146	0,3886	0,1664	0,3462	0,1999	-0,2755	0,2112	-0,2955	0,3066
0,0061	0,4172	0,3078	-0,2777	-0,0155	0,0635	-0,1623	0,3919	0,3447	-0,0876	0,0012	0,2893	0,2361	0,1634	0,0733	-0,3571	-0,2105
0,1344	0,0527	0,2114	-0,4307	-0,4209	0,2312	0,2156	0,0953	-0,2155	0,0032	-0,4044	-0,2931	0,1757	-0,3228	0,0090	0,1775	-0,0133
0,2479	-0,3534	-0,1603	-0,1322	0,0810	-0,0232	0,3500	0,2265	0,1764	0,3116	0,2125	-0,4930	0,3862	0,1166	0,0171	0,0149	-0,0995
-0,0720	0,4117	0,1625	-0,2397	-0,1901	-0,2738	-0,4097	-0,0893	0,1442	0,1502	-0,1615	0,4001	-0,2431	-0,1746	0,1808	0,1861	-0,2474
0,0691	0,1171	0,4608	0,3237	-0,2832	-0,3310	-0,0625	-0,1479	0,3116	0,0770	-0,1212	-0,1543	0,1493	0,2008	0,0599	0,1654	0,4623
0,2541	-0,1611	0,3828	0,0550	0,1221	0,0117	-0,1380	0,4183	-0,5375	0,1524	-0,0004	0,0365	-0,3410	0,1742	0,0610	0,1770	0,2359
0,3994	0,0102	0,0122	-0,0507	0,0940	-0,2730	0,2182	-0,0903	0,2652	-0,0692	-0,2166	-0,2486	-0,4199	0,1461	0,4456	0,0286	-0,3430
0,2713	0,1647	-0,3172	-0,1537	0,2122	0,0042	-0,4462	0,0579	0,0015	-0,3754	-0,2489	0,1373	0,0390	-0,0504	0,2084	0,3160	0,3980

3. CONTRIBUCION RELATIVA DE LAS ASIGNATURAS

Es posible también estudiar el porcentaje de la varianza de cada asignatura explicada en los diferentes componentes, pues estos valores brindarán información complementaria que permitirá mejorar la interpretación de los componentes. En la Table XXVII se presenta el porcentaje de la varianza total de cada variable que es explicada por los primeros cuatro componentes principales, los que forman la primera onda en la figura 19; este porcentaje ha sido calculado aplicando la fórmula [36] y los datos de las Tablas XXV y XXVI. En la Tabla XXVII se presenta el total de la variación de cada variable explicada por los referidos componentes.

Tabla XXVII. Proporción de la varianza total por asignatura explicada en los cuatro primeros componentes

Asignatura	Porcentaje de Varianza Explicada				
	1°	2°	3°	4°	Total
Anatomía y Fisiología II	69,38	0,02	0,03	0,46	69,89
Investigación Bibliográfica	61,06	0,31	0,37	9,41	71,15
Propagación de Plantas	44,64	0,18	0,01	13,31	58,14
Botánica II	34,12	4,36	13,19	0,30	51,97
Botánica I	33,35	3,51	10,37	20,87	68,10
Zoología	32,01	6,46	19,71	4,20	62,38
Agricultura	29,13	0,94	3,17	4,30	37,54
Anatomía y Fisiología I	28,08	6,18	28,70	0,54	63,50
Química General	26,45	0,16	21,55	27,18	75,34
Entomología	6,42	52,35	0,05	1,61	60,43
Producción	16,37	45,05	1,34	2,09	64,85
Extensión Agrícola	0,02	41,48	18,56	13,73	73,79
Algebra y Geometría	2,25	40,39	5,17	10,23	58,04
Dibujo	26,73	29,76	5,03	3,11	64,63
Física Aplicada	2,08	3,27	41,59	18,65	65,59
Oleoricultura	14,95	3,23	18,26	15,00	51,44
Economía Agrícola	7,86	0,66	8,75	33,02	50,29

