PESQUIMAT , Revista de la F.C.M. de la
Universidad Nacional Mayor de San Marcos
Vol. XV N°2, pp. 29-38, Lima - Perú, Marzo 2013

HISTOGRAMA CIRCULAR 3D

Erwin Kraenau Espinal[1]

Resumen: Este artículo tiene por finalidad presentar una nueva herramienta para explorar estadísticamente datos multidimensionales denominada Histograma Circular 3D. A diferencia del histograma clásico basado en coordenadas cartesianas, el que se muestra aquí, se construye a partir de un sistema de coordenadas polares y la técnica de reducción de la dimensión llamada Projection Pursuit desarrollado por Jerome H. Friedman y John Tukey en 1974. Se probó esta nueva herramienta, utilizando dos grupos de datos multidimensionales simulados. El primer conjunto de datos sigue una distribución normal multivariante y el segundo grupo simula una estructura formada por tres clusters. El Histograma Circular 3D mostró claras desviaciones de la normalidad cuando existían, así como asimetrías y con la ayuda de la Projection Pursuit se logró detectar estructuras donde las hubo.
Palabras clave: Coordenadas polares, Projection Pursuit, reducción de la dimensión, datos esferizados, estructura.

3D CIRCULAR HISTOGRAM

Abstract: This paper aims to present a new tool for statistically exploring of multidimensional data called 3D Circular Histogram. Unlike classical histogram based on cartesian coordinates, what is shown here, is constructed from a polar coordinate system and the reduction dimension technique called Projection Pursuit developed by Jerome H. Friedman and John Tukey in 1974. This new tool was tested using two groups of simulated multidimensional data. The first set of data follows a multivariate normal distribution and the second group simulates a structure formed by three clusters. The 3D Circular Histogram showed clear departures from normality when existed and asymmetries and with the help of the Projection Pursuit successfully detected where there were structures.
Keywords: Polar coordinates, Projection Pursuit, dimensionality reduction, sphered data, structure.

1. Introducción

Cuando se quiere explorar estadísticamente un conjunto de datos multidimensionales, se utilizan herramientas diseñadas para sistemas de coordenadas cartesianas, tales como el de las Componentes Principales que reducen la dimensionalidad de los datos y los proyectan en un número finito de planos o también se emplean herramientas como las Caras de Chernoff y las Curvas de Andrews entre otros, que reducen la dimensionalidad y buscan similitudes entre los datos, todas las cuales tratan de preservar la mayor cantidad de información posible.

La Física utiliza sistemas de coordenadas esféricas o cilíndricas para describir los movimientos y trayectorias de los planetas o satélites en el espacio. También lo hace la Mecánica del Continuo que emplea sistemas de coordenadas locales de todo tipo para describir por ejemplo, el esfuerzo y deformación en un medio isótropo [1]. La Estadística casi exclusivamente emplea sistemas de coordenadas cartesianas para el análisis de datos y gran parte de su teoría está basada en el supuesto de normalidad, donde muchas de sus herramientas son útiles previa comprobación de este supuesto.

Las propiedades de la distribución normal son estudiadas casi desde los inicios de la Estadística, por

[1]Departamento Académico de Estadística. UNMSM, e-mail: stoned@ec-red.com

lo que son bastante conocidas y sabemos que las superficies de nivel de esta distribución tienen un comportamiento elíptico. No es natural entonces utilizar sistemas de coordenadas cartesianas para hacer una exploración estadística de datos multidimensionales, lo más adecuado es el uso de otros sistemas de referencia como por ejemplo, los polares, cilíndricos o esféricos entre otros.
En este artículo, se pretende explorar estadísticamente un conjunto de datos multidimensionales para detectar estructuras subyacentes a este, mediante una nueva herramienta denominada Histograma Circular 3D, construida a partir de un sistema de coordenadas polares, empleando datos de dimensión reducida, conseguida a través de la Projection Pursuit.
En la Sección 2, primero se trata sobre la reducción de la dimensión para la cual se utiliza la técnica de la Projection Pursuit, necesaria para comparar el comportamiento teórico con los datos observados y en la segunda parte se detalla el algoritmo para la construcción del Histograma Circular 3D, basado en los datos de dimensión reducida previamente.
En la Sección 3, se presentan las proyecciones requeridas que detectan estructuras, así como los Histogramas Circulares 3D obtenidos a partir de dos conjuntos de datos multidimensionales simulados, donde el primer conjunto de datos sigue una distribución normal multivariante, y el segundo simula una estructura formada por tres cluster.
Para el procesamiento de los datos y la obtención de los resultados se empleó el software MATLAB (MATrix LABoratory) versión R2010a.

