UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

ANALISIS DE COMPONENTES PRINCIPALES Y OTRAS PROYECCIONES LINEALES

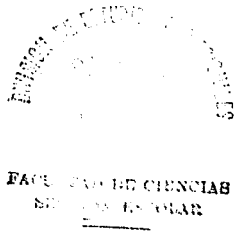
T E S I S

Que para obtener el Título de

M A T E M A T I C O

p r e s e n t a

JOSE LUIS CASTREJON CABALLERO



México, D. F. 1996

1995

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

VNIVERSIDAD NACIONAL
AVFNºMA DE
MEXICO

M. en C. Virginia Abrín Batule
Jefe de la División de Estudios Profesionales de la
Facultad de Ciencias
P r e s e n t e

Comunicamos a usted que hemos revisado el trabajo de Tesis:

Análisis de Componentes Principales y otras Proyecciones Lineales.

realizado por José Luis Castrejón Caballero.

con número de cuenta 8131805-4 **, pasante de la carrera de** Matemático.

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	Dra. Guillermina Eslava Gómez
Propietario	Mat. Margarita Elvira Chávez Cano.
Propietario	Dr. José Rodolfo Mendoza Blanco.
Suplente	M. en C. José López Estrada.
Suplente	Dr. Luis Bernardo Morales Mendoza.

Consejo Departamental de Matemáticas

M. en C. ALEJANDRO BRAVO MOJICA

A la memoria de mi padre
A mi madre

Por que me dieron lo más importante: la vida.
Por que me inculcaron los principios básicos y fundamentales de la educación.
Por la paciencia que me tuvieron.

A mis hermanos
Por todo lo que hemos compartido juntos.
Por que este trabajo también es parte de ellos.

Agradecimientos

Quiero agradecer en primer lugar el enorme apoyo brindado por la Dra. Guillermina Eslava Gómez para la realización de esta tesis. Gracias por su magnífica dirección, sus consejos, su paciencia para leer los diferentes borradores presentados, etc.

Asimismo quiero agradecer a las siguientes instituciones: la Facultad de Ciencias de la UNAM, por haber tenido la oportunidad de estudiar en ella; al Departamento de Estadística del IIMAS de la UNAM; a la Escuela Nacional de Antropología e Historia del INAH, y a la Subcomisión Mixta de Capacitación y Becas del INAH por las facilidades brindadas para la realización de esta tesis.

Gracias a la Mat. Margarita E. Chávez Cano, al Dr. José R. Mendoza Blanco, al M. en C. José López Estrada y al Dr. Luis B. Morales Mendoza por el tiempo dedicado a leer este trabajo.

Por último, agradezco a Tere quién ha estado cerca de mi durante la realización de este trabajo. Gracias por todo el impulsó brindado durante estos años.

INDICE

PRESENTACIÓN

El análisis exploratorio de datos es un conjunto de técnicas que permiten un estudio global de variables o individuos en relación a sus similitudes, diferencias o posibles estructuras de grupos. Estas técnicas suelen tener aplicación en campos de investigación tan diversos como la biología, la antropología, la geología, la medicina, etc., debido a que en muchas situaciones es de utilidad conocer si un conjunto de datos recolectados a partir de múltiples mediciones en un objeto o individuo posee alguna estructura definida, tales como conglomerados, observaciones aberrantes, etc. La estadística cuenta con una área denominada análisis multivariado que se dedica al estudio de un conjunto de variables aleatorias de manera simultánea. Esta área tradicionalmente incluye varias técnicas como análisis de componentes principales, análisis factorial, análisis de correlación canónica, análisis de discriminante, análisis de conglomerados, escalamiento multidimensional, análisis de correspondencia, etc.

En particular, existen técnicas estadísticas multivariadas basadas en transformaciones lineales cuyo objetivo es mostrar la estructura de los datos, comúnmente en una ó dos dimensiones. La técnica más conocida de este tipo, recibe el nombre de Análisis de Componentes Principales, la cual es una de las técnicas exploratorias del análisis multivariado más antigua, más conocida, y quizás la más usada. En términos de álgebra lineal la técnica consiste en buscar la descomposición espectral de una matriz, que teóricamente es positiva definida, aunque en casos prácticos no siempre es así.

Por otra parte, en años recientes, ha tenido auge una técnica denominada Proyecciones Perseguidas, del inglés *Projection Pursuit*, también basada en transformaciones lineales. Aunque la idea de realizar transformaciones lineales teóricamente tiene mucho tiempo, el desarrollo reciente de esta técnica se debe al acelerado avance que en los últimos años se ha tenido en el uso de computadoras cada vez más potentes. La técnica de Proyecciones Perseguidas tiene como objetivo buscar "proyecciones interesantes" de los datos en alguna dimensión, usualmente dos, que permita visualizar su estructura. Diferentes definiciones de lo qué es una proyección "interesante" han dado origen a plantear diferentes índices de proyección, debido a lo cual, se podría considerar que en realidad Proyecciones Perseguidas

es un conjunto de técnicas que buscan encontrar la mejores proyecciones que permitan visualizar la estructura de los datos.

Con estas ideas en mente, el objetivo de esta tesis es presentar el uso de transformaciones lineales en la estadística describiendo particularmente las técnicas de análisis de componentes principales y de proyecciones perseguidas. En este trabajo no se pretende realizar un análisis estadístico de los ejemplos presentados, sino ilustrar el uso de estas técnicas en el análisis exploratorio de datos. También utilizaremos algunos de los conceptos clásicos de la estadística, tales como media, varianza, covarianza, correlación, función de densidad, función de distribución, etc.

Aunque la instrumentación de estas técnicas en computadora nos lleva directamente al área del análisis numérico, en este trabajo no se profundizará en este aspecto utilizándose *software* ya desarrollado. En el caso del análisis de componentes principales, dado que es una técnica desarrollada en la mayoría de los paquetes computacionales estadísticos comerciales haremos uso de ellos para desarrollar las aplicaciones prácticas; en particular se utiliza el paquete SPSS. Para el caso de la técnica de proyecciones perseguidas la mayoría de los algoritmos propuestos se han desarrollado en el lenguaje de programación FORTRAN y más recientemente la empresa estadounidense *Bellcore* (*Bell Comunication Research*) (1994), ha desarrollado un programa, en lenguaje C, de gráficas dinámicas para análisis de datos llamado XGobi, implantado en el ambiente X Windows bajo el sistema operativo UNIX, el cual incluye la instrumentación de varios algoritmos propuestos por diferentes investigadores en relación a esta técnica. En esta tesis se utiliza este *software* para ejemplificar la técnica en tres casos prácticos.

En base a todo lo anterior, este trabajo de tesis está organizado de la forma siguiente:

En el capítulo 1 se presenta la notación que utilizaremos a lo largo del trabajo y algunos conceptos estadísticos generales, tales como función de densidad, función de distribución, momentos, media aritmética, varianza, covarianza, correlación, cumulantes y estimación de funciones de densidad. Asimismo se introduce el concepto de escalamiento de las variables, traducido del término inglés *Scaling*, y el de esferado de las variables, traducido del término inglés *Sphering*. En el capítulo 2 se presenta de manera extensa la técnica de Análisis de

Componentes Principales, exponiendo desde un breve bosquejo histórico, pasando por la descripción geométrica, algebraica y del calculo diferencial, así como diferentes usos que puede tener esta técnica en análisis de regresión múltiple entre otros aspectos. En el capítulo 3 se expone la técnica de Proyecciones Perseguidas concentrando la atención del capítulo en la presentación de algunos de los índices de proyección más conocidos en relación a esta técnica, propuestos por investigadores tales como Friedman y Tukey; Huber; Jones y Sibson; Jee; Friedman; Yenyukov; Eslava y Marriott; Hall; Cook, Buja y Cabrera. En el capítulo 4 se presentan tres ejemplos utilizando datos reales, el primero referente a medidas de cráneos de dos civilizaciones antiguas de México; el segundo ejemplo corresponde a la regeneración de plantas en la región del amazonas en Brasil; por último se presenta un ejemplo en el cual se utiliza la técnica para clasificar un conjunto de virus a los cuales se les han tomado algunas medidas relacionadas con restos de moléculas de aminoácidos. Algunas conclusiones sobre los temas tratados se encuentran en el capítulo 5. Finalmente, se presentan dos apéndices: el primero presenta los datos de los ejemplos estudiados y el segundo las instrucciones del paquete SPSS para el Análisis de Componentes Principales, así como una breve introducción a la aplicación de la técnica de proyecciones perseguidas utilizando el programa XGobi.

CAPÍTULO 1. INTRODUCCIÓN

En este capítulo se presenta la notación y algunos conceptos básicos que se emplearan a lo largo de la tesis, tales como función de distribución, momentos, cumulantes, esperanza, varianza, etc. Asimismo, se presentan dos procedimientos de estandarización importantes para el tratamiento de datos multivariados: el de escalamiento y el de esferado de las variables, que deben ser considerados antes de aplicar el análisis de componentes principales y el de proyecciones perseguidas respectivamente.

1.1 Notación inicial

Se considera una población como el conjunto de objetos bajo estudio, los cuales pueden ser plantas, personas, ciudades, máquinas, etc. Para denotar diferentes medidas realizadas en tales objetos se utilizará la variable x. Supondremos que tenemos p medidas diferentes, lo cual da lugar a las variables $x_1, x_2,...,x_p$. Denotaremos por N el número de elementos de la población. Una muestra de la población es un subconjunto de la población y consideraremos que su tamaño es de n elementos.

1.2 Funciones de distribución

Consideremos una variable x que mide alguna característica de un objeto. Sea $f(x)$ la proporción de veces que aparece la variable x en la población de estudio. La función $f(x)$ es llamada la función de densidad de la variable x.

Nótese que

$$\sum_{i=1}^{M} f(x_i) = 1,$$

donde M es el número de valores diferentes de x.

Aunque en la práctica es muy difícil que suceda, supongamos que x es una variable que puede tomar un número infinito de valores, entonces se deberá cumplir que la serie infinita resultante sea convergente. Es decir

$$\sum_{i=1}^{\infty} f(x_i) \quad \text{deberá converger a 1}$$

Con esta idea se puede definir una nueva función, llamada la función de distribución, que denotaremos por $F(x)$, como sigue:

$$F(x) = \sum_{i=1}^{r} f(x_i) \quad \text{para r < M}$$

Por otra parte, si x es una variable continua, entonces se pueden construir rangos o intervalos para definir la función de distribución de x. Considerando estos rangos de tamaño Δx y haciendo este número tan pequeño como sea posible, la función $F(x)$ se definiría como:

$$F(x) = \int_{-\infty}^{x} f(u)du$$

A fin de mantener la idea de lo que pasa con las variables discretas se deberá cumplir que

$$\lim_{x \to \infty} F(x) = 1.$$

Comúnmente la función $F(x)$ en el caso discreto y continuo suele escribirse respectivamente como:

$$F(x) = \sum_{i=-\infty}^{r} f_i(x)$$

$$F(x) = \int_{-\infty}^{x} f(u)du$$

la función de distribución deberá cumplir que:

$$\sum_{i=-\infty}^{\infty} f_i(x) = F(\infty) - F(-\infty) = 1$$

$$\int_{-\infty}^{\infty} f(u)du = F(\infty) - F(-\infty) = 1$$

La integral de Stieltjes tiene la característica que en el caso discreto se reduce a una suma ordinaria y en el caso continuo a la integral de Cauchy (Kendall, 1977, pags. 15-16). En consecuencia, para ambos casos de variables, discretas o continuas, la función *F(x)* se puede denotar por:

$$F(x) = \int_{-\infty}^{x} f(u)du$$

En adelante se considerará esta expresión para referirnos a la función de distribución *F(x)*.

1.3 Momentos univariados. Media aritmética y varianza

Si x es una variable utilizada para medir una característica de los elementos de una población se define el momento de orden r, μ_r^a, de los datos alrededor de un punto arbitrario a, por medio de la integral de Stieltjes:

$$\mu_r^a = \int_{-\infty}^{\infty} (x-a)^r dF$$

donde *f(x)* es la función de densidad de la variable x y $dF = f(x)dx$, cuando f es continua.

En particular el primer momento μ_1^a recibe el nombre de media aritmética de los datos alrededor del punto a y es usual denotarlo simplemente por μ^a. Es decir:

$$\mu^a = \int_{-\infty}^{\infty} (x-a)dF$$

Si $a = \mu^a$ se sustituye en la expresión para μ_r^a queda:

$$\mu_r^a = \int_{-\infty}^{\infty} (x - \mu^a)^r \, dF$$

el cual es llamado el momento central de orden r, siendo común denotarlo simplemente por μ_r. Otro caso particular es el momento central de orden dos que es conocido como la varianza de los datos alrededor de la media, la cual es una medida de dispersión de los datos. Usualmente la varianza es denota por el símbolo σ^2. Es decir

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu^a)^2 \, dF$$

A la raíz cuadrada positiva de la varianza se le llama la desviación estándar de los datos y se denota por la letra σ.

Si en lugar de una variable x se tiene una función $h(x)$, el concepto de media aritmética de esta función, denotada por $E(h(x))$ se extiende de manera natural como:

$$E(h(x)) = \int_{-\infty}^{\infty} h(x) dF$$

la letra E es usada debido a que esta expresión suele llamarse en teoría de probabilidades, el valor esperado o esperanza de la función h. En la literatura estadística es común encontrar $E(x)$ o μ para denotar la media de la variable x.

La definición de esperanza de una función se expande para el caso en que la función dependa de p variables, $h(x_1,x_2,...,x_p)$, como a continuación se observa:

$$E(h(x_1, x_2, ..., x_p)) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x_1, x_2, ..., x_p) dF$$

donde $F(x_1,x_2,...,x_p) = P(X_1 \le x_1, X_2 \le x_2,..., X_p \le x_p)$

1.4 Momentos bivariados. Covarianza y correlación

Consideremos que se han realizado dos medidas, x_1, x_2, en una población determinada. El momento bivariado $\mu_{rs}^{a_1,a_2}$ de x_1 alrededor de a_1 y x_2 alrededor de a_2 se define como:

$$\mu_{rs}^{a_1,a_2} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1-a_1)^r(x_2-a_2)^s dF$$

En esta expresión se observa que μ_{r0} representa el r-ésimo momento de x_1, y μ_{0s}, el s-ésimo momento de x_2. Si a_1 es la media de la variable x_1 y a_2 es la media de la variable x_2, entonces μ_{rs} se conoce como el momento central bivariado de las variables x_1 y x_2.

En particular μ_{11} es llamado la covarianza de las variables x_1, x_2. La expresión escrita en forma explícita es:

$$\text{cov}(x_1,x_2) = \mu_{11} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1-\mu_1)(x_2-\mu_2)dF$$

donde μ_1 y μ_2 son las medias aritméticas de x_1 y x_2 respectivamente.

Otra medida importante de relación entre dos variables es el coeficiente de correlación lineal, denotado por $\rho(x_1,x_2)$, el cual se define como:

$$\rho(x_1,x_2) = \frac{\mu_{11}}{\sqrt{\mu_{20}\mu_{02}}} = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1-\mu_1)(x_2-\mu_2)dF}{\sqrt{\int_{-\infty}^{\infty}(x_1-\mu_1)^2 dF \int_{-\infty}^{\infty}(x_2-\mu_2)^2 dF}}$$

$$= \frac{\text{cov}(x_1,x_2)}{\sigma_1\sigma_2}$$

donde σ_1 y σ_2 son las desviaciones estándar de x_1 y x_2 respectivamente y $-1 \le \rho \le 1$.

Nótese que si $x_1 = x_2$ entonces $\rho(x_1,x_2) = 1$, Asimismo, si $x_1 \neq x_2$ pero $|\rho(x_1,x_2)| \approx 1$, indica correlación alta entre las dos variables. En sentido contrario si $\rho(x_1,x_2) = 0$, indicaría no correlación entre ellas.

1.5 Cumulantes

Los momentos no son el único conjunto de constantes que describen el comportamiento de una función de distribución. Los cumulantes son otro conjunto de constantes de este tipo cuyas propiedades son teóricamente más útiles en algunas circunstancias específicas. En esta tesis serán utilizados en la construcción de un índice de proyecciones perseguidas llamado de momentos, el cual se presenta en el capítulo 3. Formalmente los cumulantes univariados k_1, $k_2,\ldots,k_r$ se definen por la identidad:

$$\exp\left\{k_1 t + \frac{k_2 t}{2!} + \cdots + \frac{k_r t^r}{r!} + \cdots\right\} = 1 + \mu_1 t + \frac{\mu_2 t^2}{2!} + \cdots + \frac{\mu_r t^r}{r!} + \cdots$$

donde los momentos son definidos alrededor de un punto a cualquiera.

Kendall (1977, pags. 69-73) desarrolla esta expresión con el objetivo de expresar a los cumulantes en términos de los momentos centrales. A continuación se presentan en forma explícita la expresión de los primeros 6 cumulantes:

$$k_1 = 0$$
$$k_2 = \mu_2$$
$$k_3 = \mu_3$$
$$k_4 = \mu_4 - 3\mu_2^2$$
$$k_5 = \mu_5 - 10\mu_3\mu_2$$
$$k_6 = \mu_6 - 15\mu_4\mu_2 - 10\mu_3^2 + 30\mu_2^3$$

De manera análoga para el caso bivariado los cumulantes están definidos por la siguiente identidad:

$$\exp\left\{\frac{k_{10}}{1!0!}t_1+\frac{k_{01}}{0!1!}t_2+\cdots+\frac{k_{rs}}{r!s!}t_1^r t_2^s+\cdots\right\}$$
$$=1+\frac{\mu_{10}}{1!0!}t_1+\frac{\mu_{01}}{0!1!}t_2+\cdots+\frac{\mu_{rs}}{r!s!}t_1^r t_2^s+\cdots$$

De nuevo Kendall (1977, pags. 85-87), da las expresiones explícitas para los cumulantes en términos de los momentos centrales bivariados. A continuación se presentan algunas de estas expresiones:

$k_{11}=\mu_{11}$
$k_{21}=\mu_{21}$
$k_{31}=\mu_{31}-3\mu_{20}\mu_{11}$
$k_{22}=\mu_{22}-\mu_{20}\mu_{02}-2\mu_{11}^2$
$k_{41}=\mu_{41}-4\mu_{30}\mu_{11}-6\mu_{21}\mu_{20}$
$k_{32}=\mu_{32}-\mu_{30}\mu_{02}-6\mu_{21}\mu_{11}-3\mu_{20}\mu_{12}$
$k_{51}=\mu_{51}-5\mu_{40}\mu_{11}-10\mu_{31}\mu_{20}-10\mu_{30}\mu_{21}+30\mu_{20}\mu_{11}$
$k_{42}=\mu_{42}-\mu_{40}\mu_{02}-8\mu_{31}\mu_{11}-4\mu_{30}\mu_{12}-6\mu_{22}\mu_{20}-6\mu_{21}^2+6\mu_{20}^2\mu_{02}+24\mu_{20}\mu_{11}^2$

1.6 Notación complementaria

Mencionamos en la sección 1.1 que utilizaremos N para denotar el tamaño de la población y p para el número de variables o medidas de los objetos. Se acostumbra denotar por X la matriz poblacional que tiene como columnas las p variables y como renglones los N casos u observaciones de la población. Por lo tanto la matriz X tiene la forma siguiente:

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix}$$

Para el caso de una muestra de tamaño n, la expresión es la misma y sólo tendremos que cambiar N por n.

En el apartado 1.4 hablamos de las covarianzas entre un par de variables; considerando además la varianza de cada una de las p variables se construye la matriz de covarianza Σ, denominada también matriz de dispersión, la cual es de la forma:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix}$$

donde σ_{ij} es la covarianza entre las variables i y j, para $i \neq j$ y $\sigma_{ii} = \sigma_i^2$. Esta matriz tiene la característica de ser simétrica y positiva definida.

Asimismo, se define la matriz de correlación P como:

$$\mathrm{P} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

la cual también es simétrica y positiva definida.

En el caso muestral denotaremos por $\bar{x}_i$ la media aritmética de la variable x_i, por c_{ij} la covarianza entre dos variables x_i, x_j y por r_{ij} la correlación. Explícitamente tenemos:

$$\bar{x}_i = \frac{\sum_{k=1}^{n} x_{ki}}{n} \quad \text{para } k = 1,\ldots,p$$

$$c_{ij} = \frac{1}{n}\sum_{k=1}^{n}\left(x_{ki} - \bar{x}_i\right)\left(x_{kj} - \bar{x}_j\right) \quad \text{para } k = 1,\ldots,p$$

$$r_{ij} = \frac{c_{ij}}{\sqrt{c_{ii}c_{jj}}}$$

Por lo tanto si la matriz de observaciones muéstrales es:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

entonces las matrices de covarianza y correlación muéstrales denotadas por $\boldsymbol{C}$ y $\boldsymbol{R}$ respectivamente están dadas por:

$$C = \begin{bmatrix} c_1^2 & c_{12} & \cdots & c_{1p} \\ c_{21} & c_2^2 & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_p^2 \end{bmatrix}$$

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

A menudo estaremos hablando de la transpuesta y de la inversa de una matriz $\boldsymbol{A}$, las cuales denotaremos por $\boldsymbol{A}'$ y $\boldsymbol{A}^{-1}$ respectivamente

En la sección 1.2 se presentó la idea de función de densidad de una variable; la función de densidad teórica más importante para datos multivariados es la Gaussiana o normal multivariada; la forma más usual de escribirla es:

$$f(x) = \frac{|\Sigma|^{-1/2}}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}$$

donde x es el vector de variables, μ es el vector de medias aritméticas y Σ es la matriz de covarianza. Esta función es muy utilizada ya que forma la base de gran parte de inferencia del análisis multivariado.

Otra función teórica a la que nos referiremos en esta tesis, es la llamada χ^2 (Ji-cuadrada). En el caso univariado esta dada por la expresión:

$$f(z) = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} \exp\left(-\frac{1}{2}z\right) z^{\frac{\nu}{2}-1} \qquad z \geq 0, \quad \nu > 0$$

donde z es una variable que denota la suma de los cuadrados de ν variables con distribución Gaussiana estándar (media = 0 y varianza = 1) y $\Gamma(x)$ es la función matemática gamma.

1.7 Estimación de densidades

En la sección 1.2 introducimos el concepto de función de densidad de una variable. En los casos prácticos en los que se trabaja con una muestra de la población no se conoce la función de densidad poblacional de un conjunto de datos, debido a lo cual es necesario estimar dicha función. Existen diferentes métodos de estimación; particularmente en esta tesis, en el capítulo 3, nos referiremos a estimaciones de funciones de densidad a partir de funciones núcleo (*kernel function*) entre otras. La teoría al respecto se puede consultar en Silverman (1986). Es común denotar por $\hat{f}(x)$ la estimación de la función de densidad $f(x)$.

1.8 Escalamiento de las variables

Cuando se esta realizando una investigación donde se miden diferentes características de un sujeto, en general las variables que se utilizan son de diferente tipo y tienen diferentes escalas de medición. Por ejemplo, en un estudio donde se quiera caracterizar a un grupo de personas adultas, con respecto a dos variables: la estatura medida en metros y su peso medido en kilogramos, se podría fácilmente cometer el error de creer que la varianza de la variable peso

será mucho mayor que la de la variable estatura debido a que las unidades son más grandes en un variable que en otra. Este hecho nos llevaría a conclusiones erróneas al emplear técnicas de análisis multivariado, en particular el análisis de componentes principales. Esta técnica no es invariante con respecto a la escala, como si lo es por ejemplo la técnica del análisis de regresión.

En el capítulo 2 se verá a detalle que la técnica de análisis de componentes principales se basa en las varianzas y covarianzas de las variables. Debido a esto, antes de emplear la técnica es recomendable escalar las variables originales.

Existen diferentes formas de escalamiento de las variables, siendo la más usada la estandarización, la cual consiste en que a cada variable se le resta la media y se divide entre su desviación estándar. Esto implica que las variables escaladas o estandarizadas tendrán media cero y varianza unitaria.

Es decir, si originalmente se tienen p variables $x_1, x_2,...,x_p$ entonces se define un nuevo conjunto de p variables mediante la transformación:

$$z_i = \frac{x_i - \bar{x}_i}{\sqrt{\text{var}(x_i)}} \qquad i=1,...,p;$$

estas nuevas variables serán la base para realizar el análisis de componentes principales. Nótese que la covarianza entre dos variables estandarizadas z_i, z_j es igual a la correlación entre las variables originales x_i, x_j tal como se demuestra a continuación:

$$\text{cov}(z_i, z_j) = \text{cov}\left(\frac{x_i - \bar{x}_i}{\sqrt{\text{var}(x_i)}}, \frac{x_j - \bar{x}_j}{\sqrt{\text{var}(x_j)}}\right) = E\left(\frac{x_i - \bar{x}_i}{\sqrt{\text{var}(x_i)}} \frac{x_j - \bar{x}_j}{\sqrt{\text{var}(x_j)}}\right)$$

$$= \frac{1}{\sqrt{\text{var}(x_i)}\sqrt{\text{var}(x_j)}} E\left(\left(x_i - \bar{x}_i\right)\left(x_j - \bar{x}_j\right)\right) = \frac{\text{cov}(x_i, x_j)}{\sqrt{\text{var}(x_i)}\sqrt{\text{var}(x_j)}}$$

$$= corr(x_i, x_j)$$

Otra forma de escalar a las variables es la de estandarizar a las variables por el rango en lugar de la desviación estándar, lo cual no parece tener mucha utilidad; una forma más es

cuando se usan estimaciones ponderadas de la varianza a fin de evitar posibles distorsiones debido a la existencia de observaciones aberrantes.

1.9 Esferado de las variables

"*Sphering*", término que puede traducirse como esferado de las variables, es el proceso de transformación de los datos originales para que la varianza sea la misma en cualquier dirección. Es decir si se tienen p variables al aplicar el proceso de esferado todos los puntos en el espacio p, se concentran en promedio en una hiper-esfera en esa dimensión, lo cual es equivalente a decir que las variables esferadas tienen como matriz de varianza-covarianza la matriz identidad I.

En términos algebraicos el esferado de las variables es el proceso de encontrar combinaciones lineales $y_1, y_2,...,y_p$ de las variables originales $x_1, x_2,...,x_p$ que sean no correlacionadas y con varianza unitaria. Este método de esferar a las variables es recomendable que se realice antes de aplicar una técnica de Proyecciones Perseguidas, a fin de evitar interpretaciones erróneas en las proyecciones debida a las posibles escalas de medición de las variables originales.

Revisando la literatura de álgebra lineal (Strang, 1982), en la parte concerniente a transformaciones lineales, se afirma que si A es una matriz simétrica positiva definida, entonces existe una matriz L, no singular, triangular inferior que cumple la igualdad $A=LL'$. Esta factorización de la matriz A es llamada la descomposición de Cholesky. La unicidad de la factorización se da cuando los elementos de la diagonal son positivos.

En el contexto estadístico, se tienen una matriz X de dimensión $p\times n$, p variables y n observaciones, y una matriz de varianza-covarianza Σ en donde su inversa Σ^{-1} es simétrica y positiva definida, por lo que se asegura que existe una matriz L que cumple $\Sigma^{-1} = LL'$, con L una matriz no singular, triangular inferior.

Si Y representan la matriz de los datos esferados, entonces $Y=L'X$, obliga a que la matriz de dispersión sea la matriz identidad. Esta afirmación se prueba a continuación:

$$\mathrm{Var}(Y) = L'\mathrm{Var}(X)L = L'(LL')^{-1}L = L'L'^{-1}L^{-1}L = I$$

Por lo tanto se puede decir que la transformación $Y = L'X$ satisface los requerimientos para el esferado de las variables.

En el caso de una muestra la matriz de varianza-covarianza Σ se estima por la matriz C, lo que simbólicamente se representa como:

$$\hat{\Sigma} = C$$

Este método presenta serios problemas si los datos originales presentan alguna estructura de agrupamiento. Al aplicar el proceso de esferado de las variables a este tipo de datos, la estructura natural de los datos se distorsiona y se complica aún más la posible identificación de la estructura de grupos. Por ejemplo, si los datos originales están agrupados en tres grupos con una estructura definida, al aplicar el proceso de esferado estos grupos se distorsionarían (ver figura 1).

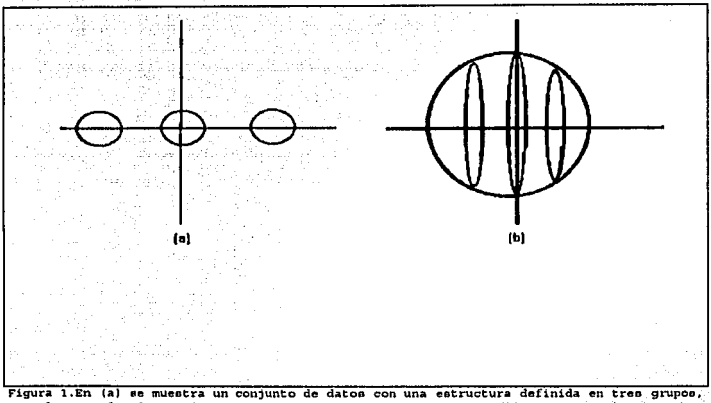


Figura 1.En (a) se muestra un conjunto de datos con una estructura definida en tres grupos, que al ser esferados se distorsionan, quedando como en indica en (b).

Esta situación ha sido fuertemente criticada por Gower (Jones and Sibson, discussion, 1987). La distorsión de la información puede ser reducida si se considera, que al utilizar los componentes principales en la transformación se introducen variables que no generen información, por lo que al considerar únicamente los componentes que absorban la mayor

parte de la variación de los datos originales y desechando los variables redundantes el problema se reducirá, y los grupos podrán ser diferenciados aún en esta situación. Geométricamente podría decirse que los datos originales son transformados de tal manera, que si originalmente tenían forma elíptica, ahora tienen forma esférica.

Para la interpretación de resultados de la aplicación posterior de cualquier técnica o índice de proyección deberá tomar en cuenta el proceso de esferado de las variables, realizado previamente y esto no siempre es fácil; el esferado de las variables originales obscurece la interpretación en términos de las variables originales.

CAPÍTULO 2. ANÁLISIS DE COMPONENTES PRINCIPALES

En este capítulo se describe una de las técnicas más antigua del análisis multivariado de datos, se presentan tres enfoques diferentes: geométrico, algebraico y del cálculo diferencial. Se presentan algunas pruebas inferenciales que se pueden realizar suponiendo distribución Gaussiana en los datos y se comenta el caso particular de cuando se tienen variables discretas. Se comenta sobre la utilidad de esta técnica en regresión lineal múltiple cuando se tiene el problema de multicolinealidad..

2.1 Introducción y breve reseña histórica

El Análisis de Componentes Principales es una técnica que permite observar posibles estructuras, si es que existen, en un conjunto de datos multivariados obtenidos de una población, cuya distribución de probabilidades no necesita ser conocida. Es una técnica algebraica-geométrica que no requiere un modelo estadístico para explorar y/o describir estructuras o comportamiento de los datos. Es decir, no existe un objetivo inferencial, se trata de una técnica exploratoria que consiste en obtener el máximo de información con el mínimo posible de hipótesis. Sin embargo si se supone que la población tiene una distribución Gaussiana, la muestra observada podrá ser utilizada para efectuar inferencia estadística a partir de pruebas de hipótesis que contribuyan a conocer la estructura de la población original.

Los primeros trabajos que se conocen, relacionados con la técnica de Análisis de Componentes Principales, se le atribuyen a Karl Pearson (1901) quien publicó un trabajo sobre el ajuste de puntos en un espacio multidimensional a una línea o a un plano. Este enfoque fue retomado por Hotelling (1933), quien fue el primero en introducir el Análisis de Componentes Principales tal como se conoce actualmente. El trabajo de Pearson se centraba en aquellas combinaciones lineales de variables originales para los cuales la varianza no explicada fuera mínima. Estas combinaciones forman un plano que es una 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 trabajo de Hotelling se centraba en el análisis de las componentes que sintetizan la máxima variabilidad del sistema de observaciones; quizás a esto se le deba el calificativo de "Principal". Por inspección de estas 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. Jolliffe (1986, pag. 5-7) presenta un desarrollo histórico más extenso de esta técnica.

Probablemente en la actualidad, de las técnicas de análisis multivariado, la de Análisis de Componentes principales sea la más conocida y por lo tanto más usada principalmente como una técnica exploratoria. Los objetivos más importantes del Análisis de componentes Principales son:

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

- Ayudar a la reducción de la dimensionalidad del problema que se está estudiando, como paso previo para futuros análisis.

- Ayudar a detectar algunas de las variables originales que aportan poca información.

2.2 Descripción Geométrica

Dado un conjunto de variables originales x_i, de un conjunto de objetos de una población, se desea encontrar nuevas variables y_i que sean rotaciones de las anteriores, para explicar mejor la variación del conjunto de datos. Estas nuevas variables ordenadas se llaman las componentes principales. Supongamos que utilizamos las primeros k componentes principales, entonces se quiere que el subespacio que ellos generan contenga el "mejor" panorama de visualización de los datos en la dimensión k. Frecuentemente las primeros componentes principales son utilizados para revelar posibles estructuras en los datos.

A fin de ilustrar geométricamente esta técnica consideremos el caso particular en el que se tienen observaciones en dos variables, x_1, x_2, de un conjunto de individuos; es decir $p=2$. Supongamos que el comportamiento de tales observaciones presentan una considerable variación en ambas variables y que la varianza mayor la tiene la variable x_2. La idea es

realizar una rotación de los ejes hasta lograr que el primero de ellos asimile la máxima varianza; esta rotación en realidad significa un cambio de base, que debe cumplir además que los nuevos ejes sean ortogonales. Si los nuevos ejes están definidos por las variables y_1, y_2 la gráfica de las observaciones se vería como en la figura 1.

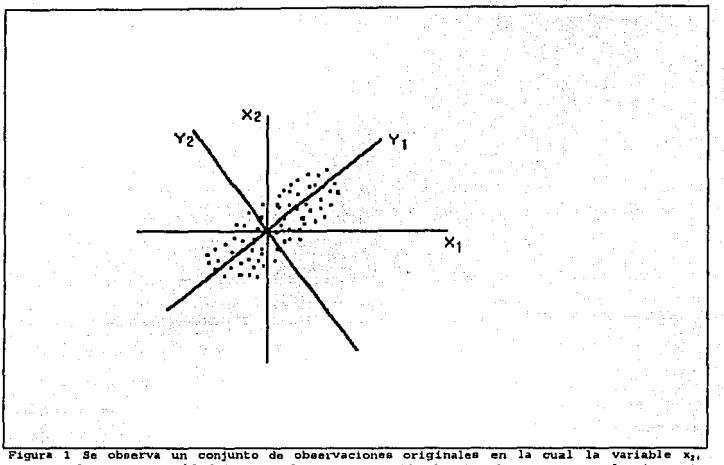


Figura 1 Se observa un conjunto de observaciones originales en la cual la variable x_1, presenta la mayor variablidad; se realiza una rotación de ejes de manera que el primer eje asimile la máxima variabilidad de los datos.

A las nuevas variables y_1, y_2 producto de estas transformaciones de las variables originales se les llama las componentes principales y se puede decir que el primero bastaría para explicar el comportamiento de los datos en un momento dado. En el caso más general, si se tiene un conjunto de p-variables que presentan una cierta correlación, entonces las primeros componentes principales serán aquellos que absorban la máxima varianza de los datos. Algunas veces el primero o los dos primeros componentes principales, ordenados con respecto a la asimilación de varianza, serán suficientes para mostrar en una dimensión o en un plano posibles estructuras y/o comportamiento de las observaciones. En términos estrictamente geométricos, las componentes principales definen los ejes principales de hiper-

elipsoides en un espacio de dimensión p, representándolos en forma decreciente de acuerdo a su magnitud. Otro tipo de interpretaciones de las componentes principales, que permiten aplicar esta técnica en investigación aplicada han sido escritos por Rao (1964, pags. 329-358).

2.3 Descripción Algebraica

Supongamos una muestra de una población de objetos con N observaciones de p variables, las cuales representaremos con el vector $x = (x_1, x_2,...,x_p)$. El problema consiste en encontrar p variables: $y_1, y_2,...,y_p$ que sean combinaciones lineales de las variables x', y que a la vez no estén correlacionadas linealmente. Es decir $\text{cov}(y_i, y_j)=0$ para $i \neq j$. Por lo tanto queremos encontrar p^2 constantes l_{ij}, $i,j=1,...,p$ que satisfagan las siguientes ecuaciones:

$$
\begin{aligned}
y_1 &= l_{11}x_1 + l_{12}x_2 + \cdots + l_{1p}x_p \\
y_2 &= l_{21}x_1 + l_{22}x_2 + \cdots + l_{2p}x_p \\
&\vdots \\
y_p &= l_{p1}x_1 + l_{p2}x_2 + \cdots + l_{pp}x_p
\end{aligned}
$$

este sistema de ecuaciones lineales se puede reescribir como:

$$y_i = \sum_{j=1}^{p} l_{ij}x_j \quad \text{para } i = 1,..., p$$

Supongamos, sin perdida de generalidad, que cada una de las variables originales están medidas alrededor de su media, es decir su media es cero, lo que implica que la media de las variables y's también sea cero. Podemos establecer la condición de no correlación de la siguiente manera:

$$\operatorname{cov}(y_i, y_j) = E(y_i y_j) = E\left(\sum_{k=1}^{p} l_{ik} x_k \sum_{m=1}^{p} l_{jm} x_m\right) = \sum_{k,m=1}^{p} l_{ik} l_{jm} E(x_k x_m)$$

$$= \sum_{k,m=1}^{p} l_{ik} l_{jm} c_{km} = 0$$

esto es valido para $i \neq j$ y donde c_{ij} denota la covarianza de x_i con x_j, y $C_{p,p}$ la matriz de covarianza muestral.

