UNIVERSIDAD DE CARABOBO
FACULTAD EXPERIMENTAL DE CIENCIAS Y TECNOLOGÍA
DEPARTAMENTO DE MATEMÁTICAS

COMPARACIÓN DE ESTADÍSTICOS DE BONDAD DE AJUSTE A NORMALIDAD MULTIVARIADA.

Trabajo presentado por La Profesora Mirba Romero
ante la Universidad de Carabobo
para ascender a la Categoría de
Profesor Asistente.

Valencia, Edo. Carabobo
Mayo, 2011.

RESUMEN

En el presente trabajo se hace un estudio comparativo entre algunos estadísticos de sesgo y curtosis y los estadísticos estudiados por Bowman y Foster, Henze y Wagner y Manzotti y Quiroz, como métodos de bondad de ajuste para normalidad multivariada, incluyendo algunos de los estadísticos estudiados recientemente por Meklin y Mundfrom. Se discuten los resultados asintóticos de cada estadístico en los casos en que están disponibles. Mediante el método Monte Carlo se estudia la convergencia de los cuantiles de cada estadístico con respecto a sus cuantiles asintóticos.

Se presentan resultados de simulación para evaluar la potencia de los métodos. Se discute la complejidad computacional de los algoritmos y se compara el desempeño para diferentes tamaños muestrales y diferentes dimensiones.

Finalmente se dan recomendaciones en cuanto a las condiciones en que cada estadístico es preferible respecto a los demás.

Palabras claves: Sesgo, curtosis, bondad de ajuste, normalidad multivariada, distribuciones asintóticas.

ÍNDICE GENERAL

RESUMEN II

ÍNDICE GENERAL III

ÍNDICE DE TABLAS V

ÍNDICE DE FIGURAS VIII

INTRODUCCIÓN 1

1. DISTRIBUCIÓN NORMAL MULTIVARIADA 4

1.1. Definiciones y propiedades 4

2. PRUEBAS DE NORMALIDAD MULTIVARIADA 9

2.1. Pruebas de bondad de ajuste 9

2.1.1. Estadísticos de sesgo y curtosis de Mardia 12

2.1.2. Estadístico de sesgo y curtosis de Srivastava 14

2.1.3. Estadístico de sesgo de Balakrishnan, Brito y Quiroz 15

2.1.4. Estadístico basado en la función característica empírica 17

2.1.5. Estadístico de estimación de densidad de Bowman y Foster 18

2.1.6. Estadístico de esféricos armónicos y funciones radiales 20

3. CUANTILES MONTE CARLO 24

3.1. Resultados y análisis del estudio de simulación 25

3.2. Complejidad computacional y complejidad de los algoritmos 37

4. POTENCIA MONTE CARLO 40

4.1. Resultados y análisis de la potencia Monte Carlo 52

5. CONCLUSIONES Y RECOMENDACIONES **71**

5.1. Conclusiones 71

5.2. Recomendaciones 73

REFERENCIAS **75**

ÍNDICE DE TABLAS

3.1. Cuantiles Monte Carlo para el estadístico de sesgo de Mardia ($\widetilde{\mathbf{b_{1,k}}}$), en dimensión k y tamaño muestral n. 26
3.2. Cuantiles Monte Carlo para el estadístico de curtosis de Mardia ($\widetilde{\mathbf{b_{2,k}}}$), en dimensión k y tamaño muestral n. 27
3.3. Cuantiles Monte Carlo para el estadístico de sesgo de Srivastava ($\widetilde{\bar{\mathbf{b}}^{\mathbf{2}}_{\mathbf{1,k}}}$), en dimensión k y tamaño muestral n. 28
3.4. Cuantiles Monte Carlo para el estadístico de curtosis de Srivastava ($\widetilde{\bar{\mathbf{b}}_{\mathbf{2,k}}}$), en dimensión k y tamaño muestral n. 29
3.5. Cuantiles Monte Carlo para el estadístico de Brito, Balakrishnan y Quiros ($Q_{n,2}$), en dimensión k y tamaño muestral n. 30
3.6. Cuantiles Monte Carlo para el estadístico de Henze y Wagner ($\mathbf{T}_{n,\beta}$), en dimensión $k = 2$ y tamaño muestral n. 31
3.7. Cuantiles Monte Carlo para el estadístico de Henze y Wagner ($\mathbf{T}_{n,\beta}$), en dimensión $k = 5$ y tamaño muestral n. 32
3.8. Cuantiles Monte Carlo para el estadístico de Henze y Wagner ($\mathbf{T}_{n,\beta}$), en dimensión $k = 8$ y tamaño muestral n. 33
3.9. Cuantiles Monte Carlo para el estadístico de Henze y Wagner ($\mathbf{T}_{n,\beta}$), en dimensión $k = 10$ y tamaño muestral n. 34
3.10. Cuantiles Monte Carlo para el estadístico de Bowman y Foster ($\widetilde{\mathbf{J}}^2$), en dimensión k y tamaño muestral n 35
3.11. Cuantiles Monte Carlo para el estadístico de esféricos armónicos y funciones radiales de Manzotti y Quiroz ($Z^2_{2,n}$), en dimensión k y tamaño muestral n. 36

4.1. Potencia Monte Carlo contra la distribución Lognormal en dimensión 2 a un nivel de significancia α=0.05 52
4.2. Potencia Monte Carlo contra la distribución Lognormal en dimensión 5 a un nivel de significancia α=0.05 53

4.3. Potencia Monte Carlo contra la distribución Logística en dimensión 2 a un nivel de significancia α=0.05 54
4.4. Potencia Monte Carlo contra la distribución Logística en dimensión 5 a un nivel de significancia α=0.05 55
4.5. Potencia Monte Carlo contra la distribución Normal Contaminada en dimensión 2 a un nivel de significancia α=0.05 56
4.6. Potencia Monte Carlo contra la distribución Normal Contaminada en dimensión 5 a un nivel de significancia α=0.05 57
4.7. Potencia Monte Carlo contra la distribución Burr-Pareto-Logística en dimensión 2 a un nivel de significancia α=0.05 58
4.8. Potencia Monte Carlo contra la distribución Burr-Pareto-Logística en dimensión 5 a un nivel de significancia α=0.05 59
4.9. Potencia Monte Carlo contra la distribución Weibull en dimensión 2 a un nivel de significancia α=0.05 60
4.10. Potencia Monte Carlo contra la distribución Weibull en dimensión 5 a un nivel de significancia α=0.05 61
4.11. Potencia Monte Carlo contra la distribución Seno Hiperbólico en dimensión 2 a un nivel de significancia α=0.05 62
4.12. Potencia Monte Carlo contra la distribución Seno Hiperbólico en dimensión 5 a un nivel de significancia α=0.05 63
4.13. Potencia Monte Carlo contra la distribución uniformemente distribuida sobre el cubo unitario en dimensión 2 a un nivel de significancia α=0.05 . . 64
4.14. Potencia Monte Carlo contra la distribución uniformemente distribuida sobre el cubo unitario en dimensión 5 a un nivel de significancia α=0.05 . . 64
4.15. Potencia Monte Carlo contra la distribución uniformemente distribuida sobre la bola unitaria en dimensión 2 a un nivel de significancia α=0.05 . . . 65
4.16. Potencia Monte Carlo contra la distribución uniformemente distribuida sobre la bola unitaria en dimensión 5 a un nivel de significancia α=0.05 . . . 65
4.17. Potencia Monte Carlo contra la distribución Chi cuadrado en dimensión 2 a un nivel de significancia α=0.05 66
4.18. Potencia Monte Carlo contra la distribución Chi cuadrado en dimensión 5 a un nivel de significancia α=0.05 67
4.19. Potencia Monte Carlo contra la distribución Normal sesgada en dimensión 2 a un nivel de significancia α=0.05 68