Se observa que las seis materias con altos coeficientes en el primer componente se encuentran representadas en él con más del 30% de su variabilidad total. Las asignaturas en la Tabla XXVII se han reordenado de modo que van de mayor a menor variabilidad explicada dentro del primer componente en las nueve asignaturas más representadas; lo mismo se ha hecho para el segundo componente (en la Tabla se trazó una línea de puntos horizontal para separar ambos grupos de materias). La única asignatura con ponderaciones similares en ambos componentes es Dibujo, ubicada en esta Tabla en la posición décimocuarta ya que fue ordenada según el valor del coeficiente del segundo componente. Se observará que las asignaturas Física Aplicada y Oleoricultura ubicadas en las posiciones décimoquinta y décimosexta, respectivamente, y que tienen coeficientes bajos en los dos primeros componentes, se encuentran representadas con elevada ponderación en el tercer componente. La asignatura Economía Agrícola, ubicada en el decimoséptimo lugar, tiene bajos coe-

ficientes en los tres primeros componentes y alto en el cuarto. Esta es una forma alternativa de presentar resultados que permite interpretarlos fácilmente por simple inspección.

En la figura 20 se ha graficado la correlación de cada asignatura con los dos primeros componentes principales. El número identifica la asignatura según el orden en que se presentaron en la Tabla XXIV. Aquellas asignaturas que más se acercan al contorno circular son las que están más completamente sintetizadas en estos dos primeros componentes, en tanto que aquellas que se ubican cerca del origen de coordenadas son las que menos intervienen en estos dos componentes. En la figura 21 se ha efectuado la representación para el tercer y cuarto componentes. Pueden observarse las diferencias en la ubicación de las asignaturas respecto al centro del círculo.

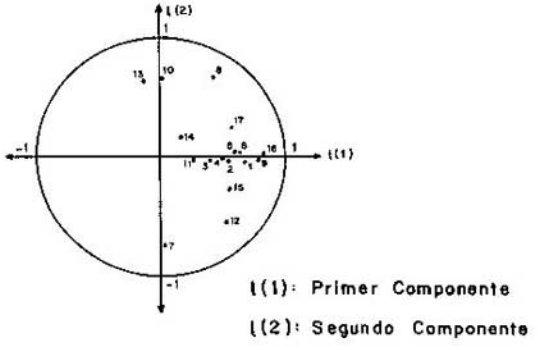


Fig. 20. Correlación de las variables originales con los dos primeros componentes, caso pensum de Agronomía. Los números identifican las asignaturas conforme al código de la Tabla XXIV. Cuanto más cerca se encuentre del centro del círculo menor será su influencia en los dos primeros componentes.

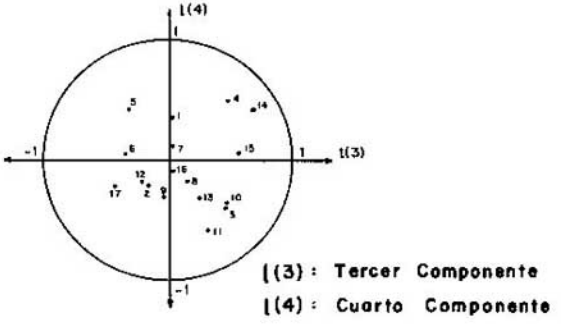
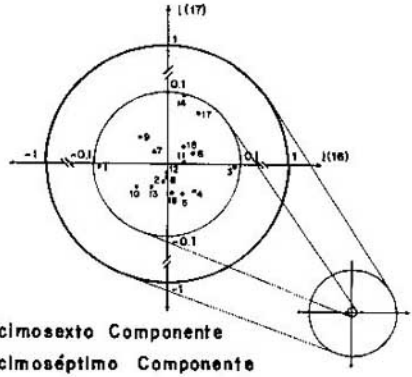


Figura 21. Correlación de las variables originales con el tercer y cuarto componentes, caso pensum de Agronomía. Los números identifican las asignaturas conforme al código de la Tabla XXIV. Cuanto más cerca del círculo de radio unitario se encuentre más influencia tendrá en estos dos componentes.