Se puede observar que se tienen $\frac{p(p-1)}{2}$ restricciones para los vectores que contienen a las constantes l's, lo cual implica que se tengan varias soluciones que satisfacen la condición de independencia requerida. Por lo tanto, a fin de tener una solución única, pediremos que los vectores l's sean ortonormales, es decir se debe cumplir que:

$$\sum_{i=1}^{p} l_{ij} l_{ik} = \begin{cases} 0, j \neq k \\ 1, j = k \end{cases}.$$

Reescribiendo el problema en notación matricial se tiene:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1p} \\ l_{21} & l_{22} & \cdots & l_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ l_{p1} & l_{p2} & \cdots & l_{pp} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$$

En forma compacta se tiene:

$$y = lx$$

las condiciones de ortonormalidad se escriben como $ll' = I$ donde I es la matriz identidad de dimensión p. Esto da pauta para despejar x en términos de y, de la forma:

$$y = lx \Rightarrow l'y = l'l\,x = Ix = x$$

es decir $x = l'y$ lo que implica que la no correlación de las y's se puede expresar como:

$$E(yy') = E((lx)(lx)') = E(lxx'l') = lE(xx')l' = \Lambda,$$

Por lo tanto se tiene la igualdad

$$lCl' = \Lambda,$$

donde Λ es una matriz con ceros fuera de la diagonal y números $\lambda_1, \lambda_2, \ldots, \lambda_p$ en la diagonal. Si esta igualdad se premultiplica por l', queda la expresión:

$$Cl' = l'\Lambda,$$

escribiéndola en forma explícita:

$$\begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pp} \end{pmatrix} \begin{pmatrix} l_{11} & l_{21} & \cdots & l_{p1} \\ l_{12} & l_{22} & \cdots & l_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ l_{1p} & l_{2p} & \cdots & l_{pp} \end{pmatrix} =$$

$$\begin{pmatrix} l_{11} & l_{21} & \cdots & l_{p1} \\ l_{12} & l_{22} & \cdots & l_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ l_{1p} & l_{2p} & \cdots & l_{pp} \end{pmatrix} \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix}$$

Al multiplicar los renglones de C por la primera columna de l se tienen p ecuaciones, como se ve a continuación:

$$\begin{aligned} c_{11}l_{11} + c_{12}l_{12} + \cdots + c_{1p}l_{1p} &= l_{11}\lambda_1 \\ c_{21}l_{11} + c_{22}l_{12} + \cdots + c_{2p}l_{1p} &= l_{12}\lambda_1 \\ \vdots \qquad\quad \vdots \qquad\qquad \vdots \qquad\quad \vdots & \\ c_{p1}l_{11} + c_{p2}l_{12} + \cdots + c_{pp}l_{1p} &= l_{1p}\lambda_1 \end{aligned}$$

homogeneizando estas ecuaciones se tiene:

$$(c_{11}-\lambda_1)l_{11} + c_{12}l_{12} + \cdots + c_{1p}l_{1p} = 0$$
$$c_{21}l_{11} + (c_{22}-\lambda_1)l_{12} + \cdots + c_{2p}l_{1p} = 0$$
$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$
$$c_{p1}l_{11} + c_{p2}l_{12} + \cdots + (c_{pp}-\lambda_1)l_{1p} = 0$$

Si las columnas de la matriz l se denotan como $l_1, l_2,...,l_p$ entonces el sistema se puede escribir como:

$$Cl_i = \lambda_i l_i$$

o bien

$$(C - \lambda_i I)l_i = 0$$

Un resultado de álgebra lineal dice que este sistema tiene solución diferente de la trivial si y solo si el determinante del sistema homogéneo es cero, es decir

$$\begin{vmatrix} c_{11}-\lambda_1 & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22}-\lambda_1 & \cdots & c_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pp}-\lambda_1 \end{vmatrix} = 0$$

Escribiendo lo anterior en notación matricial se tiene:

$$|C-\lambda_1 I|=0.$$

Repitiendo este proceso para el resto de la multiplicación de las matrices, se obtienen las p^2 ecuaciones:

$$Cl_1 = \lambda_1 l_1, \quad Cl_2 = \lambda_2 l_2, \ldots, \quad Cl_p = \lambda_p l_p$$

lo que implica plantear las ecuaciones:

$$|C-\lambda_i I| = 0 \quad \text{para } i = 1,\ldots,p$$

compactando la notación se puede escribir en general:

$$|\boldsymbol{C}-\lambda \boldsymbol{I}| = 0$$

De acuerdo al álgebra lineal los valores λ's se llaman los valores característicos de la matriz $\boldsymbol{C}$ y los vectores $\boldsymbol{l}$'s correspondientes son los vectores característicos o de la misma matriz.

Una vez que se han encontrado los valores característicos $\lambda_1, \lambda_2, \ldots, \lambda_p$ se sustituyen en las ecuaciones correspondientes y junto con la condición de ortonormalidad $\sum_{i=1}^{p} l_{ii}^2 = 1$, se resuelven los p sistemas de ecuaciones, determinando los vectores característicos, que en otras palabras son los coeficientes de las combinaciones lineales que darán las nuevas variables y's con las propiedades deseadas.

En base a las condiciones de ortonormalidad y debido a que la matriz $\boldsymbol{C}$ es simétrica y semipositiva definida, se implica la existencia y unicidad de las soluciones, salvo en casos especiales, por ejemplo donde $\boldsymbol{C} = \boldsymbol{I}$, es decir:

$$\text{cov}(x_i, x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

con lo cual se tiene la ecuación característica

$$|\boldsymbol{I} - \lambda \boldsymbol{I}| = 0$$

lo cual implica que $\lambda_1 = \lambda_2 = \cdots = \lambda_p = 1$; en este caso se tiene un número infinito de soluciones, o bien no es necesario transformar las variables ya que tienen correlación cero.

Las combinaciones lineales obtenidas de esta manera, denotadas por $y_1, y_2, \ldots, y_p$ se denominan *las componentes principales* y los valores característicos corresponden a sus varianzas.

Es costumbre ordenar a estas variables de acuerdo a su tamaño, la primera corresponde a la que tiene mayor varianza y la última a la que tiene menor varianza. Usualmente se les da el nombre de *primera, segunda,..., p-ésima componente principal.* En símbolos se tiene

$$var(y_i) = \lambda_i; \quad \text{con} \quad \lambda_1 > \lambda_2 > \cdots > \lambda_p$$

Por lo tanto, se puede decir que del total de la variabilidad de los datos, la componente principal *i* explica un porcentaje equivalente al valor de la expresión:

$$\frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i}$$

Cuando las dos primeros componentes explican una gran parte de las varianza total, la representación gráfica de estas variables será una gráfica resumen para explorar la estructura del conjunto de observaciones. En caso de que las dos componentes no expliquen la mayor parte de la varianza las gráficas de pares de componentes ayudan a explicar el comportamiento de los datos.

Un aspecto importante a resaltar, es que las componentes principales dependen de la escala de medición de cada una de las variables. Es decir, si se tienen variables que se miden en diferentes unidades, las primeros componentes reflejarán aquellas variables, en donde las unidades de medida sean las más grandes.

Como ya se vio en la sección 1.6, la forma de evitar este problema, es el escalamiento de las variables, por lo que en lugar de trabajar con los datos originales se trabaja con los datos estandarizados. La estandarización más comúnmente usada para cada variable consiste en restarle su media y dividirla por su desviación estándar. Esto trae como consecuencia que la matriz de covarianza sea sustituida por la matriz de correlación de las x'_n y que la interpretación de los resultados se complique un poco. La mayoría de los paquetes estadísticos, tienen la opción de poder utilizar cualquiera de las dos matrices.

Para finalizar esta sección se hacen algunos comentarios que remarcan algunos aspectos, respecto a la técnica de Análisis de Componentes Principales:

1. Dado que la matriz C corresponde a una matriz de covarianza, es positiva definida y simétrica, implica que, teóricamente, sus valores característicos deberán ser positivos y no complejos. Sin embargo, en la práctica cuando esta matriz se estima de manera diferente a la de la suma de cuadrados medios, o existen algunos errores numéricos, o bien casos faltantes, se pueden obtener valores característicos negativos o complejos. Una posible solución a la

problemática de los casos faltantes consiste en sustituirlos por la media o la moda de los datos que se tienen con respecto a la variable de la que se trate. Otras solución más drástica consiste en eliminar el caso completo del análisis. La decisión sobre la posible solución deberá ser discutido con el investigador del área de donde provienen los datos.

2. En el caso en que dos valores característicos o más sean iguales la solución no es única.

3. La suma de los cuadrados de las distancias de todos los puntos a su punto medio es proporcional a la suma de los valores característicos, que a su vez es igual a la suma de las varianzas originales, la varianza total S, y se dice comúnmente que la componente *i-ésima* explica una proporción $\frac{\lambda_i}{S}$ del total de la variabilidad de los datos.

4. Un aspecto importante en el uso de esta técnica es que no es invariante respecto a la escala de medición. Diferentes escalas de medición en las variables originales producirán diferentes transformaciones o componentes principales. Lo más común es trabajar con las variables originales estandarizadas a tener media cero y varianza unitaria, por lo que se trabaja con la matriz de correlación en lugar de la matriz de covarianza.

2.4 Descripción desde el punto de vista del Calculo Diferencial

Como vimos anteriormente se parte de un conjunto de variables $x_1, x_2,...,x_p$ que determinan las medidas de un sujeto y se desean encontrar transformaciones $y_1, ..., y_p$ de tal manera que se plantea el sistema:

$$\begin{aligned} y_1 &= l_{11}x_1 + l_{12}x_2 + \cdots + l_{1p}x_p \\ y_2 &= l_{21}x_1 + l_{22}x_2 + \cdots + l_{2p}x_p \\ &\vdots \\ y_p &= l_{p1}x_1 + l_{p2}x_2 + \cdots + l_{pp}x_p \end{aligned}$$

con la restricción de ortonormalidad para los coeficientes *l's*:

$$\sum_{i=1}^{p} l_{ij} l_{ik} = \begin{cases} 0, j \neq k \\ 1, j = k \end{cases}$$

Considerando a los vectores

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_p)' \text{ y } \boldsymbol{l_i} = (l_{i1}, l_{i2}, \ldots, l_{ip})'$$

se puede escribir la *i-ésima* componente principal como:

$$y_i = \boldsymbol{l_i}' \boldsymbol{x}$$

Sin perdida de generalidad podemos suponer que en la componente principal y_1 se maximiza la varianza. Es decir:

$$\text{var}(y_1) = \text{var}(\boldsymbol{l_1}' \boldsymbol{x}) = \boldsymbol{l_1}' \boldsymbol{C} \boldsymbol{l_1},$$

donde $\boldsymbol{C}$, como siempre, es la matriz de varianza-covarianza muestral de $\boldsymbol{X}$. La restricción de ortonormalidad de las *l's* se puede escribir en notación vectorial como:

$$\boldsymbol{l_1}' \boldsymbol{l_1} = 1$$

Usando la técnica de multiplicadores de Lagrange, queda la expresión por maximizar:

$$\boldsymbol{l_1}' \boldsymbol{C} \boldsymbol{l_1} - \lambda(\boldsymbol{l_1}' \boldsymbol{l_1} - 1)$$

con λ un multiplicador de Lagrange. Derivando con respecto a $\boldsymbol{l_1}$ e igualando a cero se tiene

$$2\boldsymbol{C}\boldsymbol{l_1} - 2\lambda \boldsymbol{l_1} = 0 \Rightarrow \boldsymbol{C}\boldsymbol{l_1} - \lambda \boldsymbol{l_1} = 0 \Rightarrow \boldsymbol{C}\boldsymbol{l_1} = \lambda \boldsymbol{l_1}$$

Observando la última expresión y de acuerdo a la teoría del álgebra lineal se sabe que λ es un valor característico de la matriz $\boldsymbol{C}$ y que $\boldsymbol{l_1}$ es su correspondiente vector característico. Recordando que la expresión $\boldsymbol{l_1}' \boldsymbol{C} \boldsymbol{l_1}$ es la varianza de la variable y_1, y realizando la sustitución de $\boldsymbol{C}\boldsymbol{l_1} = \lambda \boldsymbol{l_1}$ se tiene:

$$\boldsymbol{l_1}' \boldsymbol{C} \boldsymbol{l_1} = \boldsymbol{l'}_1 \lambda \boldsymbol{l_1} = \lambda \boldsymbol{l'}_1 \boldsymbol{l_1} = \lambda$$

es decir, λ es el valor de la varianza máxima. Entonces la cantidad que maximiza a $l_1'Cl_1$ es λ_1 quien es el valor más grande de los valores característicos.

En general la *k-ésima* Componente Principal de x es y_k y cumple que $\text{var}(y_k) = \lambda_k$. Se realizará la prueba de esta afirmación para el caso $k = 2$, para $k \geq 3$, aunque es ligeramente más complicada se puede probar de forma muy similar.

Para el segundo componente principal $y_2 = l_2'x$ su varianza $l_2'Cl_2$ se debe maximizar con la restricción de normalidad $l_2'l_2=1$ y con una segunda restricción que es la no correlación entre las dos componentes principales, es decir

$$\text{cov}(y_1, y_2) = 0$$

lo cual implica que:

$$\text{cov}(l_1'x, l_2'x) = l_1'Cl_2 = l_2'Cl_1 = l_2'\lambda_1 l_1 = \lambda_1 l_2'l_1 = \lambda_1 l_1'l_2 = 0$$

en consecuencia se tienen las siguientes ecuaciones que pueden servir como restricción:

$$l_1'Cl_2 = 0 \qquad l_2'Cl_1 = 0$$
$$l_1'l_2 = 0 \qquad l_2'l_1 = 0$$

escogiendo arbitrariamente alguna de ellas, la última por ejemplo, se tiene que la cantidad a maximizar es:

$$l_1'Cl_2 - \lambda(l_2'l_2 - 1) - \mu l_2'l_1$$

donde λ y μ son multiplicadores de Lagrange. Derivando con respecto a l_2 e igualando a cero, queda:

$$Cl_2 - \lambda l_2 - \mu l_1 = 0$$

multiplicando esta expresión por la izquierda por el vector l_1' se tiene:

$$l_1'Cl_2 - \lambda l_1'l_2 - \mu l_1'l_1 = 0$$

como los dos primeros términos son cero y $l_1'l_1 = 1$ implica que μ=0, por lo cual se tiene la expresión

$$Cl_2 - \lambda l_2 = 0 \Rightarrow Cl_2 = \lambda l_2$$

Es decir, se tiene otro valor característico para C, y como debe ser el más grande, tenemos que $\lambda=\lambda_2$ (no puede ser igual que λ_1, porque implicaría que $l_2 = l_1$ y $l_2'l_1 = 0$ lo cual contradice la condición de normalidad) con su correspondiente vector característico l_2.

Repitiendo este proceso hasta la *p-ésima* componente principal se encontrarán los vectores característicos l_3, l_4,..., l_p de C, correspondientes a los valores característicos λ_1, λ_2, ..., λ_p ordenados en forma decreciente.

Como se ve, se ha llegado a los mismos resultados, los valores característicos y los vectores característicos de la matriz C.

2.5 Inferencia estadística sobre los Componentes Principales

Al aplicar la técnica de Análisis de Componentes Principales, a menudo surgen dos problemas que describimos a continuación.

a) Saber escoger cuáles o cuántos de las componentes principales, que se obtuvieron aportan información y cuales no; es decir, donde esta la división de las componentes que se van a utilizar para la interpretación y las que se van a desechar.

b) Si algunas variables originales están altamente correlacionadas su aportación es redundante, por lo que sería conveniente desecharlas y realizar un análisis adicional de componentes principales en base a un subconjunto de las variables originales.

El primer problema ha sido ampliamente discutido por varios autores. Particularmente Krzanowski y Marriott (1994a pags. 82-84) presentan un resumen de algunas opciones de criterios que se han planteado en esta discusión, las cuales son las siguientes:

i) Si la proporción de varianza explicada por las primeras dos o tres componentes es de 90% o más se pueden desechar las demás componentes. La decisión es subjetiva y puede ser fácilmente engañosa: una componente puede explicar mucha de la varianza pero mucha de la información interesante está contenida en otras componentes.

ii) Si algunas componentes tienen varianza por debajo de un cierto nivel hay que rechazarlos. Esta aseveración no tiene lógica alguna, por ejemplo, consideremos que las componentes principales se calcularon utilizando variables estandarizadas, si una variable es "casi" independiente del resto aparecerá como una componente con varianza ligeramente menor que 1, pero no hay razón para suponer que no aporte información.

iii) Realizar una gráfica de i contra λ_i con los valores característicos ordenados en forma decreciente. La gráfica caerá hacia cero severamente para las primeros componentes principales y después más lentamente. Sin embargo puede ser que estén influyendo otros errores aleatorios en las variables. Esta gráfica suele llamarse, sobre todo en la literatura en psicología, *"Scree Plot"*.

iv) Realizar alguna prueba estadística suponiendo distribución Gaussiana en los datos originales.

Se presentan a continuación dos de las pruebas estadísticas más conocidas que suponen verdadero este supuesto.

A) Mardia et al (1979 pags. 233-234) describen un estadístico para probar la hipótesis de que la proporción de la variación explicada por las primeras k componentes es igual que un cierto valor W. La hipótesis nula se plantea como:

$$\text{Ho: } \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} = W \quad \text{con } k < p$$

El estimador muestral de W, $\hat{W}$, tiene una distribución Gaussiana con media W y variauza

$$\text{var}(\hat{W}) = \frac{2traza(\Sigma)}{(n-1)\left[traza(\Sigma)\right]^2}(W^2 - 2aW + a^2)$$

donde Σ es la matriz de varianza-covarianza poblacional y

$$a = \frac{\sum_{i=1}^{k}\lambda_i^2}{\sum_{i=1}^{p}\lambda_i^2} \text{ con } k < p$$

Se puede estimar var($\hat{W}$), utilizando la matriz de varianzas-covarianzas muestral C con las n observaciones y los valores característicos de esta matriz, es decir:

$$\hat{\Sigma} = C \quad \text{y} \quad traza(\hat{\Sigma}) = \sum_{i=1}^{p} l_i$$

Donde las l_i son los valores característicos muéstrales. Por lo tanto tenemos

$$\hat{\text{var}}(\hat{W}) = \frac{2traza(C)}{(n-1)\left[traza(C)\right]^2}(W^2 - 2aW + a^2)$$

En consecuencia, el estadístico

$$z = \frac{\hat{W} - W}{\sqrt{\hat{\text{var}}(\hat{W})}} \sim N(0,1)$$

puede utilizarse para construir, si se desea, intervalos de confianza para W, como sigue:

$$\hat{W} - Z_{(1-\alpha/2)}\sqrt{\hat{\text{var}}(\hat{W})} \le W \le \hat{W} + Z_{(1-\alpha/2)}\sqrt{\hat{\text{var}}(\hat{W})}$$

donde Z denota el valor de la abscisa en una distribución Gaussiana, con una probabilidad de 1-α/2.

B) Se quiere probar que los últimos *p-k* valores característicos son iguales, es decir las últimas *p-k* componentes principales tienen la misma varianza. Barttlet propuso una prueba, que en su honor lleva su nombre, para homogeneidad de varianzas, también conocida como de esferecidad debido a que implica que en las últimas *p-k* dimensiones los datos están dispersos en una hiperesfera y por lo tanto, el incluir una de las componentes en el análisis debería implicar la inclusión de todas las restantes. La hipótesis estadística quedaría como:

$$\text{Ho: } \lambda_p = \lambda_{p-1} = \cdots = \lambda_{k+1}$$

El estadístico para esta prueba de hipótesis se obtiene por el método de la razón de verosimilitud, considerando que la expresión

$$-2\log(L) = np[a\text{-}1\text{-}\log(g)]$$

se distribuye aproximadamente como una χ^2 y donde:

L es la razón de verosimilitud suponiendo distribución Gaussiana multivariada

n es el tamaño de la muestra

p el número total de variables observadas

a y *g* son la media aritmética y geométrica de los valores característicos respectivamente

Considerando únicamente los últimos *p-k* valores característicos se tiene

$$a_0 = \frac{\sum_{i=k+1}^{p} \lambda_i}{p-k} \quad \text{y} \quad g_0 = \left(\prod_{i=k+1}^{p} \lambda_i \right)^{\frac{1}{p-k}}$$

que son las expresiones para la media aritmética y la media geométrica de los últimos *p-k* valores característicos respectivamente. Por lo tanto el estadístico de prueba queda como:

$$-2\log(L) = np[a_0\text{-}1\text{-}\log(g_0)]$$

La propuesta de Bartlett, citado por Mardia (1979 pag. 236) es una aproximación que tiene la forma:

$$\left(n-\frac{2p+11}{6}\right)(p-k)\log\left(\frac{a_o}{g_o}\right) \sim \chi^2_{(p-k+2)(p-k-1)/2}$$

La distribución de esta estadística es asintótica y muchos paquetes estadísticos la reportan cuando se utiliza la rutina de componentes principales.

Por otra parte Rao (1964) describe, la manera como el análisis de componentes principales puede ser utilizado para probar diferencias entre los valores medios de diferentes grupos de individuos a través del análisis de dispersión, el cual es una generalización del análisis de varianza.

2.6 Componentes Principales y Análisis de Regresión

Consideremos un problema de regresión múltiple, en el cual algunas variables independientes presentan el problema de multicolinealidad, es decir existe dependencia lineal entre ellas. Este hecho trae como consecuencia que teóricamente los coeficientes de regresión no puedan ser calculados ya que implicaría encontrar la inversa de una matriz singular. Sin embargo debido a errores de redondeo, muchas computadoras si los calculan y si no se tiene conciencia de este problema se llega a conclusiones erróneas.

Un método que ha demostrado ser eficiente para evitar el problema de multicolinealidad, es el de sustituir las variables originales por las primeros componentes principales, las cuales por construcción están no correlacionados entre sí. Además puede ser que la explicación de la variable dependiente sea más fácil de interpretar ya que el número de variables independientes es mucho menor, lo que evita posible redundancia en la información. Jolliffe (1986 pags. 129-147) presenta una amplia discusión al respecto, demostrando además que los estimadores del modelo de regresión usando las componentes principales no son

insesgados, a diferencia de los estimadores cuando se tiene el modelo de regresión con las variables originales. También muestra que al incluir en el modelo los últimos componentes principales aumenta la varianza de los estimadores, pero, si hay una alta correlación de estos con la variable dependiente se disminuye el sesgo.

Por lo tanto, comparando el método de mínimos cuadrados con el de las componentes principales para la regresión se concluye que en problemas de multicolinealidad el segundo método es preferible, tanto para estimar los parámetros, como para seleccionar variables ya que por construcción se trabaja con variables no correlacionadas.

2.7 Análisis de Componentes Principales para variables discretas

En la construcción de las componentes principales no se particulariza explícitamente en el tipo de variables que se están manejando. Sin embargo, en la mayoría de las investigaciones en que existe la necesidad de utilizar la técnica del Análisis de Componentes Principales, las variables empleadas son de diferente tipo: continuas, categóricas y/o dicotómicas. En un caso extremo, si se trabaja con variables continuas, se puede suponer que las variables tienen una distribución Gaussiana multivariada, y aplicar por lo tanto la parte de inferencia vista en la sección 2.5. En el caso de únicamente variables dicotómicas, aunque su interpretación es más complicada, la robusticidad del método permite lograr su objetivo: el de encontrar un número pequeño de variables que expliquen el comportamiento de las variables originales. Alternativas similares al Análisis de Componentes Principales para manejar este tipo de variables son descritas por Gower (1966 pags. 325-338) y Cox (1972 pags. 113-120).

Por otro lado, si el estudio utiliza únicamente variables ordinales y dicotómicas el método de componentes principales no podrá ser utilizado para obtener resultados confiables. Una alternativa a esta situación es el llevar las variables ordinales a variables dicotómicas. Para ilustrar, supóngase que se tienen dos variables x_1 y x_2, de manera que la primera es dicotómica (toma solo valores 0 ó 1) y que la segunda es ordinal (toma los valores 1,2,3). La variable x_2 se puede reparametrizar como:

$$v_1 = \begin{cases} 1 & si \quad x_2 = 1 \\ 0 & si \quad x_2 \neq 1 \end{cases} \qquad v_2 = \begin{cases} 1 & si \quad x_2 = 2 \\ 0 & si \quad x_2 \neq 2 \end{cases}$$

de manera que el análisis de componentes principales será realizado con las variables dicotómicas x_1, v_1 y v_2.

En resumen se puede decir que la técnica de componentes principales puede ser utilizada cuando: todas la variables son continuas; la mayoría de las variables son continuas y algunas son ordinales ó todas las variables son dicotómicas. En el caso de que se tengan variables dicotómicas con ordinales o solo dicotómicas, será necesario reparametrizar las variables ordinales a fin de manejar únicamente variables dicotómicas.

2.8 Comentarios

-La técnica del Análisis de Componentes Principales es muy útil como técnica exploratoria y ayuda a resumir y describir gráficamente el comportamiento de un conjunto de observaciones, que están en una dimensión mayor a tres, incluyendo las posibles observaciones aberrantes o atípicas si es que existen; esto permitirá, una vez que se identifiquen, estudiar las posibles causas de su comportamiento "raro".

-Aunque es una técnica antigua, en los últimos años, con el desarrollo de las computadoras, su aprovechamiento ha sido canalizado en gran medida en diferentes campos de investigación como la biología, la antropología, la economía, etc.

-La escala en la que se miden las variables juega un papel importante, ya que esta técnica no es invariante bajo transformaciones, solo en el caso de que todas las variables se midan en la misma escala será similar utilizar la matriz de covarianza que la de correlación. En caso contrario, se recomienda utilizar la matriz de correlación.

-La interpretación de los resultados dependerá del problema específico del que se trate, tal como se ilustrará con los ejemplos del capítulo 4. En el apéndice II se dan las instrucciones a seguir para la utilización del paquete estadístico SPSS con la técnica de Análisis de Componentes Principales.

CAPÍTULO 3. PROYECCIONES PERSEGUIDAS

En este capítulo se presenta la idea de proyecciones perseguidas, como una técnica para explorar datos multivariados; el objetivo del capítulo es describir los diferentes índices o funciones de proyección propuestos, y desarrollados por diferentes autores.

3.1 Introducción

"*Projection pursuit*", términos que pueden traducirse como proyecciones perseguidas, es el nombre asignado por Friedman y Tukey (1974) a las proyecciones que sirven para mostrar en una, dos y en algunos casos tres dimensiones, a través de proyecciones lineales, un conjunto de datos que originalmente se encuentra en dimensión p. El procedimiento fue sugerido originalmente por Kruskal (1972) y puede ser resumido mediante el siguiente esquema:

1.- Elegir la dimensión $k < p$, en donde se realizará la proyección; usualmente se toma k=2.

2.- Elegir un criterio, que generalmente está asociado a una función objetivo ó índice, el cual se va a optimizar. Se escoge dependiendo de los objetivos que queremos que cumpla. Este paso es el más importante y más adelante analizaremos algunos de los criterios más conocidos.

3.- Evaluar el criterio. La importancia de este paso va en relación al cuarto, debido a que en esta evaluación se buscan los posibles puntos críticos de la función objetivo.

4.- Comenzar con la mejor, o una selección de la mejor, proyección encontrada en el paso anterior y usar un programa de computo óptimo para encontrar el máximo o mínimo local. Tradicionalmente se utilizan métodos del análisis numérico implantados en alguna rutina escrita en algún lenguaje de programación, usualmente Fortran.

5.- Presentar gráficamente la proyección seleccionada e interpretarla.

Considerando estos criterios y por lo expuesto en el capítulo 2, los componentes principales pueden ser considerados como un caso particular de proyecciones perseguidas, donde la

función objetivo se define por la proporción de varianza que explica cada uno de los componentes. En este caso se maximiza la función objetivo.

La idea de proyecciones perseguidas desde el punto de vista geométrico es la siguiente: se define una función de interés, también llamado índice, ya sea en $\Re$, $\Re^2$ o bien en $\Re^3$, que esté en relación a un criterio específico. Se pretende encontrar la proyección lineal que optimiza esta función o índice. La optimización deberá tener por objetivo, encontrar proyecciones que muestren las posibles estructuras en los datos, si es que existen.

Para ilustrar esta situación consideremos cuatro conjuntos de datos con distribución Gaussiana esférica, centrados en los vértices de un tetraedro regular (figura 1). En este caso existen tres posibles proyecciones que muestran la estructura agrupada.

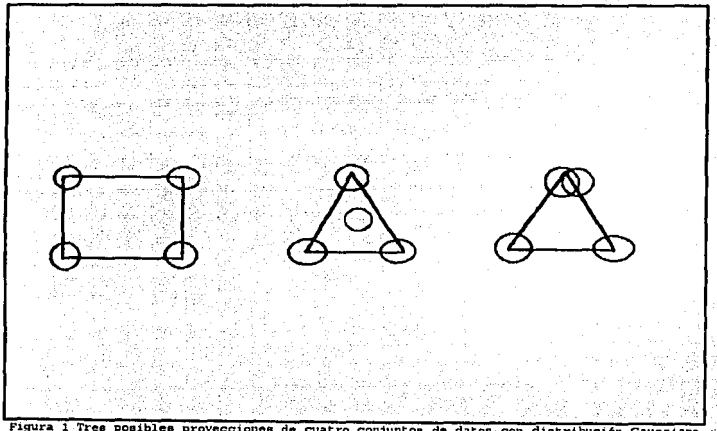


Figura 1 Tres posibles proyecciones de cuatro conjuntos de datos con distribución Gaussiana esférica centradas en los vértices de un tetraedo regular.

Se puede observar en esta figura que la primera y la segunda proyección presentan una mejor idea de la estructura de los datos, mientras que la tercera sobrepone dos grupos de datos, por lo que no queda muy claro que exista separación de los grupos. Por lo tanto, si el

criterio es buscar estructuras de agrupamiento, las proyecciones buscadas serán la primera ó la segunda.

De acuerdo a lo expuesto hasta el momento, podemos decir que el paso importante para aplicar la técnica de proyecciones perseguidas es proponer un índice o función de proyección. Como se mencionó el primero en publicar trabajos en relación a la búsqueda de índices que permitan distinguir, en una dimensión acorde (una o dos), conglomerados de puntos que originalmente se encuentran en una dimensión alta fue Kruskal (1972). Él definió un *índice de condensación*, basado en los coeficientes de la distancia entre puntos. Sin embargo su algoritmo no se llevó a la práctica. Pocos años después Friedman y Tukey (1974), quienes utilizaron por primera vez el término *"Projection Pursuit"* para definir este tipo de proyecciones, desarrollaron un algoritmo para detectar características sobresalientes de los datos. Aplicaron este algoritmo para la separación de una mezcla de 15 distribuciones esféricas bivariadas. La separación fue efectuada al aplicar el algoritmo aislando uno o dos grupos iterativamente. Tiempo después, Huber (1985) examinó diferentes índices teóricos, planteando muchos campos de aplicación en los cuales estos índices pueden ser desarrollados. Casi simultáneamente se publicaron trabajos de Jones y Sibson (1987) en donde se proponía un criterio basado en la minimización de la entropía de Shannon. Otro índice propuesto por estos investigadores esta basado en el tercero y cuarto momentos de los datos proyectados. Por otra parte Jee (1985) comparó cuatro índices para Proyecciones Perseguidas y en particular estudió uno basado también en la entropía de Shannon pero considerando estimaciones de las funciones de densidad por medio de histogramas. Friedman (1987) retoma sus investigaciones sobre proyecciones perseguidas, publicando un artículo donde plantea un criterio basado, en una medida de no-normalidad de los datos proyectados; su idea consistió en darle un mayor peso a las diferencias de la parte central de las distribuciones que a las colas, estimando las funciones de densidad de los puntos proyectados a través de desarrollos en términos de polinomios de Legendre. Por su parte Yenyukov (1988), propone un índice basado en la entropía de orden β, $\beta>0$, con un desarrollo basado en las separaciones grandes de las estadísticas de orden. Por otro lado Eslava y Marriott (1994) proponen dos índices basados en las coordenadas polares de los

datos proyectados en un plano, el "*Polar Nearest Neighbour*" (*PNN*), que puede ser traducido como los puntos vecinos en coordenadas polares más cercanos y el de la distancia radial media ($\overline{R}$), los cuales tienen el objetivo de detectar y visualizar estructuras agrupadas de datos. Por otra parte, Hall (1989) retoma la propuesta de Friedman argumentando teóricamente cierta inestabilidad en el índice propuesto por lo que propone una variante de él estimando las funciones de densidad a partir de polinomios de Hermite. En el mismo sentido son lo trabajos de Cook, Buja y Cabrera (1993) quienes publicaron un artículo donde analizan los índice basados en polinomios ortogonales propuestos por Friedman y Hall, y proponen una variante del índice de Hall basado también en el desarrollo de polinomios de Hermite para estimar las funciones de densidad. Por otra parte estos autores desarrollaron un programa en lenguaje C, llamado XGobi, en el que se han programado diez índices de proyecciones perseguidas entre las que se encuentran: el de Friedman y Tukey, el de la entropia, el de Friedman (llamado Legendre) el de Hall (llamado Hermite) y el que proponen ellos mismos (llamado Natural Hermite), que entre otros índices más presentamos en este trabajo de tesis.

Recientemente Nason (1995) ha desarrollado un índice para proyecciones en tres dimensiones, generalizando el índice de momentos propuesto por Jones y Sibson (1987).

En otro orden de ideas y haciendo la similitud con Componentes Principales, antes de aplicar algún índice de proyecciones perseguidas debe considerase la escala en la que originalmente se miden las variables. Desde el índice propuesto por Friedman y Tukey se nota esta preocupación, lo cual motivó la introducción de un término de dispersión en el índice con el objetivo de contrarestar el efecto de la escala. En la actualidad, lo común es considerar primero el proceso de esferado de las variables, descrito en la sección 1.9, cuyo objetivo es el de centrar y distribuir los puntos en una hiper-esfera, a fin de tener varianza unitaria en cualquier dirección, antes de la aplicación de algún índice de proyección perseguida.

En las siguientes secciones se presentan algunos de los índices más conocidos actualmente.

3.2 Kruskal

Kruskal fue el primero en proponer transformaciones lineales de un conjunto de datos a fin de conocer sus estructuras. El propuso un *índice de condensación* el cual esta basado en coeficientes de variación del valor de la distancia entre los puntos. Este índice es estandarizado por el coeficiente de variación de la distancia entre puntos generados por una distribución Gaussiana. El índice dio resultados insatisfactorios al aplicarse a datos artificiales, a pesar de haber estandarizado las variables por el coeficiente de variación. Sin embargo sus trabajos sugirieron una mayor investigación al respecto, lo que hizo que se propusiera la idea de esferar los datos para usar este índice.

3.3 Friedman y Tukey

El enfoque para proyecciones perseguidas dado por estos investigadores consiste en encontrar transformaciones lineales que den a conocer aspectos sobresalientes de conjuntos de datos multidimensionales. El algoritmo propuesto está basado en la distancia entre puntos y la varianza de los datos proyectados. Este algoritmo se desarrolló para el caso de proyecciones en una y dos dimensiones sin limitaciones teóricas sobre la dimensionalidad de los espacios de proyección; las limitaciones son prácticas ya que los procedimientos de optimización del índice son difíciles de calcular para el caso de más de dos dimensiones. En el caso de una dimensión, proyectando en una línea de dirección k, el índice se define como:

$$I(k) = s(k)d(k)$$

donde $s(k)$ es la desviación estándar de los datos proyectados en la línea de dirección k y $d(k)$ es la "densidad local" después de proyectarlos sobre k. Así $s(k)$ está dada por la expresión:

$$s(k) = \left[\sum_{i=qn}^{(1-q)n} (z_i(k) = \bar{z}(k))^2 / (1-2q)n \right]^{1/2}$$

donde

$$\bar{z}(k) = \sum_{i=qn}^{(1-q)n} z_i(k) / (1-2q)n$$

con n = total de puntos

$z_i(k)$ = proyección del *i-ésimo* punto x_i de la dimensión original p en la línea de dirección k

q = fracción de puntos (cercano a cero) en los extremos de la línea que son omitidos del cálculo de la desviación estándar.

Por otra parte

$$d(k) = \sum_{i=1}^{n}\sum_{j=1}^{n} f(r_{ij}) L(R - r_{ij})$$

con $r_{ij} = \left| z_i(k) - z_j(k) \right|$ y $L(x) = \begin{cases} x & para \quad x > 0 \\ 0 & para \quad x \le 0 \end{cases}$

$f(r)$ es una función monótona decreciente en el intervalo $(0,R)$; R es elegido de tal forma que sea el promedio de puntos contenidos en la ventana definida por la función $L(x)$, es decir, es una fracción de n, incrementándose mas lento que el mismo n.

Para el caso de dos dimensiones, el índice es denotado por $I(k,l)$, donde k y l representan la dirección del plano de proyección; el índice se define como:

$$I(k,l) = s(k)s(l)d(k,l)$$

con s y d definidos igual que cuando se proyecta en una dimensión, pero con

$$r_{ij} = \left\{ \left[z_i(k) - z_j(k) \right]^2 + \left[z_i(l) - z_j(l) \right]^2 \right\}^{1/2}$$

Friedman y Tukey aplicaron este algoritmo a dos conjunto de datos artificiales de mezclas de distribuciones Gaussianas centrados en los vértices de simplejos regulares y a dos conjuntos

de datos reales. Particularmente evaluaron el comportamiento del algoritmo en una mezcla de 15 distribuciones con media en los vértices de un simplejo en dimensión 14; todas las distribuciones tenían varianza unitaria. El algoritmo se desarrolla iterativamente y observaron que solo se separaban uno o dos grupos de datos en cada paso. Esto sirvió para que los autores notaran la problemática de llevar este algoritmo en la práctica. Originalmente los autores desarrollaron un programa Fortran para su índice y en la actualidad también se encuentra desarrollado en el programa XGobi.

3.4 P. J. Huber

El enfoque dado por Huber a las proyecciones perseguidas, es el de una metodología para encontrar proyecciones interesantes de datos originales en una dimensión alta, p. Parte de algunos argumentos heurísticos en los que toma como base el hecho de que una proyección es menos interesante si la distribución de sus datos es parecida a la Gaussiana; para fortalecer esta afirmación toma como validos los siguientes argumentos:

i) una distribución multivariada es Gaussiana si todas sus proyecciones en dimensiones menores son Gaussianas.

ii) si la proyección menos Gaussiana está relativamente cerca de una Gaussiana, no es útil observar otras proyecciones.

iii) para muchas nubes de puntos en una dimensión grande las proyecciones en una dimensión menor tienen distribución aproximadamente Gaussiana.