2. Metodología

2.1. Projection Pursuit

Freidman y Tukey en el año 1974, describieron la Projection Pursuit como un camino de búsqueda para explorar una estructura multidimensional de datos no lineal, examinando muchas posibles proyecciones bidimensionales. La idea es que la proyección bidimensional ortogonal de los datos revele la estructura original de estos [2].
El análisis de datos utilizando la Projection Pursuit logra encontrar muchas proyecciones, pero la calidad de la proyección se mide a través de un índice. En muchos casos el interés es la no normalidad, entonces el índice de proyección cuantifica la desviación de la normalidad. El índice es conocido como el *índice chi-cuadrado* y es desarrollado en la metodología de Posse [3].
Este método consiste básicamente de dos partes:

1. Un índice de la Projection Pursuit que mide el grado de la estructura (o desvío de la normalidad).
2. Un método para encontrar la proyección que produce el mayor valor del índice.

Posse utilizó una búsqueda aleatoria para localizar el punto óptimo global del índice de proyección y lo combina con la remoción de estructuras de Freidman para que se logre una sucesión "interesante" de proyecciones bidimensionales. Cada proyección encontrada muestra una estructura que es menos importante (en términos del índice de proyección) que la anterior [3].

2.1.1. Notación

- $\mathbf{X}$ es una matriz $n \times d$, donde cada fila $\mathbf{x}_i$, corresponde a una observación d-dimensional y n es el número de datos.
- $\mathbf{Z}$ es la versión esferizada de $\mathbf{X}$.
- $\widehat{\boldsymbol{\mu}}$ el vector de medias muestrales de tamaño d, donde $\widehat{\boldsymbol{\mu}} = \frac{1}{n}\mathbf{X}^t\mathbf{1}$.
- $\widehat{\boldsymbol{\Sigma}}$ es la matriz de covarianza muestral, donde $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\mathbf{X}^t(\mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}^t)\mathbf{X}$.
- $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ son vectores ortonormales tal que ($\boldsymbol{\alpha}^t\boldsymbol{\alpha} = \boldsymbol{\beta}^t\boldsymbol{\beta} = 1$ y $\boldsymbol{\alpha}^t\boldsymbol{\beta} = 0$) y a su vez son vectores d-dimensionales que generan el plano de proyección.
- $P(\boldsymbol{\alpha}, \boldsymbol{\beta})$ es el plano de proyección generado por $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$.
- $z_i^{\boldsymbol{\alpha}}$, $z_i^{\boldsymbol{\beta}}$ son las proyecciones esferizadas de las observaciones, donde

$$z_i^{\alpha} = \mathbf{z}_i^t \boldsymbol{\alpha},\ z_i^{\beta} = \mathbf{z}_i^t \boldsymbol{\beta}$$

$(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ denota el plano donde el índice es máximo.

- $PI_{\chi^2}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ es el índice de proyección chi-cuadrado evaluado, utilizando los datos proyectados sobre el plano generado por $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$.
- ϕ_2 es la función de densidad normal bivariada.
- c_k es la probabilidad evaluada sobre la k-ésima región utilizando la función de densidad normal estándar bivariada. El valor se obtiene de $c_k = \iint\limits_{B_k} \phi_2\, dz_1\, dz_2$.
- B_k es una región en el plano de proyección (ver Figura 1).
- I_{B_k} es la función indicadora de la región B_k.
- $\eta_j = \pi j/36$, $j = 0, ..., 8$ es el ángulo por el cual el dato es rotado en el plano antes de ser asignado a una de las regiones B_k.
- $\boldsymbol{\alpha}(\eta_j)$ y $\boldsymbol{\beta}(\eta_j)$, donde

$$\boldsymbol{\alpha}(\eta_j) = \boldsymbol{\alpha}\, cos\, \eta_j - \boldsymbol{\beta}\, sen\, \eta_j$$
$$\boldsymbol{\beta}(\eta_j) = \boldsymbol{\alpha}\, sen\, \eta_j + \boldsymbol{\beta}\, cos\, \eta_j$$