Como se señaló en el punto 5 del capítulo 3 no sólo es importante la inspección de los primeros componentes principales, sino también de los últimos, ya que si bien no contribuyen con cantidades importantes a la síntesis de la variabilidad total, ofrecen información útil acerca de las variables que generan variaciones espúreas en el conjunto de datos y pueden utilizarse para detectar marginales. En un análisis de este tipo, los marginales pueden ser alumnos que se destacan como muy buenos en general, muy malos, o muy buenos en alguna de las asignaturas en particular. En la figura 22 se representa la correlación de las asignaturas con los dos últimos componentes principales (el décimosexto y decimoséptimo). Para efectuar los cálculos se aplicó la fórmula [35] a los datos consignados en las Tablas XXV y XXVI. Se observará que ninguna asignatura ha quedado representada cerca del círculo de valor unitario, por lo cual no hay una asignatura que pueda descartarse por carecer de influencia marcada en el conjunto. Esto también puede ser confirmado si se observan los valores de la varianza sintetizada en los primeros cuatro componentes presentados en la Tabla XXVII, ya que todas las asignaturas contribuyen en total con porcentajes superiores al 50%, salvo una que tiene 37,54%.



l(16): Décimosexto Componente

l(17): Decimoséptimo Componente

Fig. 22. Correlación de las variables originales con los dos últimos componentes, caso pensum de Agronomía. Los números identifican las asignaturas conforme al código de la Tabla XXIV. La escala del centro del círculo ha sido aumentada a fin de mejorar la representación gráfica de la ubicación de las especies. Todas se encuentran dentro de un círculo de radio 0,10, lo cual indica una contribución muy pequeña a estos últimos componentes.

BIBLIOGRAFIA

(1) ALLEN, D. M. The relationship between variable selection and data augmentation and a method for prediction, *Technometrics*, **16** (1): 125-127 (1974).

(2) ANDERSON, T. W. An introduction to multivariate analysis, Wiley, Nueva York, N. Y., 2a. ed., 675 págs. (1984).

(3) ARNOLD, S. F. The theory of linear models and multivariate analysis, Wiley, Nueva York, N. Y., 474 págs. (1981).

(4) BARGMANN, R. E. Exploratory techniques involving artificial variables. En: Multivariate analysis II, KRISHNAIAH, P. R. (ed.), Academic, Nueva York, N. Y., págs. 567-580 (1969).

(5) BARNETT, V. y LEWIS, T. Outliers in statistical data, Wiley, Nueva York, N. Y., 365 págs. (1978).

(6) CATTELL, R. B. The Scree test for the number of factors, *Multivar. Behav. Res.*, **1**:245-276 (1966).

(7) COOLEY, A. L. y LOHNES, P. Multivariate data analysis, Wiley, Nueva York, N. Y., 364 págs. (1971).

(8) CHATFIELD, C. y COLLINS, A. J. Introduction to multivariate analysis, Chapman & Hall, Nueva York, N. Y., 246 págs. (1980).

(9) CHATTERJEE, S. y PRICE, B. Regression analysis by examples, Wiley, Nueva York, N. Y., 228 págs. (1977).

(10) GIRI, N. C. Multivariate statistical inference, Academic, Nueva York, N. Y., 319 págs. (1977).

(11) GNANADESIKAN, R. Methods for statistical data analysis of multivariate observations, Wiley, Nueva York, N. Y., 298 págs. (1977).