En base a lo anterior, el problema consiste en buscar una medida de la distancia entre la distribución original de los datos proyectados y la distribución Gaussiana. Con estas ideas en mente Huber propone los siguientes índices:

a) El primer índice llamado "cumulantes estandarizados absolutos", definido como

$$Q = \frac{|k_m(z)|}{(k_2(z))^{m/2}}, \text{ con } m>2,$$

donde z es la proyección de los datos originales X en dimensión p y k_m es el cumulante de orden m, dada por la m-ésima derivada del logaritmo natural de la función característica definida por:

$$k_m = \left(\frac{d}{idt}\right)^m \ln\left[E(e^{itz})\right]$$

Si $m = 3$, Q es el valor absoluto de la asimetría y si $m = 4$, es el valor absoluto de la kurtosis.

b) El siguiente índice es llamado la información estandarizada de Fisher, el cual es definido como:

$$Q = \sigma_z^2 \int \left(\frac{f'(z)}{f(z)}\right)^2 f(z)dz - 1$$

donde σ_z^2 es la varianza de los datos proyectados z y $f(z)$ es su función de densidad

c) el último índice corresponde a la estandarización negativa de la entropía de Shannon definido como:

$$Q = \int \ln[f(z)] f(z)dz + \ln\left[\sigma_z \sqrt{2\pi e}\right]$$

basado en que si $g(x)$ es la distribución Gaussiana estándar, entonces se cumple:

$$\int \ln[g(x)] g(x)dx = -\ln\left[\sigma_x \sqrt{2\pi e}\right]$$

Los tres índices toman su valor mínimo de cero cuando los datos tienen distribución Gaussiana.

Huber presenta la discusión teórica para justificar estos índices, pero no realiza ningún tipo de desarrollo computacional. Por otra parte en su documento presenta algunas extensiones de las proyecciones perseguidas tales como proyecciones perseguidas y regresión. proyecciones perseguidas y estimación de densidades, proyecciones perseguidas y series de tiempo entre otras.

3.5 M.C. Jones y R. Sibson

Estos autores enfocaron su trabajo de proyecciones perseguidas como la búsqueda de proyecciones lineales, vistas como funciones de distribución que difieran lo más posible de la distribución Gaussiana. En consecuencia, su interés se concentra en la medida de la diferencia entre la función de distribución de los datos proyectados y la Gaussiana. Para medir esta diferencia analizan las limitaciones prácticas que tiene el índice basado en el negativo de la entropía de Shannon propuesto por Huber, presentado en esta tesis en la sección 3.4

Para la estimación de la densidad, *f(z)*, de los datos proyectados, utilizan el concepto de *"kernel density estimate"*, que podría traducirse como estimación de densidad del núcleo, la cual está dada por la expresión:

$$\hat{f}(z) = \frac{1}{nw}\sum_{i=1}^{n}\varphi\left(\frac{z-z_i}{w}\right)$$

donde φ es la función *kernel* , $w > 0$ es un parámetro suavizador o ancho de ventana, z_i (i=1,...n) son los datos proyectados. φ es una función no negativa y se aproxima a cero cuando su argumento tiende a más o menos infinito. Un caso típico de estas funciones es la densidad Gaussiana.

En base a esto la estimación del índice de entropía, en el caso de que la proyección sea en dimensión uno, se calculará por la expresión:

$$\int \hat{f}(z)\ln(\hat{f}(z))dz$$

la cual se calcula empleando alguna técnica de integración numérica:

Si empíricamente se conociera la función de distribución de los datos proyectados, supongamos que es $F_n(x)$, entonces el índice se calcula como:

$$\int \ln(\hat{f}(z))dF_n(z) = \frac{1}{n}\sum_{i=1}^{n}\ln(\hat{f}(z_i))$$

En casos prácticos lo más común es que la Gaussiana estándar sea utilizada como la función φ con ancho de ventana $w = n^{-0.2}$ (Silverman, 1986, pags. 44-48), el cual es el valor que minimiza el cuadrado medio del error y la varianza vale uno en cualquier dirección debido a que los datos son previamente esferados.

Ahora bien, si la proyección se realiza en un plano, el índice de entropía puede calcularse usando la estimación de una función de densidad bivariada pero su desarrollo no es tan fácil. Por lo tanto Jones y Sibson proponen una aproximación, a partir de un índice basado en el tercer y cuarto momentos centrales de los datos proyectados.

Con esta propuesta, el índice de momentos en el caso de la proyección en una dimensión es:

$$I_m = \frac{1}{12}\left[k_3^2 + \frac{1}{4}k_4^2\right]$$

donde $k_3 = \mu_3$, $k_4 = \mu_4$ y μ_i es el *i-ésimo* momento central de los datos proyectados, definido en la sección 1.3 de esta tesis, y k_i es el *i-ésimo* cumulante (ver sección 1.5)

En el caso de dos dimensiones, el índice de momentos, I_m, esta dado por la expresión:

$$I_m = \frac{1}{12}\left\{\left(k_{30}^2 + 3k_{21}^2 + 3k_{12}^2 + k_{03}^2\right) + \frac{1}{4}\left(k_{40}^2 + 4k_{31}^2 + 6k_{22}^2 + 4k_{13}^2 + k_{04}^2\right)\right\}$$

donde k_{rs} es el cumulante bivariado de orden (r,s) dado por $k_{rs}=\mu_{rs}$; para $r+s=3$ se tiene que $k_{30}=\mu_{30}$, $k_{03} = \mu_{03}$, $k_{21} = \mu_{21}$, $k_{12}=\mu_{12}$, $k_{22} = \mu_{22} - 1$ y μ_{rs} es el (r,s) momento bivariado central de los datos proyectados, los cuales se definieron en la sección 1.4 de este trabajo.

Los índices de momentos toman su valor mínimo cero para el caso en que los datos proyectados tengan distribución Gaussiana.

Por otra parte, estos autores dedican gran importancia al proceso de esferado de las variables, mencionando que deberá realizarse antes de aplicar cualquier índice de proyección, dando algunas alternativas de desarrollo, mismas que fueron presentadas en la sección 1.9 de esta tesis.

Los investigadores aplicaron los índices a un conjunto de datos reales, mostrando que la técnica de proyecciones perseguidas puede revelar más claramente la estructura de los datos que otras técnicas multivariadas, tales como componentes principales y análisis de conglomerados.

Una propuesta reciente del índice de momentos en tres dimensiones ha sido presentada por Nason (1995). El autor utiliza conjuntos de datos artificiales centrados en los vértices de un tetraedro para demostrar que este índice puede ser útil para mostrar conglomerados. Por otra parte aplica el índice para analizar imágenes multiespectrales.

3.6 R. J. Jee

El autor retoma el estudio de proyecciones perseguidas como la búsqueda de proyecciones interesantes. Para él una proyección es más interesante cuando más difiera de la distribución Gaussiana. Propone cuatro índices basadas en medidas de no-normalidad, retomando trabajos de algunos de los investigadores mencionados:

i) La información estandarizada de Fisher presentado en la sección 3.4-(b).

ii) El negativo de la entropía de Shannon estandarizada de forma que el índice evaluado para la función Gaussiana vale cero. Ver sección 3.2-(c)

iii) La diferencia L_1 entre dos funciones medida por:

$$L_1 = \int |f(z) - \phi(z)| dz$$

donde f es la función de densidad de la población de los datos proyectados z y ϕ es la función de densidad Gaussiana con media y varianza igual que la función f.

iv) La diferencia entre dos funciones de densidad, medida por la métrica de Hellinger, L_H, definida por:

$$L_H = \int \left[f^{1/2}(z) - \phi^{1/2}(z) \right]^2 dz$$

Estos índices alcanzan el valor mínimo cero solo en caso de que la distribución f sea Gaussiana. Al igual que en otros métodos de proyecciones perseguidas, los datos deberán ser esferados antes de aplicar el algoritmo.

Jee hace una comparación entre estas cuatro medidas en modelos poblacionales basados en mezclas de distribuciones Gaussianas tanto para proyectar datos de dos a una dimensión, así como para datos en tres dimensiones en el plano. A continuación se describen algunas características de cada uno de estos casos particulares analizados.

a) Proyecciones en la línea.

a1) Cuando el modelo poblacional $f(x)$, esta definido como una mezcla de tres distribuciones multivariadas Gaussianas esféricas, cada una con media diferente, la matriz de varianza de cada componente es $\frac{1}{4}I$, con I la matriz identidad. Es decir:

$$f(x) = \frac{1}{3}\sum_{i=1}^{3} \phi(x;\mu_i,\frac{1}{4}I)$$

Jee estudio el caso particular cuando las medias de cada componente esta dada por los vectores: (-2,-2), (2,2) y (2,-2); en este caso el índice basado en la información estandarizada de Fisher favorece la proyección en que las tres distribuciones son separadas. Los otros índices son maximizados por proyecciones en que dos distribuciones se agrupan en una y la tercer distribución es separada a partir de las otros dos (ver figura 2).

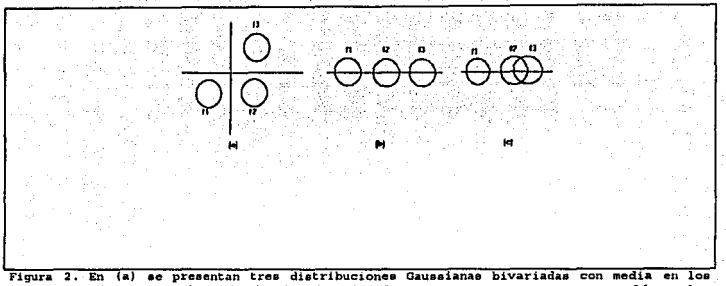


Figura 2. En (a) se presentan tres distribuciones Gaussianas bivariadas con media en los vectores en los puntos (-2,-2), (2,2) y (2,-2). Al proyectar estos puntos en una línea, los tres componentes se separan sólo en el caso del índice basado en la estandarización de la información de Fisher, (b). En (c) se utilizaron los otros índices, el de la entropía, el L_I y el L_{II}, los cuales encuman dos de las distribuciones.

a2) Se analiza el caso en que el modelo poblacional se compone de dos distribuciones Gaussianas esféricas y la mezcla nuevamente tiene proporciones iguales, caracterizadas porque una distribución tiene como media el vector (-2,0) y matriz de varianza la identidad; la otra tiene media el vector (3,0) y matriz de dispersión la matriz diagonal(2,2). La optimización lograda en estos casos es similar en los índices de Fisher, Shannon, y L_I, presentando en todos los casos un comportamiento unimodal de la distribución. Sólo el índice L_{II}, muestra el comportamiento bimodal de la población.

b) Proyección en el plano

El autor aplica los índices de información de Fisher y el de la entropía negativa de Shannon considerando a la población como una mezcla de cuatro distribuciones Gaussianas esféricas de igual proporción. Cada componente de la mezcla tiene matriz de varianza la identidad y son localizadas en los vértices de un simplejo con lados de longitud 6 unidades. Jee sigue dos caminos para optimizar los índices en el plano: el primero consiste en optimizar mediante el empleo de funciones de densidad bivariadas; el segundo consiste en primer lugar optimizar el índice en una dimensión, y luego optimizar en una segunda dirección caracterizada por ser ortogonal a la solución unidimensional. Aunque las dos formas de optimizar producen resultados numéricamente diferentes, su comportamiento es muy similar.

El índice de la información de Fisher da como resultado una proyección que muestra los cuatro grupos de datos. El índice de la entropía negativa de Shannon, nuevamente como en el caso univariado, muestra una distribución grande entre dos pequeñas.

En el caso teórico se utiliza el hecho de conocer la función de densidad de la población. En la práctica es necesario estimarla de alguna manera. Así como Jones y Sibson (1987) utilizan un estimador de densidad de núcleo (*kernel density*), en este caso Jee desarrolla un estimador basado en histogramas de acuerdo a las ideas de Terrel y Scott (Terrel y Scott 1985). Para la aplicación a conjuntos de datos reales, el ancho h de los intervalos de los histogramas se considera igual. Es decir, si se tiene un conjunto de puntos proyectados $\{z_1,...,z_n\}$ con función de densidad f, su estimador por medio de histogramas es:

$$\hat{f}(z) = \frac{\#\left\{z_j | t_{i-1} \le z_j \le t_i\right\}}{nh} \quad \text{para} \quad t_{i-1} \le z \le t_i$$

donde los t_i son los extremos de los intervalos

Entonces el estimador del índice de la entropía negativa de Shannon es:

$$\hat{I} = \int \hat{f}(z) \ln(\hat{f}(z)) dz$$

Por su parte, el índice estandarizado de la información de Fisher esta dado por la expresión:

$$\hat{I} = \frac{8}{h^2}\left[1 - \frac{1}{n}\sum_{i=1}^{n}\sqrt{(n_i + \frac{1}{4})(n_{i-1} + \frac{1}{4})}\right]$$

donde n_i es el número de puntos en el *i-ésimo* intervalo.

Estos estimadores son evaluados para diferentes tamaños de muestras con conjuntos de datos simulados, comparándolos con los resultados del índice teórico. Se observó en las gráficas que el índice de entropía aparentemente presenta una rápida convergencia.
Los índices se aplicaron a dos conjuntos de datos reales. En uno de ellos los índices de proyecciones perseguidas revelan estructuras en los datos que no son detectados por componentes principales. En el otro las proyecciones perseguidas no revelan más información que la mostrada por componentes principales.

3.7 J.H. Friedman

Continuando con sus investigaciones en torno a las proyecciones perseguidas, Friedman, retoma la misma idea de proyecciones perseguidas como la búsqueda de proyecciones interesantes, definidas como aquellas donde la función de densidad de los puntos proyectados difiere más de la función de densidad Gaussiana. Sin embargo, en el desarrollo del algoritmo, estas diferencias son más importantes si se dan en el centro de las distribuciones que en las colas. Al igual que en los criterios presentados anteriormente los datos deberán, previamente al proceso, ser esferados. Denotando como siempre los datos proyectados por la variable z, Friedman propone la transformación de estos datos por medio de la expresión:

$$T(z) = 2\Phi(z) - 1$$

donde Φ es la función de densidad Gaussiana estándarizada acumulada, definida como:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt$$

Con esta transformación se logra que T tome valores en el intervalo [-1,1] y como z tiene distribución Gaussiana, entonces T tiene una distribución uniforme. Por lo tanto, la no-

normalidad de los puntos proyectados z, implica no-uniformidad de T. La medida de no uniformidad tomada por Friedman es:

$$\int_{-1}^{1}\left[f_T(T)-\frac{1}{2}\right]^2 dT = \int_{-1}^{1} f_T^2(T)dT - \frac{1}{2}$$

donde $f_T(T)$ es la función de densidad de T, la cual explícitamente esta dada por la expresión:

$$f_T(T) = \frac{1}{2}\left\{\frac{f_\alpha\left[\Phi^{-1}\left(\frac{T+1}{2}\right)\right]}{\phi\left(\Phi^{-1}\left(\frac{T+1}{2}\right)\right)}\right\}$$

con f_α la función de densidad de z, ϕ es la función de densidad Gaussiana y Φ^{-1} es la inversa de la función de densidad Gaussiana acumulada.

Una estimación de esta función de densidad teórica se construye mediante la expansión de $f_T(T)$ por polinomios de Legendre, truncando la suma al orden J. Con base a esto la propuesta del índice es la siguiente:

$$I(\alpha) = \frac{1}{2}\sum_{j=1}^{J}(2j+1)E_T^2\left[P_j(T)\right],$$

donde los polinomios de Legendre están definidos por:

$$P_0(T) = 1$$
$$P_1(T) = T$$

Para $j > 2$ se define la expresión para $P_j(T)$ como:

$$P_j(T) = \frac{1}{j}\left\{(2j-1)TP_{j-1}(T) - (j-1)P_{j-2}(T)\right\}$$

El índice alcanza su valor mínimo cero, cuando T tiene una distribución uniforme, pero como se esta usando un índice aproximado, en realidad la distribución uniforme no es la única que minimiza el índice. Para el caso de una muestra, los valores esperados, $E_T^2\left[P_j(T)\right]$ son calculados por los correspondientes promedios muéstrales; en consecuencia el estimador del índice esta dado por:

$$\hat{I}(\alpha)=\frac{1}{2}\sum_{j=1}^{J}(2j+1)\left\{\frac{1}{n}\sum_{i=1}^{n}P_j(2\Phi(\alpha' x_i)-1)\right\}^2$$

donde $z_i = \alpha' x_i$ es la proyección del *i-ésimo* punto. El índice es maximizado con respecto a la q-componente de α sujeto a $\alpha'\alpha=1$.

Si la proyección se realiza en dos dimensiones, el índice es:

$$I(\alpha,\beta)=\frac{1}{4}\sum_{j=1}^{J}(2j+1)E^2\left[P_j(T_1)\right]+\frac{1}{4}\sum_{k=1}^{J}(2k+1)E^2\left[P_k(T_2)\right]+$$

$$\frac{1}{4}\sum_{j=1}^{J}\sum_{k=1}^{J-j}(2j+1)(2k+1)E^2\left[P_j(T_1)P_k(T_2)\right]$$

La versión muestral, al igual que en el caso univariado es obtenido por la sustitución de promedios muéstrales, para las esperanzas involucradas

Empíricamente J se selecciona en el intervalo [4,8], aunque no se debe de tomar como una guía infalible, debido a que mientras mayor sea el tamaño de muestra, más grande deberá ser el valor de J.

Friedman aplico este algoritmo a dos conjuntos de datos simulados y a tres de datos reales.

El índice propuesto por Friedman ha dado origen a trabajos posteriores de otros investigadores. En particular Hall (1989, pags. 589-605) presenta un índice teórico pues demuestra que el de Friedman presenta cierta inestabilidad en las colas de las distribuciones.

Por otra parte Cook, Buja y Cabrera (1993, pags. 225-250), proponen nuevos índices basados en la expansión de polinomios ortogonales. Este índice ha sido programado en el programa XGobi, con el nombre de índice de Legendre.

3.8 L S. Yenyukov

El autor interpreta las proyecciones perseguidas como una metodología para probar la presencia de estructuras agrupadas. Su punto de partida es la medida de la entropía de orden β (β>0). La estimación del índice se basa en las separaciones grandes entre las estadísticas de orden de los datos proyectados. En la práctica, el índice tiende a seleccionar proyecciones que muestren un grupo grande con varios grupos pequeños o bien simples observaciones que están a su alrededor. El índice esta definido en base a la expresión:

$$I_\beta = \sigma_z^\beta \int f(z)^{1+\beta} dz, \qquad \beta>0$$

y el estimador para el caso muestral es:

$$\hat{I}_\beta = \hat{\sigma}_z^\beta \left(\frac{2r}{n}\right)^\beta \sum_{i=1}^{n} \left\{ z_{(i+r)} - z_{(i-r)} \right\}^{-\beta}$$

donde $z_{(i)}$ es la estadística de orden de los datos proyectados z_i $(i=1,...,n)$ y r es un número entero tal que, $r<n/2$; se sugiere que tome el valor $c\sqrt{n}$ para $c \in (0.5,1.5)$; además $(i+r)_+ = \min\{n, i+r\}$; $(i-r)_+ = \max\{1, i-r\}$; por otra parte $\hat{\sigma}_z^2$ es un estimador robusto para la varianza de los datos proyectados.

Yenyukov explora el comportamiento del índice para β=1, considerando el modelo teórico de la mezcla de distribuciones Gaussianas de la siguiente forma:

$$f_1(x) = \frac{2}{3}N(0,1) + \frac{1}{3}N(a,1)$$
$$f_2(x) = \frac{1}{3}\{N(-b,1) + N(0,1) + N(b,1)\}$$

los valores del índice para estos dos casos es:

$$I_1 = (1+\frac{2}{9}a^2)^{-1/2}(4\pi)^{-1/2}\frac{1}{9}\{5+4\exp(-\frac{1}{4}a^2)\}$$
$$I_1 = (1+\frac{2}{3}b^2)^{-1/2}(4\pi)^{-1/2}\frac{1}{9}\{3+4\exp(-\frac{1}{4}b^2)+2\exp(-b^2)\}$$

para f_1 y f_2 respectivamente. El índice para f_1 es tan grande como para f_2 si las medias satisfacen la desigualdad:

$$\frac{b\sqrt{27}}{5} < a < b\sqrt{3}$$

Por lo tanto, funcionan mejor las proyecciones bimodales que las proyecciones trimodales cuando se usa el índice basado en la entropía de orden 1. Por ejemplo en la mezcla de cuatro distribuciones esféricas Gaussianas con centro en los vértices de un simplejo regular, como en la figura 1 de este capítulo, el índice de entropía de orden dos, es maximizado cuando dos grupos están sobrepuestos.

3.9 Eslava y Marriott

Estos investigadores proponen dos índices basados en las coordenadas polares de los puntos proyectados en un plano. Suponen que el proceso de esferado de las variables se ha realizado previamente, por lo que el centroide de los puntos proyectados será la proyección del centroide de los datos en la dimensión p, y la proyección sobre cualquier radio tiene la misma varianza. La selección de ejes para las proyecciones es irrelevante ya que cada punto se definirá en coordenadas polares (r,θ). El radio es invariante bajo rotaciones en el plano, lo cual implica que las estadísticas estarán basadas en la distribución de θ en un círculo. Por lo

tanto, las proyecciones interesantes tomarán como base la medida de no-uniformidad, en la distribución de θ. Esto sugiere un criterio.

Una medida conveniente de no-uniformidad en el círculo es la distancia media, mínima entre los vecinos más cercanos en coordenadas polares (PNN), (en inglés *mean Polar Nearest Neighbour distance*), distancia que se define como :

$$P = \frac{1}{m}\sum_{i=1}^{m} min\left\{|\theta_i - \theta_{i-1}|, |\theta_i - \theta_{i+1}|\right\}$$

donde

$$tan\theta_i = \frac{z_{i2}}{z_{i1}}$$

y (z_{i1}, z_{i2}) son las coordenadas cartesianas para el *i-ésimo* punto proyectado. $i=1,...,m$ y $\theta_0=\theta_m$, $\theta_{m+1}=\theta_1$, con $\theta_1,...,\theta_m$ ordenados. El número de puntos considerados, m, serán aquellos que cumplan que su distancia radial desde el origen sea más grande que una proporción α de la distancia radial máxima r_{max}. Es decir el *i-ésimo* punto será considerado sólo cuando se cumpla que $r_i \geq \alpha r_{max}$ donde $r_{max} = máx\{r_1,...r_n\}$ y $0 \leq \alpha < 1$. Este se puede calcular muy fácilmente, ya que los valores de θ son ordenados rápidamente en una dimensión. El número de grupos a formar puede ser de cualquier tamaño, desde dos hacia arriba, pero generalmente con más de cinco grupos, este criterio tiende a sobreponerlos. El máximo valor se da para una distribución uniforme alrededor del círculo.

Algunas limitaciones de entrada son: cuando se tienen observaciones que de antemano sean interesantes y que se encuentran en un anillo, no serán detectadas.

Otro criterio propuesto se basa en la varianza de la distancia radial $\overline{R}$ definido como el promedio de las distancias de cada punto desde la media, es decir:

$$\overline{R} = \frac{1}{n}\sum_{i=1}^{n} r_i$$

donde $r_i = \sqrt{z_{i1}^2 + z_{i2}^2}$ para $i = 1,...n$ y (z_{i1}, z_{i2}) son las coordenadas cartesianas para el *i-ésimo* punto.

El máximo de $\overline{R}$ corresponde al mínimo de la varianza de la distancia radial: como se ve a continuación:

$$\text{var}(r) = \frac{1}{n}\sum_{i=1}^{n}\left(r_i - \overline{R}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2 - \overline{R}^2 = 2 - \overline{R}^2$$

se ha considerado que los datos fueron esferados.

La práctica ha sugerido que se obtiene una gran ventaja si los dos índices descritos anteriormente se utilizan simultáneamente. Es decir, se calculan ambos y aquellas proyecciones que optimizan uno de ellos o ambos simultáneamente son interesantes.

3.10 P. Hall

El autor presenta una discusión teórica referente al índice propuesto por Friedman (1987), presentada en la sección 3.7 de esta tesis, retomando el estudio de índices de proyección mediante la estimación de las funciones de densidad por medio de funciones ortogonales. Critica el índice de Friedman, el cual se basa en el desarrollo de las funciones de densidad en términos de polinomios de Legendre, mostrando que éste presenta una gran inestabilidad en las colas de la función de densidad, y por lo tanto argumenta que no es una buena medida en general de la desviación de la densidad Gaussiana que tienen los datos proyectados.

Por lo tanto, Hall propone otro índice de medida de desviación de los datos proyectados de la densidad Gaussiana mediante el índice siguiente:

$$I = \int_{-\infty}^{\infty} \{f(z) - \phi(z)\}^2 \, dz$$

donde $f(z)$ es la función de densidad de los datos proyectados y $\phi(z)$ es la función de densidad Gaussiana.

Hall propone el desarrollo de las funciones de densidad por medio de los polinomios de Hermite, por lo que el índice queda como:

$$I = \sum_{i=0}^{\infty} a_i^2 - \frac{\sqrt{2}}{\sqrt[4]{\pi}} a_0 + \frac{1}{2\sqrt{\pi}}$$

donde a_i es el valor esperado de las funciones de Hermite $h_i(z)$.

La estimación del índice para el caso muestral en el que tenemos n observaciones esta dada por:

$$\hat{I}_m = \sum_{i=0}^{m} \hat{a}_i^2 - \frac{\sqrt{2}}{\sqrt[4]{\pi}} \hat{a}_0 + \frac{1}{2\sqrt{\pi}}$$

donde

$$\hat{a}_i = \frac{1}{n} \sum_{i=1}^{n} h_i$$

El autor prueba que el empleo de polinomios de Hermite ayuda a ponderar de manera pesada las colas de la distribución de los datos por el término e^{-x^2} lo que permite solucionar la principal objeción al índice propuesto por Friedman. No se reporta la puesta en práctica de este índice y por lo tanto no se sabe de resultados obtenidos por el autor. Actualmente este índice se encuentra desarrollado en el programa XGobi bajo el nombre de índice de Hermite.

3.11 Cook, Buja y Cabrera

El punto de partida para el trabajo presentado por los autores parte del índice propuesto por Friedman (1987), descrito en esta tesis en la sección 3.7, en el cual para estimar la función de densidad de los datos proyectados se realiza una expansión en términos de los polinomios de Legendre, basados en la distribución uniforme. Los investigadores parten por lo tanto de la idea que, una proyección interesante de los datos será aquella en la que su función de densidad este lo más alejada de la densidad Gaussiana, denotada por $\phi(z)$.

Los autores generalizan el índice propuesto por Friedman, considerando una transformación, estrictamente monótona $T: \Re \to \Re$ de la variable aleatoria z, tal que $y = T(z)$. Denotan como $f(z)$ a la función de densidad de los datos proyectados y como $g(y)$ a la función de densidad de los datos transformados bajo T. Considerando que la versión nula de la densidad de $f(z)$ es $\phi(z)$, y denotando por $\psi(y)$ a la versión nula de la densidad de $g(y)$, definen una familia de índices como:

$$I = \int_{\Re} \{g(y) - \psi(y)\}^2 \psi(y) dy$$

Considerando la transformación $T(z) = 2\phi(z) - 1$ se tiene el índice propuesto por Friedman, llamado el índice de *Legendre*, debido a los polinomios utilizados en su estimación.
Encontrando la transformación inversa, para obtener el índice en términos de los datos originales proyectados, la transformación inversa queda como:

$$I = \int_{\Re} \left\{ \frac{f(z)}{T'(z)} - \frac{\phi(z)}{T'(z)} \right\}^2 \phi(z) dz$$

$$= \int_{\Re} \{f(z) - \phi(z)\}^2 \frac{\phi(z)}{(T'(z))^2} dz$$

Considerando la transformación propuesta por Friedman este índice se transforma en:

$$\int_{\Re} \{f(z) - \phi(z)\}^2 \frac{1}{4\phi(z)} dz$$

Se observa que, irónicamente, el mapeo propuesto por Friedman para reducir las fluctuaciones de la distribución en las colas hace exactamente lo contrario, ya que el término $1/\phi(z)$ pondera hacia arriba las observaciones en las colas, dejando el índice de Legendre muy sensible a las diferencias de la normalidad en las colas de $f(z)$. Sin embargo el problema

es más teórico que práctico ya que la estimación es realizada por una expansión finita de funciones. Este problema ya había sido detectado por Hall (1989), lo que motivó que propusiera un índice alternativo el cual parte de considerar la igualdad

$$\frac{\phi(z)}{(T'(z))^2} = 1$$

Despejando $T'(z)$ queda:

$$T'(z) = \sqrt{\phi(z)}$$

Por lo cual podemos considerar a $T(z)$ como una aproximación a la distribución Gaussiana con varianza 2. Es decir:

$$T(z) \propto \phi_{\sigma=\sqrt{2}}(z),$$

por lo tanto el nuevo índice, como se vio en la sección anterior, queda definido por:

$$I^H = \int_{\Re} \{f(z) - \phi(z)\}^2 dz$$

el cual es llamado el índice de *Hermite* porque esta basado en los polinomios de Hermite para la expansión de las funciones de densidad.

Bajo la idea original de Friedman de dar más peso al centro de la distribución que a las colas y yendo un paso más allá, los autores proponen el uso de la transformación $T(z) = z$, que es más natural que la de Hall, lo cual da lugar a la propuesta del siguiente índice:

$$I^N = \int_{\Re} \{f(z) - \phi(z)\}^2 \phi(z) dz$$

Los autores lo denominan índice *Natural de Hermite*, debido a que al igual que el de Hall, esta basado en la expansión de $f(z)$ por polinomios de Hermite.

La optimización de este índice es realizado a partir de la expansión por polinomios de Hermite de $f(z)$ y $\phi(z)$ como sigue:

$$f(z) = \sum_{i=0}^{\infty} a_i p_i(z) \quad \text{y} \quad \phi(z) = \sum_{i=0}^{\infty} b_i p_i(z)$$

donde los $p_i(x)$ son polinomios ortogonales con respecto de $\phi(x)$.

Sustituyendo estas expresiones en el índice Natural de Hermite se obtiene:

$$I^N = \sum_{i=0}^{\infty} (a_i - b_i)^2$$

donde

$$a_i = \int_{\Re} f(z) p_i(z) \phi(z) dz$$

$$b_i = \int_{\Re} \phi(z) p_i(z) \phi(z) dz$$

Los coeficientes b_i fueron calculados analiticamente por Abramowitz y Stegun (1972 pag. 778) quedando como:

$$b_{2i} = \frac{(-1)^i ((2i)!)^{1/2}}{\sqrt{\pi}\, i!} \frac{1}{2^{2i+1}}, \; b_{2i+1} = 0; \; i = 0,1,2,...$$

Como los a_i dependen de $f(z)$, deberán estimarse para poder estimar I^N. Reinterpretando a_i como una esperanza matemática se tiene

$$a_i = E_f\left(p_i(z)\phi(z)\right)$$

lo cual conduce a la siguiente estimación obvia del muestreo:

$$\hat{a}_i = \frac{1}{n}\sum_{j=1}^{n} p_i(z_j)\phi(z_j)$$

Por lo tanto la estimación de I^N truncando a M términos es:

$$\hat{I}_M^N = \sum_{i=0}^{M}(\hat{a}_i - b_i)^2$$

Los autores desarrollaron este índice en un programa especial de Proyecciones Perseguidas llamado *XGobi*, y de acuerdo al objetivo de buscar posibles estructuras en los datos, concluyeron que los valores más interesantes para M son cero ó uno.

Bajo estas condiciones encuentran los valores óptimos para el caso de uno y dos términos del índice estimado.

De manera similar al de una dimensión, estimaron el índice Natural de Hermite en dos dimensiones, truncando polinomios hasta M términos como sigue:

$$\hat{I}_M^N = \sum_{i,j\geq 0, i+j\leq M}(\hat{a}_{ij} - b_{ij})^2$$

Este índice está desarrollado en el programa XGobi con el nombre de índice Natural de Hermite.

3.12 Comentarios

-El objetivo principal de la técnica de proyecciones perseguidas es la de encontrar proyecciones que ayuden a conocer la estructura subyacente en los datos multivariados, particularmente, por ejemplo, conglomerados de puntos.

-Tradicionalmente las proyecciones más interesantes son las que se realizan en dos dimensiones. Sin embargo con el desarrollo de computadoras más poderosas no se descarta la posibilidad de realizar estas proyecciones en dimensión tres, tal como lo indica el trabajo de Nason (1995) quien ha desarrollado un algoritmo para encontrar proyecciones en tres dimensiones generalizando el índice de momentos propuesto por Jones y Sibson(1987). Las proyecciones en una dimensión más que un interés práctico tienen un interés teórico debido a que su estudio puede ayudar en la selección de criterios convenientes en el caso de dos ó tres dimensiones.

-El aspecto más importante de la técnica de proyecciones perseguidas es el de la selección de un criterio conveniente de lo que constituye una proyección interesante. La tendencia general es definir que una proyección es interesante mientras más difiera de la distribución Gaussiana. Sin embargo, existen muchas maneras de que un conjunto de datos difiera de la distribución Gaussiana, por ejemplo: a) la curva pueden ser muy empinada (leptocúrtica); b) la curva puede ser asimétrica; c) la curva puede ser relativamente plana (platicúrtica); d) Las colas de la curva pueden ser muy pesadas (posible presencia de observaciones aberrantes); e) Puede que la distribución no sea unimodal sino multimodal. Este último caso es particularmente interesante si el objetivo de los índices de proyección es detectar estructuras agrupadas en los datos, por ejemplo el índice propuesto por Eslava y Marriott. Por lo tanto, muchos de los índices expuestos se basan en proponer alguna medida de no-normalidad de los datos proyectados. Lo siguiente es la manera de estimar las funciones de densidad de los datos proyectados, que como hemos visto existen diferentes maneras. Todo lo anterior hace que se tenga una gran variedad de proyecciones de los datos.

- Considerando que el método de proyecciones perseguidas se basa en la optimización de los índices de proyección, deberá considerarse que no resulta fácil encontrar puntos óptimos

globales y muchas veces solo se encuentran puntos óptimos locales aun sin saberlo. Por lo tanto se recomienda realizar más de una proyección a fin de tener un panorama más amplio de la estructura de los datos.

- Es recomendable aplicar las técnicas de proyecciones perseguidas iterativamente. Es decir, puede ser que al aplicar una vez la técnica se consiga observar algunos conglomerados de puntos. Será necesario aplicar la técnica en cada uno de los conglomerados a fin de observar posibles estructuras subyacentes en los conglomerados.
- En un principio, un punto importante para aplicar la técnica de proyecciones perseguidas era producir un algoritmo que pudiera desarrollarse fácilmente en una computadora ya que muchos de los programas realizados involucraban una gran cantidad de cálculos; esto dificultaba que se pudiera trabajar con conjunto grandes de datos debido a las restricciones que presentaban las computadoras. Generalmente los algoritmos se programaron en lenguaje Fortran.

En la actualidad se cuenta con un *software* computacional llamado XGobi desarrollado por D. Swayne, D. Cook y A. Buja en el instituto Bellcore (1994) donde se han desarrollado diez índices de proyecciones perseguidas, de los cuales cinco han sido descritos en este capítulo: el propuesto por Friedman y Tukey (1974), el índice de entropía propuesto por Jones y Sibson (1987), el índice de Legendre propuesto por Friedman (1987), el índice de Hermite propuesto por Hall (1989) y el índice llamado Natural de Hermite propuesto por Cook, Buja y Cabrera (1993). Dos índices más son modificaciones del de Tukey y Friedman y el de la entropía a fin de acelerar el proceso de optimización de los índices. Los otros tres índices desarrollados, que sólo mencionamos sin dar detalles, son: el índice *Holes* llamado así porque responde a proyecciones que contienen muy pocos puntos en el centro; el índice *Central Mass* que esta diseñado para proyecciones con alta concentración de puntos en el centro; y finalmente el índice *Skewness*, el cual responde a proyecciones con distribución de datos asimétricos. En el capítulo 4 se presentan 3 ejemplos de aplicación en los cuales para facilidad nos referiremos a estos índices con el nombre asignado en XGobi.

Este *software* trabaja en computadoras que funcionan con sistema operativo UNIX, en un ambiente X-Windows. Para mayor información hay que comunicarse a la dirección electrónica: dfs@bellcore.com del instituto bellcore.

En los ejemplos presentados en el capítulo 4 se muestran algunos gráficas de proyecciones, utilizando este programa.

-Es importante recalcar el proceso de esferar las variables que deberá realizarse previamente antes de aplicar la proyección de los puntos. El objetivo es garantizar la invarianza de la escala, evitando aspectos que no tengan que ver con la naturaleza elipsoide de los datos. Desde la propuesta de Friedman y Tukey se tenía esta preocupación, debido a lo cual su índice contiene un término de dispersión.

- A continuación se presenta la tabla 1, a manera de guía para consulta rápida, en la cual se resumen los índices teóricos de proyección presentados, la referencia bibliográfica donde se publicaron y el *software* en el que se desarrollaron, si es el caso.