4.20. Potencia Monte Carlo contra la distribución Normal sesgada en dimensión 5 a un nivel de significancia α=0.05 68
4.21. Potencia Monte Carlo contra la distribución Logística clásica en dimensión 2 a un nivel de significancia α=0.05 69
4.22. Potencia Monte Carlo contra la alternativa Logística clásica en dimensión 5 a un nivel de significancia α=0.05 69

ÍNDICE DE FIGURAS

4.1. Gráfica de la Distribución lognormal multivariada en dimensión $k = 2$ con parámetros $(\sigma_1, \sigma_2, \rho)$= (0.5,0.5,0), (0.05,0.5,0.8) y (0.25,0.25,-0.5) y en dimensión $k = 5$ (la primera coordenada contra la segunda coordenada) con parámetros $(\sigma_1, \sigma_2, \rho_1, \rho_2)$=(0.5,0.5,0,0), (0.05,0.5,0.5,-0.5) y (0.25,0.25,0.25,-0.5). 42

4.2. Gráfica de la Distribución logística multivariada en dimensión $k = 2$ y en dimensión $k = 5$ (la tercera coordenada contra la quinta coordenada) con parámetro α=0.5 y 2. 43

4.3. Gráfica de la Distribución Normal contaminada en dimensión $k = 2$ con ε =0.05, 0.1 y $\mu = (3, 3, \ldots 3)^T$ y en dimensión $k = 5$ (la tercera coordenada contra la quinta coordenada) con ε =0.05, 0.1 y $\mu = (3, 3, \ldots 4)^T$. 44

4.4. Gráfica de la Distribución Burr-Pareto-Logística en dimensión $k = 2$ y en dimensión $k = 5$ (la primera coordenada contra la segunda coordenada) con parámetro α=0.25 y 0.5. 45

4.5. Gráfica de la Distribución Weibull en dimensión $k = 2$ y en dimensión $k = 5$ (la primera coordenada contra la segunda coordenada) para los parámetros (α,β)=(1,1), (1,2) y (1,2.5). 46

4.6. Gráfica de la Distribución Seno hiperbólico inverso normal mltivariado en dimensión $k = 2$ con parámetros $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$=(0,2,0.1,1,0.8) y (0,2,0.25,0.25,0.5) y en dimensión $k = 5$ (la primera coordenada contra la segunda coordenada) con parámetros $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho_1, \rho_2)$=(0,2,0.1,1,0,0) y (0,2,0.25,0.25,0,0). 47

4.7. Gráfica de la Distribución uniformemente distribuida sobre el cubo unitario y sobre la bola unitaria en dimensión k =2 48

4.8. Gráfica de la Distribución chi-cuadrado con d= 3 y 6 grados de libertad, en dimensión k=2 49

4.9. Gráfica de la Distribución normal sesgada en dimensión k=2 con parámetros λ_0= 1 , -1 y λ_1= (1,5), (2,2) y en dimensión k=5 con parámetros λ_0= 1 , -1 y λ_1=(1,2,3,4,5), (2,2,2,2,2). 50
4.10. Gráfica de la Distribución logística clásica en dimensión k=2. 51

INTRODUCCIÓN

A menudo el estudio de datos provenientes de pruebas clínicas, investigación de mercadotécnia de empresas, sociología y experimentos psicológicos involucran datos de respuestas multivariadas. El aumento en las capacidades de cómputo, tanto en la velocidad de procesamiento como en sofisticación de la base de datos, pone a la orden del analista, datos más complejos en cuanto a la cantidad de variables disponibles.

Muchos de los procedimientos requeridos para analizar tales datos, incluyendo MANOVA, análisis discriminante y regresión multivariada, asumen normalidad multivariada. Con frecuencia es necesario que todas las variables que intervienen en un análisis multivariado sean normales, aunque ello no garantiza la normalidad multivariada. El recíproco siempre es cierto, es decir, la normalidad multivariada implica la normalidad de cada variable.

Hasta hace unos treinta (30) años existían muy pocas pruebas de bondad de ajuste para contrastar normalidad multivariada, y las que había estaban altamente restringidas, ya que algunas eran extremadamente complejas para implementar y otras requerían mucho trabajo numérico, porque no se disponía, como hoy, de recursos computacionales de fácil acceso para implementar dichas pruebas.

Hoy, se tiene un gran número de estadísticos disponibles en la literatura y recursos computacionales de alta capacidad de cómputo a bajo costo. Estas ventajas deben ser razón suficiente para promover el empleo rutinario de medidas de sesgo y curtosis y pruebas de bondad de ajuste, por parte de los usuarios de estadística multivariada.

En la literatura se encuentran algunos artículos que comparan diversos métodos. Entre algunos de los estudios de comparación más reciente se tienen:

Romeu y Ozturk (1993), hacen un estudio comparativo de pruebas de bondad de ajuste para normalidad multivariadad y concluyen que los estadísticos de sesgo y curtosis de Mardia son los que mejor se comportan en términos de potencia, contra las alternativas consideradas. Manzotti y Quiroz (2001), estudian dos estadísticos de bondad de ajuste para la hipótesis nula de normalidad multivariada, basados en los promedios sobre la muestra estandarizada de esféricos armónicos multivariados, de funciones radiales y de sus productos, incluyendo en la comparación al estadístico de Bowman y Foster (1993) y al estadístico BHEP (Baringhaus, Henze, Epps y Pulley) de función característica empírica estudiado por Baringhaus y Henze (1988), concluyendo que estos tres estadísticos, y en particular el de Bowman y Foster, son potentes contra una amplia gama de alternativas. Balakrishnan y Quiroz (2004) realizan un estudio sobre una noción vectorial de sesgo y su uso en pruebas para simetría multivariada, observando que dicho estadístico puede también ser utilizado en pruebas de bondad de ajuste a la normalidad multivariada. Meklin y Mundfrom (2005)) hacen un estudio de comparación Monte Carlo del error tipo I y tipo II para pruebas de la normalidad multivariante y recomiendan el estadístico de Henze y Zirkle (1990). Dicha recomendación la hacen en función del los resultados del error tipo I y de la potencia contra un grupo diverso de trece (13) alternativas. Farrel y otros (2006) realizan un estudio sobre pruebas para normalidad multivariada donde se hace una comparación similar a la de Meklin y Mundfrom (2005)) llegando a conclusiones que también son similares a las de Meklin y Mundfrom (2005)), basándose en un número más pequeño de pruebas. Sin embargo, a pesar de toda la literatura existente, parece haber lugar para un nuevo estudio comparativo, pues algunos métodos potencialmente muy útiles, tales como el de Kotz y otros (2000), el de Bowman y Foster (1993) y el de Balakrishnan y Quiroz (2004) no han sido suficientemente estudiados en las comparaciones más divulgadas.

Por lo expuesto anteriormente, se hace necesario actualizar la comparación de estadísticos de bondad de ajuste a normalidad multivariada, incluyendo las pruebas que no fueron consideradas en estudios previos, mencionadas arriba, para orientar la escogencia por parte de los analistas de datos.

Este trabajo se estructura de la siguiente manera En el capítulo 1 se da la teoría básica sobre la distribución normal multivariada. En el capítulo 2 se expone la teoría básica sobre pruebas de normalidad multivariada. y se describen brevemente las características de los estadísticos considerados. En el capítulo 3 se presentan los resultados de la implementa-

ción computacional (a través del método Monte Carlo) de cada uno de los estadísticos estudiados, incluyendo sus cuantiles aproximados bajo la hipótesis nula de normalidad multivariada y se presentan los resultados de teoría asintótica, existentes en la literatura que proporcionan cuantiles límites. En el capítulo 4 se calcula la potencia de cada estadístico contra un conjunto diverso de alternativas y se establecen comparaciones tomando en cuenta la literatura. Se evalúa el desempeño de cada estadístico para la hipótesis nula de normalidad multivariada, para distintas dimensiones y distintos tamaños muestrales.