(12) GNANADESIKAN, R. y WILK, M. B. Data analytic methods in multivariate statistical analysis. En: Multivariate analysis II, KRISHNAIAH, P. R. (ed.), Academic, Nueva York, N. Y., págs. 593-638 (1969).

(13) GNANADESIKAN, R. y KETTENRING, J. R. Robust estimates, residuals and outlier detection with multiresponse data, *Biometrics*, **28**(1): 81-124 (1972).

(14) GREENBERG, E. Minimum variance properties of principal component regression, *J. Am. Stat. Assoc.*, **70**(349):194-197 (1975).

(15) GUNST, R. F., WEBSTER, J. T. y MASON, R. L. A comparison of least squares and latent root regression estimators, *Technometrics*, **18** (1):75-83 (1976).

(16) HARRIS, R. A primer of multivariate statistics, Academic, Nueva York, N. Y., 332 págs. (1967).

(17) HAWKINS, D. M. The detection of errors in multivariate data using principal components, *J. Am. Stat. Assoc.*, **69**(346):340-344 (1974).

(18) HENRIQUEZ, W. Análisis de calificaciones por componentes principales, Tesis de Maestría, Posgrado en Estadística, Universidad Central de Venezuela, Maracay, Venezuela (1985).

(19) HOCKING, R. R. The analysis and selection of variables in linear regression, *Biometrics*, **32**(1):1-49 (1976).

(20) HOTELLING, H. Analysis of a complex of statistical variables into principal components, *J. Educ. Psychol.*, **24**:417-441 y 498-520 (1933).

(21) KENDALL, M.G. A course in multivariate analysis, Griffin, Londres, 152 págs. (1957).

(22) KENDALL, M. G. Multivariate analysis, Griffin, Londres, 2a. ed., 210 págs. (1980).

(23) MARDIA, K. V., KENT, J. T. y BIBBY, J. M. Multivariate analysis, Academic, Londres, 521 págs. (1979).

(24) MARQUARDT, D. W. Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation, *Technometrics*, **12** (3):591-612 (1970).

(25) MATTEUCCI, S., COLMA, A. y PLA, L. Análisis regional de la vegetación y el ambiente del Estado Falcón: La vegetación, ediciones del Departamento de Investigación, IUTC, Coro, Venezuela, 292 págs. + 1 mapa (1979).

(26) MATTEUCCI, S. y COLMA, A. Metodología para el estudio de la vegetación, monografía científica No. 22, serie de biología, Secretaría General de la Organización de los Estados Americanos, Washington, D. C., 168 págs. (1982).

(27) McCABE, G.P. Principal variables, *Technometrics*, **26** (2):137-144 (1984).

(28) MORRISON, D. F. Multivariate statistical methods, McGraw, Nueva York, N. Y., 2a. ed., 515 págs. (1976).

(29) MUIRHEAD, R. J. Aspects of multivariate statistical theory, Wiley, Nueva York, N. Y., 673 págs. (1982).

(30) PEARSON, K. On lines and planes of closed fit to system of point in space, *Phil. Mag.*, 6:559-572 (1901).

(31) RAO, C.R. Linear statistical inference and its applications, Wiley, Nueva York, N. Y., 625 págs. (1973).

(32) SEAL, H. Multivariate statistical analysis for biologists, Methuen, Londres, 238 págs. (1964).

(33) SRIVASTAVA, M.S. y CARTER, E. M. An introduction to applied multivariate statistics, North Holland, Nueva York, N. Y., 394 págs. (1983).

(34) STAT-MOD. User Manual, Blue Lakes LTD, 121 págs. (1982).

(35) UNEFM-FCA. VELAZQUEZ, G. y PLA, L. (coordinadores). Diagnóstico lechero de los Distritos Mauroa, Federación y Zamora del Estado Falcón, Informe Final, 3 vols., 234 págs. (1984).

(36) WEISBERG, S. Applied linear regression, Wiley, Nueva York, N. Y., 283 págs. (1980).