Tabla 1. Indices de proyección por autor

Número	Autor/Año/Referencia	Fórmula de Indice	Términos en el indice	Software
1	Kruskal, J.H. (1972). Linear transformation of multivariate data to reveral clustering. In Multidimensional scaling: theory and applications in the behavioural sciences. Vol. 1(eds. R.N. Shepard), London: Seminar Press	Indice de Condensación		No se implementó
2	Friedman, J.H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. IEEE Trans. Comput., 23.	a) I(k) = s(k)d(k) b) I(k,) = s(k)s( )d(k,)	$s(k)=\left[\sum_{i=qn}^{(1-q)n}(z_i(k)-\bar{z}(k))^2/(1-2q)n\right]^{1/2}$ $d(k)=\sum_{i=1}^{n}\sum_{j=1}^{n}f(r_{ij})L(R-r_{ij})$	Programa Fortran
3	Huber, P. J. (1985). Projection pursuit (with discussion). Ann. Statist., 13	a) $Q=\frac{\lvert k_m(z)\rvert}{(k_2(z))^{m/2}}$ b) $Q=\sigma_z^2\int\left(\frac{f'(z)}{f(z)}\right)^2 f(z)dz-1$ c) $Q=\int\ln[f(z)]f(z)dz+\ln[\sigma_z\sqrt{2\pi e}]$	Los $k_m(z)$ son los cumulantes de orden m.	No se implementaron
4	Jones, M.C. and Sibson, R. (1987). What is projection pursuit? (with discussion). Journal of the Royal Statistical Society, serie A No. 150.	a) $I=\int f(z)\ln(f(z))dz$ b) $I_m=\frac{1}{12}\left[k_3^2+\frac{1}{4}k_4^2\right]$ $I_m=\frac{1}{12}\left\{\left(k_{30}^2+3k_{21}^2+3k_{12}^2+k_{03}^2\right)+\frac{1}{4}\left(k_{40}^2+4k_{31}^2+6k_{22}^2+4k_{13}^2+k_{04}^2\right)\right\}$	Los k's son los cumulantes de los datos.	Programa Fortran

5	Jee, R. (1985). A study on projection pursuit methods. Ph. D. thesis Rice University, U.S.A	a) $Q=\sigma_z^2\int\left(\frac{f(z)}{f'(z)}\right)f(z)dz-1$ b) $Q=\int \ln[f(z)]f(z)dz+\ln\left[\sigma_z\sqrt{2\pi e}\right]$ c) $L_1=\int\lvert f(z)-\phi(z)\rvert dz$ d) $L_H=\int\left[f^{1/2}(z)-\phi^{1/2}(z)\right]^2dz$	$\phi(z)$ es la densidad normal estándar.	Programa Fortran
6	Friedman, J. H. (1987). Exploratory projection pursuit. Journal American Statistics. Ass., 82	$I(\alpha)=\frac{1}{2}\sum_{j=1}^{J}(2j+1)E^2\left[P_j(T)\right]$ $I(\alpha,\beta)=\frac{1}{4}\sum_{j=1}^{J}(2j+1)E^2\left[P_j(T_1)\right]+\frac{1}{4}\sum_{k=1}^{J}(2k+1)E^2\left[P_k(T_2)\right]+$ $\frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{J-j}(2j+1)(2k+1)E^2\left[P_j(T_1)P_k(T_2)\right]$	Los $P_i(T)$ son los polinomios de Legendre.	Programa Fortran
7	Yenyukov, I.S. (1988). Detecting structure by means of projection pursuit. In Compstat 1988 (eds. D. Edwards and E. Raun). Physica-Verlag Heidelberg for IASC.	$I_\beta=\sigma_z^\beta\int f(z)^{1+\beta}dz$	Para $\beta>0$	No se implemento

8	Eslava G. and Marriott F.H.C. (1994). Some Criteria for projection pursuit. Statistics and Computing, 4, pags. 13-20.	a) $P = \frac{1}{m}\sum_{i=1}^{m} min\left\{\left\|\theta_i - \theta_{i-1}\right\|, \left\|\theta_i - \theta_{i+1}\right\|\right\}$ b) $\overline{R} = \frac{1}{n}\sum_{i=1}^{n} r_i$	(r_i, θ_i) son las coordenadas polares del i-ésimo punto proyectado	Programa Fortran
9	Hall P. (1989) Polynomial Projection Pursuit. The Annals of Statistics, 17, pags.(589-605)	$I = \int_{-\infty}^{\infty} \{f(z) - \phi(z)\}^2 dz$	$\phi(z)$ es la densidad normal estándar.	No se implemento
10	Cook., Buja y Cabrera (1993). Projection Pursuit Based on Orthonormal Function Expansions. Journal of Computational and Graphical Statistics, Vol. 2, No. 3.	$I^N = \int_{\Re} \{f(z) - \phi(z)\}^2 \phi(z) dz$	$\phi(z)$ es la densidad normal estándar.	XGobi

CAPÍTULO 4. APLICACIONES

Con el objetivo de ilustrar las técnicas de análisis de componentes principales y de proyecciones perseguidas, en este capítulo se presentan tres ejemplos de datos reales. El primer ejemplo corresponde a medidas en el cráneo de dos poblaciones del México antiguo; el segundo corresponde a medidas tomadas en una región del Amazonas para analizar la manera que se da la regeneración de plantas; finalmente se presenta un ejemplo relacionado con un conjunto de virus.

4.1 Cráneos

Algunas civilizaciones antiguas en México tenían la costumbre de predestinar a niños recién nacidos para ser sacerdotes, guerreros, gobernantes, etc. Una forma física de hacer tal distinción consistía en que, al nacer los niños y hasta los 30 ó 45 días de nacidos, se les aplicaba cierta deformación craneana mediante la colocación de tablas amarradas con vendas alrededor de los cráneos de los recién nacidos; se hacía en esta etapa de la vida de los niños debido a que físicamente se realiza la conformación de los huesos craneanos.

Los datos que se analizan en esta sección corresponden a muestras de dos poblaciones: Cholula y Cueva de la Candelaria. La información fue recolectada por Díaz Leñero (1995) de las bodegas del Museo Nacional de Antropología. Estudios antropológicos han permitido conocer que en la civilización asentada en el valle de Cholula (Puebla-Tlaxcala), se tenía la costumbre de deformación craneana, contrario a lo que se sabe de la civilización asentada en la Cueva de la Candelaria (Coahuila) donde no se tiene conocimiento sobre costumbres similares.

El objetivo de la investigación es establecer si tal deformación en realidad afectó las medidas craneanas. Estadísticamente se traduce en establecer posibles diferencias entre las dos poblaciones de cráneos.

Aunque existen médicos y antropólogos que definen una gran medida en los cráneos (craneometría) se han realizado esfuerzos importantes para estandarizar mundialmente las medidas de todo el cuerpo humano, y en particular del cráneo. En base a esto, el

Departamento de Antropología Física del Instituto Nacional de Antropología e Historia ha diseñado un instrumento para recolectar la información referente a las medidas del cráneo humano.

En cada cráneo se manejan 46 medidas subdivididas en cráneo cerebral, cráneo facial y mandíbula. Sin embargo, en reuniones con antropólogos físicos nos dimos cuenta que, para el objetivo de la investigación, muchas de estas medidas son redundantes debido a que algunas o miden lo mismo sólo que por partes, o bien son medidas altamente correlacionadas. Por otra parte, muchas de las medidas de la cédula corresponden a diferentes índices obtenidos a partir de algunas medidas base. Por lo tanto, conjuntamente con investigadores del área de antropología física, se determinó utilizar cinco variables para el análisis estadístico, las cuales se consideran básicas para determinar las características de un cráneo. A continuación se presentan estas cinco variables y su significado físico.

X1 = DIAMETRO ANTERO-POSTERIOR MÁXIMO

X2 = DIAMETRO TRANSVERSO MÁXIMO

X3 = DIAMETRO BREGMA BASION

X4 = DIAMETRO BIZIGOMATICO

X5 = DIAMETRO NASIO-PROSTION

La figura 1 ilustra cada una de estas medidas en los cráneos. Los datos originales se presentan en el Apéndice I de esta tesis.

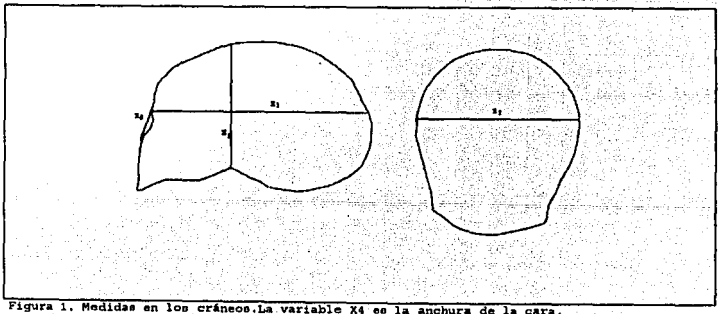


Figura 1. Medidas en los cráneos.La variable X4 es la anchura de la cara.

Debido al estado de conservación de los cráneos sólo se obtuvo una muestra compuesta de 55 casos, correspondiente a los cráneos que estaban menos deteriorados, de los cuales los primeros 30 casos corresponden a la población de Cholula y los restantes 25 a La Cueva de la Candelaria. Sin embargo, existen cráneos en los que fue imposible tomar alguna de las medidas, por lo que tenemos 11 datos faltantes repartidos como sigue: un dato corresponde a un caso de la población de La Candelaria en la variable X2; tres datos son de la población de Cholula en la variable X4; siete datos son de la población de La Candelaria en la variable X4.; se decidió, conjuntamente con los antropólogos, sustituir estos datos faltantes por la media de la variable y de la población a la cual corresponde el caso.

En la figura 2 se presentan las gráficas de las variables individuales tomadas de dos en dos, en las cuales se observa cierta separación de los grupos de cráneos en algunos casos, por ejemplo en la gráfica de X1 vs X2, y en otros la separación no es tan clara, por ejemplo en la gráfica de X3 vs X5.

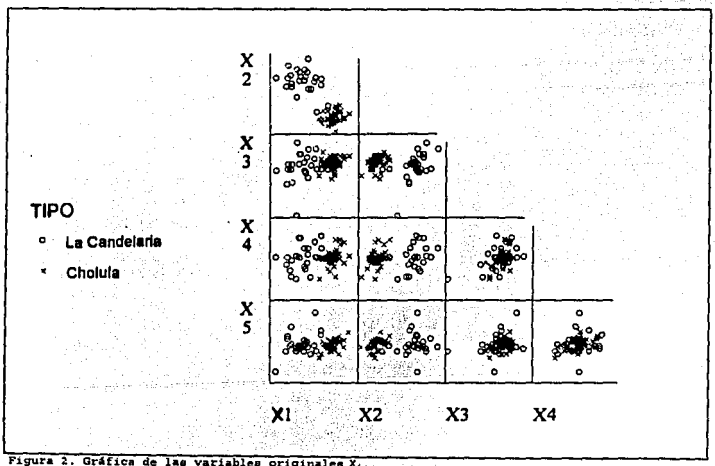


Figura 2. Gráfica de las variables originales X_i.

4.1.1 Análisis de Componentes Principales

En primer lugar se realizó un análisis de componentes principales, con el cual se obtuvieron los siguientes resultados, mediante el uso del programa SPSS.

La matriz de correlación entre las variables originales es:

	X1	X2	X3	X4	X5
X1	1				
X2	-.80755	1			
X3	.26677	.03477	1		
X4	.17286	.17048	.49538	1	
X5	.19296	.09855	.32068	.26245	1

El determinante de esta matriz de correlación es de .1247842

En la matriz observamos una alta correlación entre las variables X1 y X2 que corresponden a la longitud del diámetro antero posterior máximo y al diámetro transverso máximo. Probablemente este hecho origina que el valor del determinante sea pequeño.

Por otra parte, en la tabla 1 se presentan los valores característicos de la matriz de correlación, así como el porcentaje de varianza que explica cada componente.

Componente	Valor caract.	% de Var.	% Acum.
1	1.95047	39.0	39.0
2	1.68630	33.7	72.7
3	.77169	15.4	88.2
4	.49152	9.8	98.0
5	.10002	2.0	100.0

Tabla 1. Valores característicos y porcentaje de varianza explicado por cada componente.

Se puede observar que es suficiente considerar tres componentes principales para explicar cerca del 90% de la variación total de los datos. En la figura 3 se presenta una gráfica de las componentes principales ordenados en forma decreciente, contra su valor, es decir la gráfica denominada *Scree Plot*. En esta gráfica se observa que la caída más pronunciada se tiene

hasta las primeras tres componentes. Este criterio gráfico se utiliza comúnmente en el área de Psicología.

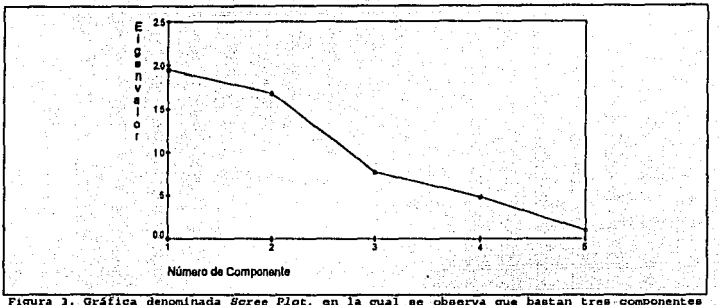


Figura 3. Gráfica denominada *Scree Plot*, en la cual se observa que bastan tres componentes principales, para obtener la máxima variabilidad de los datos.

En la tabla 2 se presentan los valores de las correlaciones entre las variables originales y cada componente principal.

	Componente 1	Componente 2	Componente 3	Componente 4	Componente 5
X1	.84272	-.48563	.00687	.07383	.22023
X2	-.54717	.80880	.02285	-.01272	.21393
X3	.65486	.49158	-.23143	-.52459	-.02749
X4	.52938	.60141	-.38656	.45349	-.05459
X5	.48147	.43928	.75374	.07122	-.04493

Tabla 2. Correlación entre los componentes principales y las variables originales.

Debido a que en muchas ocasiones las dos primeros componentes absorben la mayor variación de los datos, es costumbre resumir en un gráfico las correlaciones de las variables originales con las dos primeras componentes principales. Este gráfico suele llamarse comúnmente círculo de correlaciones. En el caso de este ejemplo, la figura 4 muestra la gráfica correspondiente en la cual observamos que existe una alta correlación de las

variables X1 y X2 con las dos componentes principales, teniendo una mayor correlación la variable X2 con la primer componente y la variable X1 con la segunda componente. La variable que presenta menor correlación con las dos primeras componentes principales es la variable X5.

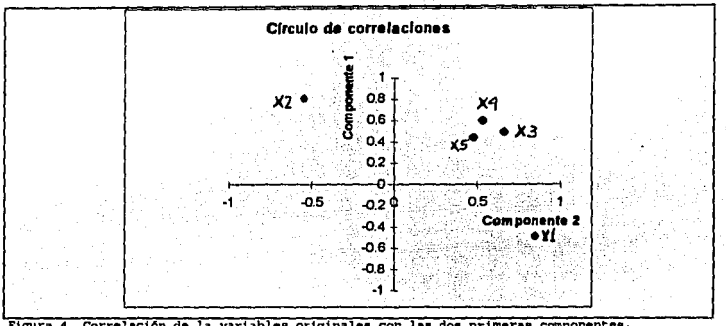


Figura 4. Correlación de la variables originales con las dos primeras componentes.

Por otra parte, una gráfica que permite observar la estructura de los datos en un plano es la correspondiente a la de los valores de pares de componentes principales para los casos que componen la muestra. La figura 5 muestra la gráfica de los dos primeros componentes principales, los cuales explican aproximadamente el 73 % de la variabilidad de los datos, en la cual se observa la separación de los dos grupos de cráneos. Esta gráfica nos proporciona elementos para afirmar que la hipótesis antropológica de que la deformación intencional de los cráneos si afecta las medidas craneales es verdadera. Por otra parte se observa un punto que se sale del comportamiento general de los demás.

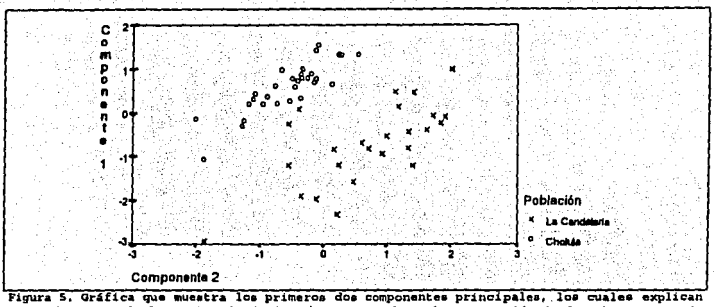


Figura 5. Gráfica que muestra los primeros dos componentes principales, los cuales explican cerca del 73 % de la variación de los datos. Es clara la separación de los dos grupos de cráneos.

Muchas veces no basta con representar gráficamente las observaciones en términos de las dos primeras componentes, puede ser que algunas gráficas de las componentes restantes nos revelen información adicional del comportamiento de los datos. En el caso del ejemplo presentado, dado que las primeras tres componentes explican cerca del 90% de la variación total de los datos, en la figura 6 se presentan las gráficas de las posibles combinaciones de estas componentes principales. En este caso, la gráfica que mejor separa los grupos corresponden a las componentes 1 y 2. La gráfica correspondiente a las componentes 1 y 3 aunque muestra los dos grupos separados, la diferenciación no es tan clara. Por otra parte la gráfica correspondiente a las componentes 2 y 3 no permite observar ninguna estructura de grupos en los datos.

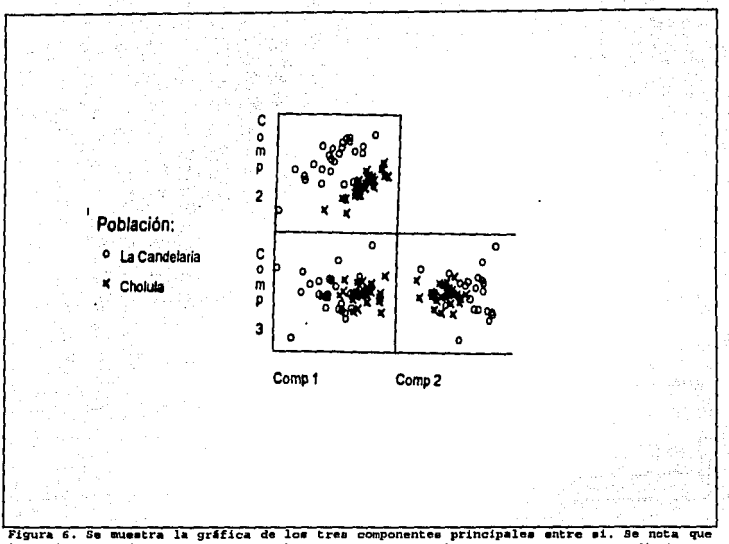


Figura 6. Se muestra la gráfica de los tres componentes principales entre sí. Se nota que las gráficas de la primera y segunda componente contra la tercera componente no distinguen los grupos, debido a la poca varianza que explica cada par.

4.1.2 Proyecciones Perseguidas

Como mencionamos antes, en el programa XGobi se han implementado algunos de los índices expuestos en el capítulo 3, además de otros no tratados en esta tesis. Se utilizó este *software* a fin de encontrar las proyecciones de los datos originales del ejemplo así como la gráfica del comportamiento del índice, cuando es optimizado. Dada la característica de la técnica (exploratoria) y debido a que se tienen varios índices de proyección, el número de gráficas es muy alto, sin embargo, tanto en este ejemplo como en los dos restantes se

decidió incluir todas las gráficas. Para fines de presentación y análisis de los ejemplos se deben seleccionar aquellas gráficas que se consideren aporten mayor información.

Las gráficas siguientes representan las diferentes proyecciones encontradas con los índices, decidiendo respetar el nombre del índice asignado por los autores del programa, teniendo la equivalencia con respecto a los índices presentados en el mencionado capítulo como sigue: Friedman-Tukey es la propuesta original de Friedman y Tukey(1977); entropía es la propuesta presentada por Jones y Sibson (1987); Legendre es la propuesta de Friedman (1987); Hermite es el índice propuesto por Hall(1989); Natural de Hermite es el índice desarrollado por Cook y colaboradores(1993). Las gráficas obtenidas son las siguientes:

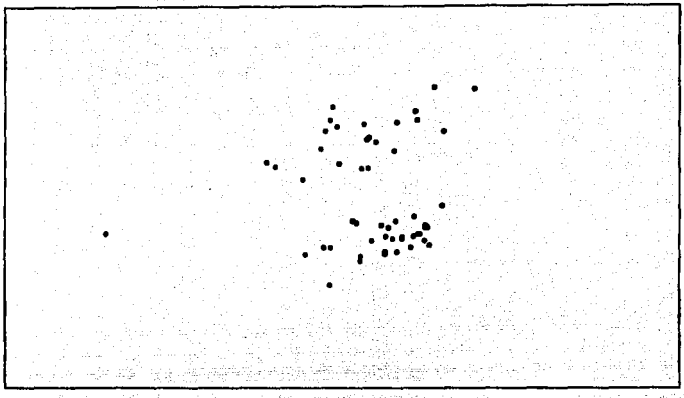


Figura 7a. Ejemplo: Cráneos. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Friedman-Tukey.

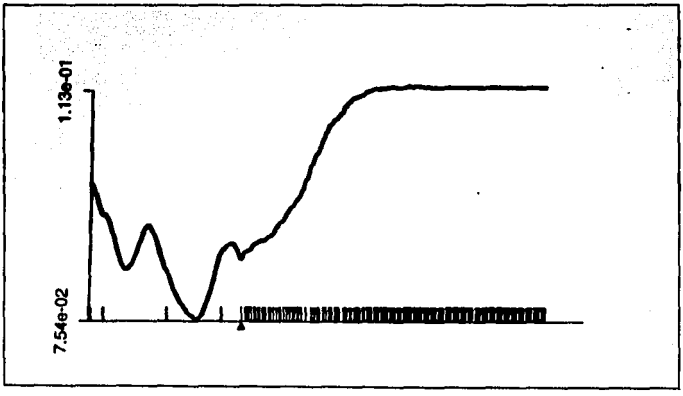


Figura 7b. Ejemplo: Cráneos. Comportamiento del índice de Friedman-Tukey al proyectar en dos dimensiones.

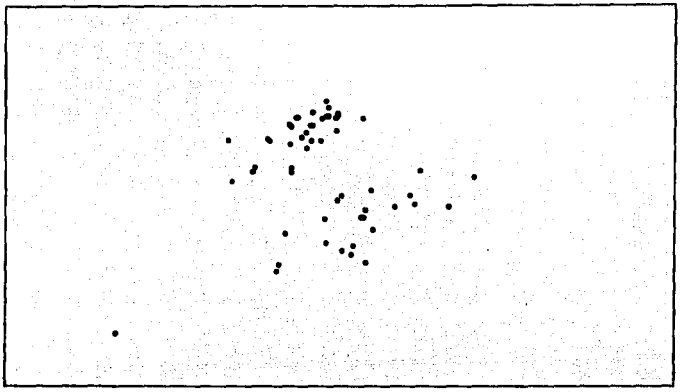


Figura 8a. Ejemplo: Cráneos. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Friedman-Tukey Modificado.

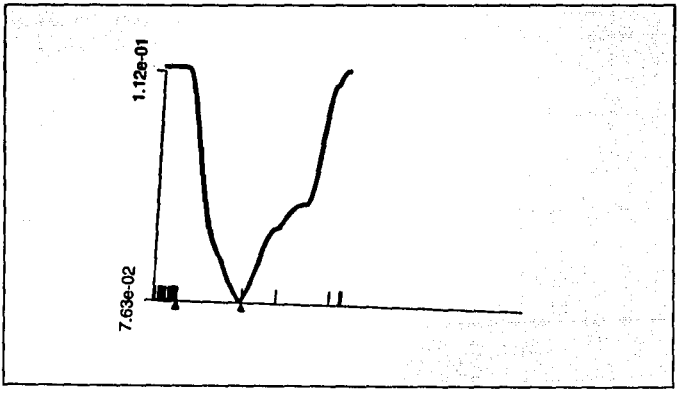


Figura 8b. Ejemplo: Cráneos. Comportamiento del índice de Friedman-Tukey Modificado al proyectar en dos dimensiones.

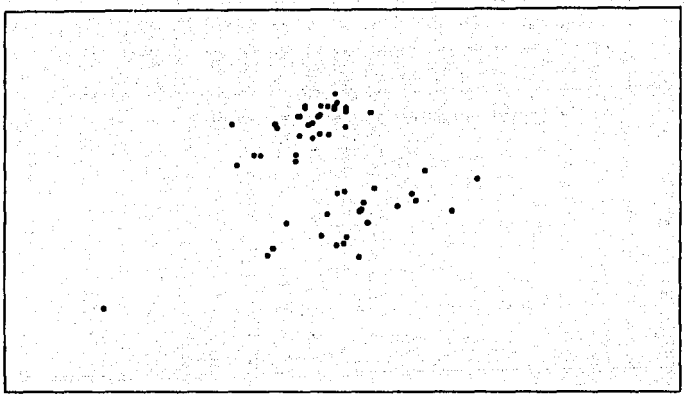


Figura 9a. Ejemplo: Cráneos. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Entropía.

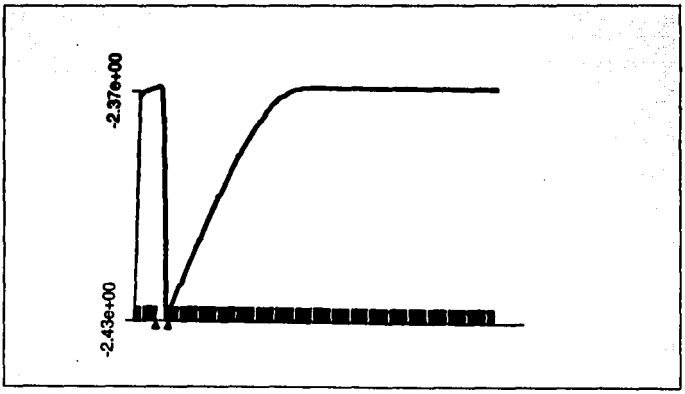


Figura 9b. Ejemplo: Cráneos. Comportamiento del índice de Entropía al proyectar en dos dimensiones.

Figura 10a. Ejemplo: Cráneos. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Entropía Modificado.

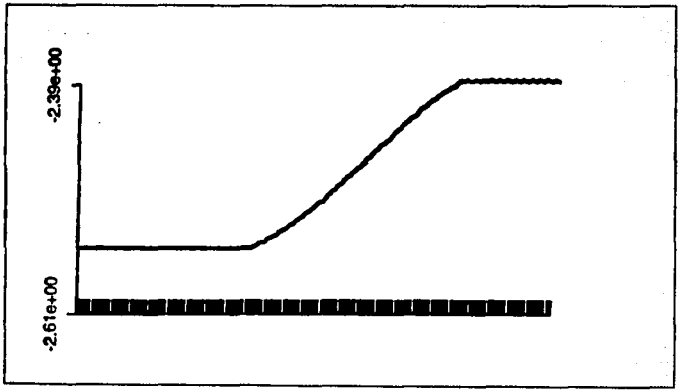


Figura 10b. Ejemplo: Cráneos. Comportamiento del índice de Entropía Modificado al proyectar en dos dimensiones.

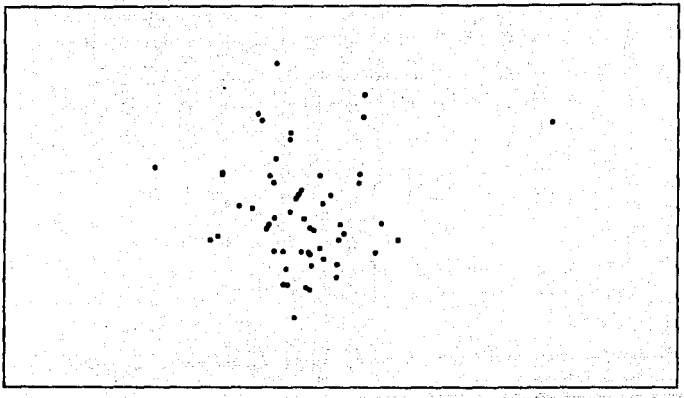


Figura 11a. Ejemplo: Cráneos. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Legendre usando polinomios de grado 1.

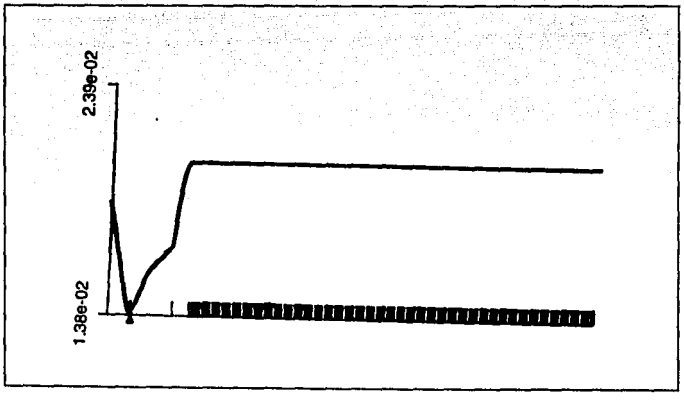


Figura 11b. Ejemplo: Cráneos. Comportamiento del índice de Legendre usando polinomios de grado 1.

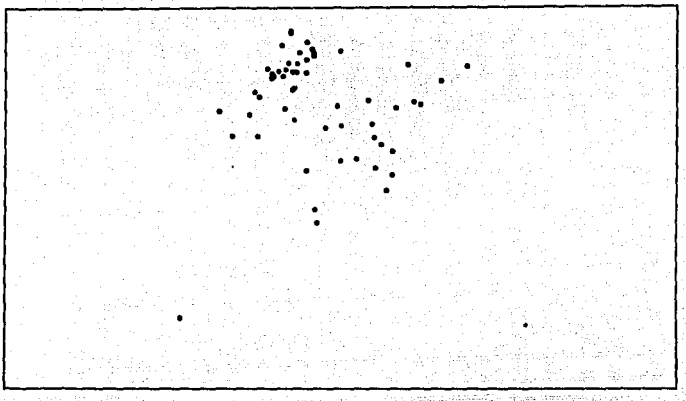


Figura 12a. Ejemplo: Cráneos. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Legendre usando polinomios de grado 2.

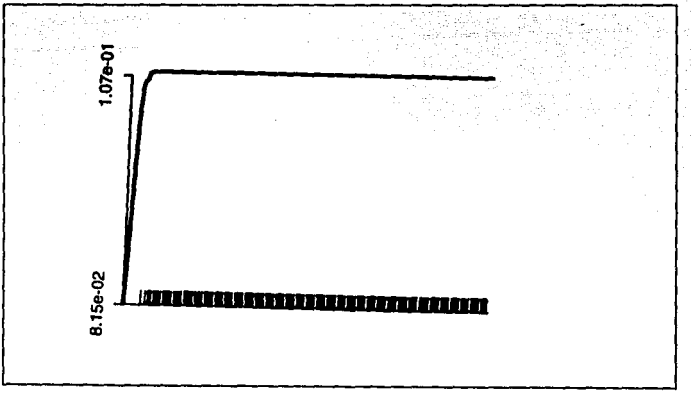


Figura 12b. Ejemplo: Cráneos. Comportamiento del índice de Legendre usando polinomos de grado 2.

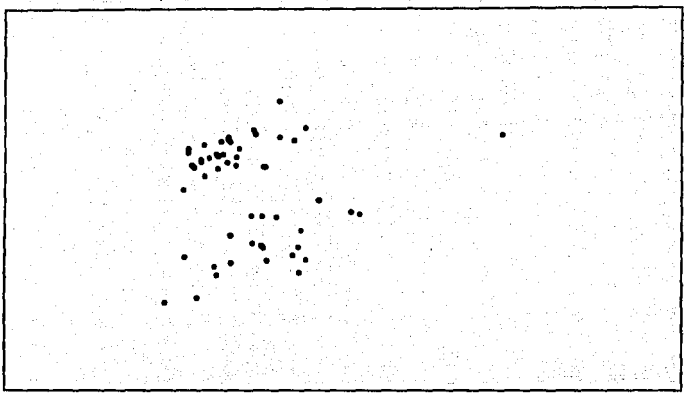


Figura 13a. Ejemplo: Cráneos. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Legendre con un polinomios de grado 7.

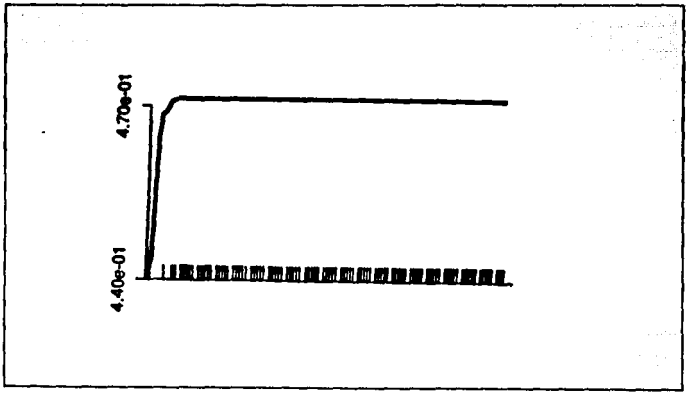


Figura 13b. Ejemplo: Cráneos. Comportamiento del índice de Legendre usando polinomios de grado 7.

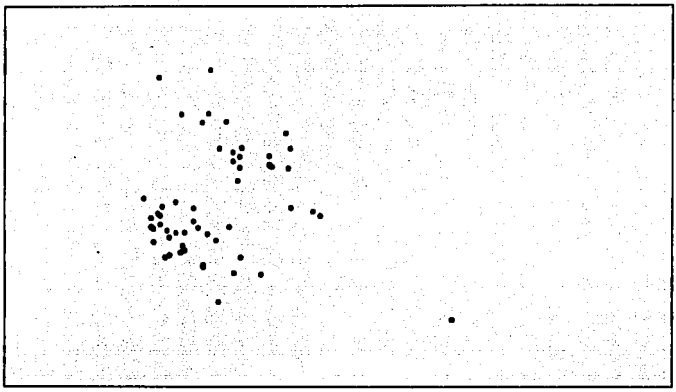


Figura 14a. Ejemplo: Cráneos. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Legendre, usando polinomios de grado 8.

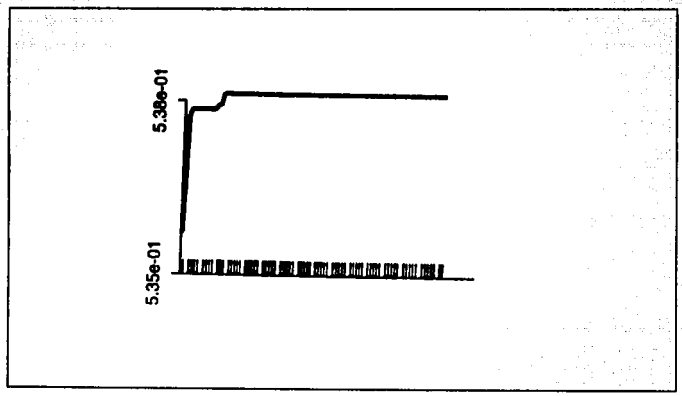


Figura 14b. Ejemplo: Cráneos. Comportamiento del índice de Legendre, usando polinomios de grado 8.

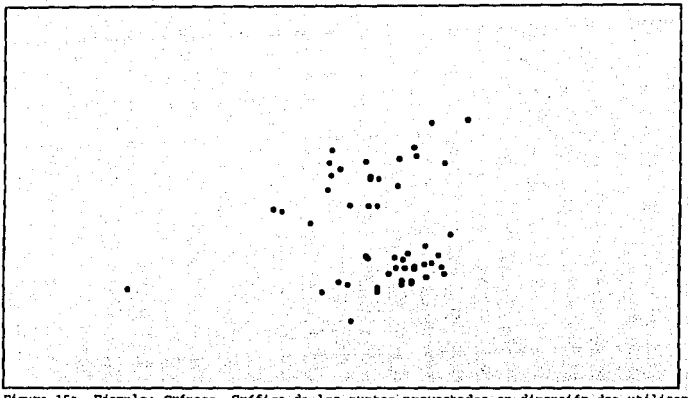


Figura 15a. Ejemplo: Cráneos. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Hermite, utilizando polinomios de grado 0.

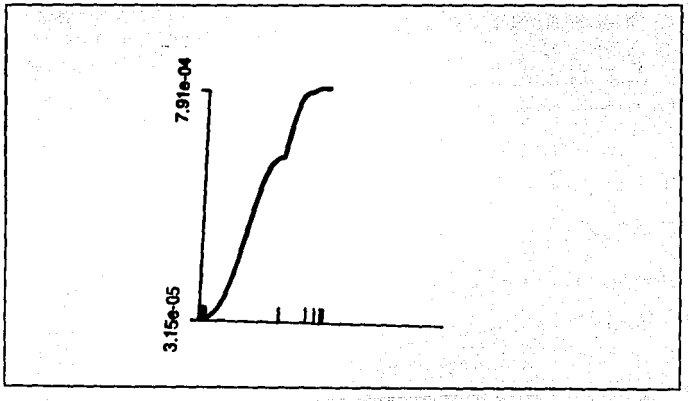


Figura 15b. Ejemplo: Cráneos. Comportamiento del índice de Hermite, utilizando polinomios de grado cero.

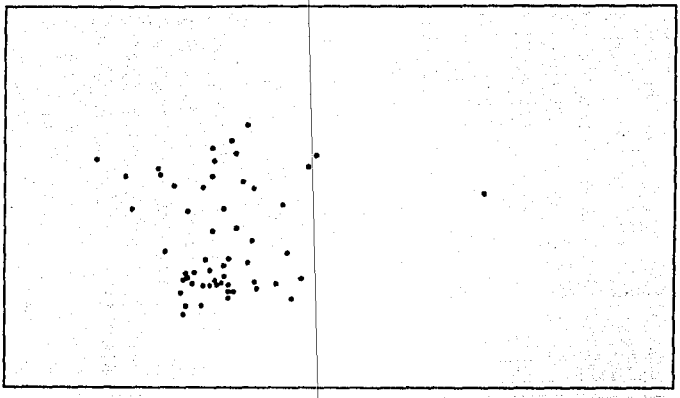


Figura 16a. Ejemplo: Cráneos. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Hermite, con polinomios de grado 1.

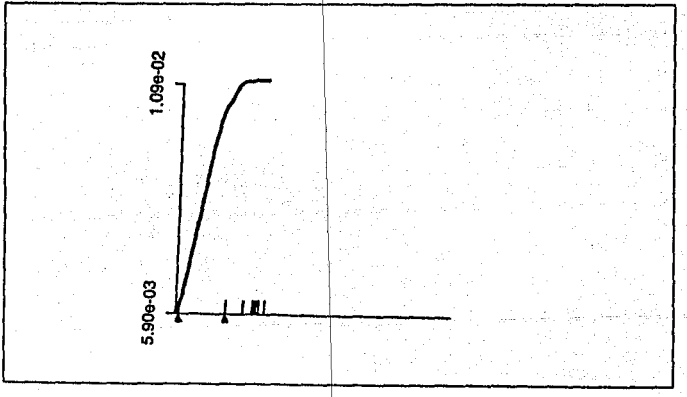


Figura 16b. Ejemplo: Cráneos. Comportamiento del índice de Hermite, usando polinomios de grado 1.