Por último, se dan las conclusiones y recomendaciones para el analista de datos, en cuanto a cual estadístico resulta preferible en cada caso (dimensión, tamaño muestral, tipo de alternativa, etc).

Capítulo 1

DISTRIBUCIÓN NORMAL MULTIVARIADA

Una gran parte de los métodos estadísticos univariados y multivariados se basan en los supuestos de normalidad uni y multivariada, respectivamente. Con frecuencia es necesario que todas las variables que intervienen en un análisis multivariado sean normales, aunque ello no garantice la normalidad multivariada. El recíproco siempre es cierto, es decir, la normalidad multivariada implica la normalidad de cada variable. Muchas variables consideradas en estudios presentados en la literatura, poseen de manera natural una distribución normal. Este es el caso cuando se consideran dimensiones físicas (por ejemplo talla) de seres vivos o de piezas construidas por el hombre. Sin embargo, otras variables importantes y que se estudian con frecuencia hoy en día, tales como tiempos de vida de pacientes o de dispositivos o variables asociadas al flujo de información por internet no son naturalmente normales y resulta importante que el analista establezca la normalidad de los vectores estudiados antes de aplicar ciertos procedimientos estadísticos. A continuación se presenta brevemente la teoría básica de la distribución normal multivariada Rencher (2002), Johnson y Wichern. (200), Johnson (1987):

1.1. Definiciones y propiedades

Definición 1. *Un vector aleatorio* $\mathbf{X}_i$*, (con* $i = 1, 2, \ldots, n$*) tiene una distribución normal multivariada si y sólo si cada combinación lineal de las* k*-componentes de* $\mathbf{X}_i$ *tiene una distribución normal univariada. Se escribe* $\mathbf{X}_i \backsim \mathcal{N}_k(\mu, \mathbf{\Sigma})$*, donde* μ *es el vector de*

medias $k \times 1$ y $\mathbf{\Sigma}$ es la matriz de covarianza $k \times k$. Si la matriz $\mathbf{\Sigma}$ es singular, entonces con probabilidad uno la distribución de $\mathbf{X}_i$ está en un subespacio de $\mathbb{R}^k$. Si la matriz $\mathbf{\Sigma}$ tiene rango completo k, entonces la función de densidad de $\mathbf{X}_i$ es

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right\}, \quad para \quad \mathbf{x} \in \mathbb{R}^k. \tag{1.1}$$

Si $\mathbf{\Sigma}$ es semidefinida positiva, pero no definida positiva, entonces $\mathbf{X}$ tiene una distribución degenerada.

Una distribución es normal multivariada si y sólo si $t^T\mathbf{X}$ tiene una distribución normal univariada para todo vector $t \in \mathbb{R}^k$.

Propiedades 1.1.1. *Sea $\mathbf{X}_i \backsim \mathcal{N}_k(\mu, \mathbf{\Sigma})$ con $\mathbf{\Sigma}$ definida positiva , entonces*

a) La distribución es simétrica alrededor de μ, en el sentido que se describe más adelante.

b) Como mencionamos anteriormente, cualquier combinación lineal fija de $\mathbf{X}_i$, digamos $t^T\mathbf{X}_i$ con $t \in \mathbb{R}^k$, también se distribuye normal, es decir, $t^T\mathbf{X}_i \backsim \mathcal{N}(t^T\mu, t^T\Sigma t)$

c) Sea $\delta \in \mathbb{R}^k$, un vector fijo, entonces $\mathbf{X}_i + \delta \backsim \mathcal{N}_k(\mu + \delta, \mathbf{\Sigma})$.

d) La densidad tiene un único máximo en μ:
Como $\mathbf{\Sigma}$ es definida positiva, el término $(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)$ es siempre positivo y la función de densidad será máxima cuando dicho término sea igual a cero, lo cual ocurre si $\mathbf{X}_i = \mu$.

e) Al cortar la densidad de $\mathbf{X}_i \backsim \mathcal{N}_k(\mu, \mathbf{\Sigma})$ con hiperplanos paralelos al definido por las k variables que definen el vector aleatorio $\mathbf{X}_i$, se obtienen las curvas de nivel cuya ecuación es:

$$(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) = C^2 \tag{1.2}$$

Las curvas de nivel son elipsoides y definen una medida de la distancia de un punto al centro de la distribución.
El sólido elipsoidal de los valores de $\mathbf{x}$ que satisfacen

$$(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) \leq \mathcal{X}_k^2(\alpha) \tag{1.3}$$

tiene probabilidad $1-\alpha$ para la distribución $\mathcal{N}_k(\mu, \mathbf{\Sigma})$, siendo $\mathcal{X}_k^2(\alpha)$ el cuantil $1-\alpha$ de la distribución chi-cuadrado con k grados de libertad. Esto equivale a decir que

el cuadrado de la distancia de Mahalanobis $D^2 = (\mathbf{X}_i - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{X}_i - \mu)$ *sigue una distribución chi-cuadrado con* k *grados de libertad.*

Estos se demuestra considerando lo siguiente:

Como $\mathbf{\Sigma}$ *es definida positiva, existe una matriz* $\mathbf{A}$ *cuadrada y simétrica que verifica que* $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^{\mathbf{T}}$ *(por ejemplo la factorización de Cholesky de* $\mathbf{\Sigma}$*).*

Si $\mathbf{X}_i \backsim \mathcal{N}_k(\mu, \mathbf{\Sigma})$*, entonces* $\mathbf{Z} = \mathbf{A}^{-1}(\mathbf{X}_i - \mu) \backsim \mathcal{N}_k(\mathbf{0}, \mathbf{I})$*,*

donde las componentes de $\mathbf{Z}$ *son variables aleatorias independientes normal estándar, es decir,* $Z_i \backsim \mathcal{N}_k(0, 1)$*. Luego* $D^2 = \mathbf{Z}^T\mathbf{Z} = \sum_i Z_i^2 \backsim \chi_k^2$*.*

y por lo tanto,

$(\mathbf{X}_i - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{X}_i - \mu) \backsim \chi_k^2$

f) Los semi-ejes del elipsoide que contiene probabilidad $(1-\alpha)$ *está dado por los autovalores* (λ_i) *y autovectores* $(\mathbf{e}_i)$ *de* $\mathbf{\Sigma}$*, de tal forma que los semi-ejes son* $\pm\sqrt{\lambda_i}\mathbf{e}_i$*.*

g) Si $\mathbf{X}_i \backsim \mathcal{N}_k(\mu, \mathbf{\Sigma})$ *y* $\mathbf{\Sigma}$ *es desconocido, se reemplaza a* $\mathbf{\Sigma}$ *por su estimador insesgado* $\mathbf{S}$ *(matriz de covarianza muestral), en el cálculo de la distancia de Mahalanobis y se obtiene el estadístico* T^2*-Hotelling Johnson (1987) , el cual está definida por :*

$T^2 = n(\mathbf{X}_i - \mu_0)^T \mathbf{S}^{-1}(\mathbf{X}_i - \mu_0)$*, con* $n \geq k + 1$*.*

Bajo la hipótesis nula de que $\mu = \mu_0$

$$\frac{n-k}{k(n-1)} T^2 \backsim F \tag{1.4}$$

con k *y* $n-k$ *grados de libertad. Si* $k = 1$*, la* T^2 *es un estadístico t-Student elevado al cuadrado.*

La matriz de covarianza muestral $\mathbf{S}$ *y el vector de media muestral* $\overline{\mathbf{X}}$*, se calculan como:*