- c es un escalar que determina el tamaño de la vecindad alrededor de $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ y es utilizado en la búsqueda de planos con el objeto de encontrar mejores valores del índice de proyección.
- $\mathbf{v}$ es un vector uniformemente distribuido sobre la esfera unitaria d-dimensional.
- γ especifica el número de iteraciones sin un incremento en el índice de proyección al mismo tiempo que el tamaño de la vecindad es dividida.
- m representa el número de búsquedas aleatorias para encontrar el mejor plano.

2.1.2. Índice de Proyección La Projection Pursuit se realiza a través de la proyección de las variables en diferentes planos para encontrar el más "interesante" según el índice de proyección chi-cuadrado. Este procedimiento se realiza a través de dos etapas:

1. **Búsqueda de la no-normalidad de los datos.** El plano es dividido en 48 regiones o cajas B_k distribuidos en anillos (Figura 1), cada uno con ancho angular de $\pi/4$ radianes y de ancho radial $\sqrt{2\,ln\,6}/5$, lo cual garantiza que cada región tenga aproximadamente la misma probabilidad $(1/48)$ para la distribución normal bivariante. El índice de proyección chi-cuadrado está dado por

$$PI_{\chi^2}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \tfrac{1}{9} \sum_{j=1}^{8} \sum_{k=1}^{48} \tfrac{1}{c_k} \left[\tfrac{1}{n} \sum_{i=1}^{n} I_{B_k} \left(z_i^{\boldsymbol{\alpha}(\eta_j)}, z_i^{\boldsymbol{\beta}(\eta_j)} \right) - c_k \right]^2 \quad (1)$$

   Una de las ventajas del uso del índice chi-cuadrado es que no es muy afectado por los datos discordantes y busca la proyección que da como resultado el mayor de los índices chi-cuadrado.

2. **Búsqueda de la estructura.** El algoritmo inicializa aleatoriamente los vectores $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$, para crear un mejor primer plano $(\boldsymbol{\alpha}^*$ y $\boldsymbol{\beta}^*)$, luego se genera dos planos vecinos dados por las ecuaciones (2) y (3), en donde se evalúa el índice chi-cuadrado para ellos, si uno de ellos presenta una mejoría en el índice éste será el nuevo mejor plano, de lo contrario se generan dos nuevos planos vecinos.

$$\mathbf{a}_1 = \frac{\boldsymbol{\alpha} * + c\,\mathbf{v}}{||\boldsymbol{\alpha} * + c\,\mathbf{v}||} \quad \mathbf{b}_1 = \frac{\boldsymbol{\beta} * - (\mathbf{a}_1^t \boldsymbol{\beta}^*)\mathbf{a}_1}{||\boldsymbol{\beta} * - (\mathbf{a}_1^t \boldsymbol{\beta}^*)\mathbf{a}_1||} \quad (2)$$

$$\mathbf{a}_2 = \frac{\boldsymbol{\alpha} * -c\,\mathbf{v}}{||\boldsymbol{\alpha} * -c\,\mathbf{v}||} \quad \mathbf{b}_2 = \frac{\boldsymbol{\beta} * -(\mathbf{a}_2^t \boldsymbol{\beta}^*)\mathbf{a}_2}{||\boldsymbol{\beta} * -(\mathbf{a}_2^t \boldsymbol{\beta}^*)\mathbf{a}_2||} \tag{3}$$

Si después de cierto número de iteraciones no ha habido mejora entonces se reduce el tamaño de la vecindad de búsqueda a través de la disminución del parámetro c.