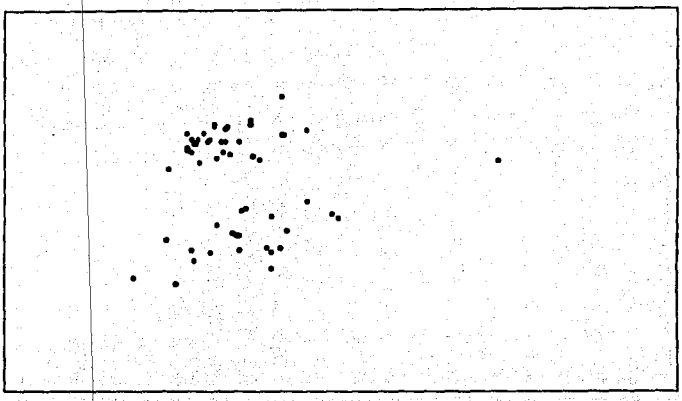


Figura 17a. Ejemplo: Cráneos. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Hermite, con polinomios de grado 7.

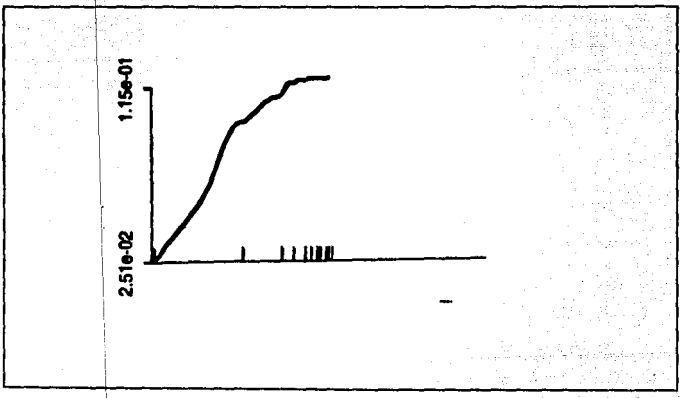


Figura 17b. Ejemplo: Cráneos. Comportamiento del índice de Hermite, usando polinomios de grado 7.

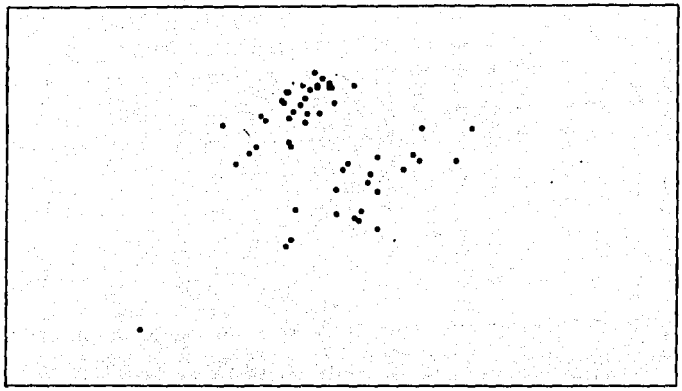


Figura 18a. Ejemplo: Cráneos. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Hermite, con polinomios de grado 8.

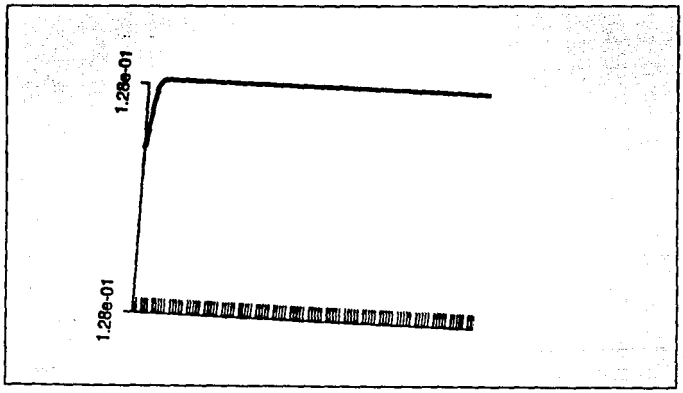


Figura 18b. Ejemplo: Cráneos. Comportamiento del índice de Hermite, usando polinomios de grado 8.

Figura 19a. Ejemplo: Cráneos. Gráfica de los puntos proyectados en dimensión dos utilizando el índice Natural de Hermite, utilizando polinomios de grado 0.

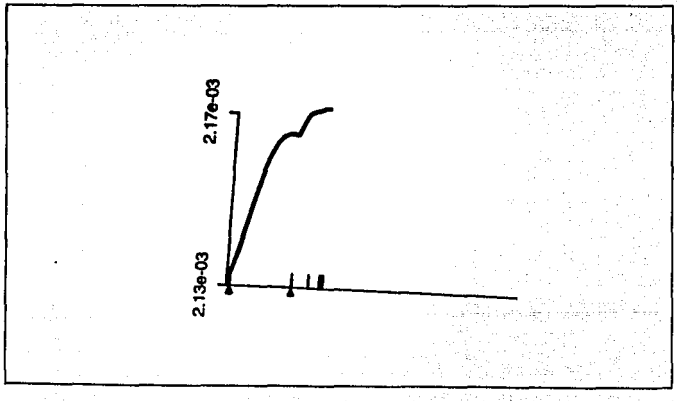


Figura 19b. Ejemplo: Cráneos. Comportamiento del índice Natural de Hermite, utilizando polinomios de grado cero.

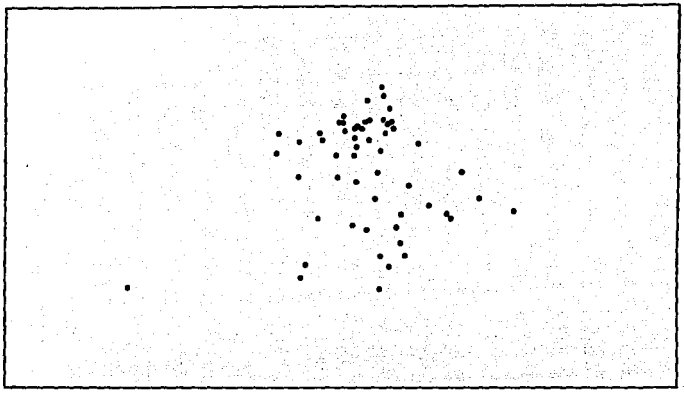


Figura 20a. Ejemplo: Cráneos. Gráfica de los puntos proyectados en dimensión dos utilizando el índice Natural de Hermite, con polinomios de grado 1.

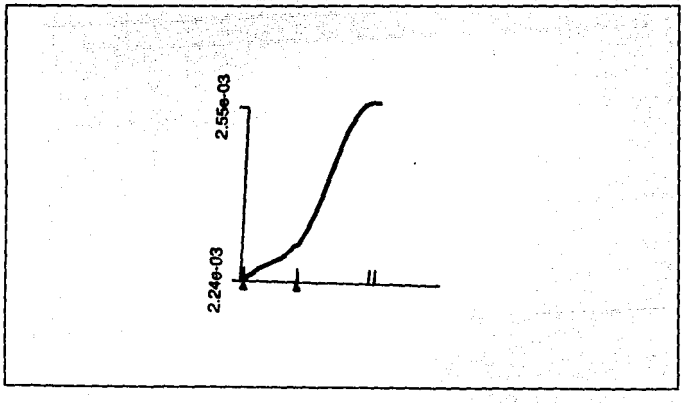


Figura 20b. Ejemplo: Cráneos. Comportamiento del índice Natural de Hermite, usando polinomios de grado 1.

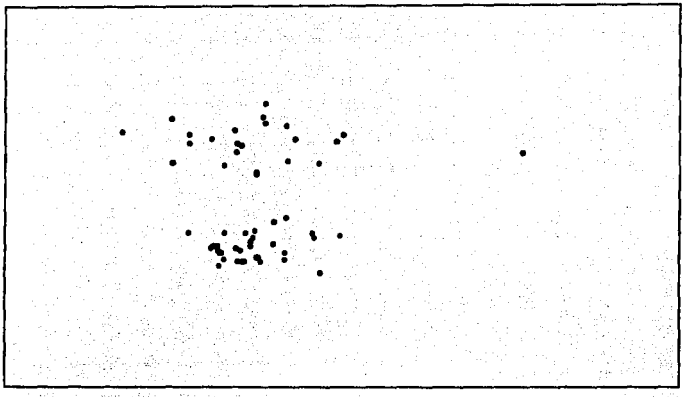


Figura 21a. Ejemplo: Cráneos. Gráfica de los puntos proyectados en dimensión dos utilizando el índice Natural de Hermite, con polinomios de grado 7.

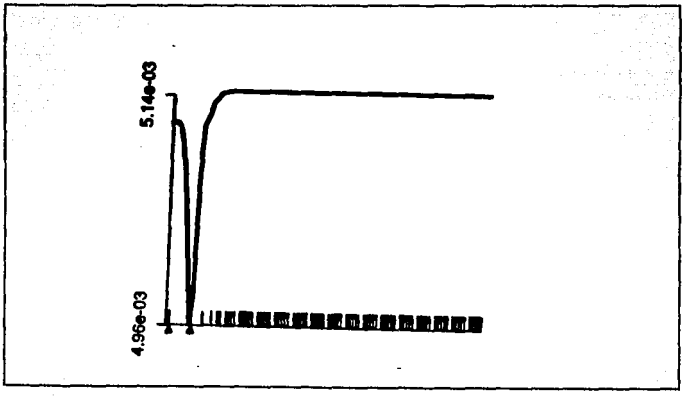


Figura 21b. Ejemplo: Cráneos. Comportamiento del índice Natural de Hermite, usando polinomios de grado 7.

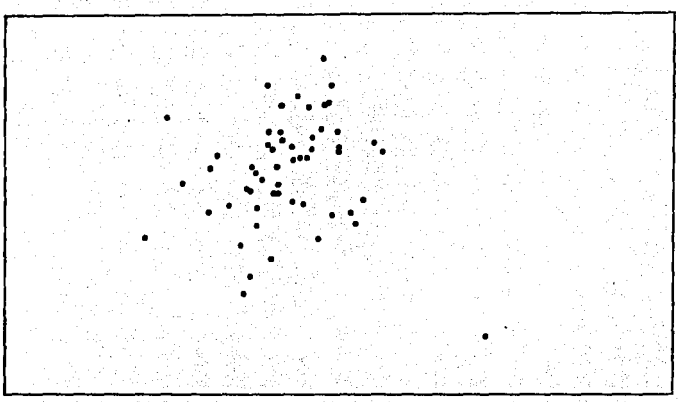


Figura 22a. Ejemplo: Cráneos. Gráfica de los puntos proyectados en dimensión dos utilizando el índice *Central Mass*.

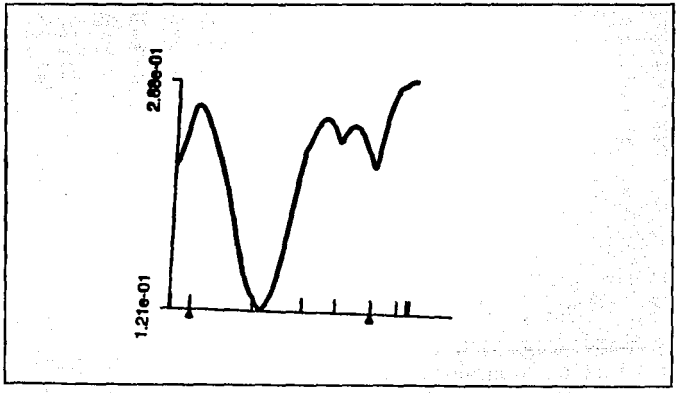


Figura 22b. Ejemplo: Cráneos. Comportamiento del índice *Central Mass* al proyectar en dos dimensiones.

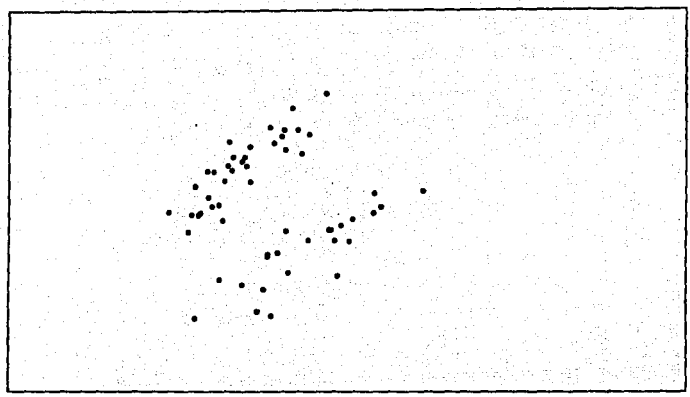


Figura 23a. Ejemplo: Cráneos. Gráfica de los puntos proyectados en dimensión dos utilizando el índice *Holes*.

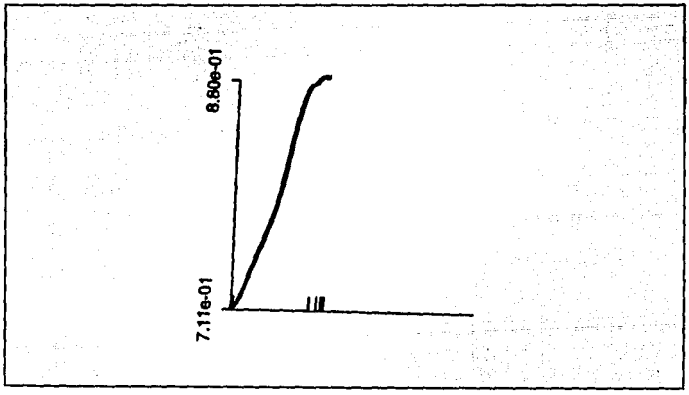


Figura 23b. Ejemplo: Cráneos. Comportamiento del índice *Holes* al proyectar en dos dimensiones.

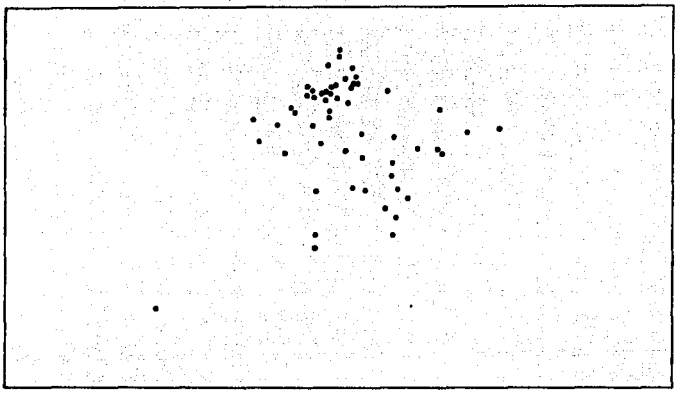


Figura 24a. Ejemplo: Cráneos. Gráfica de los puntos proyectados en dimensión dos utilizando el índice *Skewness*.

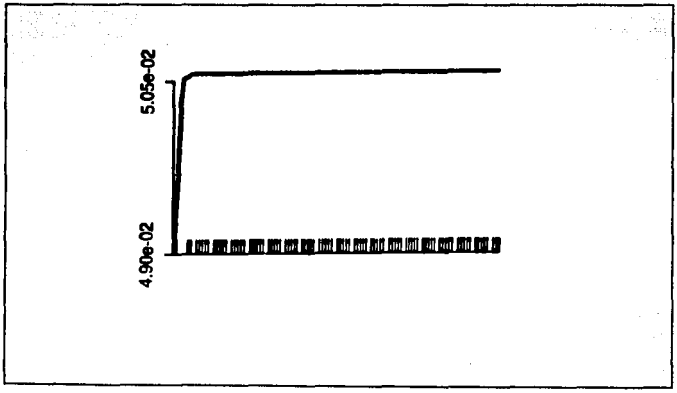


Figura 24b. Ejemplo: Cráneos. Comportamiento del índice *Skewness*.

Las proyecciones que permiten distinguir mejor la separación de los dos grupos de cráneos son los correspondientes a los índices siguientes: Friedman-Tukey, Friedman-Tukey modificado, Entropía, Entropía modificado, Legendre con polinomios de grado 8, Hermite con polinomios de grado 0, 7 y 8; Natural de Hermite con polinomios de grado 8 y el índice llamado *Holes*. En la mayoría de las gráficas se observa un punto atípico, correspondiente al caso 34 (ver apéndice I). A fin de observar la influencia de este punto en las diferentes proyecciones, se omitió y se volvieron a optimizar algunos de los índices. Se vio que en algunos de los índices como el de Friedman-Tukey o el de Entropía, este punto no tenia demasiada influencia, sin embargo en algunos otros como el Natural de Hermite se obtuvieron proyecciones diferentes. En las figuras 25 y 26 se presentan las gráficas correspondientes a las proyecciones obtenidas sin el punto atípico utilizando los índices de Entropía y Natural de Hermite respectivamente.

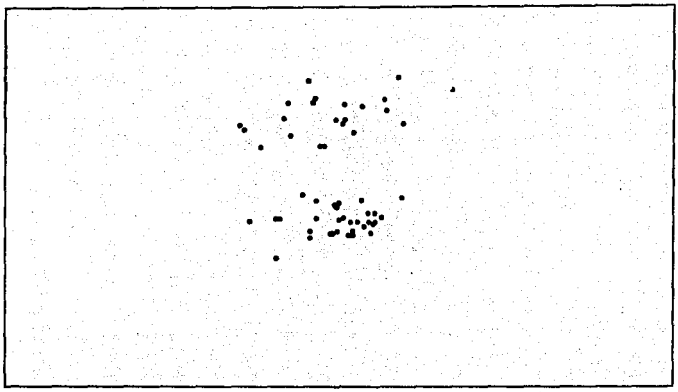


Figura 25a. Ejemplo: Cráneos. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Entropía sin el punto atípico.

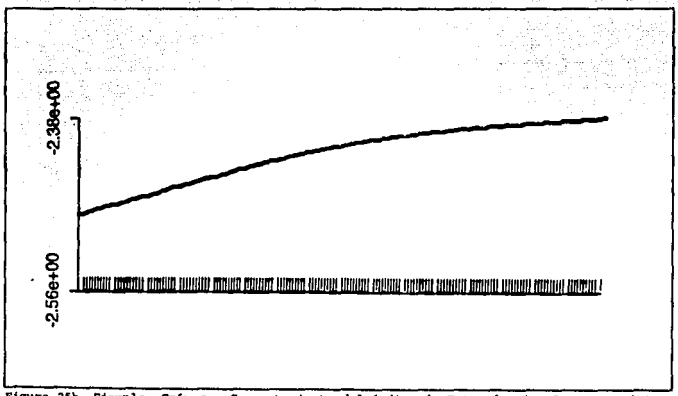


Figura 25b. Ejemplo: Cráneos. Comportamiento del índice de Entropía sin el punto atípico.

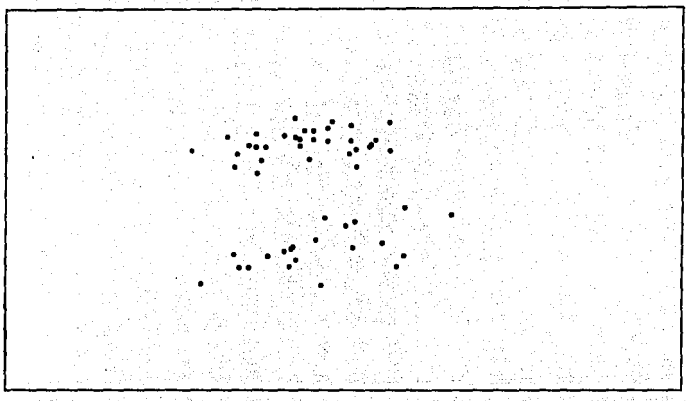


Figura 26a. Ejemplo: Cráneos. Gráfica de los puntos proyectados en dimensión dos utilizando el índice Natural de Hermite con polinomios de grado 0 sin el punto atípico.

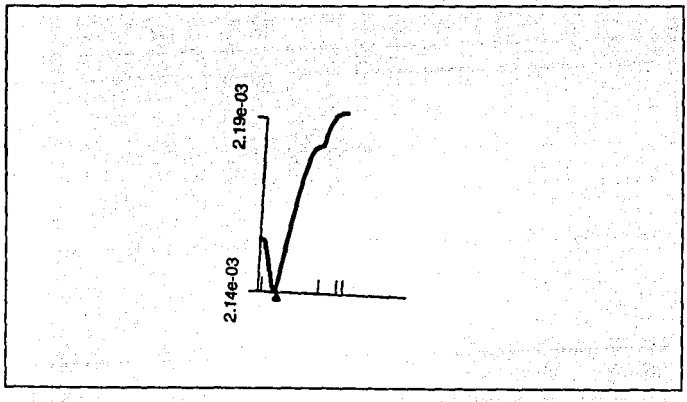


Figura 26b. Ejemplo: Cráneos. Comportamiento del índice Natural de Hermite usando polinomios de grado 0 sin el punto atípico.

En la tabla 3 se presentan, a manera de resumen, los valores de los índices que optimizaron cada una de las proyecciones presentadas.

Nombre del Índice	Característica	Valor óptimo
Friedman-Tukey	Original	1.13e-01
Friedman-Tukey	Modificado	1.12e-01
Entropía	Original	-2.37e+00
Entropía	Modificado	-2.39e+00
Legendre	Polinomios de grado 1	2.39e-02
Legendre	Polinomios de grado 2	1.07e-01
Legendre	Polinomios de grado 7	4.70e-01
Legendre	Polinomios de grado 8	5.38e-01
Hermite	Polinomios de grado 0	7.19e-04
Hermite	Polinomios de grado 1	1.09e-02
Hermite	Polinomios de grado 7	1.15e-01
Hermite	Polinomios de grado 8	1.28e-01
Natural de Hermite	Polinomios de grado 0	2.17e-03
Natural de Hermite	Polinomios de grado 1	2.55e-03
Natural de Hermite	Polinomios de grado 7	5.14e-03
Holes	No se presenta teóricamente	8.80e-01
Central Mass	No se presenta teóricamente	2.88e-01
Skewness	No se presenta teóricamente	5.05e-02
Entropía	Sin punto atípico	-2.38e+00
Natural de Hermite	Polinomios de grado 0, sin punto atípico	2.19e-03

Tabla 3. Ejemplo: Cráneos. Valor que optimiza el índice de proyección.

En el caso de los índices que se basan en desarrollos de polinomios para estimar las funciones de densidad, se estuvo experimentando con diferentes grados de polinomios, observando que después de un cierto exponente, las proyecciones de los puntos eran similares, debido a lo cual se decidió presentar los dos índices con polinomios de grado más bajo (1y 2 ó 0 y 1) y dos de grado alto (7 y 8).

4.2 Plantas

En esta sección se retoma un ejemplo reportado por Eslava Gómez G. (1993), donde se emplea un subconjunto de datos que son parte de un amplio estudio realizado por Dantas (1989) en la región del Amazonas en Brasil. El objeto del estudio es analizar la forma en que se regenera la vegetación nativa después de que utilizar procesos de poda y quema en una cierta área. Los datos fueron recolectados en los años 1979, 1980, 1981 y 1986, en cuatro sitios diferentes en Para, Brasil. Las medidas representan el número de plantas de una familia particular en alguno de los cuatro sitios y en un tamaño específico. Los datos originales se encuentran en el apéndice I de este trabajo.

Debido a que el interés se centra en tratar de localizar grupos homogéneos de familias de acuerdo a su forma de regeneración, más que por su abundancia, se decidió trabajar con proporciones, y no con el número de plantas observadas. Se puede pensar que cada observación en los cuatro años representa un "perfil". Por ejemplo para la primera observación Annonaceae que corresponde a 4, 3, 1 y 34 su "perfil" es (4/42, 3/42, 1/42, 34/42) ya que 4+3+1+34=42.

Si tratáramos de dibujar los "perfiles" en una sola gráfica, los "perfiles" se sobreponen y son difíciles de distinguir. Por ejemplo, en la figura 27 se presenta una gráfica de los perfiles de las primeras 10 familias de plantas.

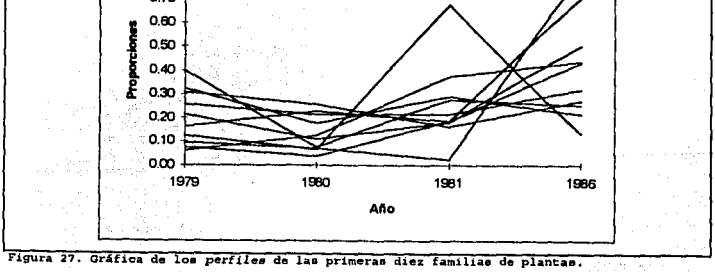


Figura 27. Gráfica de los *perfiles* de las primeras diez familias de plantas.

Este caso es un ejemplo donde es necesario emplear alguna técnica del análisis multivariado para distinguir posibles estructuras de grupos en los datos. En particular alguna proyección lineal de los datos mediante la técnica de análisis de componentes principales o mediante algún índice de proyecciones perseguidas.

Debido a que las variables son proporciones, es claro que teniendo las tres primeras observaciones, la cuarta es una combinación lineal de estas tres. Por lo tanto, en virtud de esta dependencia lineal, se decidió trabajar sólo con las tres primeras proporciones debido a que este conjunto transmite la misma información en términos de varianza.

4.2.1 Componentes Principales

Considerando que las variables observadas son los años y los casos las diferentes familias de plantas, al realizar el análisis de componentes principales se obtuvo en primer lugar la matriz de correlación siguiente:

	1979	1980	1981
1979	1.00000		
1980	-.13850	1.00000	
1981	-.23274	.12869	1.00000

El determinante de esta matriz de correlación tiene el valor de 0.92. Se observa que las corrrelaciones entre las variables son bajas.

En la tabla 4 se muestran los *eigenvalores* de la matriz de correlación así como el porcentaje de varianza explicada. Se puede observar que los dos primeros componentes explican cerca del 75% de la variación de los datos.

Componente	Eigenvalor	Pct de Var	Pct Acum.
1	1.33828	44.6	44.6
2	.89476	29.8	74.4
3	.76695	25.6	100.0

Tabla 4. Eigenvalores obtenidos y su porcentaje de varianza explicada.

Por otra parte, en la tabla 5 se presentan los valores de las correlaciones entre las variables originales y los componentes principales, observando una alta correlación de lo acontecido en el año 1977 y 1981 con el primer componente, mientras que el año 1980 presenta una correlación alta con el segundo componente.

Año	Componente 1	Componente 2	Componente 3
1977	-.71896	.29430	.62967
1980	.56418	.82474	.03876
1981	.70928	-.35770	.60743

Tabla 5. Correlaciones de las variables originales con los componentes principales.

Los argumentos anteriores se representan gráficamente en el denominado círculo de correlaciones, presentado en la figura 28 y en el cual observamos que el proceso de lo que sucede en el año 1977 con las familias de plantas es inverso a lo que sucede en el año de 1981. También observamos la alta correlación de las variables originales con los dos primeros componentes principales.

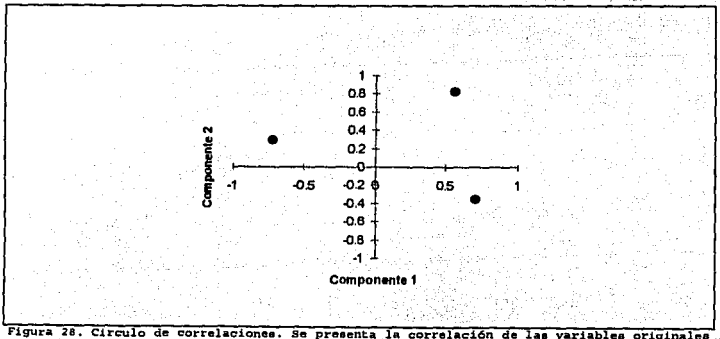


Figura 28. Círculo de correlaciones. Se presenta la correlación de las variables originales (años) con los primeros dos componentes principales.

En la figura 29 se presenta la gráfica de los dos primeros componentes principales calculados a partir de las tres proporciones tomadas como variables originales. En este caso,

observamos que los componentes principales no resultan ser de gran ayuda para detectar posibles estructuras agrupadas en los datos.

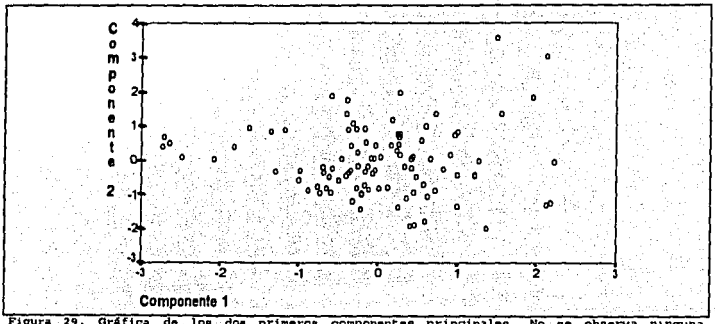


Figura 29. Gráfica de los dos primeros componentes principales. No se observa ninguna estructura de grupos en los datos.

4.2.2 Proyecciones Perseguidas

Para las proyecciones perseguidas utilizamos, otra vez, el programa XGoby, para encontrar diferentes proyecciones de los puntos, tomando como base los datos esferados (es una opción del programa). Se incluye además una gráfica de proyección utilizando el índice de la distancia radial media, PNN, propuesto por Eslava y Marriott. Ver Eslava (1989, pag. 119). Las representaciones gráficas obtenidas son las siguientes:

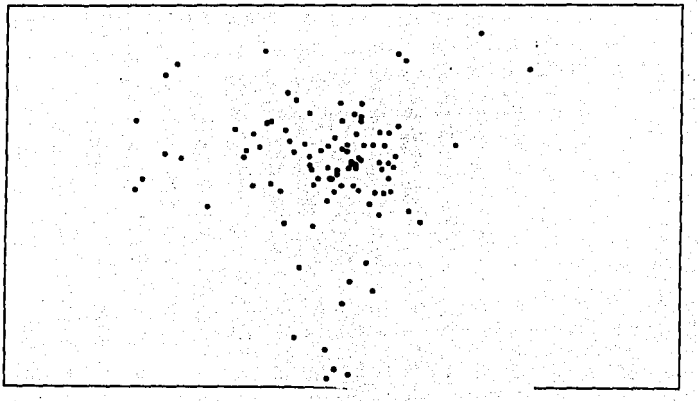


Figura 30a. Ejemplo: Plantas. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Friedman-Tukey.

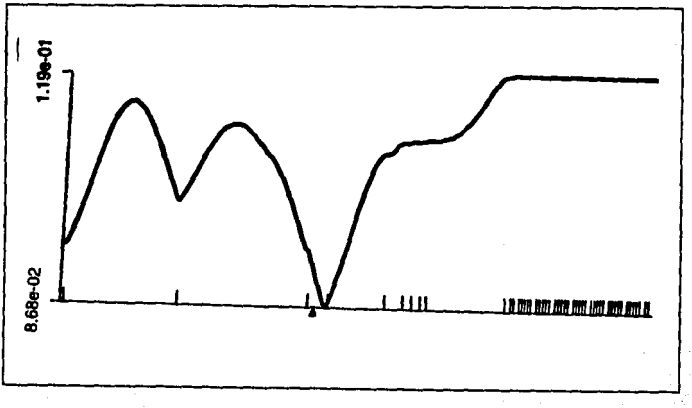


Figura 31b.Ejemplo: Plantas. Comportamiento del índice de Friedman-Tukey al proyectar en dos dimensiones.

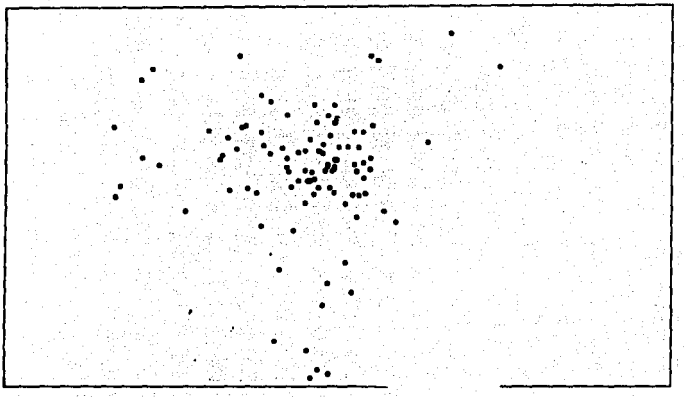


Figura 31a. Ejemplo: Plantas. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Friedman-Tukey Modificado.

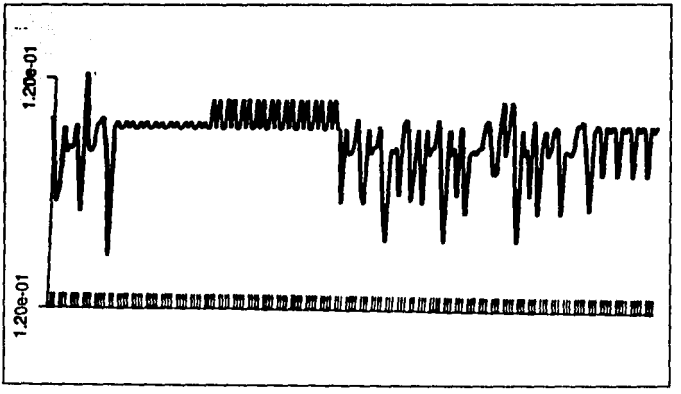


Figura 31b. Ejemplo: Plantas. Comportamiento del índice de Friedman-Tukey Modificado al proyectar en dos dimensiones.

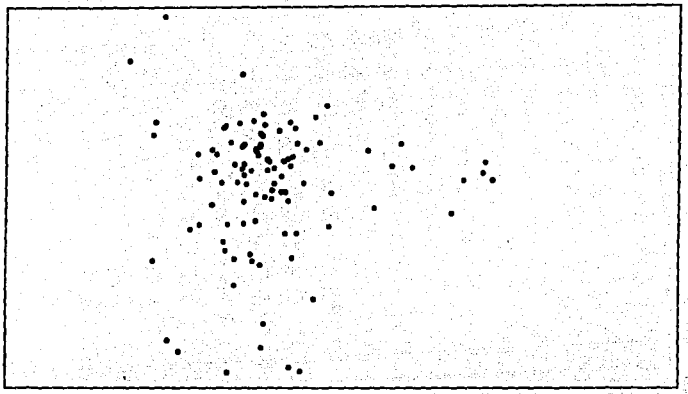


Figura 32a. Ejemplo: Plantas. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Entropía.

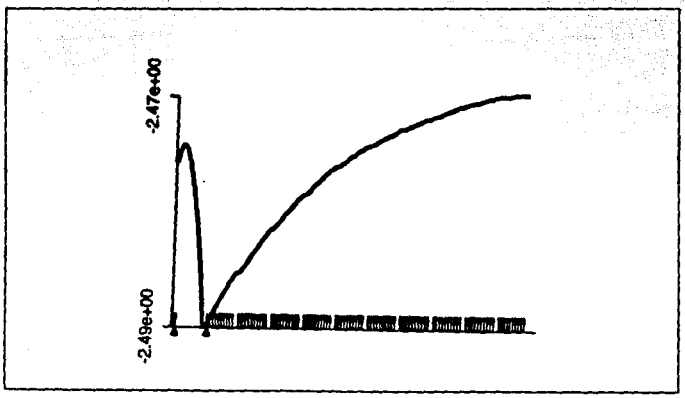


Figura 32b. Ejemplo: Plantas. Comportamiento del índice de Entropía al proyectar en dos dimensiones.

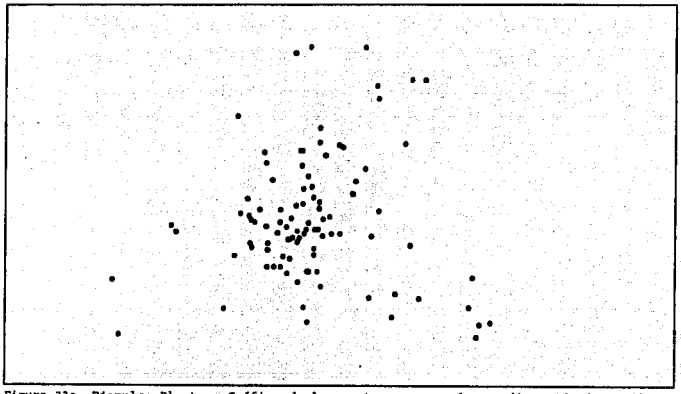


Figura 33a. Ejemplo: Plantas. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Entropia Modificado.

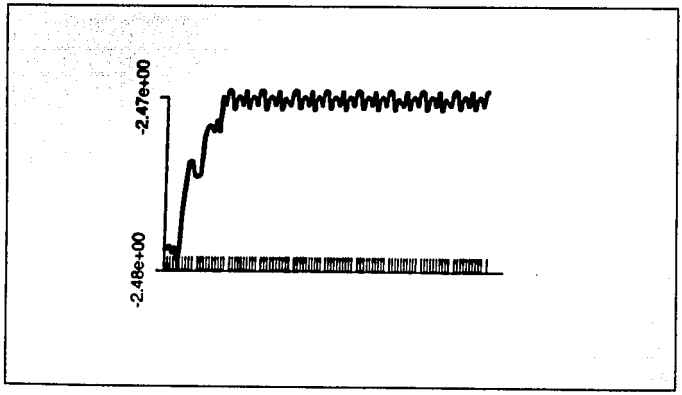


Figura 33b. Ejemplo: Plantas. Comportamiento del índice de Entropia Modificado al proyectar en dos dimensiones.