$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T$ *y* $\overline{\mathbf{X}} = [\overline{\mathbf{X}}_\mathbf{1}, \overline{\mathbf{X}}_\mathbf{2}, \ldots, \overline{\mathbf{X}}_\mathbf{k}]^T$ *donde* $\overline{\mathbf{X}}_\mathbf{j} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$ *con* $j = 1, 2, \ldots, k$

Teorema 1.1.1. ***Descomposición Espectral:***

Sea $\mathbf{\Sigma}$ *una matriz real simétrica definida no-negativa, entonces* $\mathbf{\Sigma} = \sum_{i=1}^{k} \lambda_i \mathbf{e}_i \mathbf{e}_i^T$, *donde* $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_k \geq 0$ *son los autovalores de la matriz* $\mathbf{\Sigma}$ *y* $\mathbf{e}_1$, $\mathbf{e}_2$,... ,$\mathbf{e}_k$ *son los correspondientes autovectores. Además*

$$\mathbf{\Sigma} = \mathbf{B} \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_k \end{pmatrix} \mathbf{B}^T \tag{1.5}$$

donde $\mathbf{B}$ *es la matriz ortogonal* $(\mathbf{e}_1 \mid \mathbf{e}_2 \ldots \mid \mathbf{e}_k)_{k\times k}$, *con* $\mathbf{e}_i^T \mathbf{e}_i = 1$ *y* $\mathbf{e}_j^T \mathbf{e}_i = 0$ *para* $i \neq j$. *También se tiene,*

$$\mathbf{B}^T \mathbf{\Sigma} \mathbf{B} = \mathbf{B}^T \mathbf{B} \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_k \end{pmatrix} \mathbf{B}^T \mathbf{B} = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_k \end{pmatrix}$$

De lo anterior se obtiene fácilmente

Proposición 1.1.1. *La raíz cuadrada de la matriz* $\mathbf{\Sigma}$ *es*

$$\mathbf{\Sigma}^{1/2} = \mathbf{B} \begin{pmatrix} \sqrt{\lambda_1} & & & \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ & & & \sqrt{\lambda_k} \end{pmatrix} \mathbf{B}^T = \sum_{i=1}^{k} \sqrt{\lambda_i} \mathbf{e}_i \mathbf{e}_i^T$$

similarmente,

$$\mathbf{\Sigma}^{-1/2} = \sum_{i=1}^{k} \frac{1}{\sqrt{\lambda_i}} \mathbf{e}_i \mathbf{e}_i^T = \mathbf{B} \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}} & & & \\ & \frac{1}{\sqrt{\lambda_2}} & & \\ & & \ddots & \\ & & & \frac{1}{\sqrt{\lambda_k}} \end{pmatrix} \mathbf{B}^T$$

y $\quad \mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2} = \mathbf{\Sigma}$; $\quad \mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{-1/2} = \mathbf{\Sigma}^{-1}$; $\quad \mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}\mathbf{\Sigma}^{-1/2} = \mathbf{I}_{k\times k}$

donde $\mathbf{I}_{k\times k}$ *es la matriz identidad de orden k.*

Teorema 1.1.2. ***Ley de los grandes Números:***

Sean $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ *una muestra aleatoria i.i.d de una población con vector de media* μ*, entonces* $\overline{\mathbf{X}}$ *es un estimador consistente de* μ*, es decir, para cualquier* $\epsilon > 0$*,*

$$Pr\left(\|\overline{\mathbf{X}} - \mu\| > \epsilon\right) \longrightarrow 0 \quad \textit{cuando } n \longrightarrow \infty$$

donde $\|\overline{\mathbf{X}} - \mu\|^2 = (\overline{\mathbf{X}} - \mu)^T(\overline{\mathbf{X}} - \mu) = \sum_{i=1}^{k}(\overline{\mathbf{X}}_\mathbf{i} - \mu_i)^2$.

Como una consecuencia inmediata de la ley de los grandes numeros, se tiene que cada $\overline{\mathbf{X}}_\mathbf{i}$ converge en probabilidad a μ_i, para $i = 1, 2, \ldots, k$, entonces

$$\overline{\mathbf{X}} \quad \text{converge} \quad \text{en} \quad \text{probabilidad} \quad \text{a} \quad \mu$$

Teorema 1.1.3. ***Teorema Central del Límite Multivariado:***

Sean $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ *una muestra aleatoria i.i.d de una población con vector de medias* μ *y matriz de covarianza* $\mathbf{\Sigma}$*, entonces*

$$\sqrt{n}(\overline{\mathbf{X}} - \mu) \xrightarrow{d} \mathcal{N}_k(\mathbf{0}, \mathbf{\Sigma}) \quad \textit{cuando} \quad n \longrightarrow \infty.$$

La validez de la aproximación que dá este teorema, para muestras de tamaño medio (digamos $n = 100$) depende de la simetría de la distribución P que genera los datos. Si P es simétrica o cercana a simétrica, la aproximación dada por el teorema será válida para $n = 100$ e incluso para muestras más pequeñas. Pero si P es marcadamente asimétrica, se necesitarán muestras de varios centenares de datos para que la aproximación se cumpla. El Teorema Central del Límite, por otro lado, significa (es equivalente a) que si se tiene cualquier función continua y acotada $h : \mathbb{R}^k \longrightarrow \mathbb{R}$, entonces

$$Eh(\sqrt{n}(\overline{\mathbf{X}} - \mu)) \longrightarrow Eh(\mathbf{A}Z)$$

donde $Z \backsim \mathcal{N}_k(0, 1)$, $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^\mathbf{T}$ y E es el operador esperanza.

Capítulo 2

PRUEBAS DE NORMALIDAD MULTIVARIADA

Antes de utilizar cualquier modelado estadístico, es crucial verificar si los datos satisfacen las suposiciones distribucionales subyacentes. Para la mayoría de los análisis estadísticos multivariados es importante que los datos sigan una distribución normal multivariada, sino exactamente, por lo menos en forma aproximada. En la práctica se usa la distribución normal multivariada por las siguientes razones

1. La distribución normal sirve en algunos casos como modelo auténtico de la población.
2. La distribución muestral de muchos estadísticos es aproximadamente normal.
3. Muchos de los procedimientos estadísticos estándar del análisis multivariante, incluyendo análisis multivariante de varianza (MANOVA), análisis discriminante, correlación canónica, análisis de componentes principales, regresión, estimación, etc., suponen que los datos siguen una distribución normal multivariada.

Cuando no se cumplen estos supuestos distribucionales los métodos de análisis multivarientes pueden arrojar resultados erróneos.

2.1. Pruebas de bondad de ajuste

Una prueba de bondad de ajuste es un método para verificar si el conjunto de datos se ajusta adecuadamente a una determinada distribución (por ejemplo, la distribución normal). Dichas pruebas están basadas en la hipótesis nula de que los datos provienen de

la distribución teórica supuesta. Para formular la hipótesis nula deben tenerse en cuenta los siguientes aspectos

a) La naturaleza de los datos a analizar (el mecanismo que produce los datos puede sugerir el tipo de distribución apropiada).

b) Histograma. En el caso univariado, la forma que tome el histograma de frecuencia es quizás la mejor indicación del tipo de distribución a considerar. En el caso multivariado, pueden considerarse histogramas de proyecciones $t^T\mathbf{X}$ para $t \in \mathbb{R}$.

Entre las pruebas univariadas de bondad de ajuste, se tienen dos pruebas clásicas La chi-cuadrado y la de Smirnov-Kolmogorov. Ambas pruebas miden el grado de ajuste que existe entre la distribución obtenida a partir de la muestra y la distribución teórica que se supone debe seguir esa muestra. La prueba de Smirnov-Kolmogorov se aplica solo a variable continuas y se basa en la comparación de la función de distribución acumulada de los datos observados con respecto a la de una distribución dada, midiendo la máxima distancia entre ambas curvas. La prueba Chi-cuadrado se usa tanto para variables continuas como para variables discretas y puede aplicarse aún con muestras relativamente pequeñas, pues, esta prueba se basa en la comparación entre la frecuencia observada en un intervalo de clase y la frecuencia esperada en dicho intervalo, calculada de acuerdo con la hipótesis nula formulada, determinando así si las frecuencias observadas en la muestra están lo suficientemente cerca de las frecuencias esperadas bajo la hipótesis nula.