2.1.3. Búsqueda de la Projection Pursuit

La Projection Pursuit se realiza a través de la proyección de las variables en diferentes planos para encontrar el de mayor interés según el índice de proyección chi-cuadrado [3]. El diagrama del proceso simplificado es mostrado en la Figura 2.

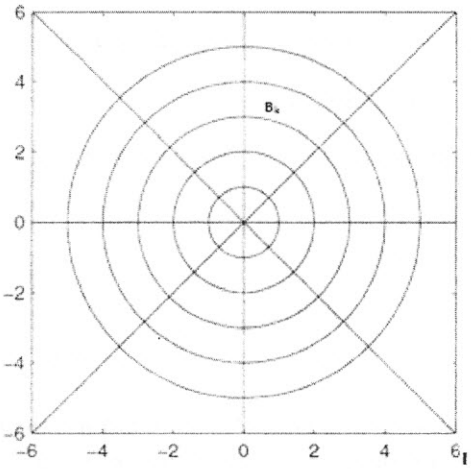

Figura 1: Disposición de las regiones B_k

El algoritmo para encontrar este índice de proyección es el siguiente:

a. Esferizar los datos usando la siguiente transformación

$$\mathbf{z}_i = \boldsymbol{\Lambda}^{-1/2}\mathbf{Q}^t(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})$$

donde las columnas de $\mathbf{Q}$ son los autovectores de $\widehat{\boldsymbol{\Sigma}}$, $\boldsymbol{\Lambda}$ es una matriz diagonal que corresponde a los autovalores de $\widehat{\boldsymbol{\Sigma}}$, y $\mathbf{x}_i$ es la i-ésima observación.

b. Proponer un plano inicial aleatorio, $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$. Este es el plano más actual $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$.

c. Evaluar el índice de proyección $PI_{\chi^2}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ expresado en la ecuación (1), para el plano inicial.

d. Generar dos planos candidatos $(\mathbf{a}_1, \mathbf{b}_1)$ y $(\mathbf{a}_2, \mathbf{b}_2)$ de acuerdo a las ecuaciones (2) y (3).

e. Evaluar el índice de proyección de esos planos $PI_{\chi^2}(\mathbf{a}_1, \mathbf{b}_1)$ y $PI_{\chi^2}(\mathbf{a}_2, \mathbf{b}_2)$ utilizando (1).

f. Si uno de los planos candidatos produce un índice con un valor más alto que el índice de proyección, entonces ese plano se convierte en el mejor plano actual $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$.

g. Repitir los pasos del (d) al (f) mientras hayan mejoras en el índice de proyección.

h. Si el índice no mejora en γ iteraciones, entonces disminuir el valor de c a la mitad.

i. Repitir los pasos (d) al (h) hasta que c sea un número tan pequeño como se haya elegido.

2.2. Histograma 3D

El diseño del Histograma Circular 3D, comienza reduciendo la dimensión de los datos a través de la Projection Pursuit (descrita en la Sección 2.1), para luego en base a estos datos transformados construir el Histograma Circular 3D. El algoritmo que lo produce es el siguiente:

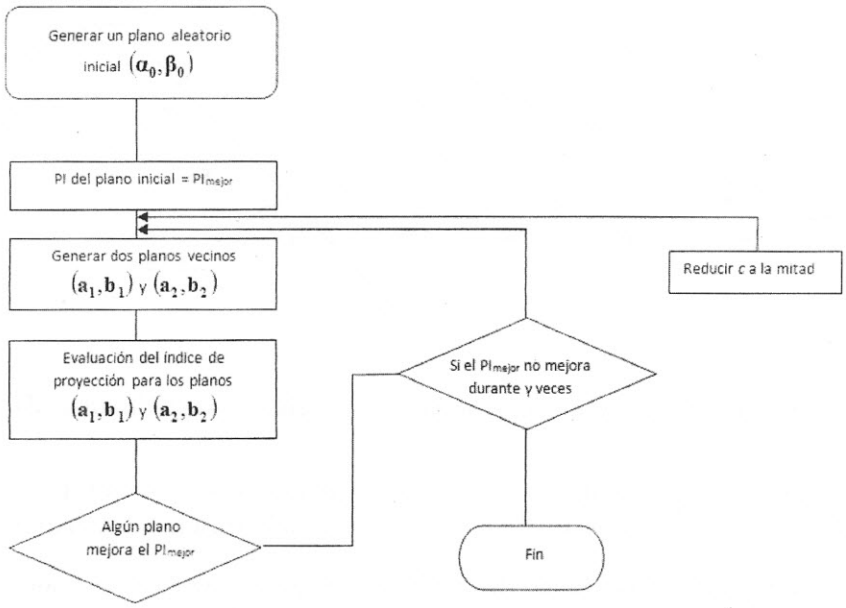

Figura 2: Proceso de búsqueda del Índice de Proyección

a. Particionar el plano polar en cuñas del mismo ángulo y anillos circulares del mismo ancho. Cada cuña tiene por defecto un ángulo de $\pi/6$ radianes como se muestra en la Figura 3, pudiendo este ángulo variarse. El número de regiones puede aumentarse o disminuirse, esto se logra haciendo

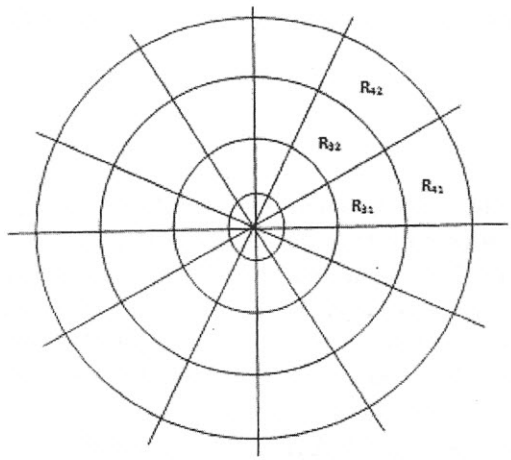

Figura 3: Base del Histograma Circular 3D

variar el ancho del anillo que por defecto tiene una unidad de longitud (como se observa en el lado izquierdo de la Figura 4 denotado por cc). Dependiendo de la cantidad de regiones definidas en la partición del plano polar, se refinará el Histograma Circular 3D.

b. Nombrar cada región por R_{ij}, esto indicará que pertenece al i-ésimo anillo y a la j-ésima cuña, como se muestra en la Figura 3 y en el lado derecho de la Figura 4.

c. Construir las matrices $\mathbf{XX}$ y $\mathbf{YY}$, que contienen las coordenadas cartesianas de los vértices de cada región R_{ij} en los ejes X e Y, respectivamente. Estas tendrán la siguiente forma:

$$\mathbf{XX} = \begin{bmatrix} cc\,(i-1)\cos\theta_1 & cc\,(i)\cos\theta_2 & cc\,(i-1)\cos(\theta-h) \\ cc\,(i-1)\cos\theta_1 & cc\,(i)\cos\theta_2 & cc\,(i-1)\cos(\theta-h) \end{bmatrix}$$

$$\mathbf{YY} = \begin{bmatrix} cc\,(i-1)\,sen\,\theta_1 & cc\,(i)\,sen\,\theta_2 & cc\,(i-1)\,sen\,(\theta-h) \\ cc\,(i-1)\,sen\,\theta_1 & cc\,(i)\,sen\,\theta_2 & cc\,(i-1)\,sen\,(\theta-h) \end{bmatrix}$$

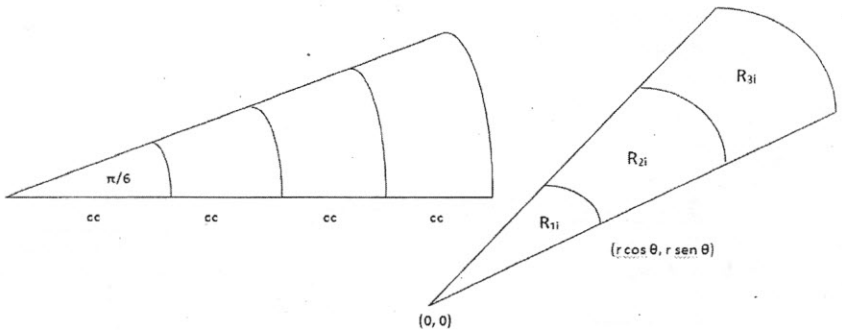

Figura 4: Cuñas y regiones de la base del Histograma Circular 3D

donde θ_1 y θ_2 son los ángulos iniciales y finales, respectivamente, que forma cada cuña con el eje polar. El ángulo θ toma valores desde h hasta 2π variando h radianes (ángulo de cada cuña).