AGRADECIMIENTOS

Al profesor José Joaquín Villasmil, quien debió ser coautor de esta monografía, no sólo porque es un MAESTRO y como tal es capaz de transmitir sus amplios y siempre actualizados conocimientos, sino porque me ha servido de guía en mi formación estadística y, sobre todo, porque es un amigo muy exigente.

A mis colegas, Silvia Matteucci y Aída Colma, a quienes he acompañado desde mi graduación más o menos asiduamente en las tareas de investigación y las que con su ejemplo, dedicación y cariño me han familiarizado con una disciplina de trabajo, estudio y discusión, no siempre bien aprendida.

COLECCION DE MONOGRAFIAS CIENTIFICAS

Publicadas

Serie de matemática

N° 1. La Revolución en las Matemáticas Escolares, por el Consejo Nacional de Maestros de Matemáticas de los Estados Unidos de América.
N° 2. Espacios Vectoriales y Geometría Analítica, por Luis A. Santaló.
N° 3. Estructuras Algebraicas I, por Enzo R. Gentile.
N° 4. Historia de las Ideas Modernas en la Matemática, por José Babini.
N° 5. Algebra Lineal, por Orlando E. Villamayor.
N° 6. Algebra Linear e Geometria Euclideana, por Alexandre Augusto Martins Rodrigues.
N° 7. El Concepto de Número, por César A. Trejo.
N° 8. Funciones de Variable Compleja, por José I. Nieto.
N° 9. Introducción a la Topología General, por Juan Horváth.
N° 10. Funções Reais, por Djairo G. de Figueiredo.
N° 11. Probabilidad e Inferencia Estadística, por Luis A. Santaló.
N° 12. Estructuras Algebraicas II (Algebra Lineal), por Enzo R. Gentile.
N° 13. La Revolución en las Matemáticas Escolares (Segunda Fase), por Howard F. Fehr, John Camp y Howard Kellog.
N° 14. Estructuras Algebraicas III (Grupos Finitos), por Horacio H. O'Brien.
N° 15. Introducción a la Teoría de Grafos, por Fausto A. Toranzos.
N° 16. Estructuras Algebraicas IV (Algebra Multilineal), por Artibano Micali y Orlando E. Villamayor.
N° 17. Introdução à Análise Funcional: Espaços de Banach e Cálculo Diferencial, por Leopoldo Nachbin.
N° 18. Introducción a la Integral de Lebesgue en la Recta, por Juan Antonio Gatica.
N° 19. Introducción a los Espacios de Hilbert, por José I. Nieto.
N° 20. Elementos de Biomatemática, por Alejandro B. Engel.
N° 21. Introducción a la Computación, por Jaime Michelow.
N° 22. Estructuras Algebraicas V (Teoría de Cuerpos), por Héctor A. Merklen.
N° 23. Estructuras Algebraicas VI (Formas Cuadráticas), por Francisco M. Piscoya.
N° 24. Estructuras Algebraicas VII (Estructuras de Algebras), por Artibano Micali.
N° 25. Aritmética Elemental, por Enzo R. Gentile.
N° 26. Algebra Elemental, por Leopoldo Nachbin.
N° 27. Análisis Multivariado-Método de Componentes Principales, por Laura E. Pla.

Serie de física

N°. 1. Concepto Moderno del Núcleo, por D. Allan Bromley.
N° 2. Panorama de la Astronomía Moderna, por Félix Cernuschi y Sayd Codina.
N° 3. La Estructura Electrónica de los Sólidos, por Leopoldo M. Falicov.
N° 4. Física de Partículas, por Igor Saavedra.
N° 5. Experimento, Razonamiento y Creación en Física, por Félix Cernuschi.