Entre las principales carácteristicas que pueden extraerse de una distribución se tienen el sesgo y la curtosis, las cuales, suelen denominarse medidas de forma porque se relacionan con la representación gráfica de los datos.

En el caso univariado, el sesgo es una medida de asimetría, que permite establecer en cual dirección la cola de la distribución es más pesada (tiende a cero más lentamente) y se define como el tercer momento respecto a la media (es decir, $E[(X-\mu)^3]$) y la curtosis es una medida de concentración central o de apuntamiento, indica si las frecuencias están concentradas en el centro y se define como el cuarto momento respecto a la media (es decir, $E[(X-\mu)^4]$) . La distribución se llama leptocúrtica (indica alejamiento de la normalidad) si presenta un elevado grado de concentración alrededor de los valores centrales de la variable, platicúrtica (indica una distribución relativamente plana) si presenta un reducido grado de concentración alrededor de los valores centrales de la variable y es mesocúrtica

(un caso es cuando la distribución es normal) si presenta un grado de concentración medio alrededor de los valores centrales de la variable.

Las medidas de sesgo y curtosis univariadas pueden ser usadas para probar normalidad univariada. Esas medidas son generalizadas para probar la hipótesis nula de normalidad multivariadas.

En Malkovich y Afifi (1973), definen medidas de sesgo y curtosis multivariados, para una muestra aleatoria $\mathbf{X}$ i.i.d, de la siguiente manera:
Para $t \in \Omega_k = \{x \in \mathbb{R}^k : \|x\| = 1\}$ (esfera unitaria k-dimensional), se tiene

$$\mathbf{X} \backsim \mathcal{N}_k(\mu, \mathbf{\Sigma}) \Longleftrightarrow t^T\mathbf{X} \backsim \mathcal{N}_k(t^T\mu, t^T\mathbf{\Sigma}t^T)$$

Por ello, según estos autores, la distribución de una variable aleatoria $\mathbf{X}$ tiene sesgo multivariado si

$$\beta_1(t) = \frac{[E\{(t^T\mathbf{X} - t^TE(\mathbf{X}))^3\}]^2}{[Var(t^T\mathbf{X})]^3} > 0$$

para algún t. La desigualdad anterior se satisface si y sólo si

$$\beta_1^* = \max_{t\in\Omega_k} \beta_1(t) > 0.$$

β_1^* es la medida de sesgo multivariado propuesta por Malkovich y Afifi. Por otra parte, la distribución de una variable aleatoria $\mathbf{X}$ se dice que tiene curtosis multivariada, si

$$[\beta_2(t)]^2 = \left[\frac{E\{(t^T\mathbf{X} - t^TE(\mathbf{X}))^4\}^2}{\{Var(t^T\mathbf{X})\}^2}\right] > 9$$

para algún t. Entonces

$$(\beta_2^*)^2 = \max_{t\in\Omega_k} [\beta_2(t) - 3]^2 > 0.$$

$(\beta_2^*)^2$ es la medida de curtosis multivariado.

Una prueba de tipo chi cuadrado adaptada al caso normal multivarido es la Moore y Stubblbine (1981), quienes usan para su estadístico celdas dependientes de los datos acotadas por hiperelipses, las cuales son superficies de función de densidad de probabilidad constante para la distribución normal con parámetros estimados de los datos.

Mudholkar, MacDermott y Srivastava (1992) propusieron una prueba que es una adaptación al caso multivariado de la prueba de Lin y Mudholkar (1980) para normalidad

univariada. Einmahl y Mason (1992) han discutido acerca de las pruebas de bondad de ajuste basadas en cuantiles chi cuadrado generalizados por Farrel y otros (2006).

Los estadísticos que vamos a considerar en base a la efectividad reportada en otros estudios (Meklin y Mundfrom (2005)), Romeu y Ozturk (1993),Farrel y otros (2006)) son los siguientes:

2.1.1. Estadísticos de sesgo y curtosis de Mardia

El trabajo de Mardia (1970) introduce definiciones de medidas de sesgo y curtosis multivariadas para ser usadas en pruebas de normalidad multivariada. Estas medidas han dado origen a una gran variedad de extensiones y generalizaciones. Mardia y otros (1979) y Decarlos (1997) indicaron que las pruebas de hipótesis que implican vectores de media son más sensibles a los efectos de asimetría, mientras que las pruebas que implican matrices de varianza-covarianza son más sensibles a la curtosis, ver por ejemplo Meklin y Mundfrom (2005)). En una serie de publicaciones, Mardia (1970, 1974, 1975) define y discute las propiedades de dos medidas invariantes afín, dadas por (Kotz y otros (2000)), las cuales están dadas por:

$$\beta_{1,k} = E[\{\mathbf{X}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{Y}-\mu)\}^3] \qquad \beta_{2,k} = E[\{(\mathbf{X}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{X}-\mu)\}^2]$$

donde $\mathbf{X}$ una variable aleatoria k-dimensional, con vector de media μ y matriz de covarianza $\mathbf{\Sigma}$ y $\mathbf{Y}$ es una variable aleatoria independiente y con la misma distribución que $\mathbf{X}$. Para el caso univariado, los coeficientes de sesgo y curtosis $\beta_{1,1}$ y $\beta_{2,1}$ se escriben como β_1 y β_2. Para la distribución normal multivariada se tiene que $\beta_{2,k} = k(k+2)$. Además, para cualquier variable aleatoria $\mathbf{X}$ centralmente simétrica, se tiene que $\beta_{1,k} = 0$.
Las medidas de sesgo y curtosis dadas se pueden escribir de manera equivalente, como:

$$\beta_{1,k} = E[(\tilde{\mathbf{X}}\tilde{\mathbf{Y}})^3] \qquad y \qquad \beta_{2,k} = E[(\|\tilde{\mathbf{X}}\|)^4]$$

donde $\tilde{\mathbf{X}} = \mathbf{\Sigma}^{-1/2}(\mathbf{X}-\mu)$ y $\tilde{\mathbf{Y}} = \mathbf{\Sigma}^{-1/2}(\mathbf{Y}-\mu)$ son vectores aleatorios (estandarizados) centrados y reducidos (según Mahalanobis) y $\|\cdot\|$ es la norma Euclidiana.

Mardia propuso una versión muestral (para una muestra $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$) de esas medidas:

$$b_{1,k} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{(\mathbf{X}_i - \overline{\mathbf{X}})^T\mathbf{S}^{-1}(\mathbf{X}_j - \overline{\mathbf{X}})\}^3 \tag{2.1}$$

y

$$b_{2,k} = \frac{1}{n}\sum_{i=1}^{n}\{(\mathbf{X}_i - \overline{\mathbf{X}})^T\mathbf{S}^{-1}(\mathbf{X}_i - \overline{\mathbf{X}})\}^2 \tag{2.2}$$

donde $\overline{\mathbf{X}}$ y $\mathbf{S}$ son el vector de media muestral y la matriz de covarianza muestral, respectivamente. Si $n \geq k+1$, entonces $\mathbf{S}$ es una matriz no singular casi siempre (con probabilidad 1).