d. Asignar un valor al grosor del anillo denotado por *cc* que por defecto es uno.

e. Proyectar los datos de dimensión reducida mediante el algoritmo de la Projection Pursuit descrito en la Sección 2.1 sobre la partición generada por las matrices **XX** y **YY** del paso (c).

f. Contabilizar los datos transformados que caen en cada región R_{ij}, obteniéndose así las frecuencias absolutas.

g. Hacer el trazado de sólidos rectos con bases formadas por las regiones R_{ij} con coordenadas en **XX** y **YY**, cuyos volúmenes son proporcionales a las frecuencias absolutas correspondientes, obtenidas en el paso (f).

3. Resultados

Para mostrar la utilidad del Histograma Circular 3D, se emplearon datos multidimensionales simulados que previamente fueron esferizados para trabajar en el plano polar.
El primer conjunto de datos simulado sigue una distribución normal multivariante. En la Figura 5 se muestra el resultado de aplicar el algoritmo de la Sección 2.1, para obtener la proyección buscada. Esta figura no presenta desvíos de la normalidad (estructuras) ya que la nube de puntos forma prácticamente un círculo. Este plano polar donde se han proyectado los datos simulados (Figura 5), es la base del Histograma Circular 3D que en sus dos versiones son mostrados en la Figura 6 y la Figura 7, el primero con estilo de superficie (sentencia *surf* del MATLAB) y el segundo con estilo de malla (sentencia *mesh* del MATLAB). Obsérvese la simetría de este Histograma Circular 3D y en cada anillo las alturas de las barras son más o menos las mismas. Se aprecia claramente la concentración de los datos alrededor del origen polar en la Figura 6 y Figura 7, siendo una de las propiedades de la distribución normal que las frecuencias mayores se dan en los valores cercanos a la media o vector de medias. Esto muestra visualmente que es muy probable que los datos sigan una distribución normal multivariante.
El segundo conjunto de datos simula una estructura con tres clusters. La Figura 8 muestra el resultado de aplicar el algoritmo de la Sección 2.1, para obtener la proyección buscada. En esta figura se distingue claramente tres clusters, mostrando un notable desvío de la normalidad (estructura). El plano polar donde se ha proyectado la estructura simulada (Figura 8), es la base del Histograma Circular 3D mostrado en la Figura 9. Obsérvese la no simetría de este Histograma Circular 3D y en cada anillo las alturas de las barras son muy diferentes. No hay concentración de los datos alrededor del origen polar como se esperaría bajo el supuesto de normalidad. Adicionalmente, se aprecia tres grupos de datos en el Histograma Circular 3D, justo donde las barras se agrupan formando bloques.