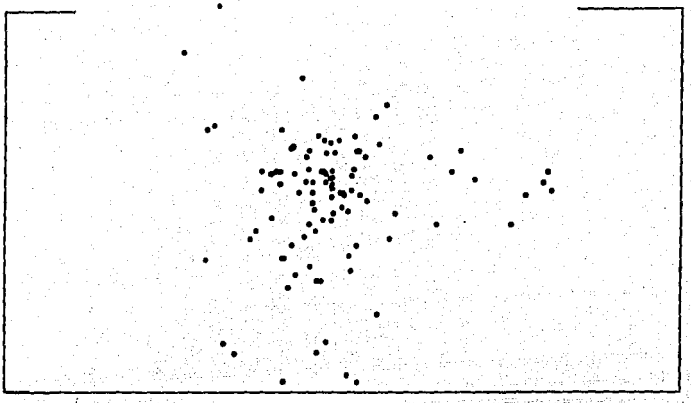


Figura 34a. Ejemplo: Plantas. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Legendre usando polinomios de grado 1.

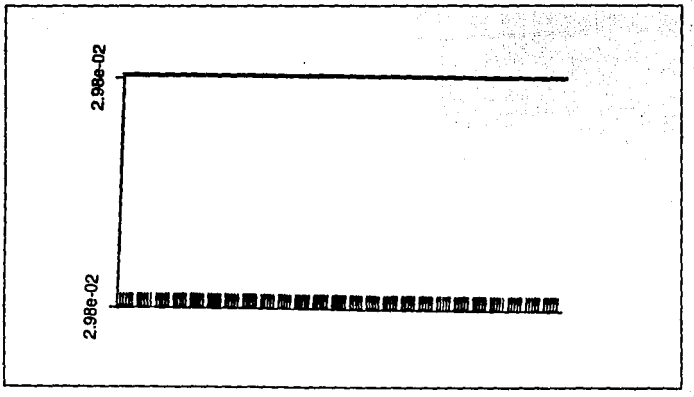


Figura 34b. Ejemplo: Plantas. Comportamiento del índice de Legendre usando polinomios de grado 1.

Figura 35a. Ejemplo: Plantas. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Legendre usando polinomios de grado 2.

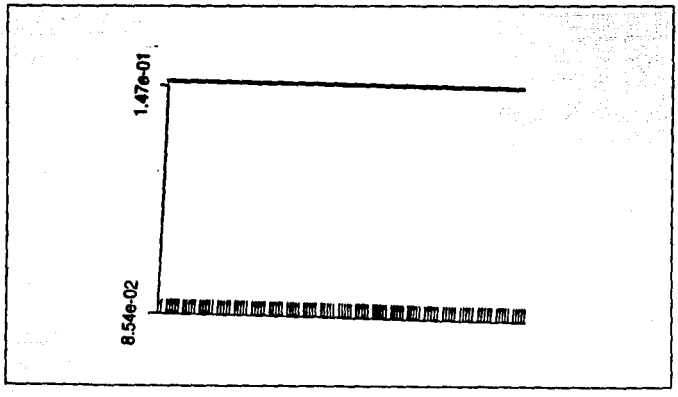


Figura 35b. Ejemplo: Plantas. Comportamiento del índice de Legendre usando polinomos de grado 2.

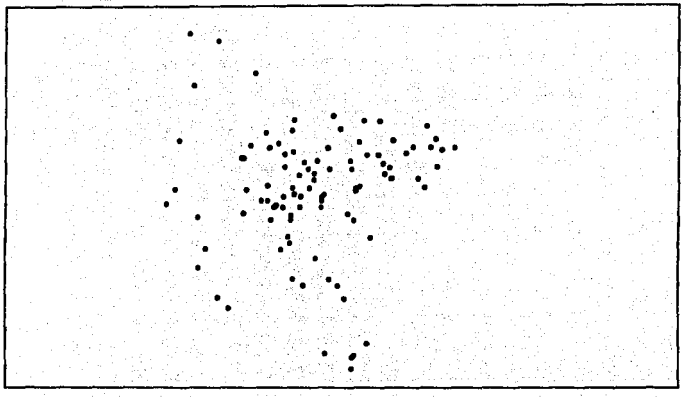


Figura 36a. Ejemplo: Plantas. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Legendre con un polinomios de grado 7.

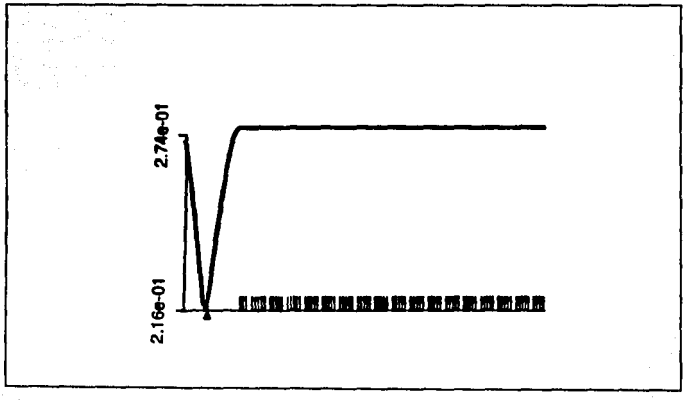


Figura 36b. Ejemplo: Plantas. Comportamiento del índice de Legendre usando polinomios de grado 7.

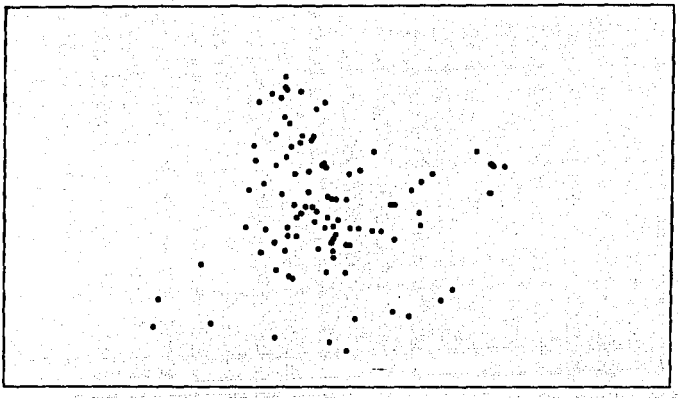


Figura 37a. Ejemplo: Plantas. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Legendre, usando polinomios de grado 8.

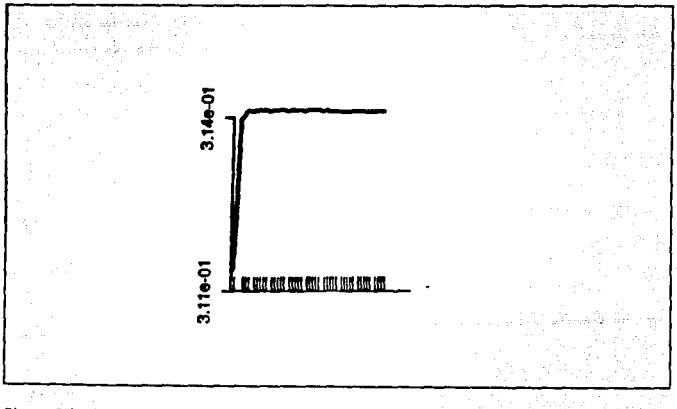


Figura 37b. Ejemplo: Plantas. Comportamiento del índice de Legendre, usando polinomios de grado 8.

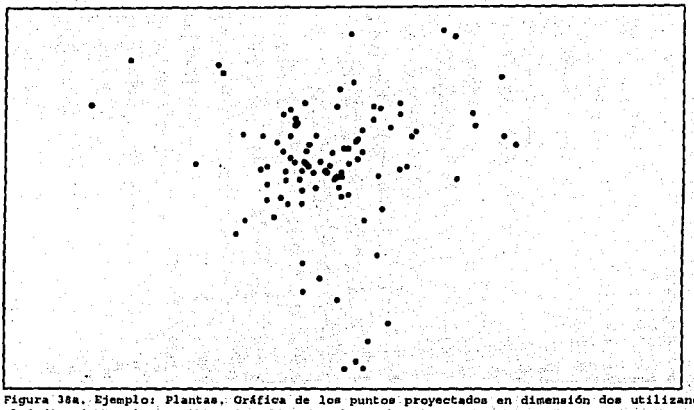


Figura 38a. Ejemplo: Plantas. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Hermite, utilizando polinomios de grado 0.

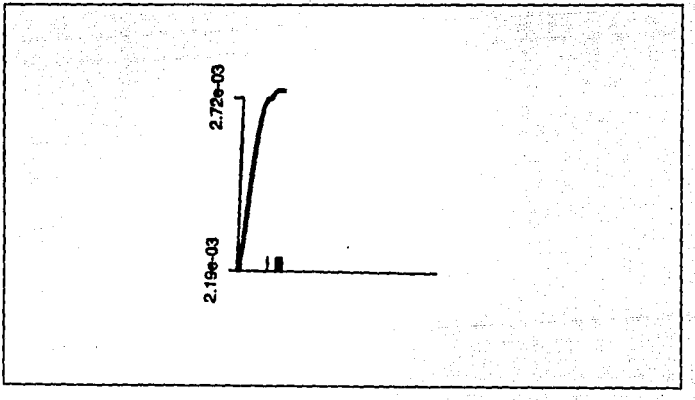


Figura 38b. Ejemplo: Plantas. Comportamiento del índice de Hermite, utilizando polinomios de grado cero.

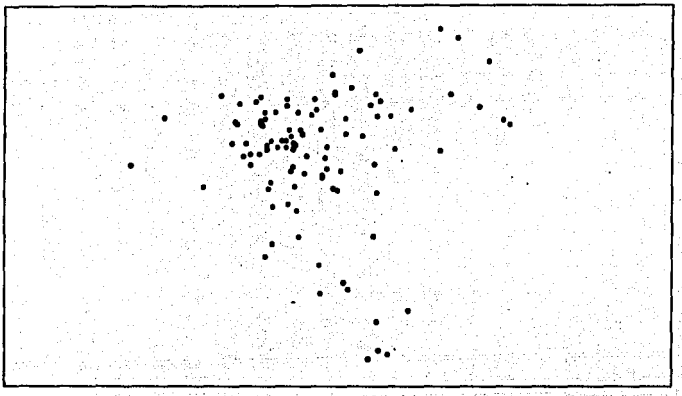


Figura 39a. Ejemplo: Plantas. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Hermite, con polinomios de grado 1.

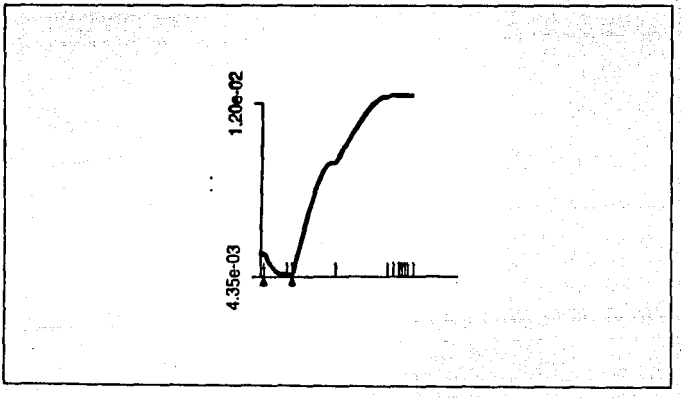


Figura 39b. Ejemplo: Plantas. Comportamiento del índice de Hermite, usando polinomios de grado 1.

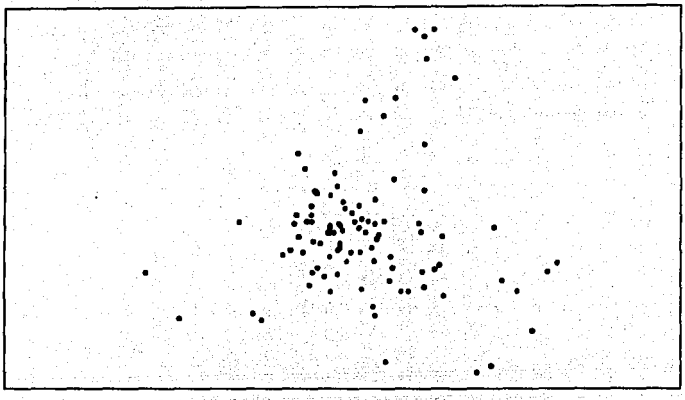


Figura 40a. Ejemplo: Plantas. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Hermite, con polinomios de grado 7.

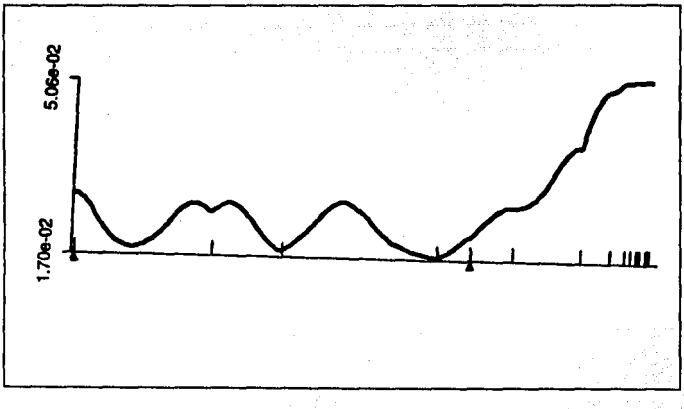


Figura 40b. Ejemplo: Plantas. Comportamiento del índice de Hermite, usando polinomios de grado 7.

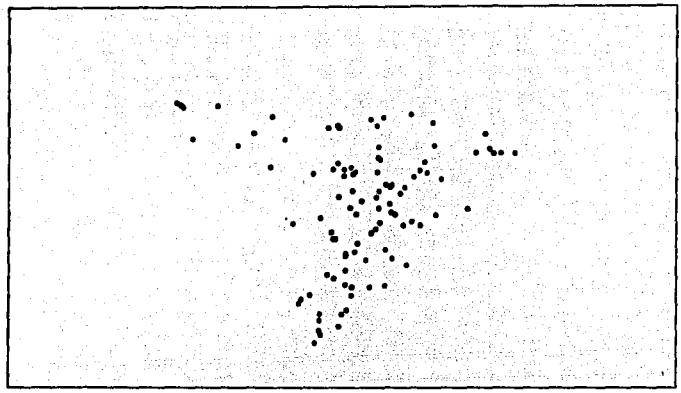


Figura 41a. Ejemplo: Plantas. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Hermite, con polinomios de grado 8.

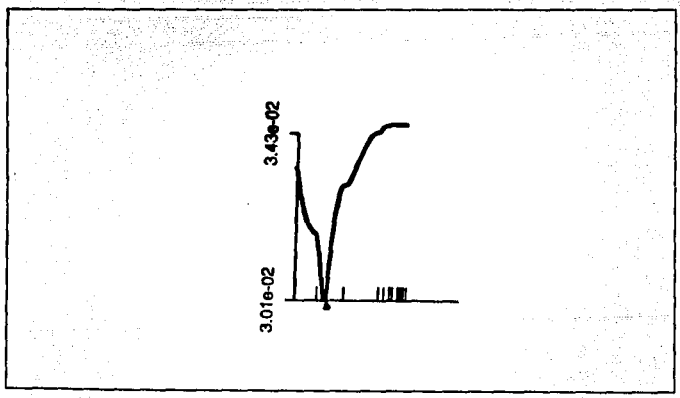


Figura 41b. Ejemplo: Plantas. Comportamiento del índice de Hermite, usando polinomios de grado 8.

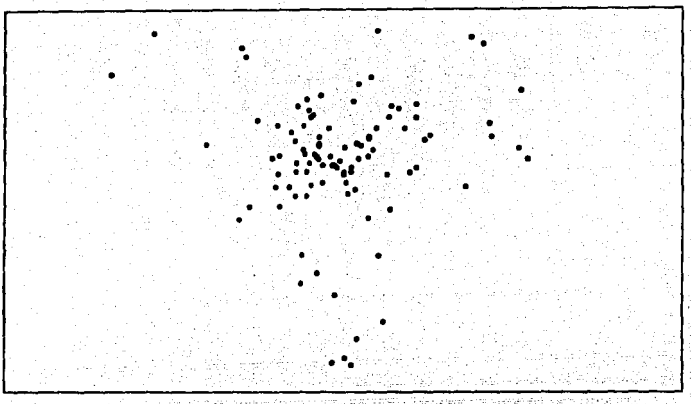


Figura 42a. Ejemplo: Plantas. Gráfica de los puntos proyectados en dimensión dos utilizando el índice Natural de Hermite, utilizando polinomios de grado 0.

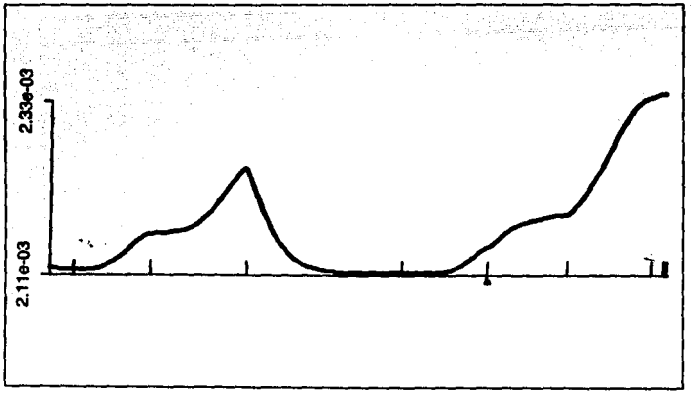


Figura 42b. Ejemplo: Plantas. Comportamiento del índice Natural de Hermite, utilizando polinomios de grado cero.

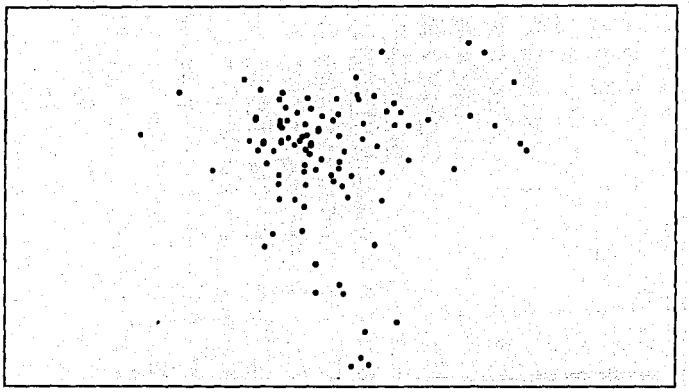


Figura 43a. Ejemplo: Plantas. Gráfica de los puntos proyectados en dimenaión dos utilizando el índice Natural de Hermite, con polinomios de grado 1.

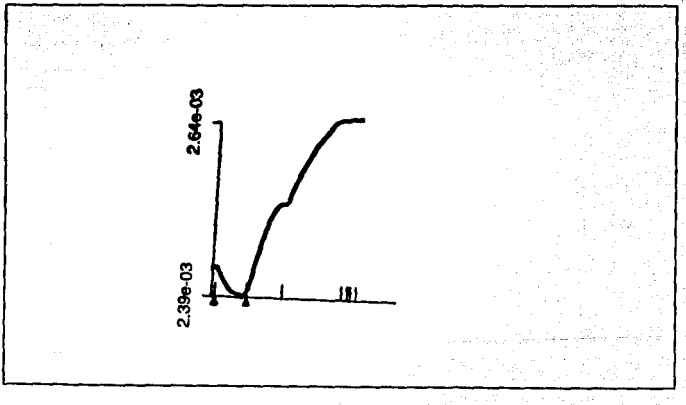


Figura 43b. Ejemplo: Plantas. Comportamiento del índice Natural de Hermite, usando polinomios de grado 1.

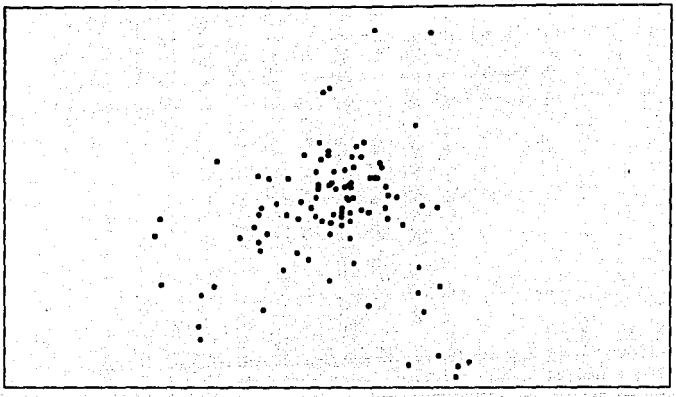


Figura 44a. Ejemplo: Plantas. Gráfica de los puntos proyectados en dimensión dos utilizando el índice Natural de Hermite, con polinomios de grado 7.

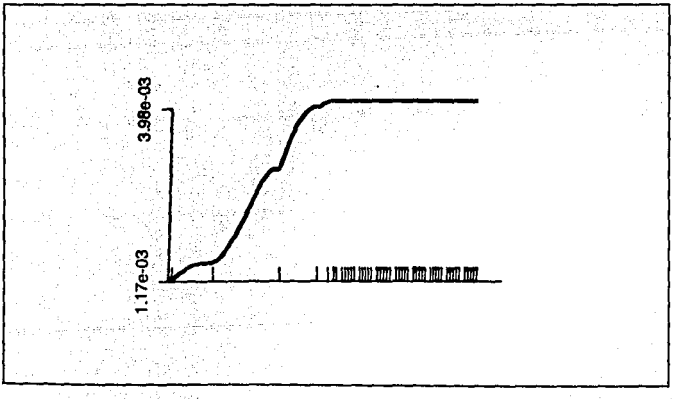


Figura 44b. Ejemplo: Plantas. Comportamiento del índice Natural de Hermite, usando polinomios de grado 7.

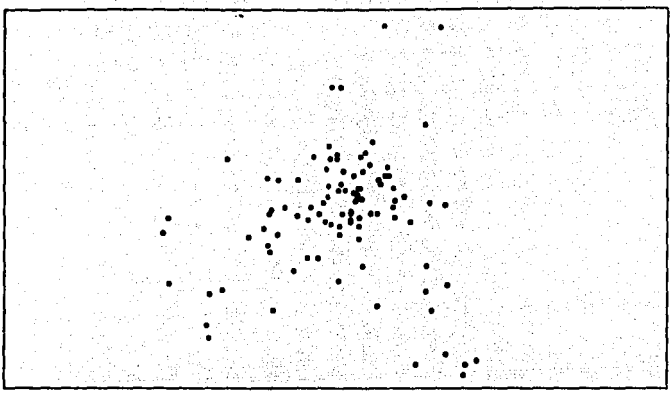


Figura 45a. Ejemplo: Plantas. Gráfica de los puntos proyectados en dimensión dos utilizando el índice *Central Mass*.

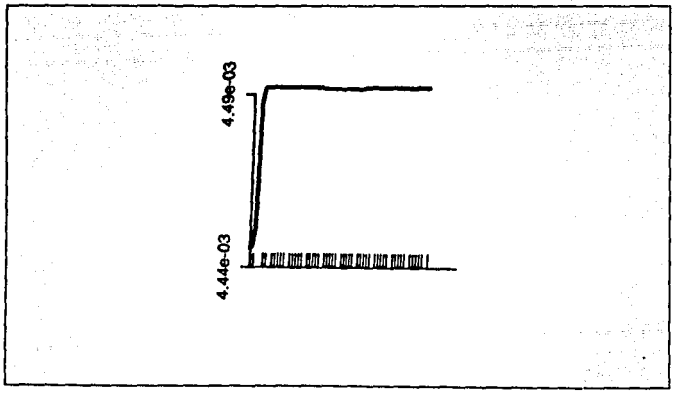


Figura 45b. Ejemplo: Plantas. Comportamiento del índice *Central Mass* al proyectar en dos dimensiones.

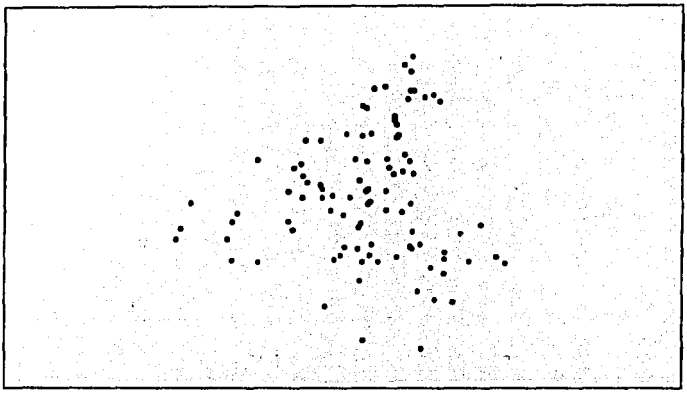


Figura 46a. Ejemplo: Plantas. Gráfica de los puntos proyectados en dimensión dos utilizando el índice *Holes*.

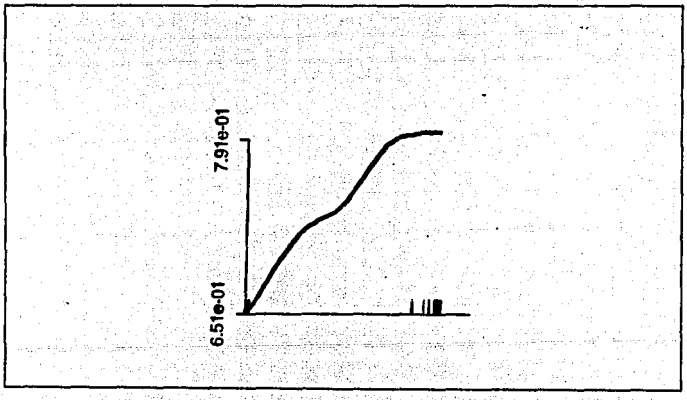


Figura 46b. Ejemplo: Plantas. Comportamiento del índice *Holes* al proyectar en dos dimensiones.

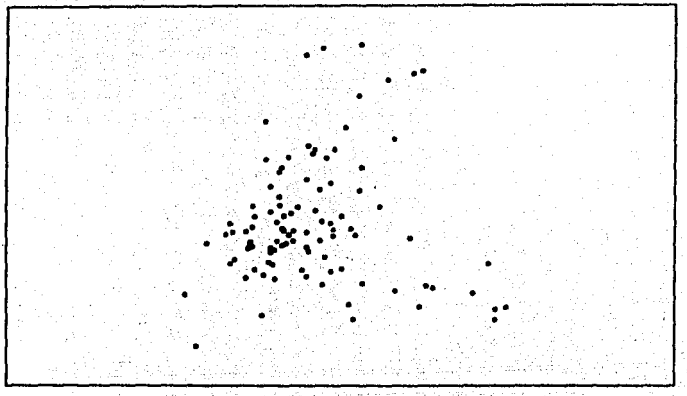


Figura 47a. Ejemplo: Plantas. Gráfica de los puntos proyectados en dimensión dos utilizando el índice *Skewness*.

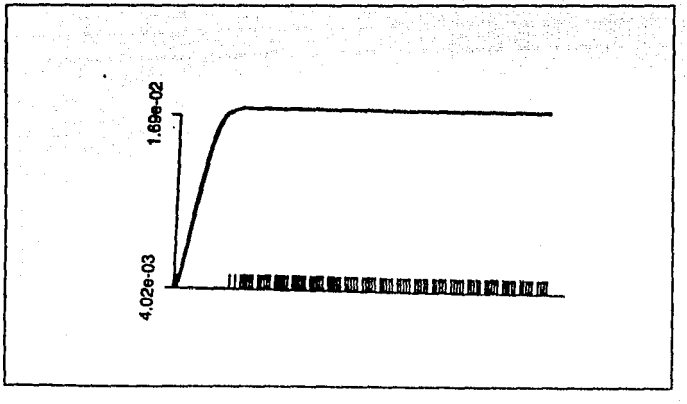


Figura 47b. Ejemplo: Plantas. Comportamiento del índice *Skewness*.

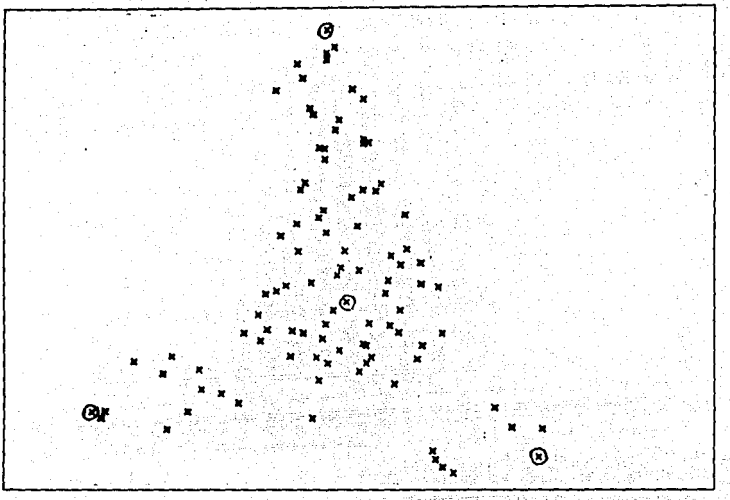


Figura 48. Ejemplo: Plantas. Gráfica de los puntos proyectados en dimensión dos utilizando el índice PNN propuesto por Eslava y Marriott.Valor que optimiza el índice PNN = 0.0327.

En la mayoría de las gráficas en aparencia no se visualizan posibles grupos en los datos. Sin embargo en la última gráfica correspondiente a la proyección de los datos optimizando el índice PNN, se observan cuatro posibles grupos.

En la tabla 6 se presentan, a manera de resumen, los valores de los índices que optimizaron la proyección.

Nombre del Índice	Característica	Valor óptimo
Friedman-Tukey	Original	1.19e-01
Friedman-Tukey	Modificado	1.20e-01
Entropía	Original	-2.47e+00
Entropía	Modificado	-2.47e+00
Legendre	Polinomios de grado 1	2.98e-02
Legendre	Polinomios de grado 2	1.47e-01
Legendre	Polinomios de grado 7	2.74e-01
Legendre	Polinomios de grado 8	3.14e-01
Hermite	Polinomios de grado 0	2.72e-03
Hermite	Polinomios de grado 1	1.20e-02
Hermite	Polinomios de grado 7	5.06e-02
Hermite	Polinomios de grado 8	3.43e-02
Natural de Hermite	Polinomios de grado 0	2.33e-03
Natural de Hermite	Polinomios de grado 1	2.64e-03
Natural de Hermite	Polinomios de grado 7	3.98e-03
Natural de Hermite	Polinomios de grado 8	4.49e-03
Holes	No se presenta teóricamente	7.91e-01
Central Mass	No se presenta teóricamente	3.55e-01
Skewness	No se presenta teóricamente	1.69e-02
PNN	m = 57, iteraciones = 49	0.0327

Tabla 6. Valor del índice de proyección. Ejemplo: Plantas

4.3 Virus

Se retoma un ejemplo presentado por Eslava y Marriott (1994), donde se compara un conjunto de datos correspondiente a 61 virus, de los cuales 48 han sido clasificados en tres grupos y 13 son no-clasificados. Los grupos de clasificación son los siguientes:

Tipo	Nombre del Grupo	# de Casos
1	Hordeirvirus	3
2	Tobravirus	6
3	Tobamovirus	39
4	No-clasificados	13

A cada virus se le tomaron 18 medidas. Cada medida expresa el número de residuos de aminoácidos de moléculas y las denotaremos por las variables X1,..., X18. Se considera que todas las variables son continuas a pesar de que algunas toman pocos valores, por ejemplo X8 que toma sólo los valores 0, 1 ó 2; o bien X10 que toma los valores de 0, 1, 2, 3, 4, ó 7. Los datos originales se encuentran en el apéndice II de esta tesis.

4.3.1 Análisis de Componentes Principales

En la tabla 7 se presentan los valores obtenidos para los primeros ocho eigenvalores de la matriz de correlación, utilizando todas las variables originales.

Componente	Eigenvalor	Pct de Var	Pct Acum
1	5.95	33.1	33.1
2	2.98	16.6	49.6
3	2.34	13.0	62.7
4	1.76	9.8	72.4
5	1.16	6.5	78.9
6	1.02	5.7	84.6
7	.61	3.4	88.0
8	.46	2.6	90.6

Tabla 7. Ocho eigenvalores de la matriz de correlación, explicando más del 90% de la variabilidad de los datos.

En la tabla 8 se presentan los valores de las correlaciones entre las variables originales y los cinco primeros componentes principales.

	Componente1	Componente 2	Componente 3	Componente 4	Componente 5
X1	.77729	.10437	.00185	.28352	-.10788
X2	-.59465	-.36358	.20668	.37053	-.03012
X3	-.16016	.62491	-.08723	.53303	.06217
X4	.66285	-.41346	.25593	-.23924	.23365
X5	.44484	.51342	.02800	-.38252	.30617
X6	.83616	.11462	.08595	.41492	.03821
X7	.64838	.47630	.02191	-.11977	-.04502
X8	.12222	-.26109	.52235	.19634	.70257
X9	.55930	-.32583	-.49801	.33409	.27689
X10	-.19494	.10006	.77890	-.18169	.13122
X11	.44068	-.60691	-.35646	.32154	.10999
X12	.75811	.01993	.09157	-.06692	-.42329
X13	.39189	-.09029	.61289	.31824	-.41956
X14	-.23082	.62984	-.59459	.02369	.13922
X15	.94227	.08526	.16999	.11594	-.02319
X16	.45304	.74938	.21919	-.10004	.13966
X17	.82601	-.20705	-.37263	-.21919	.01704
X18	.30604	-.45187	-.11633	-.63577	-.06402

Tabla 8. Correlaciones de las variables originales con los primeros cinco componentes principales.

En la figura 49 se presenta la gráfica del denominado circulo de correlaciones, en el cual se observa la correlación entre las variables originales con los dos primeros componentes principales.

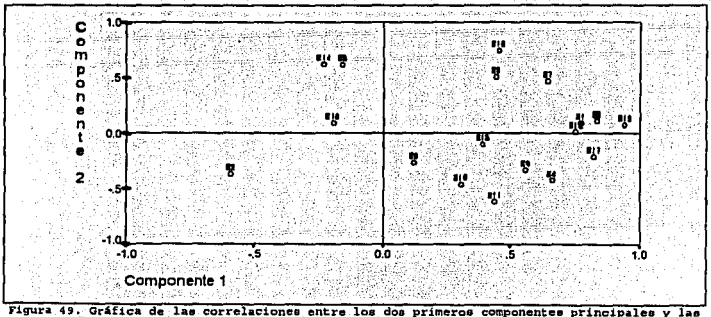


Figura 49. Gráfica de las correlaciones entre los dos primeros componentes principales y las variables originales.

En esta figura se observa, entre otros aspectos, que existe una alta correlación entre las variables X1,X6, X12, X15 y X17 con el primer componente principal. Por otra parte las

variables X8 y X10 presentan correlación baja con los dos primeros componentes principales. La interpretación real de estos resultados deberá darse conjuntamente con un investigador del área de estudio.

A fin de visualizar posibles estructuras en los datos, en la figura 50 se presenta la gráfica de las observaciones evaluados en los dos primeros componentes principales.

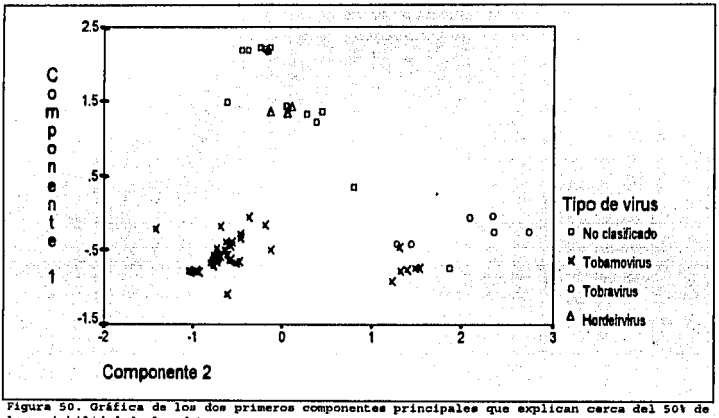


Figura 50. Gráfica de los dos primeros componentes principales que explican cerca del 50% de la variabilidad de los datos.

En la gráfica se observa que la familia de virus denominada Tobamovirus se encuentra repartido en dos conglomerados: uno que contiene solo virus de este tipo y otro en el que aparentemente forma un grupo con dos virus Tobravirus y uno no clasificado. Los restantes 4 virus de la familia Tobravirus forman a su vez un grupo. Por otra parte, el grupo de los Hordevirus forma un grupo con 4 de los virus no clasificados, por lo que se podría suponer pertenecen a esta familia. De los restantes virus no clasificados la mayoría forma un grupo, mientras que dos más parecen estar aislados de cualquier estructura.

4.3.2 Proyecciones Perseguidas

A continuación presentamos las gráficas de las diferentes proyecciones obtenidas mediante la aplicación de diferentes índices de proyecciones perseguidas, utilizando como datos esferados, cuatro de los componentes principales que explican poco más de 70% de la variabilidad de los datos.

La mayoría de las gráficas se obtuvieron mediante el programa XGobi, salvo las correspondientes a las gráficas que corresponden a los índices propuestos por Eslava y Marriott en as que se presentan las gráficas que los autores obtuvieron con estos datos.

En varias de las gráficas obtenidas sobresale un caso atípico, (ver por ejemplo el índice de Entropía en la figura 52); el programa XGobi tiene una opción para identificar los puntos por lo que supimos que corresponde al caso 11. Las figuras 72 y 73 presentan dos gráficas de las proyecciones obtenidas con los índices de Entropía y de Legendre sin el punto atípico, observando algunos cambios en dichas proyecciones con respecto a las obtenidas originalmente.

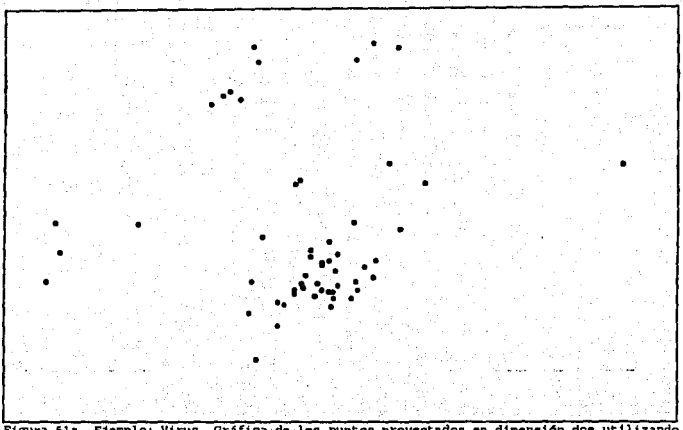


Figura 51a. Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Friedman-Tukey.