Invariancia afín y convergencia asintótica:

Los estadísticos de Mardia son invariantes con respecto a transformaciones afines de los datos (Malkovich y Afifi (1973)), es decir,

si $\mathbf{Y} = \mathbf{A}\mathbf{X} + b$, entonces los coeficientes de sesgo y curtosis de $\mathbf{Y}$ y de $\mathbf{X}$ son iguales, para cada matriz $\mathbf{A} \in \mathbb{R}^{k\times k}$ no singular y cada vector $b \in \mathbb{R}^k$.

Bajo la hipótesis nula de normalidad multivariada, Mardia (1970) estableció los siguientes resultados asintóticos para $b_{1,k}$ y $b_{2,k}$:

$$\widetilde{\mathbf{b_{1,k}}} = \frac{n}{6}b_{1,k} \longrightarrow \chi^2_\nu \quad \text{cuando } n \to \infty \tag{2.3}$$

es decir, $\widetilde{\mathbf{b_{1,k}}}$ se distribuye asintóticamente, como una distribución chi-cuadrado con $\nu = \frac{k(k+1)(k+2)}{6}$ grados de libertad.

$$\widetilde{\mathbf{b_{2,k}}} = \sqrt{n}\left\{\frac{b_{2,k} - \frac{n-1}{n+1}k(k+2)}{\sqrt{8k(k+2)}}\right\} \longrightarrow \mathcal{N}(\mathbf{0},\mathbf{I}) \quad \text{cuando } n \to \infty \tag{2.4}$$

es decir, $\widetilde{\mathbf{b_{2,k}}}$ se distribuye asintóticamente, como una distribución normal estándar. Waringhaus y Henze (1992) y Henze (1994) asumieron una distribución elíptica para $\mathbf{X}$ y establecieron los siguientes resultados asintóticos para $b_{1,k}$ y $b_{2,k}$, respectivamente:

Si $\mathbf{X}$ tiene una distribución elíptica tal que $E[\{(\mathbf{X}-\mu)^T\Sigma^{-1}(\mathbf{X}-\mu)\}^3] < \infty$, entonces $nb_{1,k}$ es asintóticamente distribuido como una combinación lineal de dos variables chi-cuadrado, una con k y la otra con $k(k-1)(k+4)/6$ grados de libertad. Para una distribución más general con $\beta_{1,k} = 0$, la distribución asintótica de $nb_{1,k}$ es también una combinación lineal de distribuciones chi-cuadrado pero con más de dos términos.

2.1.2. Estadístico de sesgo y curtosis de Srivastava

Srivastava (1984), usa componentes principales de la matriz de covarianza $\mathbf{\Sigma}$ y define medidas de sesgo y curtosis multivariado Kotz y otros (2000) de la siguiente manera: Sean $\lambda_1, \lambda_2, \ldots, \lambda_k$ los autovalores de la matriz $\mathbf{\Sigma}$ y sean $\gamma_1, \gamma_2, \ldots, \gamma_k$ las columnas de una matriz Γ tal que $\Gamma^T \mathbf{\Sigma} \Gamma = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)$. Sea $Y_i = \gamma_i^T X$ y $\theta_i = \gamma_i^T \mu$ para $i = 1, 2, \ldots, k$, entonces, Srivastava propone a $\bar{\beta}_{1,k}^2$ y a $\bar{\beta}_{2,k}$ como las medidas respectivas de sesgo y curtosis de $\mathbf{X}$, definidas por:

$$\bar{\beta}_{1,k}^2 = \frac{1}{k} \sum_{i=1}^{k} \left\{ \frac{E[(Y_i - \theta_i)^3]}{\lambda_i^{3/2}} \right\}^2 \qquad \bar{\beta}_{2,k} = \frac{1}{k} \sum_{i=1}^{k} \left\{ \frac{E[(Y_i - \theta_i)^4]}{\lambda_i^2} \right\}$$

Las versiones muestral de dichas medidas son:

$$\bar{b}_{1,k}^2 = \frac{1}{k} \sum_{i=i}^{k} \left\{ \frac{1}{n} \sum_{j=1}^{n} \frac{(\widetilde{\gamma_i}^T X_j - \widetilde{\gamma_i}^T \overline{\mathbf{X}})^3}{\widetilde{\lambda_i}^{3/2}} \right\}^2 \tag{2.5}$$

y

$$\bar{b}_{2,k} = \frac{1}{k} \sum_{i=i}^{k} \left\{ \frac{1}{n} \sum_{j=1}^{n} \frac{(\widetilde{\gamma_i}^T X_j - \widetilde{\gamma_i}^T \overline{\mathbf{X}})^4}{\widetilde{\lambda_i}^2} \right\} \tag{2.6}$$

donde $\widetilde{\lambda_1}, \widetilde{\lambda_2}, \ldots, \widetilde{\lambda_k}$ son los autovalores, $\widetilde{\gamma_1}, \widetilde{\gamma_2}, \ldots, \widetilde{\gamma_k}$ los autovectores de la matriz de covarianza muestral $\mathbf{S}$, $\overline{\mathbf{X}}$ es el vector de medias muestrales y X_j es un dato k-dimensional. $\theta_i = \gamma_i^T \overline{\mathbf{X}}$ y $Y_{ij} = \gamma_i^T X_j$.

El estadístico de sesgo de Srivastava promedia ponderadamente, los sesgos univariados en las direcciones de las componentes principales. Algo similar hace el estadístico de curtosis.

Invariancia afín y convergencia asintótica:

Los estadísticos de sesgo $\bar{b}_{1,k}^2$ y curtosis $\bar{b}_{2,k}$ son invariante ante transformaciones lineales de los datos, esto es, si $\mathbf{Y} = \mathbf{A}\mathbf{X} + b$, entonces los coeficientes de sesgo y curtosis de $\mathbf{Y}$ y de $\mathbf{X}$ son iguales, para cada matriz $\mathbf{A} \in \mathbb{R}^{k \times k}$ no singular y cada vector $b \in \mathbb{R}^k$.

Bajo la hipótesis nula de normalidad multivariada, Srivastava demostró los siguientes

resultados asintóticos:

$$\widetilde{\mathbf{\bar{b}^2_{1,k}}} = \frac{nk}{6}\bar{b}^2_{1,k} \longrightarrow \chi^2_k \quad \text{cuando } n \to \infty \tag{2.7}$$

es decir, $\widetilde{\mathbf{\bar{b}^2_{1,k}}}$ se distribuye asintóticamente como una distribución chi-cuadrado con k grados de libertad.

$$\widetilde{\mathbf{\bar{b}_{2,k}}} = \sqrt{\frac{nk}{24}}(\bar{b}_{2,k} - 3) \longrightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad \text{cuando } n \to \infty \tag{2.8}$$

es decir, $\widetilde{\mathbf{\bar{b}_{2,k}}}$ se distribuye asintóticamente como una distribución normal estándar.

2.1.3. Estadístico de sesgo de Balakrishnan, Brito y Quiroz

En (Balakrishnan y Quiroz (2004)) hacen una modificación del estadístico de sesgo dado en (Malkovich y Afifi (1973)) y estudian las propiedades de una noción de sesgo para muestras multivariadas que proporciona, además de una magnitud, una indicación de la dirección del sesgo presente en la muestra.

Para $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ vectores aleatorios $i.i.d$ en $\mathbb{R}^k$, se obtiene una muestra estandarizada $Z_i = \mathbf{S}^{-1/2}(\mathbf{X}_i - \overline{\mathbf{X}})$ a través de la transformación de Mahalanobis, donde $\mathbf{S}^{-1/2}$ es la inversa de la raíz cuadrada de la matriz de covarianza $\mathbf{S}$, para la muestra.

La medida usual de sesgo de la muestra estandarizada en la dirección $t \in \Omega_k$, está dada por

$$c_{1,n}(t) = \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (t^T Z_i)^3$$

la cual puede verse como una medida signada de sesgo, de la muestra estandarizada en la dirección t.