N° 6. Semiconductores, por George Bemski.
N° 7. Aceleradores de Partículas, por Fernando Alba Andrade.
N° 8. Física Cuántica, por Onofre Rojo y Harold V. McIntosh.
N° 9. La Radiación Cósmica, por Gastón R. Mejía y Carlos Aguirre.
N° 10. Astrofísica, por Carlos Jaschek y Mercedes C. de Jaschek.
N° 11. Ondas, por Oscar J. Bressan y Enrique Gaviola.
N° 12. El Láser, por Mario Garavaglia.
N° 13. Teoría Estadística de la Materia, por Antonio E. Rodríguez y Roberto E. Caligaris.
N° 14. Aplicações da Teoria de Grupos na Espectroscopia Raman e do Infra-Vermelho, por Jorge Humberto Nicola y Anildo Bristoti.
N° 15. Fundamentos de Cristalografía Física, por Jaime Rodríguez Lara.

Serie de química

N° 1. Cinética Química Elemental, por Harold Behrens Le Bas.
N° 2. Bioenergética, por Isaias Raw y Walter Colli.
N° 3. Macromoléculas, por Alejandro Paladini y Moisés Burachik.
N° 4. Mecanismo de las Reacciones Orgánicas, por Jorge A. Brieux.
N° 5. Elementos Encadenados, por Jacobo Gómez Lara.
N° 6. Enseñanza de la Química Experimental, por Francisco Giral.
N° 7. Fotoquímica de Gases, por Ralf-Dieter Penzhorn.
N° 8. Introducción a la Geoquímica, por Félix González-Bonorino.
N° 9. Resonancia Magnética Nuclear de Hidrógeno-1 y de Carbono-13, por Pedro Joseph-Nathan.
N° 10. Cromatografía Líquida de Alta Presión, por Harold M. McNair y Benjamín Esquivel H.
N° 11. Actividad Optica, Dispersión Rotatoria Optica y Dicroísmo Circular en Química Orgánica, por Pierre Crabbé.

N° 12. Espectroscopia Infrarroja, por Jesús Morcillo Rubio.
N° 13. Polarografía, por Alejandro J. Arvía y Jorge A. Bolzan.
N° 14. Paramagnetismo Electrónico, por Juan A. McMillan.
N° 15. Introducción a la Estereoquímica, por Juan A. Garbarino.
N° 16. Cromatografía en Papel y en Capa Delgada, por Xorge A. Domínguez.
N° 17. Introducción a la Espectrometría de Masa de Sustancias Orgánicas, por Otto R. Gottlieb y Raimundo Braz Filho.
N° 18. Cinética Química, por Rodolfo V. Caneda.
N° 19. Fuerzas Intermoleculares, por Mateo Díaz Peña.
N° 20. Físico-Química de Superfícies, por Tibor Rabockai.
N° 21. Corrosión, por José R. Galvele.
N° 22. Introducción a la Electroquímica, por Dionisio Posadas.
N° 23. Cromatografía de Gases, por Harold M. McNair.
N° 24. Cinética de Disolución de Medicamentos, por Edison Cid Cárcamo.
N° 25. Introducción a la Química de Suelos, por Elemer Bornemisza.
N° 26. Elementos de Catálisis Heterogénea, por Sergio E. Droguett.
N° 27. Introducción a la Electrocatálisis, por Alejando J. Arvía y María Cristina Giordano.
N° 28. Química de Sólidos, por Julio César Bazán.
N° 29. Química Bioinorgânica, por Henrique Eisi Toma.
N° 30. Introducción al Estudio de los Productos Naturales, por Eduardo G. Gros, Alicia B. Pomilio, Alicia M. Seldes y Gerardo Burton.

Serie de biología

N° 1. La Genética y la Revolución en las Ciencias Biológicas, por José Luis Reissig.
N° 2. Bases Ecológicas para la Explotación Agropecuaria en la América Latina, por Guillermo Mann F.
N° 3. La Taxonomía y la Revolución en las Ciencias Biológicas, por Elías R. de la Sota.