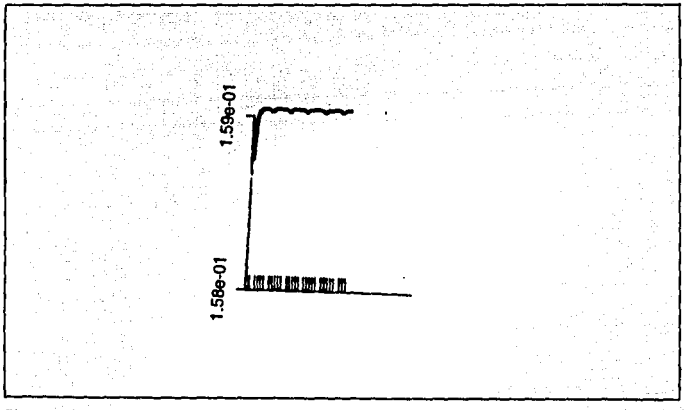


Figura 51b. Ejemplo: Virus.Comportamiento del índice de Friedman-Tukey.

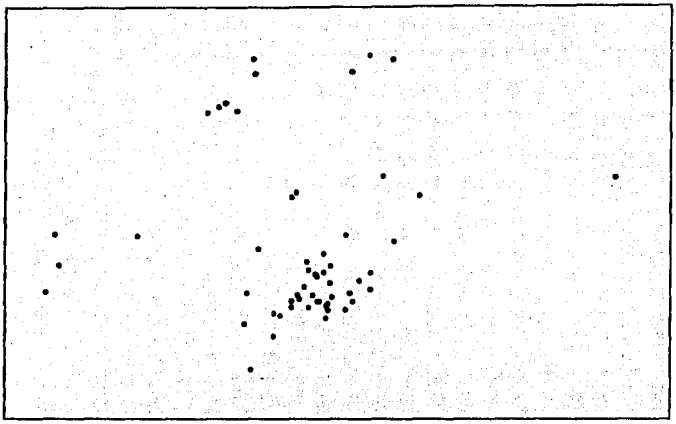


Figura 52a. Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Friedman-Tukey Modificado.

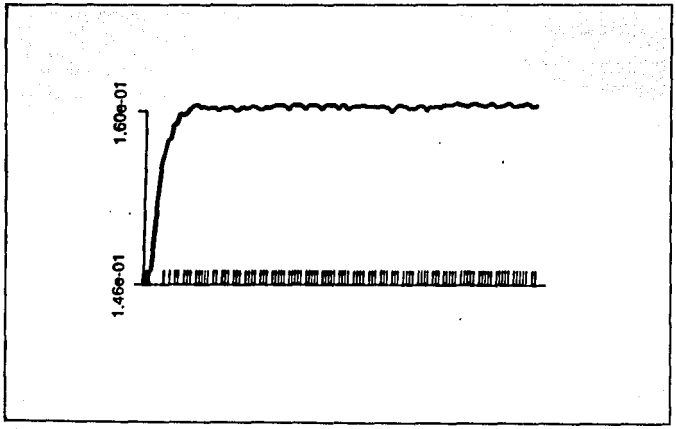


Figura 52b. Ejemplo: Virus.Comportamiento del índice de Friedman-Tukey Modificado.

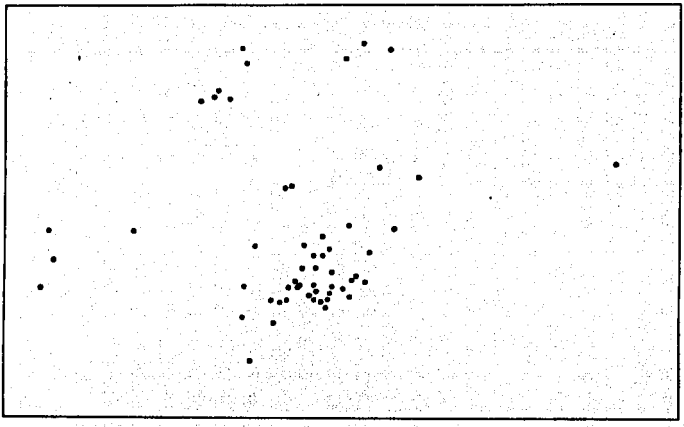


Figura 53a. Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Entropia.

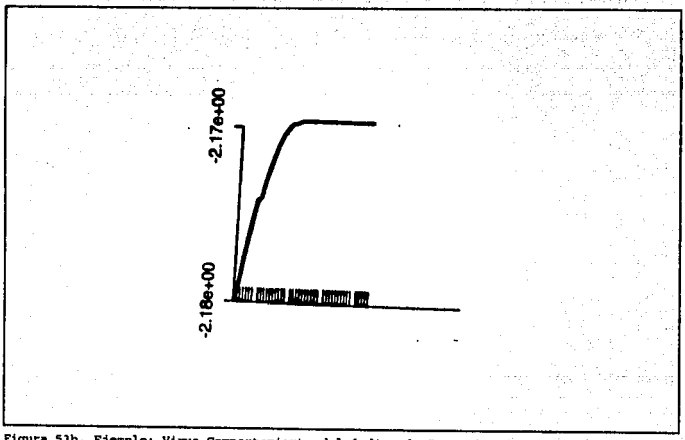


Figura 53b. Ejemplo: Virus.Comportamiento del índice de Entropia.

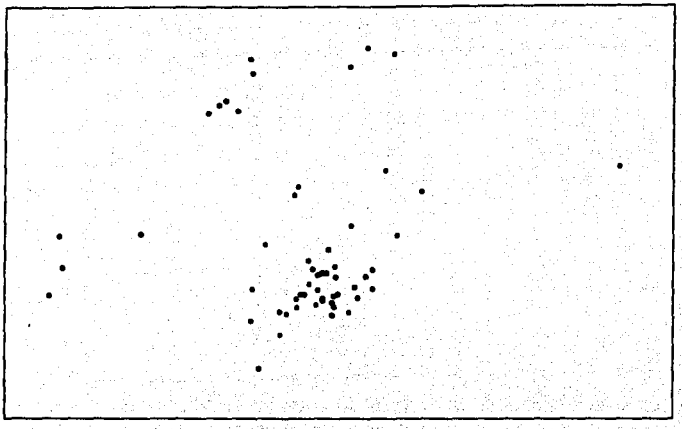


Figura 54a. Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Entropía modificado.

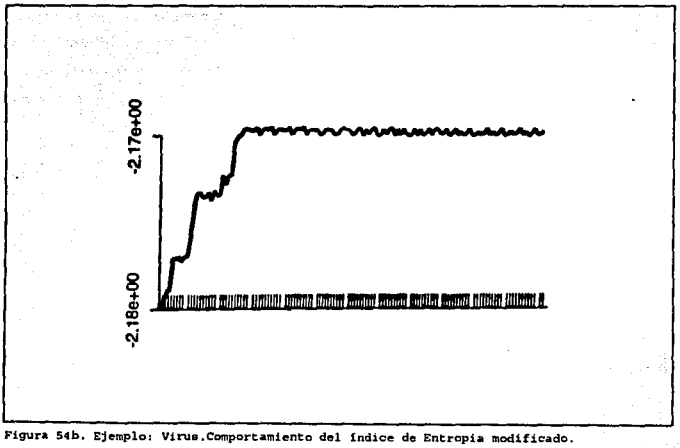


Figura 54b. Ejemplo: Virus.Comportamiento del índice de Entropía modificado.

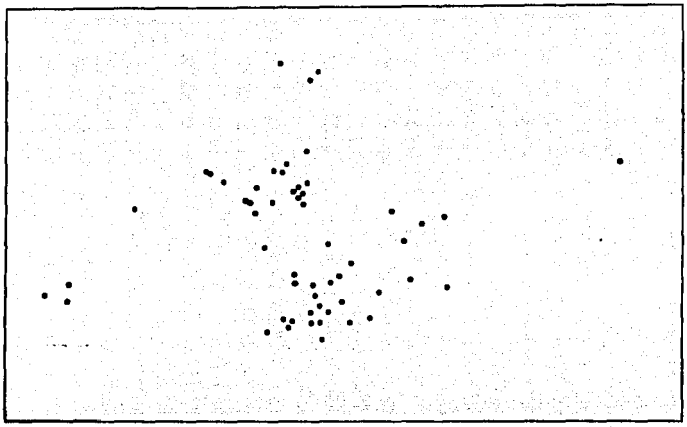


Figura 55a. Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Legendre con polinomios de orden 1.

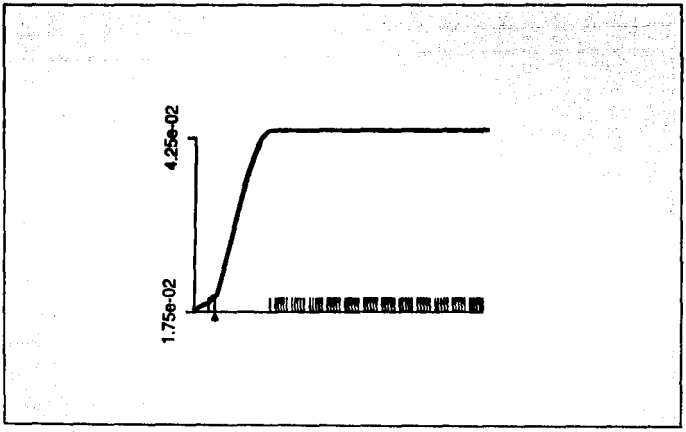


Figura 55b. Ejemplo: Virus.Comportamiento del índice de Legendre, usando polinomios de orden 1

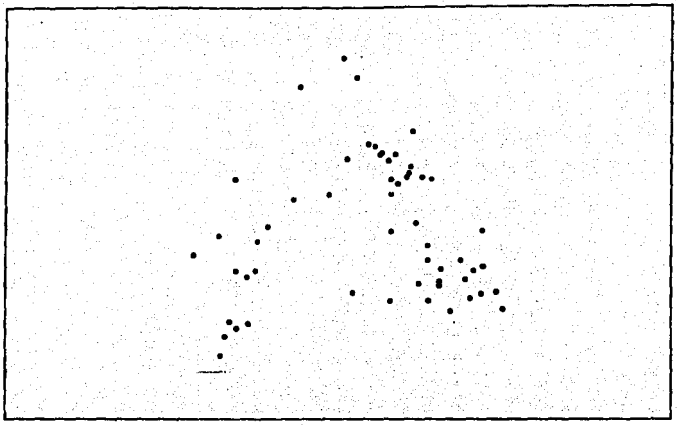


Figura 56a.Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Legendre con polinomios de grado 2.

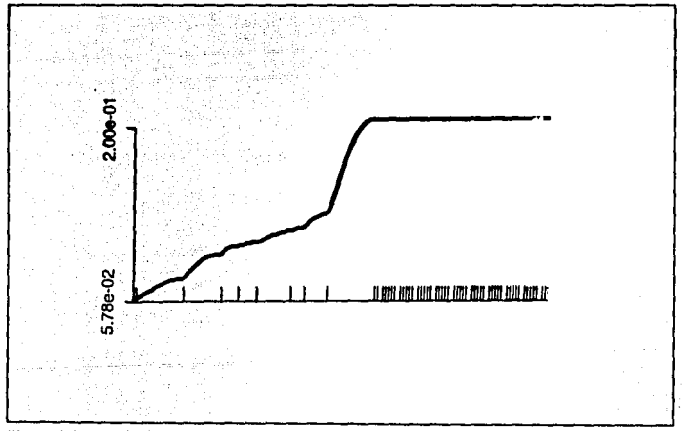


Figura 56b.Ejemplo: Virus.Comportamiento del índice de Legendre con polinomios de grado 2.

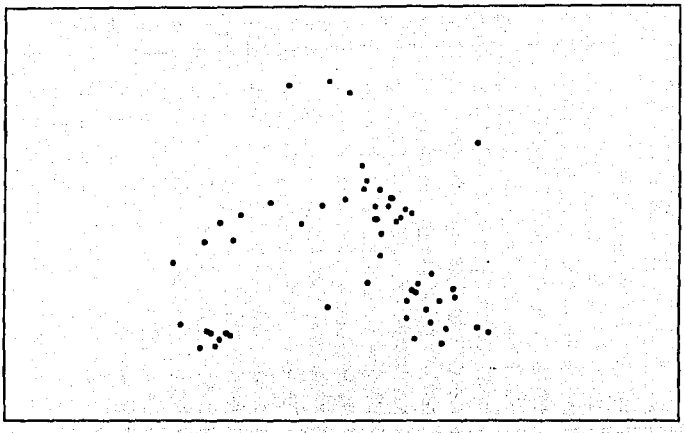


Figura 57a. Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Legendre con polinomios de grado 7.

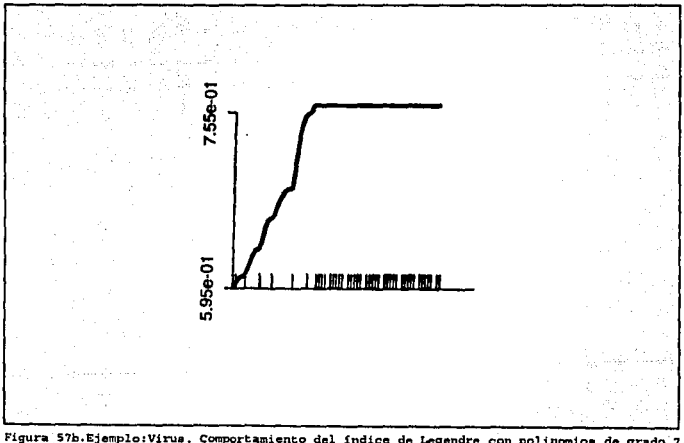


Figura 57b.Ejemplo:Virus. Comportamiento del índice de Legendre con polinomios de grado 7.

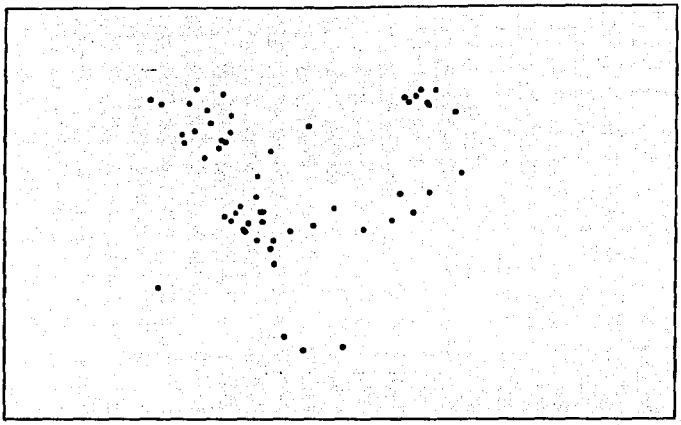


Figura 58a.Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Legendre con polinomios de grado 8.

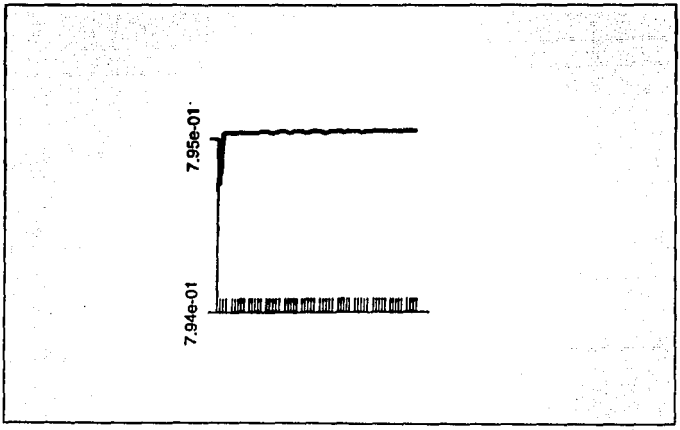


Figura 58b.Ejemplo: Virus.Comportamiento del índice de Legendre con polinomios de grado 8.

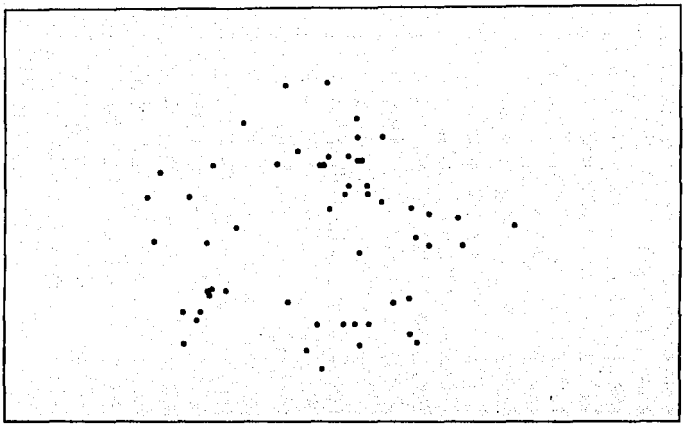


Figura 59a.Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Hermite con polinomios de grado cero.

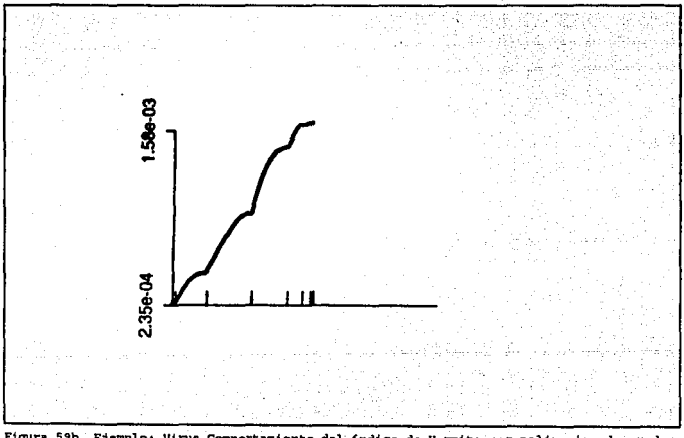


Figura 59b. Ejemplo: Virus.Comportamiento del índice de Hermite con polinomios de grado cero.

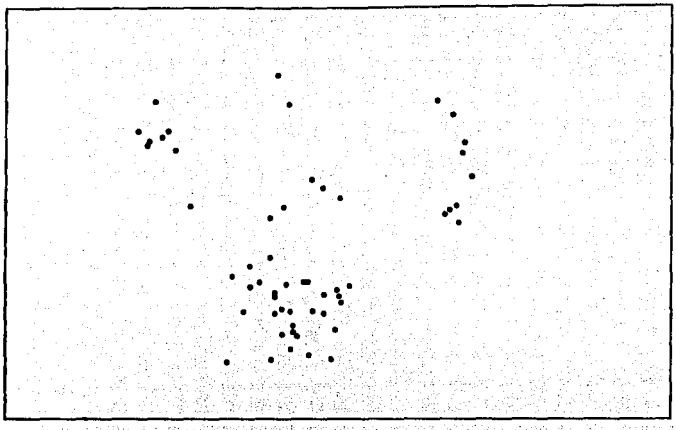


Figura 60a.Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Hermite con polinomios de grado 1.

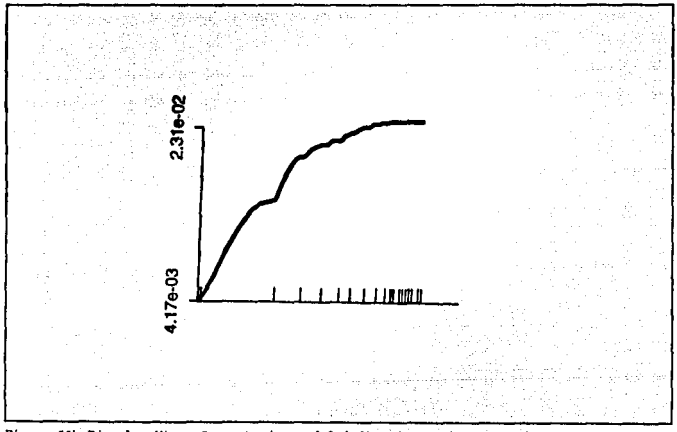


Figura 60b.Ejemplo: Virus.Comportamiento del índice de Hermite con polinomios de grado 1.

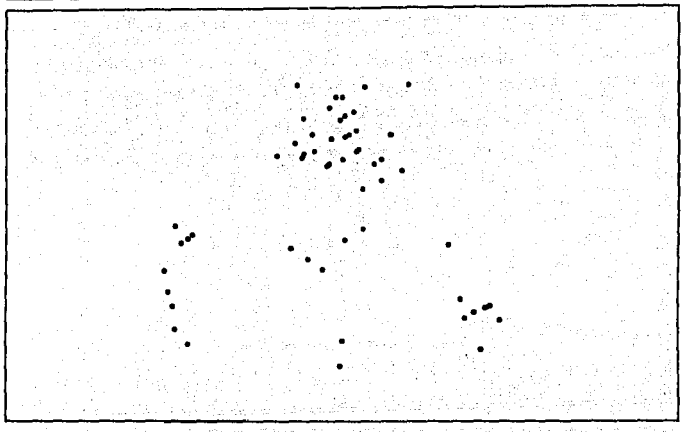


Figura 61a. Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Hermite con polinomios de grado 7.

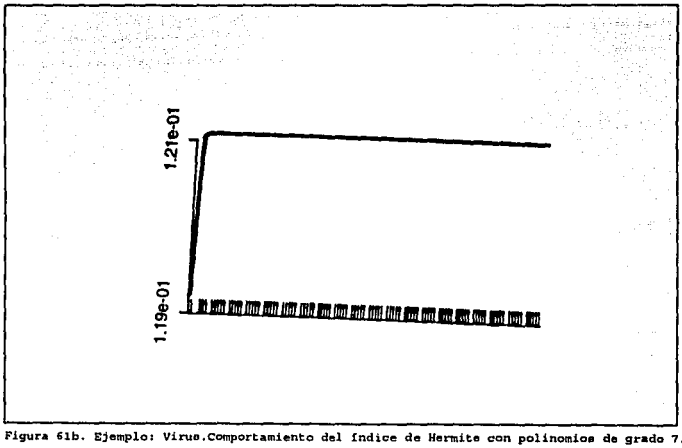


Figura 61b. Ejemplo: Virus.Comportamiento del índice de Hermite con polinomios de grado 7.

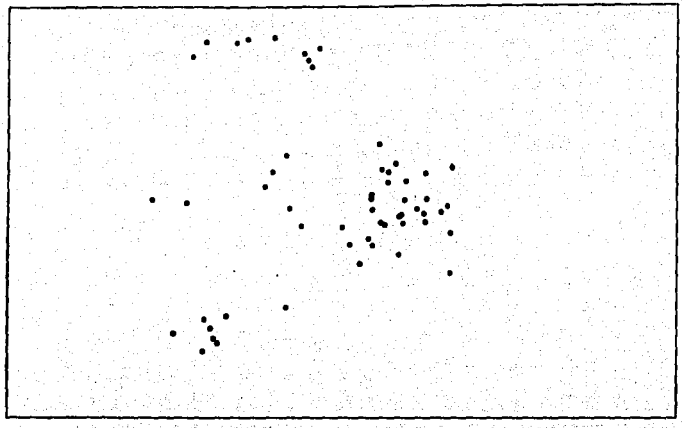


Figura 62a. Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Hermite con polinomios de grado 8.

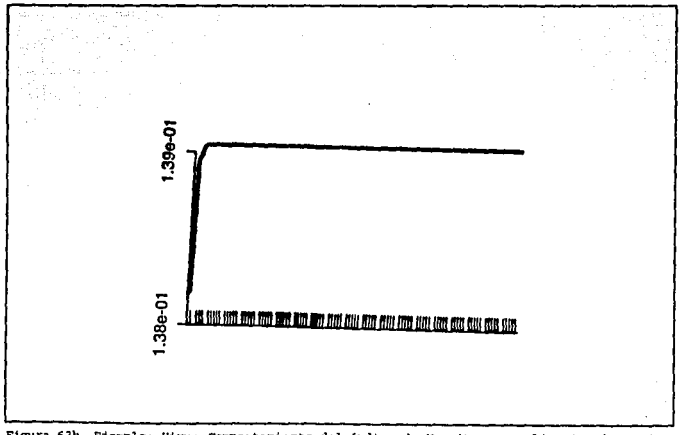


Figura 62b. Ejemplo: Virus.Comportamiento del índice de Hermite con polinomios de grado 8.

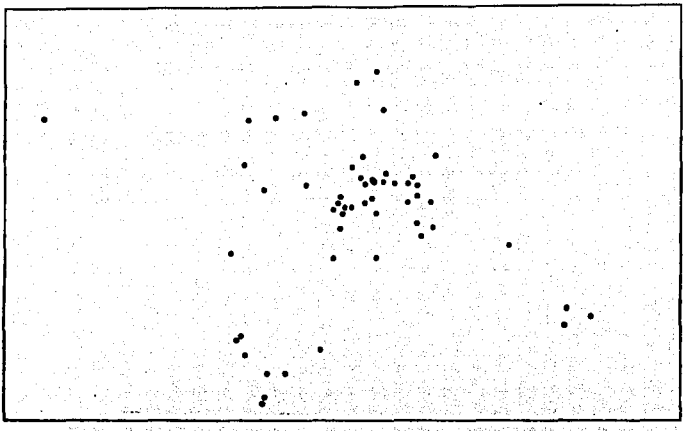


Figura 63a.Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice Natural de Hermite con polinomios de grado cero.

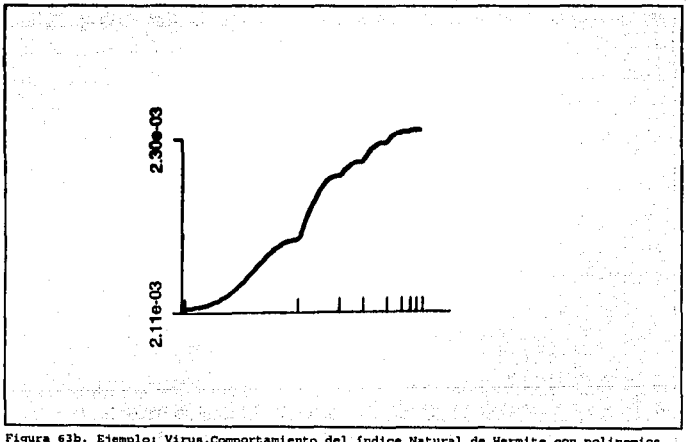


Figura 63b. Ejemplo: Virus.Comportamiento del índice Natural de Hermite con polinomios de grado cero.

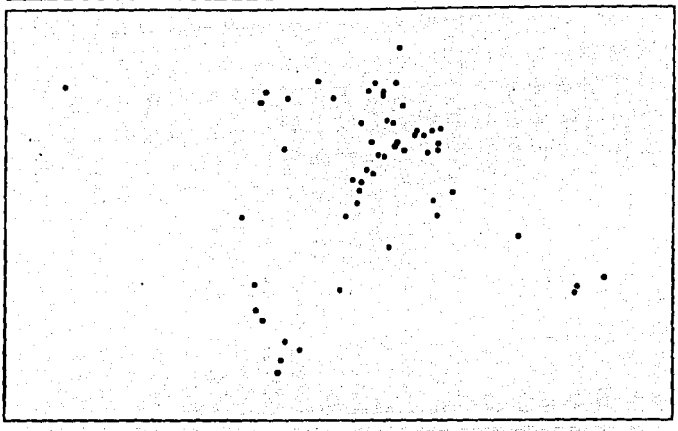


Figura 64a.Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice Natural de Hermite con polinomios de grado 1.

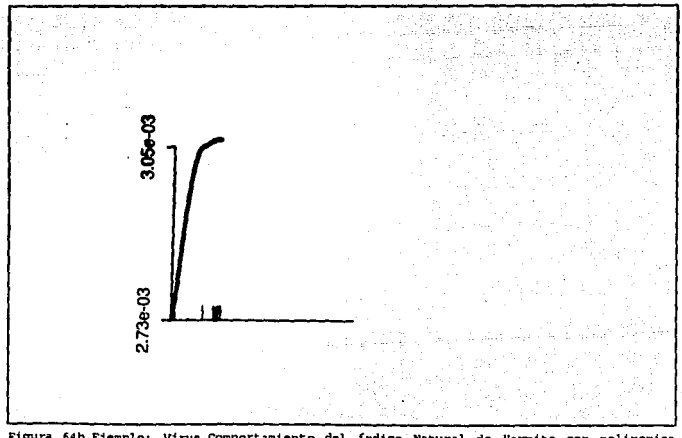


Figura 64b.Ejemplo: Virus.Comportamiento del índice Natural de Hermite con polinomios de grado 1.

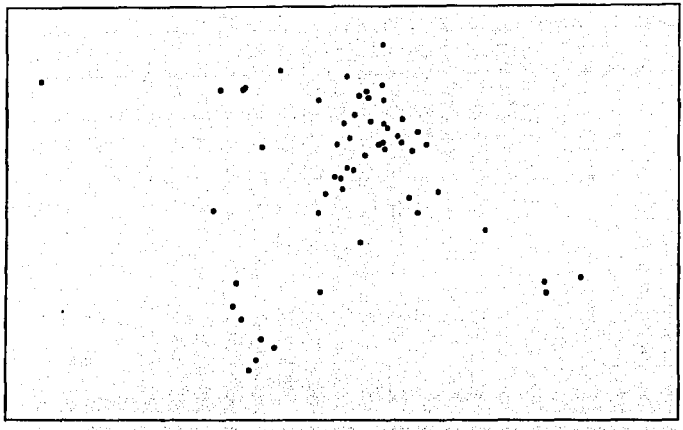


Figura 65a. Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice Natural de Hermite con polinomios de grado 7.

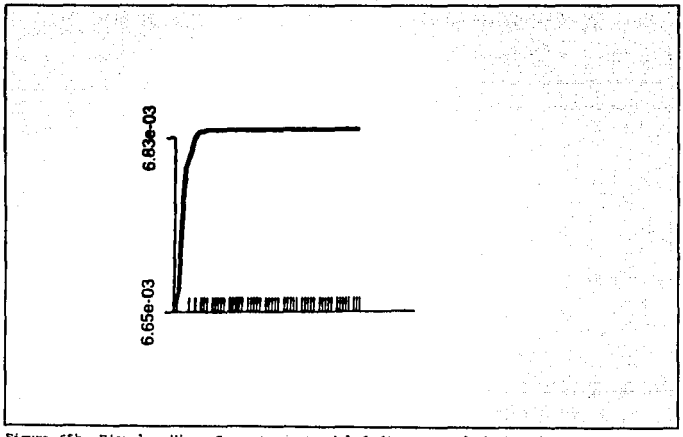


Figura 65b. Ejemplo: Virus.Comportamiento del índice Natural de Hermite con polinomios de grado 7.

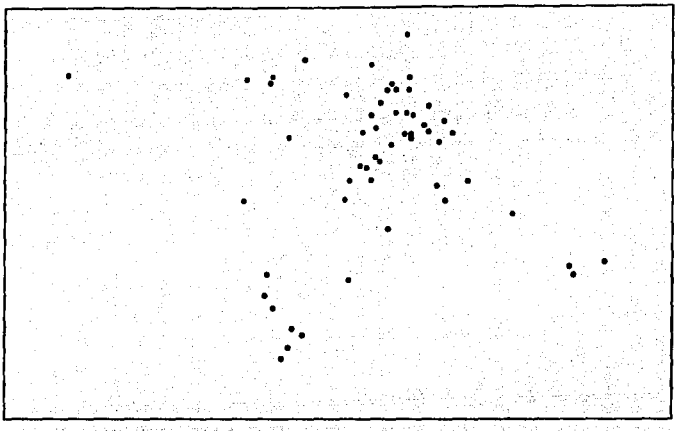


Figura 66a. Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice Natural de Hermite con polinomios de grado 8.

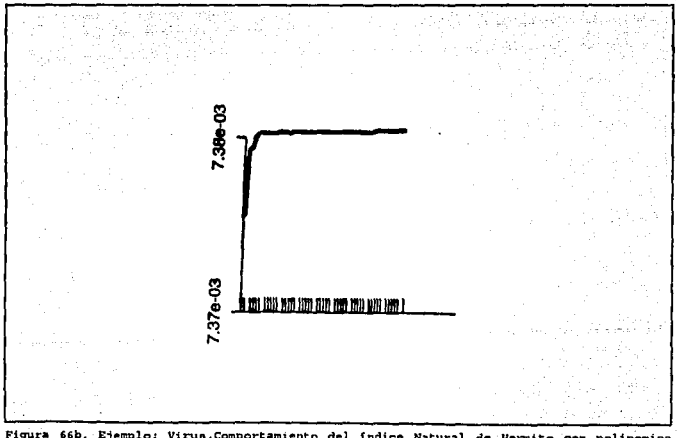


Figura 66b. Ejemplo: Virus.Comportamiento del índice Natural de Hermite con polinomios de grado 8.

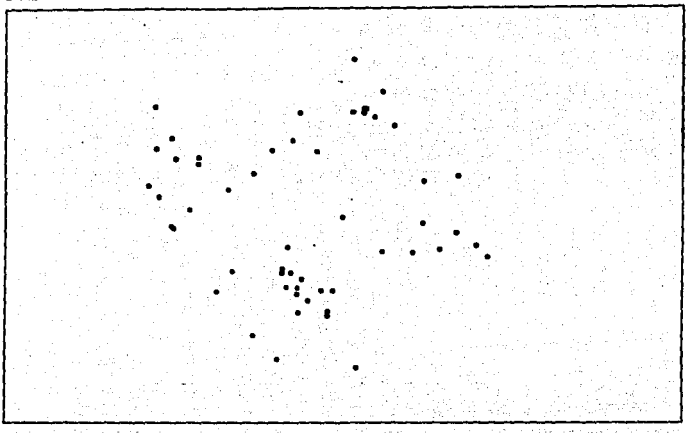


Figura 67a.Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice *Holes*.

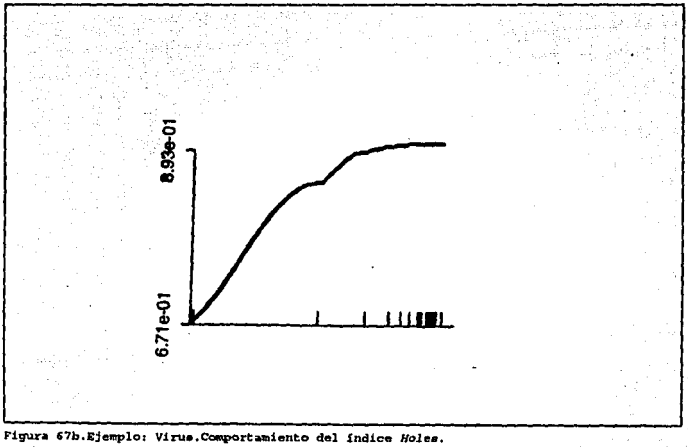


Figura 67b.Ejemplo: Virus.Comportamiento del índice *Holes*.

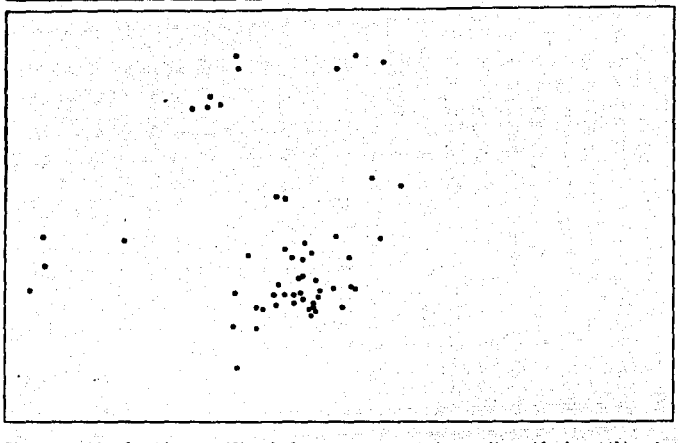


Figura 68a.Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice *Central Mass*.

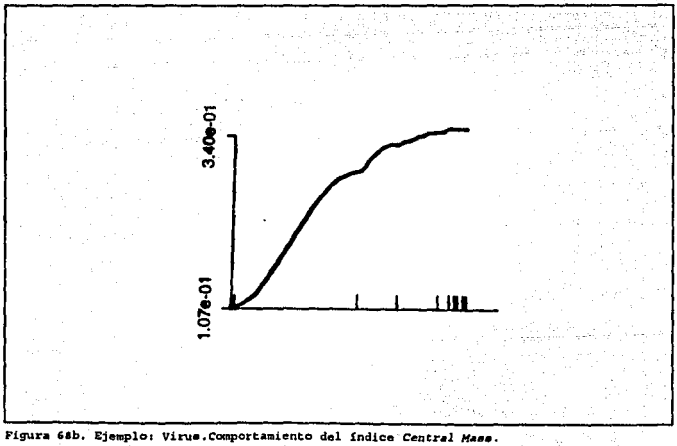


Figura 68b. Ejemplo: Virus.Comportamiento del índice *Central Mass*.

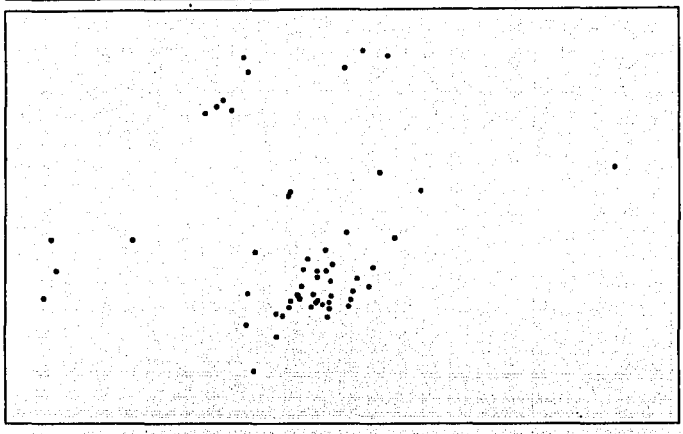


Figura 69a.Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice *Skewness*.