El vector $tc_{1,n}(t)$ provee una indicación vectorial del sesgo en la dirección t o $-t$. El estadístico vectorial de sesgo que proponen en Balakrishnan y Quiroz (2004) , está dado por:

$$T_n = \int_{\Omega_k} tc_{1,n}(t) d\lambda(t) \tag{2.9}$$

donde λ denota la medida de probabilidad rotacionalmente invariante en la esfera unitaria.

Bajo la hipótesis nula de normalidad multivariada, T_n sigue, asintóticamente, una distribución normal multivariada, con matriz de covarianza $\mathbf{D}$ diagonal y la forma cuadrática

$$Q_n = T_n^T \mathbf{D}^{-1} T_n \tag{2.10}$$

tiene una distribución chi-cuadrado con k grados de libertad. La matriz $\mathbf{D}$ de (2.10) tiene la forma $\mathbf{D} = \sigma^2 \mathbf{I}_k$, donde $\mathbf{I}_k$ es la matriz identidad y σ^2 está dado por:
$\sigma^2 = 15J_4^2 + 9J_{2,2}^2(k-1)(k+1) + 9J_2^2 + 18(k-1)J_4J_{2,2} - 18J_2(J_4 + (k-1)J_{2,2})$
con $J_4 = \dfrac{3}{k(k+2)}$; $J_2 = \dfrac{1}{k}$ y $J_{2,2} = \dfrac{1}{k(k+2)}$, $j \neq l, j \geq 1, l \leq k$.

La forma cuadrática dada en (2.10) es considerada en (Balakrishnan y Quiroz (2004)) como un estadístico de prueba para normalidad multivariada contra la hipótesis alternativa de sesgo, estos definen la versión muestral del estadístico viene dada por:

$$Q_{n,2} = \frac{\|T_n\|^2}{\sigma^2} \tag{2.11}$$

donde T_n es el estadístico dado en (2.9) calculado de la muestra dada, cuya r-ésima coordenada puede calcularse como sigue:

$$T_{n,r} = \sqrt{n} J_4 \frac{1}{n} \sum_{i \leq n} Z_{i,r}^3 + 3\sqrt{n} \sum_{j \neq r} J_{2,2} \frac{1}{n} \sum_{i \leq n} Z_{i,j}^2 Z_{i,r} \tag{2.12}$$

siendo $Z_{i,j}$ es la j-ésima coordenada del dato estandarizado Z_i.

Invariancia afín y convergencia asintótica:

Cuando una transformación afín, $\mathbf{Y} = \mathbf{MX} + c$, es aplicada al conjunto de datos originales, el valor de T_n, es rotado por una matriz ortogonal $\mathbf{U}$. Si $\mathbf{M}$ es una rotación, $\mathbf{U}$ coincidirá con $\mathbf{M}$. En este caso hay invarianza afín pero no en el sentido usual.

Balakrishnan y Quiroz (2004) probaron que bajo la hipótesis nula de normalidad multivariada:

$$Q_{n,2} \longrightarrow \chi_k^2 \quad \text{cuando } n \to \infty \tag{2.13}$$

es decir, $Q_{n,2}$ se distribuye asintóticamente con una distribución chi-cuadrado con k grados de libertad.

2.1.4. Estadístico basado en la función característica empírica

Muchas pruebas propuestas en la literatura han sido criticadas por no ser consistentes o invariantes bajo transformaciones lineales de los datos. Una prueba que es simultáneamente consistente contra cualquier distribución no normal multivariada e invariante fue propuesta por Epps y Pulley (1983):

$$\mathbf{T} = \int_{-\infty}^{\infty} |\phi_n(t) - \widehat{\phi}_0(t)|^2 dG(t) \tag{2.14}$$

donde $\phi_n(t)$ es la función característica empírica, $\widehat{\phi}_0(t)$ es la función característica de la distribución normal que es estimada usando la media y la varianza y $G(t)$ es una función de peso.

Henze y Wagner (1997) presentan una nueva aproximación de la prueba de normalidad multivariada estudiada por Baringhaus y Henze (1988) y por Henze y Zirkler (1990). El estadístico de prueba está dado de la siguiente manera:
Dada los vectores aleatorios $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ y $\beta > 0$.

$$\mathbf{T}_{n,\beta} = n\left(4\mathbb{I}_{\{S_n \text{ es singular}\}} + \mathbf{W}_{n,\beta}\mathbb{I}_{\{S_n \text{ es no singular}\}}\right) \tag{2.15}$$

donde, $\mathbb{I}_{\{.\}}$ es la función indicadora y $\mathbf{W}_{n,\beta}$ es definido como

$$\mathbf{W}_{n,\beta} = \int_{\mathbb{R}^k} |\Psi_n(t) - \widehat{\Psi}_0(t)|^2 \varphi_\beta(t) dt \tag{2.16}$$

la distancia ponderada L^2 entre la función característica empírica de las observaciones estandarizadas $(\Psi_n(t))$ y la función característica de una distribución normal estándar multivariada $(\widehat{\Psi}_0(t))$, las cuales vienen dadas por:
$\Psi_n(t) = \dfrac{1}{n}\sum_{l=1}^{n} \exp(it^T Y_l)$, donde $i^2 = -1$ y $Y_l = \mathbf{S}_n^{-1/2}(\mathbf{X}_l - \overline{\mathbf{X}}_n)$ $(l = 1, 2, \ldots, n)$ son los residuales escalados, y $\widehat{\Psi}_0(t) = \exp\left(-\dfrac{\|t\|^2}{2}\right)$.
La función de peso (función de núcleo) usada es: $\varphi_\beta(t) = (2\pi\beta^2)^{-k/2} \exp\left(-\dfrac{\|t\|^2}{2\beta^2}\right)$ y depende del parámetro β, que puede ser elegido de manera arbitraria.

Los autores obtienen una formula cerrada para el estadístico $\mathbf{W}_{n,\beta}$:

$$\begin{aligned}\mathbf{W}_{n,\beta} &= \frac{1}{n^2}\sum_{j,l=1}^{n}\exp\left(-\frac{\beta^2}{2}\|Y_j - Y_l\|^2\right) \\ &- 2(1+\beta^2)^{-k/2}\frac{1}{n}\sum_{j=1}^{n}\exp\left(-\frac{\beta^2\|Y_j\|^2}{2(1+\beta^2)}\right) + (1+2\beta^2)^{-k/2} \qquad (2.17)\end{aligned}$$

El estadístico (2.16) fue propuesto por Epps y Pulley para el caso $k = 1$, Baringhaus y Henze (1988) extendieron esta noción al caso $k > 1$, Henze y Zirkler (1990) Henze y Zirkler (1990) propusieron una extensión multivariada del estadístico de Epps, Pulley, Baringhaus y Henze (EPBH), incorporando una función de peso que es la función de densidad de una $\mathcal{N}_k(0, \beta^2\mathbf{I}_k)$, con β arbitrario.

Invariancia afín y convergencia asintótica:

Henze y Zirkler (1990) demostraron que el estadístico $\mathbf{T}_{n,\beta}$ es invariante afín, es decir, si $\mathbf{Y} = \mathbf{AX} + b$, entonces $\mathbf{T}_{n,\beta}(\mathbf{Y}) = \mathbf{T}_{n,\beta}(\mathbf{X})$ para cada matriz $\mathbf{A} \in \mathbb{R}^{k\times k}$ no singular y cada vector $b \in \mathbb{R}^k$.