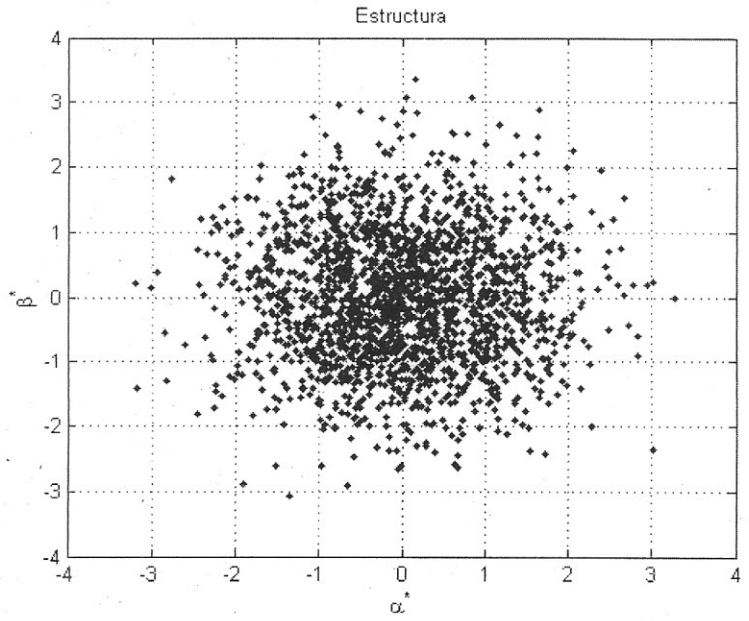

Figura 5: Estructura obtenida mediante la Projection Pursuit

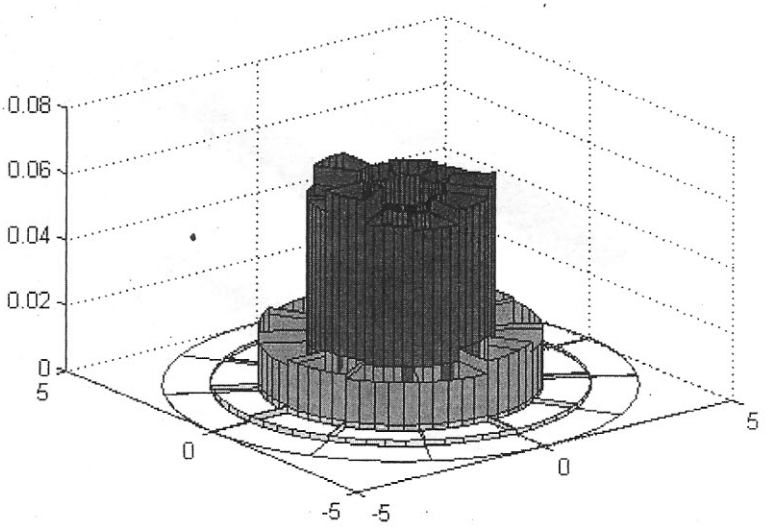

Figura 6: Histograma Circular 3D utilizando la sentencia ***surf***

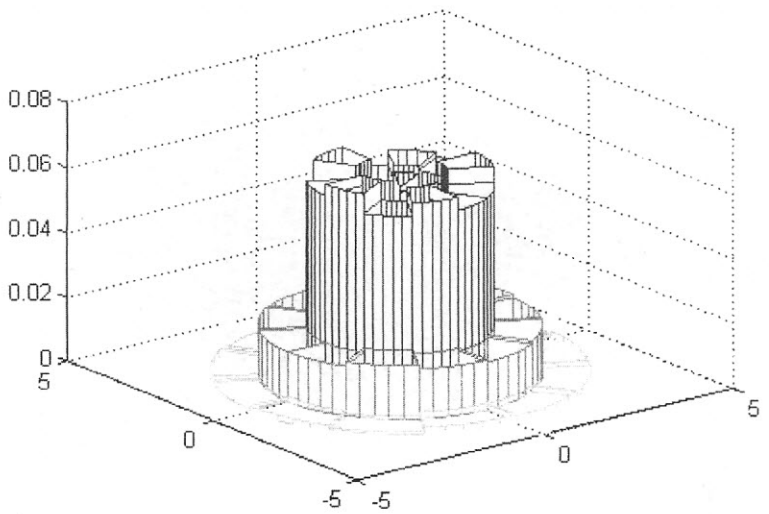

Figura 7: Histograma Circular 3D utilizando la sentencia ***mesh***

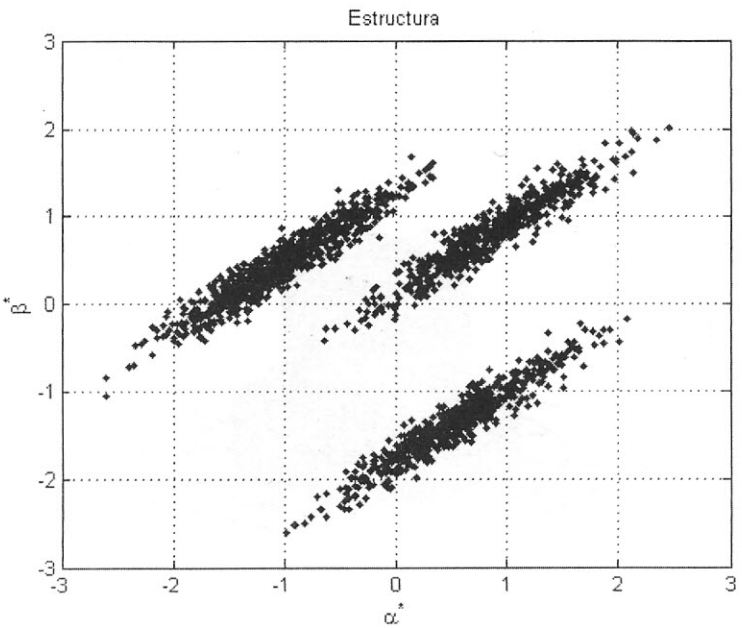

Figura 8: Estructura obtenida mediante la Projection Pursuit

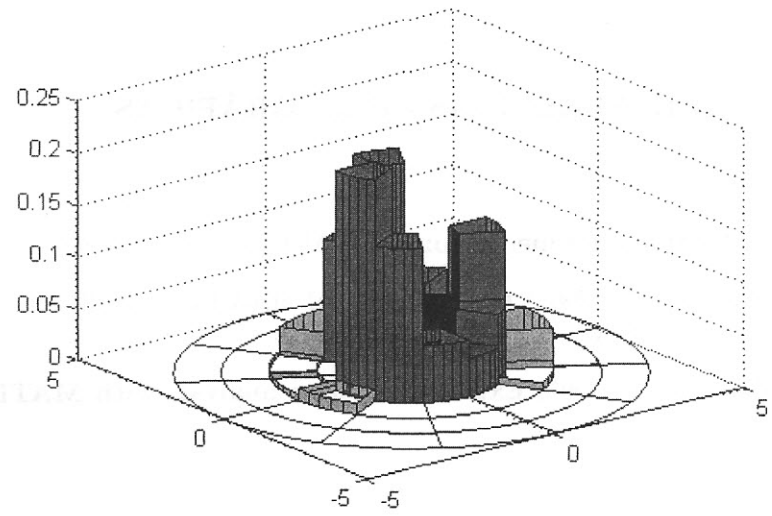

Figura 9: Histograma Circular 3D de la estructura simulada

4. Conclusiones

1. Mediante la Projection Pursuit se redujo la dimensión de los conjuntos de datos simulados, proyectándolos en un plano polar de manera que no se detectó estructuras en la primera simulación y si se detectó en el otro conjunto de datos.

2. El Histograma Circular 3D tiene como base el plano polar, que por su diseño se adecúa más al comportamiento elíptico de los contornos de la distribución normal multivariante.

3. Cuando el Histograma Circular 3D muestra más de un agrupamiento de barras en cada anillo, es indicador que existen clusters, cosa que difícilmente se puede detectar en un histograma clásico.

4. El Histograma Circular 3D ha demostrado ser una herramienta útil para el análisis visual de la simetría de datos multidimensionales.

5. El Histograma Circular 3D ha demostrado ser una herramienta adecuada para la exploración de la normalidad en datos multidimensionales.

6. El MATLAB mostró su capacidad para representar transformaciones de datos multidimensionales que se expresan matricialmente y también para la visualización espacial al ser un software que interactúa con el usuario.

REFERENCIAS BIBLIOGRÁFICAS

[1] Mase, G. (1977) **Mecánica del medio continuo.** McGraw - Hill, México.

[2] Martinez, W.; Martinez, A. (2002) **Computational statistics handbook with MATLAB.** Chapman y Hall/CRC, U.S.A..

[3] Martinez, W.; Martinez, A. (2005) **Exploratory data analysis with MATLAB.** Chapman y Hall/CRC, U.S.A..