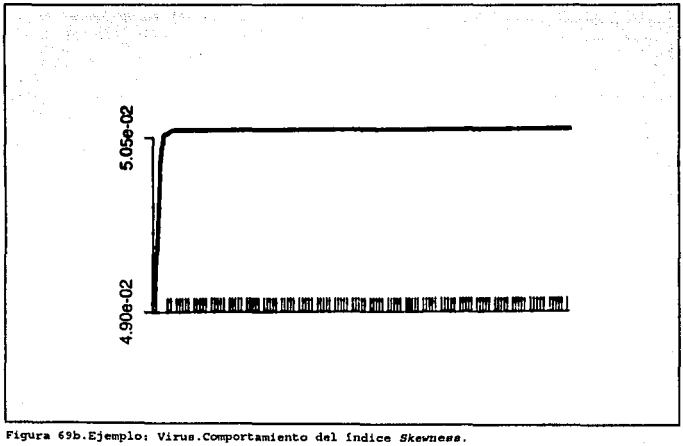


Figura 69b.Ejemplo: Virus.Comportamiento del índice *Skewness*.

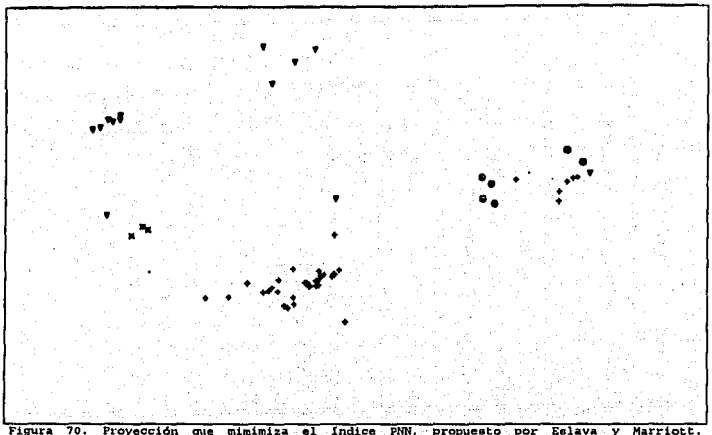


Figura 70. Proyección que mimimiza el índice PNN, propuesto por Eelava y Marriott. PNN=0.0190 (m=59).

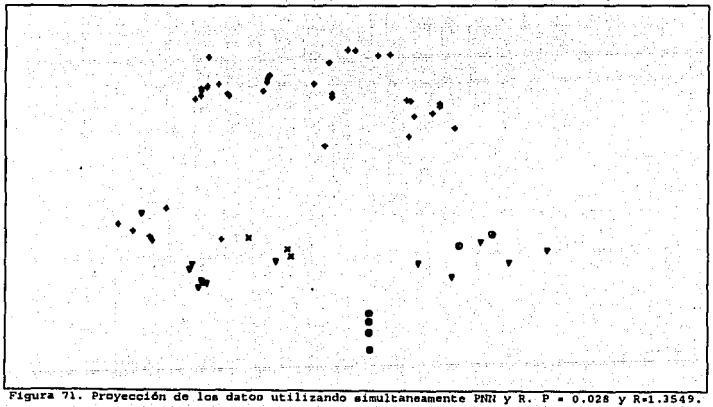


Figura 71. Proyección de los datos utilizando simultaneamente PNN y R. P = 0.028 y R=1.3549.

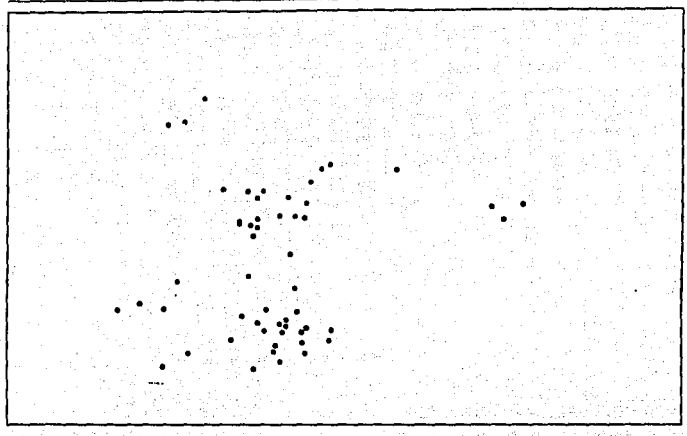


Figura 72a.Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Entropía sin el punto atípico.

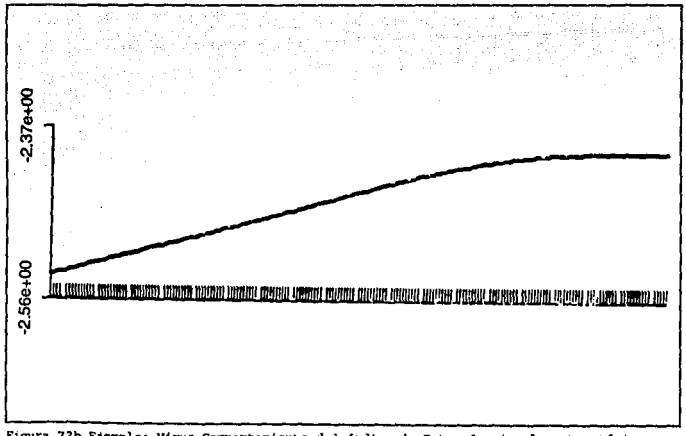


Figura 72b.Ejemplo: Virus.Comportamiento del índice de Entropía sin el punto atípico.

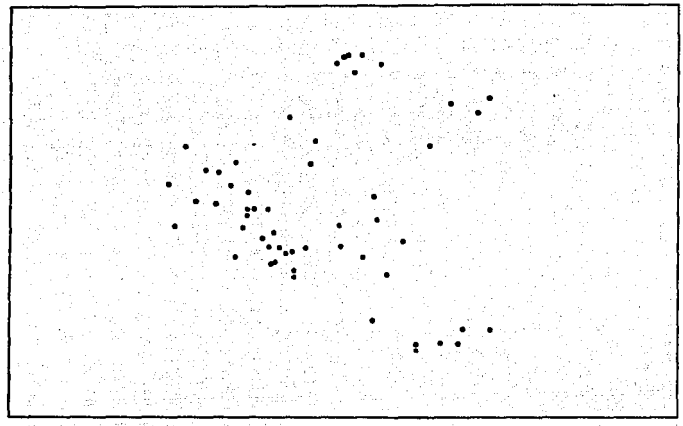


Figura 73a.Ejemplo: Virus. Gráfica de los puntos proyectados en dimensión dos utilizando el índice de Legendre con polinomios de grado uno, sin el punto atípico.

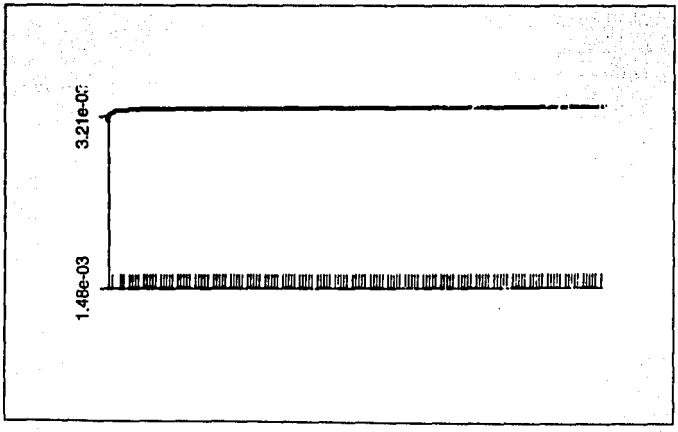


Figura 73b.Ejemplo: Virus.Comportamiento del índice de Legendre con polinomios de grado uno, sin el punto atípico.

Finalmente, en la tabla 9 se presentan, a manera de resumen, los valores de los índices que optimizaron las proyecciones.

Nombre del Índice	Característica	Valor óptimo
Friedman-Tukey	Original	1.59e-01
Friedman-Tukey	Modificado	1.60e-01
Entropía	Original	-2.17e+00
Entropía	Modificado	-2.17e+00
Legendre	Polinomios de grado 1	4.25e-02
Legendre	Polinomios de grado 2	2.00e-01
Legendre	Polinomios de grado 7	7.55e-01
Legendre	Polinomios de grado 8	7.95e-01
Hermite	Polinomios de grado 0	1.58e-03
Hermite	Polinomios de grado 1	2.31e-02
Hermite	Polinomios de grado 7	1.21e-01
Hermite	Polinomios de grado 8	1.39e-01
Natural de Hermite	Polinomios de grado 0	2.30e-03
Natural de Hermite	Polinomios de grado 1	3.05e-03
Natural de Hermite	Polinomios de grado 7	6.83e-03
Natural de Hermite	Polinomios de grado 8	7.38e-03
Holes	No se presenta teóricamente	8.93e-01
Central Mass	No se presenta teóricamente	3.40e-01
Skewness	No se presenta teóricamente	5.05e-02
PNN	m=59	0.0280
Rmean		1.3549
Entropía	Sin el punto atípico	-2.37e00
Legendre	Polinomios de grado 1, sin el punto atípico	3.21e-03

Tabla 9. Valor del índice de proyección. Ejemplo: Virus

CAPÍTULO 5. CONCLUSIONES

La representación gráfica como un medio para explorar datos multivariados es una de las ideas más explotadas en la estadística, para la búsqueda de grupos o conglomerados de observaciones, posibles puntos atípicos, etc., entre otros propósitos. El uso de transformaciones lineales para proyectar datos que originalmente se encuentran en una dimensión alta es una de las maneras más natural de realizar dicha representación gráfica. Aunque teóricamente la utilización de transformaciones lineales tiene muchos años, el enorme trabajo de cálculos que implicaba llevarlas a la práctica hicieron que no fuera sino hasta en los últimos años, en que las computadoras han tenido un gran auge, que estas ideas matemáticas han podido realizarse.

El Análisis de Componentes Principales es de las primeras técnicas que retoman la idea de encontrar nuevas variables a través de la aplicación de transformaciones lineales a las variables originales. El objetivo de estas transformaciones es que pocas de las nuevas variables tengan la mayor variabilidad posible de los datos, para que la representación geométrica de estas pocas nuevas variables sirvan para revelar la estructura subyacente en los datos multivariados. En términos matemáticos, esta técnica consiste en encontrar la descomposición espectral de la matriz de correlación de los datos originales. El uso de programas de computo estadísticos permite calcular los componentes principales. En los ejemplos presentados en el capítulo 4 observamos, en el caso de los cráneos, que realizar un análisis de componentes principales permite revelar la estructura de dos grupos de datos con lo que se puede suponer que provienen de poblaciones diferentes. Asimismo, nos permitió visualizar una observación atípica. Por otra parte, en el ejemplo de los virus, la técnica permite clasificar y reclasificar al conjunto de datos. Sin embargo en el caso del ejemplo de las plantas, la realización de un análisis de componentes principales no permite observar ninguna posible estructura de grupos en los datos.

Con la idea de que proyecciones lineales pueden revelar características de interés de los datos, aunado al acelerado desarrollo en los últimos años de las computadoras, recientemente se ha explotado en la práctica esta idea a través de una técnica denominada Proyecciones Perseguidas. La mayoría de los enfoques presentados en este trabajo se

refieren a encontrar proyecciones "interesantes" de los datos a través de la búsqueda de proyecciones que sean lo más diferentes a la normalidad de los datos. Bajo este criterio se tienen diferentes medidas o estadísticas para la no-normalidad de los datos proyectados, los cuales son usados como índices de proyección. Dado que existen varias maneras de diferir de la normalidad, implica que algunos de los índices de proyección presentados en este trabajo tiendan a sobreponer los grupos de puntos o bien mostrar un número pequeño de grupos grandes entre otros aspectos.

Otro aspecto que está involucrado en la propuesta de un índice de proyección, es la manera de estimar la función de densidad de los puntos proyectados.

Teóricamente se puede demostrar que la mayoría de los índices funcionan de manera adecuada para separar grupos sin embargo, en la práctica, dado que están involucradas cuestiones de optimización numérica y de estimación de funciones de densidad no resulta tan fácil; es decir, es común que varios de los índices encuentren numéricamente un óptimo local, lo cual podría dar pie a pensar que es un óptimo global. Por otra parte el hecho de utilizar estimaciones de las funciones de densidad, produce que las implementaciones prácticas de los índices no sean las que teóricamente se desean. Por lo tanto, es necesario realizar con cuidado las diferentes interpretaciones de las proyecciones que se realizan.

En este sentido, es recomendable realizar algunas simulaciones con conjuntos de datos artificiales creados a partir de mezclas de distribuciones Gaussianas multivariadas centradas en los vértices de un simplejo en dimensión alta y evaluar los índices de proyección propuestos.

En relación a la puesta en práctica de los índices, generalmente se implementaron en un programa escrito en fortran, sin embargo recientemente varios de los índices se han desarrollado en un programa en lenguaje C llamado XGobi.

En el caso de los ejemplos presentados en el capítulo 4, observamos lo siguiente:

Los índices de Friedman-Tukey y el de entropía son muy lentos, lo cual supone una gran cantidad de cálculos. En apariencia no encuentran un óptimo, sin embargo la función tiende a estabilizarse con el tiempo (la gráfica es paralela al eje horizontal). La modificación

realizada por los autores del programa acelera los índices, optimizándose en el mismo valor que en los índices originales.

Los índices basados en polinomios ortogonales (Legendre, Hermite y Natural de Hermite) se comportan de manera similar, dando diferentes proyecciones de los datos. Es decir, después de un cierto grado del polinomio que estima la función de densidad de los puntos, la gráfica de los índices de proyección tienden a ser paralelos en un momento dado al eje horizontal, en diferente valor dependiendo del grado del polinomio. Sin embargo las gráficas de los puntos proyectados son muy similares. En el caso del índice de Hermite, se encuentran los óptimos hasta con 6 términos, sin embargo para grados superiores tiende a estabilizarse. El índice de Legendre con cualquier número de términos la gráfica del índice tiende a la estabilización. El índice Natural de Hermite sólo alcanza su óptimo cuando se manejan el uno ó dos términos en los polinomios pero la gráfica de los puntos proyectados no muestra ninguna estructura en particular; con polinomios de exponente mayor sucede lo mismo que en los índices anteriores.

En relación a los índices restantes, dado que no se presentaron teóricamente en esta tesis, y sólo en base a las gráficas obtenidas se puede decir lo siguiente: el índice *Holes*, encuentra separaciones entre los grupos, sin embargo no detecta las observaciones aberrantes. Los índices *Central Mass* y *Skewness* no muestran la separación de los grupos.

Por lo tanto, observamos que proyecciones perseguidas es una técnica que tiene mucho por estudiarse tanto desde el punto de vista teórico, estadístico o bien del análisis numérico. Es una técnica que en la medida de su desarrollo y conocimiento general tendrá más aplicación cada vez, dado las computadoras sofisticadas con que se cuenta hoy en día.

Finalmente se puede decir, que el uso de transformaciones lineales en estadística y particularmente en análisis multivariado, es una herramienta potente para el análisis exploratorio de datos, si se requiere identificar grupos, observaciones aberrantes o estructuras no lineales que permitan extraer algunas conclusiones de un conjunto de datos determinado.

APENDICE I. DATOS

Este apéndice contiene los datos originales de los tres ejemplos expuestos en el capítulo 4.

A.I.1 Datos del ejemplo Cráneos

Las siguientes son los datos originales correspondientes al ejemplo de los cráneos.

CASO	TIPO	X1	X2	X3	X4	X5
1	1	185	136	131	132	70
2	1	181	128	129	129	65
3	1	172	131	132	130	67
4	1	169	131	125	120	58
5	1	182	138	133	133	70
6	1	182	122	125	121	66
7	1	181	135	126	126	64
8	1	171	128	135	130	61
9	1	179	129	129	128	70
10	1	178	133	137	124	63
11	1	176	126	133	125	65
12	1	176	131	134	128	72
13	1	185	130	132	130	61
14	1	179	134	123	128	63
15	1	172	137	139	.	70
16	1	182	134	133	131	71
17	1	181	135	137	130	71
18	1	189	132	135	128	71
19	1	182	133	136	130	70
20	1	186	137	137	137	70
21	1	176	137	132	.	68
22	1	182	133	134	139	72
23	1	193	135	137	131	77
24	1	185	141	134	139	74
25	1	189	131	137	138	66
26	1	179	127	133	130	67
27	1	177	131	136	133	67
28	1	183	130	132	134	71
29	1	180	129	132	132	70
30	1	182	140	136	.	69
31	2	171	156	132	135	78
32	2	147	155	121	121	68
33	2	151	167	145	131	65
34	2	151	147	101	120	63
35	2	159	165	136	128	70
36	2	147	157	132	129	81
37	2	176	141	129	.	67
38	2	154	162	138	140	65
39	2	167	136	129	.	63
40	2	135	161	128	.	48
41	2	144	170	129	.	72
42	2	144	161	132	.	74
43	2	160	175	141	135	66
44	2	146	163	129	124	67
45	2	166	161	141	.	91
46	2	158	162	127	125	68
47	2	153	165	138	140	67
48	2	157	159	133	126	71

49	2	157	167	128	135	63
50	2	165	158	135	138	67
51	2	162	153	131	121	61
52	2	171	160	129	141	75
53	2	143	155	120	127	63
54	2	160	153	124	132	66
55	2	153	.	132	.	66

Las variables utilizadas tienen el siguientes significado:

TIPO =Tipo de población de la cual fueron extraídos los cráneos:

(1 = Cholula, 2 = candelaria)

X1 = DIAMETRO ANTERO-POSTERIOR MÁXIMO

X2 = DIAMETRO TRANSVERSO MÁXIMO

X3 = DIAMETRO BREGMA BASION

X4 = DIAMETRO BIZIGOMATICO

X5 = DIAMETRO NASIO-PROSTION

. = DATOS FALTANTES

A.I.2. Datos del ejemplo Plantas

Los siguientes datos corresponden al ejemplo presentado en el capítulo 4 sobre la regeneración de plantas en la región del amazonas.

	FAMILIA	1979	1980	1981	
1986					
1	*Annonaceae*	4	3	1	34
2	*Bignoniaeceae*	16	3	11	10
3	*Boraginaceae*	1	2	6	7
4	*Compositae*	42	23	38	28
5	*Cyperaceae*	19	10	106	21
6	*Dilleniaceace*	2	1	5	19
7	*Euphorbiaceae*	17	14	14	21
8	*Flacourtiaceae*	6	3	5	14
9	*Graminae*	42	35	22	37
10	*Guttiferae*	15	21	17	40
11	*Heliconiaceae*	45	128	209	103
12	*Lecythidaceae*	8	10	5	25
13	*Leguminosae*	10	11	30	73
14	*Marantaceae*	28	2	38	76
15	*Melastomataceae*	11	1	5	15
16	*Moraceae*	56	30	33	105
17	*Myrtaceae*	1	1	1	5
18	*Pipiraceae*	141	33	20	16
19	*Rhamnaceae*	2	9	34	2
20	*Rubiaceae*	2	1	3	8
21	*Rutaceae*	1	1	0	10
22	*Sapindaceae*	1	0	1	35
23	*Solanaceae*	132	16	21	28
24	*Tiliaceae*	40	24	47	71
25	*Ulmaceae*	31	1	7	0
26	*Violaceae*	6	2	19	26
27	*Zingibera*	1	5	6	8
28	*Amaranthaceae*	2	2	5	0
29	*Annonaceae*	12	2	7	63
30	*Bignoniaceae*	10	13	15	26
31	*Celastraceae*	2	0	1	10
32	*Compositae*	53	96	63	47
33	*Cyperaceae*	19	135	100	9
34	*Dilleniaceace*	2	4	13	23
35	*Euphorbiaceae*	1	1	3	12
36	*Flacourtiaceae*	1	2	7	40
37	*Graminae*	62	96	29	51
38	*Guttiferae*	3	6	14	101
39	*Heliconiaceae*	24	34	44	59
40	*Lecythidaceae*	1	5	2	46

41	*Leguminosae*	8	107	207	185
42	*Marantaceae*	18	18	12	59
43	*Melastomataceae*	0	5	5	13
44	*Moraceae*	64	27	56	283
45	*Passifloraceae*	3	1	0	14
46	*Pipiraceae*	93	84	85	41
47	*Rhamnaceae*	3	34	15	0
48	*Rubiaceae*	2	1	4	69
49	*Solanaceae*	102	8	38	46
50	*Tiliaceae*	28	18	26	141
51	*Ulmaceae*	50	1	4	0
52	*Violaceae*	3	12	7	46
53	*Zingiberaceae*	9	4	9	178
54	*Annonaceae*	1	4	1	0
55	*Bignoniaceae*	40	41	28	23
56	*Burseraceae*	4	4	9	9
57	*Connaraceae*	1	1	4	4
58	*Dilleniaceace*	4	1	1	0
59	*Euphorbiaceae*	78	3	6	1
60	*Graminae*	103	57	99	148
61	*Heliconiaceae*	78	62	72	21
62	*Icacinaceae*	0	2	7	1
63	*Lecythidaceae*	21	23	26	37
64	*Leguminosae*	42	35	69	53
65	*Marantaceae*	3	4	9	4
66	*Moraceae*	7	2	9	6
67	*Palmae*	7	5	6	0
68	*Pipiraceae*	1	3	5	6
69	*Polygonaceae*	0	3	2	2
70	*Polypodiaceae*	34	13	35	22
71	*Rhamnaceae*	51	5	101	5
72	*Rubiaceae*	4	0	5	3
73	*Sapindaceae*	6	0	11	1
74	*Sapotaceae*	1	2	2	1
75	*Solanaceae*	5	2	12	0
76	*Ulmaceae*	19	1	1	0
77	*Violaceae*	9	4	20	17
78	*Zingiberaceae*	1	1	1	3
79	*Acanthaceae*	3	2	1	0
80	*Bignoniaceae*	68	41	39	66
81	*Caricaceae*	9	7	9	4
82	*Euphorbiaceae*	272	85	65	42
83	*Gramineae*	42	14	26	59
84	*Heliconiaceae*	243	143	115	98
85	*Lecythidaceae*	0	4	7	1
86	*Leguminosae*	11	18	14	13
87	*Marantaceae*	2	3	3	3
88	*Monimiaceae*	7	5	3	4
89	*Palmae*	12	20	22	8
90	*Pipiraceae*	4	3	9	6
91	*Polypodiaceae*	0	1	1	4
92	*Rutaceae*	11	0	1	2

93	*Sapindaceae*	6	4	3	3
94	*Solanaceae*	12	7	5	0
95	*Sterculiaceae*	1	0	3	2
96	*Ulmaceae*	8	5	2	0
97	*Violaceae*	2	8	8	22
98	*Zingiberaceae*	1	2	2	13

A.I.3. Datos del ejemplo Virus

Los siguientes son los datos originales utilizados en el ejemplo presentado en la sección 4.3 referente a los residuos por molecula de diferentes aminoacidos.

TIPO	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18
1	1	25	9	9	19	12	8	20	0	10	0	6	21	8	7	4	7	17	5
2	1	26	9	9	20	13	8	20	0	10	0	6	21	8	7	4	7	17	5
3	1	25	9	9	22	10	10	23	0	13	0	6	19	5	6	4	8	16	5
4	2	15	10	21	13	18	12	22	1	9	2	4	11	5	10	1	14	8	2
5	2	17	11	22	15	14	10	23	1	11	2	4	11	5	9	1	13	9	1
6	2	22	17	17	16	10	15	13	1	7	2	3	14	9	9	2	12	6	2
7	2	21	18	18	15	11	15	16	1	7	2	3	14	6	8	2	12	7	2
8	2	20	9	16	15	16	6	19	1	7	3	4	14	4	11	1	16	11	3
9	2	22	10	17	18	13	6	21	1	8	3	4	13	4	11	1	15	10	3
10	3	17	13	14	16	4	9	14	1	13	0	11	13	5	7	1	4	11	5
11	3	12	11	9	12	6	5	12	1	9	1	7	12	5	6	0	4	8	2
12	3	18	16	16	16	8	6	14	1	14	0	9	12	4	8	0	2	11	3
13	3	18	16	15	19	8	6	11	1	15	1	7	13	5	8	0	2	9	3
14	3	17	13	13	22	8	4	18	1	10	3	8	11	7	6	1	2	10	2
15	3	16	13	16	21	9	3	17	1	10	4	7	12	7	5	1	2	11	3
16	3	22	19	10	16	10	4	18	1	12	2	8	11	6	8	0	1	8	2
17	3	20	10	24	10	6	9	21	0	7	0	7	18	4	9	1	4	8	2
18	3	20	21	12	15	9	7	11	1	9	3	8	14	6	7	0	1	10	3
19	3	20	21	12	15	9	7	11	1	9	3	9	14	5	7	0	1	10	3
20	3	18	11	24	10	9	6	19	0	12	0	7	14	4	11	0	4	9	1
21	3	20	12	23	10	8	5	20	0	13	0	6	13	4	11	0	4	10	1
22	3	18	19	18	16	8	4	12	0	12	0	10	15	8	6	1	1	12	1
23	3	17	16	17	15	8	6	14	1	14	0	9	12	4	8	0	3	11	3
24	3	19	17	14	16	8	6	14	1	14	0	8	12	4	8	0	2	12	3
25	3	19	17	15	16	8	5	14	1	14	0	8	12	4	8	0	2	12	3
26	3	19	15	16	16	8	6	14	1	15	0	8	12	4	8	0	2	12	3
27	3	17	17	16	19	8	6	11	1	15	1	7	13	5	8	0	2	9	3
28	3	18	17	15	19	8	6	11	1	15	1	7	13	5	8	0	2	9	3
29	3	22	19	10	16	10	4	18	1	12	2	8	11	6	8	0	1	8	2
30	3	22	19	10	16	10	5	17	1	12	2	8	11	6	8	0	1	8	2
31	3	18	20	10	18	6	8	17	1	14	1	5	16	4	7	0	2	9	2
32	3	18	16	16	15	8	6	13	1	14	1	8	12	4	8	1	2	12	3
33	3	20	21	12	15	9	7	11	1	10	3	8	14	7	7	0	1	9	3
34	3	20	21	12	15	9	7	11	1	10	3	9	14	5	7	0	1	10	3
35	3	18	12	23	10	9	5	20	0	14	0	7	12	4	11	0	4	10	1

36	3	18	12	21	10	10	5	18	0	13	0	8	12	4	12	0	4	10	1
37	3	17	12	22	10	8	5	18	0	14	0	5	13	4	10	0	3	9	1
38	3	17	16	16	16	8	6	15	1	14	0	9	12	4	8	0	2	11	3
39	3	19	17	15	17	7	6	15	1	14	0	8	12	4	8	0	2	10	3
40	3	18	16	16	19	8	6	11	1	15	1	7	13	5	8	0	2	9	3
41	3	18	17	15	17	8	6	15	1	14	0	8	12	4	8	0	3	9	3
42	3	15	12	14	23	8	3	17	1	9	4	7	15	6	6	1	2	11	2
43	3	13	11	14	22	7	3	17	1	10	4	8	13	6	6	1	3	11	2
44	3	16	11	15	23	10	4	18	1	10	3	7	12	6	5	1	2	9	3
45	3	14	11	14	25	11	3	19	2	10	2	7	12	6	5	1	2	9	3
46	3	11	11	15	24	10	5	18	1	11	1	7	14	5	7	2	3	11	2
47	3	15	9	12	21	8	4	21	1	10	3	7	15	7	6	1	3	10	3
48	3	15	11	15	22	7	3	19	1	8	3	4	14	6	5	1	2	10	2
49	4	27	8	13	25	12	26	21	1	20	0	11	18	5	7	5	7	19	3
50	4	27	8	13	25	13	27	21	1	19	0	11	18	6	8	5	7	19	3
51	4	27	7	12	25	12	26	21	1	20	0	11	17	6	8	5	7	19	3
52	4	26	8	13	25	13	26	21	1	19	0	11	18	6	8	5	8	19	3
53	4	28	6	13	24	12	30	22	1	18	0	11	18	6	8	4	7	19	3
54	4	27	8	14	25	13	26	21	1	18	0	11	18	6	8	5	7	19	3
55	4	24	15	18	14	10	14	19	1	14	7	5	19	4	6	2	12	10	4
56	4	25	14	15	15	9	12	14	0	8	3	6	12	4	14	1	10	8	0
57	4	29	11	12	23	9	15	23	0	16	1	10	13	5	8	3	6	23	5
58	4	28	15	22	21	7	32	21	1	13	2	8	16	12	5	5	8	9	1
59	4	29	14	22	20	9	20	20	2	15	2	9	16	7	6	6	10	13	2
60	4	29	16	18	18	8	32	22	1	14	2	9	18	15	4	4	8	9	1
61	4	31	14	21	20	9	21	20	3	15	2	8	17	6	7	6	10	13	1

APENDICE II. INSTRUCCIONES DE USO DE PROGRAMAS

En este apéndice se presentan las instrucciones para realizar el análisis de componentes principales en el paquete estadístico SPSS, así como algunas instrucciones para trabajar el programa XGobi en lo que se refiere a las proyecciones perseguidas.

A.II.1 Análisis de Componentes principales en SPSS

Las instrucciones para que el paquete estadístico SPSS, realice el análisis de componentes principales suponiendo que las variables originales son x1, x2, x3, x4 y x5 y que se va a utilizar un archivo llamado craneos.dat son como sigue:

```
DATA LIST FILE ='CRANEO.DAT' FREE/
FACTOR
  /VARIABLES x1 x2 x3 x4 x5  /MISSING LISTWISE /ANALYSIS x1 x2 x3 x4 x5
  /PRINT INITIAL CORRELATION DET EXTRACTION
  /CRITERIA FACTORS(5) ITERATE(25)
  /EXTRACTION PC
  /ROTATION NOROTATE
  /SAVE REG(ALL) .
```

Si se quiere una gráfica de los dos primeros componentes principales se dan las siguientes instrucciones.

```
GRAPH
  /SCATTERPLOT(BIVAR)=fac2_1 WITH fac1_1
```

A.II.2 Proyecciones Perseguidas en XGobi

XGobi es un programa desarrollado por BELLCORE (1994), que se ejecuta bajo el sistema operativo UNIX en un ambiente X-Windows. Para ejecutar el programa se le debe dar la siguiente instrucción:

XGobi archivodat

donde archivodat es el nombre del archivo en código ASCII que contiene los datos del ejemplo a trabajar.

Al entrar al programa XGobi aparece en pantalla el menú principal ubicado en la parte superior, en la parte central de la pantalla aparece una ventana para realizar una gráfica del tipo X-Y. Si del menú principal se selecciona la opción llamada **Tour**, aparece del lado izquierdo un menú que entre otras opciones contiene lo siguiente: una opción con el nombre **PrnCmp**, que si se selecciona el programa trabajará con los datos esferados en lugar de los originales; aparece una opción llamada **PP Index Menu** el cual permite seleccionar un índice de proyección de un total de 10 índices. Al seleccionar un índice aparece una nueva ventana en la cual se puede apreciar el comportamiento del índice cuando esta proyectando los puntos. Si se selecciona la opción **Optimiz** el índice buscará su valor óptimo. Del lado derecho de la pantalla aparecen las variables que se están considerando para la proyección del índice. Basta con dar un *click* para seleccionar o quitar alguna de las variables.

Si se quiere guardar alguna gráfica en un archivo, deberá seleccionarse la opción **I/O Menu** y seleccionar en este menú la opción **Write**.

Los índices que pueden ser seleccionados son:

El índice del **Legendre**: es el índice propuesto por Friedman (1987).

El índice de **Hermite**: es el índice propuesto por Hall (1989).

El índice **Natural Hermite**: es el índice propuesto por Cook y colaboradores (1993).

El índice **Holes**: responde a proyecciones que contienen pocos datos en el centro.

El índice **Central Mass**: responde a proyecciones con altas concentraciones de puntos en el centro.

El índice **Skewness**: responde a proyecciones que exhiben asimetrías.

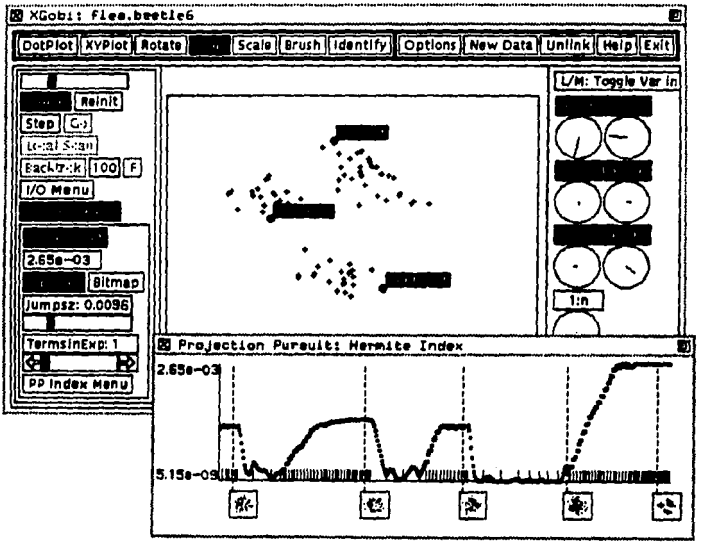
El índice **Friedman-Tukey**: es el índice original propuesto por Friedman y Tukey (1974).

El índice **Entropy**: Es el índice presentado por Jones y Sibson (1987).

El índice **Binned Friedman-Tukey**: da las mismas proyecciones que el índice Friedman-Tukey sólo que optimiza más rápido.

El índice **Binned Entropy**: da las mismas proyecciones que el índice Entropy sólo que optimiza más rápido.

Por último se presenta una pantalla del programa:



BIBLIOGRAFIA

Abramowitz and Stegun (1972), *Handbook of Mathematical Functions*, New York: Dover

Bellcore (1994). *User's Manual for Xgoby: A Dynamic Graphics Program for Data Analysis Implemented in the X Window System.*

Cook D., Buja A. and Cabrera J. (1993). *Projection Pursuit Based on Orthonormal Function Expansions.* Journal of Computational and Graphical Statistics, Vol. 2, No. 3.

Cox, D.R. (1972). *The analysis of multivariate binary data.* Appl. Statist., 21.

Dantas M. (1989). *Studies on succession in clear areas of amazonian rain forest.* D Phil thesis University of Oxford, U. K.

Diaz Leñero A.(1995). *La Deformación Craneana en Cholula, Puebla.* Un estudio morfométrico. Tesis de Licenciatura (En borrador). Escuela Nacional de Antropología e Historia. Méx., D.F.

Eslava G. G. (1989). *Projection Pursuit and other graphical methods for multivariate data.* D. Phil. thesis University of Oxford, U. K.

Eslava G. G. (1993). *On detecting clusters using Projection pursuit methods: an example.* Symposium I Anvendt Statisk, KØBENHAVN, 25-27. JANUAR 1993, UNI•C, Dinamarca, pags. 383-397.

Eslava G. and Marriott F.H.C. (1994). *Some Criteria for projection pursuit.* Statistics and Computing, 4, pags. 13-20.

Friedman, J. H. (1987). *Exploratory projection pursuit.* Journal American Statistics. Ass., vol. 82, No. 397, pags. 249-266.

Friedman, J.H. and Tukey, J. W. (1974). *A projection pursuit algorithm for exploratory data analysis.* IEEE Trans. Comput., vol. C-23, No. 9 pags.881-890.

Gower J.C. (1966). *Multivariate Analysis and multidimensional geometry.* Statistician, 17.

Hall P. (1989). *Polynomial Projection Pursuit.* The Annals of Statistics, 17, pags.589-605

Hotelling H. (1933). *Analysis of a complex of statistical variables into principal components.* J. Educ. Psychol.

Huber, P. J. (1985). *Projection pursuit (with discussion).* Ann. Statist., Vol. 13, No. 2 pags. 435-475.

Jee, R. (1985). *A study on projection pursuit methods.* Ph. D. thesis Rice University, U.S.A.

Jollife, I.T. (1986). *Principal Component Analysis.* Springer-Verlag. U.S.A.

Jones, M.C. and Sibson, R. (1987). *What is projection pursuit? (with discussion).* Journal of the Royal Statistical Society, serie A, No. 150, Part. I, pags. 1-36.

Kendall M. and Stuart A. (1977). *The advanced theory of statistics.* Volume 1. Distribution Theory. London

Kendall, M. (1980). *Multivariate Analysis.* Charles Griffin and Company. U.S.A.

Kruskal, J.H. (1972). *Linear transformation of multivariate data to reveral clustering. In Multidimensional scaling: theory and applications in the behavioural sciences.* Vol. I(eds. R.N. Shepard), London: Seminar Press

Krzanowski, W.J. (1988). *Principles of Multivariate Analysis.* Oxford University.

Krzanowski W.J. and Marriot F.H.C. (1994). *Multivariate Analysis, Part I (Distributions, Ordination and Inference).* Edward Arnold. New York, U.S.A.

Manly Bryan F.J. (1986). *Multivariate Statistical Methods.* Chapman and Hall. London.

Mardia K., Kent J. and Bibby J. (1979). *Multivariate Analysis.* Academic Press, London, New York.

Marriott F.H.C. (1974). *The Interpretation of Multiple Observations.* Academic Press Inc. London.

Nason G. (1995). *Three-dimensional Projection Pursuit.* Applied Statistics, The Royal Statistical Society Serie C, 44, No. 4, pags. 411-430.

Pearson K. (1901). *On lines and planes of closest fit to systems of points in space.* The London, Edinburgh and Dublin Philosophical Magagazine and Journal of Science. Sixth Series, 2, 559-572.

Rao D.R. (1964). *The use and interpretation of principal components in applied research.* Sankhya, Series A, 26, 329-359.

Silverman B. W. (1986). *Density Estimation for Statitics and Data Analysis.* Chapman and Hall, London.

Strang G. (1982). *Algebra Lineal y sus Aplicaciones.* Fondo Educativo Interamericano, México, D.F.

Yenyukov, I.S. (1988). *Detecting structure by means of projection pursuit.* In Compstat 1988 (eds. D. Edwards and E. Raun). Physica-Verlag Heidelberg for IASC, pags 47-58.