N° 4. Principios Básicos para la Enseñanza de la Biología, por Oswaldo Frota-Pessoa.
N° 5. A Vida da Célula, por Renato Basile.
N° 6. Microorganismos, por J. M. Gutiérrez-Vázquez.
N° 7. Principios Generales de Microbiología, por Norberto J. Palleroni.
N° 8. Los Virus, por Enriqueta Pizarro-Suárez y Gamba.
N° 9. Introducción a la Ecología del Bentos Marino, por Manuel Vegas Vélez.
N° 10. Biosíntesis de Proteínas y el Código Genético, por Jorge E. Allende.
N° 11. Fundamentos de Inmunología e Inmunoquímica, por Félix Córdoba Alva y Sergio Estrada Parra.
N° 12. Bacteriófagos, por Romilio Espejo T.
N° 13. Biogeografía de América Latina, por Angel L. Cabrera y Abraham Willink.
N° 14. Relación Hospedante-Parásito. Mecanismo de Patogenicidad de los Microorganismos, por Manuel Rodríguez Leiva.
N° 15. Genética de Poblaciones Humanas, por Francisco Rothhammer.
N° 16. Introducción a la Ecofisiología Vegetal, por Ernesto Medina.
N° 17. Aspectos de Biología Celular y la Transformación Maligna, por Manuel Rieber.
N° 18. Transporte a Través de la Membrana Celular, por P. J. Garrahan y A. F. Rega.
N° 19. Duplicación Cromosómica y Heterocromatina a Nivel Molecular y Citológico, por Néstor O. Bianchi.
N° 20. Citogenética Básica y Biología de los Cromosomas, por Francisco A. Sáez y Horacio Cardoso.
N° 21. Ecología de Poblaciones Animales, por Jorge E. Rabinovich.
N° 22. Metodología para el Estudio de la Vegetación, por Silvia D. Matteucci y Aída Colma.
N° 23. Los Sistemas Ecológicos y la Humanidad, por Ariel E. Lugo y Gregory L. Morris.
N° 24. A Germinação das Sementes, Por Luiz Gouvêa Labouriau.
N° 25. Introducción a la Farmacocinética, por Edison Cid Cárcamo.
N° 26. Introducción a la Teoría y Práctica de la Taxonomía Numérica, por Jorge Víctor Crisci y María Fernanda López Armengol.
N° 27. ¿Qué es la Diferenciación Celular?, por Roberto B. García y Susana Pereyra Alfonso.
N° 28. Limnología Sanitaria, Estudio de la Polución de Aguas Continentales, por Samuel Murgel Branco.
N° 29. Etología: El Estudio Biológico del Comportamiento Animal, por Raúl Vaz-Ferreira.
N° 30. Fotosíntesis, por Carlos S. Andreo y Rubén H. Vallejos.
N° 31. Pesca y Piscicultura en Aguas Continentales de América Latina, por Argentino A. Bonetto y Hugo P. Castello.
N° 32. Fundamentos de Genética Biométrica y sus Aplicaciones al Mejoramiento Genético, por Jorge A. Mariotti.

En preparación

Serie de matemática

Geometrías Finitas, por Oscar Barriga.
Computadoras y Procesamiento de Datos, por Julio Villanueva y Oscar Harasic.

Serie de física

Teoría de Fluidos en Equilibrio, por Antonio E. Rodríguez y Roberto E. Caligaris.

Funcional y Adaptativa, por Elías R. de la Sota.
Origen y Anatomía del Cromosoma Eucarionte, por Nestor O. Bianchi.
Limnología Básica, por Jese Galizia Tundisi.
El Plancton de las Aguas Continentales, por Aída González de Infante.

Nota: Las personas interesadas en adquirir estas monografías deben dirigirse a la Oficina de Ventas y Promoción, Departamento de Información Pública, Organización de los Estados Americanos, Washington, D.C., 20006-4499 o a las Oficinas de la OEA en el país respectivo.