Henze y Zirkler (1990) probaron que la distribución límite de $\mathbf{T}_{n,\beta}$ bajo la hipótesis nula de normalidad multivariada corresponde a la combinación lineal $\sum_{j\geq 1}\lambda_j(\beta)N_j^2$, donde $N_1, N_2, \ldots$ son variables aleatorias normal estándar $i.i.d$ y $(\lambda_j(\beta))_{j\geq 1}$ es la sucesión de autovalores asociadas al operador integral B dado en el Teorema 3.1 de Henze y Zirkler (1990), definido por

$$Bq(s) = \int K(s,t)q(t)\varphi_\beta(t)dt$$

donde
$K(s,t) = exp\left(-\frac{\|s-t\|^2}{2}\right) - \left\{1 + s^Tt + \frac{(s^Tt)^2}{2}\right\}exp\left(-\frac{\|s\|^2 + \|t\|^2}{2}\right)$, con $s, t \in \mathbb{R}^k$.

Pero, no es tan sencillo resolver la ecuación $Bq(s) = \lambda q(s)$ y encontrar una forma explícita para $\lambda_j(\beta)$ (ver Henze y Wagner (1997)).

2.1.5. Estadístico de estimación de densidad de Bowman y Foster

El método de estimación por núcleos es una técnica bien conocida. La estimación de densidades por núcleos adaptables es un método cuya idea básica es considerar el parámetro de suavidad variable para diferentes datos muestrales, según la densidad de

observaciones presentes en un entorno de los mismos. Zonas con baja densidad de observaciones, por ejemplo en densidades con largas colas, permiten un parámetro de suavidad mayor que al mismo tiempo evite distorsiones en las estimaciones resultantes.

El estimador de densidad por núcleo fijo, para datos univariados está definido por:

$$\widehat{f}(x) = n^{-1} \sum_{i=1}^{n} h^{-1} W\left(\frac{x - X_i}{h}\right) \tag{2.18}$$

donde $W(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ es la función de densidad de probabilidad de la distribución normal estándar, referida como la función del núcleo, $X_1, X_2, \ldots X_n$ es la muestra aleatoria de la distribución desconocida y h es el parámetro de suavidad (ancho de ventana). Una extensión simple multivariada de (2.18) está dada por:

$$\widehat{f}(\mathbf{x}) = n^{-1} \sum_{i=1}^{n} h^{-k} W_k\{h^{-1}(\mathbf{x} - \mathbf{X}_i)\}$$

donde W_k es una función de densidad simétrica k-dimensional (W tiene media cero), h es el parámetro de suavidad (ancho de ventana).

Un estadístico de error cuadrático medio integrado basado en la estimación de densidad por núcleos, puede ser construido por analogía con las funciones de Cramer-Von Mises para distribuciones. Este estadístico tiene la forma:

$$\int_{-\infty}^{\infty} \{F(x) - F_n(x)\}^2 w(x) dF(x)$$

donde F es la función de distribución, F_n es la función de distribución empírica de una muestra aleatoria de la distribución F y $w(x)$ es una función de peso.

Un enfoque análogo con estimador de densidad de núcleo fijo conduce al estadístico:

$$\int_{-\infty}^{\infty} \left\{\widehat{f}(x) - \mathcal{N}(x, (1 + h^2))\right\}^2 dx$$

donde $\mathcal{N}(x, (1 + h^2))$ denota la densidad normal univariada en x con media 0 y varianza $1+h^2$ (ver ecuación (1.1)). La razón para restar $\mathcal{N}(x, (1+h^2))$ es que $1+h^2$ es la esperanza del estimador por núcleos.

El correspondiente estadístico multivariado propuesto en (Bowman y Foster (1993)) viene dado por:

$$\mathbf{J}^2 = \int \left\{ \widehat{f}(x) - \mathcal{N}_k(x, (1+h^2)\mathbf{I}_k) \right\}^2 dx \tag{2.19}$$

Cuando $\widehat{f}$ es construido de núcleos normales, entonces integrando (2.19) se obtiene:

$$\begin{aligned} \mathbf{J}^2 = \mathcal{N}_k(\mathbf{0}, 2(1+h^2)\mathbf{I}_k) - \frac{2}{n}\sum_i \mathcal{N}_k(x_i, (1+2h^2)\mathbf{I}_k) \\ + \frac{1}{n}\mathcal{N}_k(\mathbf{0}, 2h^2\mathbf{I}_k) + \frac{2}{n^2}\sum_{i<j} \mathcal{N}_k(x_i - x_j, 2h^2\mathbf{I}_k) \end{aligned} \tag{2.20}$$

donde $\mathbf{I}_k$ es la matriz identidad de orden k y n es el tamaño de la muestra. Según Bowman y Foster, una buena elección del ancho de ventana, para datos multivariados, es:

$$h = \left\{ \frac{4}{n(k+2)} \right\}^{\frac{1}{k+4}}. \tag{2.21}$$

Invariancia afín y convergencia asintótica:

Bowman y Foster (1993) probaron que el estadístico es invariante bajo transformaciones lineales de los datos, es decir, si $\mathbf{Y} = \mathbf{A}\mathbf{X} + b$, entonces $\mathbf{J}^2(\mathbf{Y}) = \mathbf{J}^2(\mathbf{X})$, para cada matriz $\mathbf{A} \in \mathbb{R}^{k\times k}$ no singular y cada vector $b \in \mathbb{R}^k$.

Bajo la hipótesis nula de normalidad multivariada, Bowman y Foster (1993) sugieren usar el ancho de ventana (2.21) para el estadístico $\mathbf{J}^2$, para determinar la convergencia del estimador de núcleos.

Para el momento de la elaboración de este trabajo no se sabía nada acerca de la distribución asintótica del estadístico.

2.1.6. Estadístico de esféricos armónicos y funciones radiales

En (Manzotti y Quiroz (2001)) se estudian dos estadísticos para pruebas de bondad de ajuste de normalidad multivariada, de los cuales sólo se va a estudiar el estadístico basado en promedios sobre la muestra estandarizada de esféricos armónicos y radiales multivariados.

Sean $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ vectores aleatorios $i.i.d$ de una ley de probabilidad P en $\mathbb{R}^k$. Para q entero positivo, sean $f_1, f_2, \ldots, f_q$ funciones reales en $L^2(\mathcal{N}(\mathbf{0}, \mathbf{I}_k))$, donde $\mathbf{I}_k$ es la matriz identidad $k \times k$.
Sea $\mathbf{V}$ una matriz $q \times q$ con entradas

$$v(i,j) = E(f_i f_j) - E(f_i)E(f_j)$$

donde las esperanzas son tomadas con respecto a la distribución $\mathcal{N}(\mathbf{0}, \mathbf{I}_k)$. Sean $\underline{f} = (f_1, f_2, \ldots, f_q)^T$, $\nu_n(f_j) = \dfrac{1}{\sqrt{n}} \displaystyle\sum_{i=1}^{n} (f_j(Y_i) - E(f_j))$ y $\nu_n(\underline{f}) = (\nu_n(f_1), (f_2), \ldots, \nu_n(f_q))^T$, donde $Y_i = T_n(X_i)$ son vectores residuales escalados y $T_n(x) = \mathbf{S}^{-1/2}(x - \overline{\mathbf{X}})$, $x \in \mathbb{R}^k$ es la transformación de Mahalanobis.

La forma general del estadístico de bondad de ajuste estudiado por Manzotti y Quiroz (2001) es:

$$Z_n^2 = \nu_n(\underline{f})^T \mathbf{V}^{-1} \nu_n(\underline{f}) \tag{2.22}$$

donde las funciones f_j involucran funciones armónicas y funciones radiales.

Un esférico armónico de grado j es la restricción a Ω_k (esfera unitaria) de un polinomio homogéneo $p(x)$ de grado j en $\mathbb{R}^k$, tal que $\displaystyle\sum_{i=1}^{k} \frac{\partial^2 p}{\partial x_i^2} \equiv 0$ en $\mathbb{R}^k$.

En dimensión $k = 2$ los esféricos armónicos coinciden con las funciones trigonométricas sobre el circulo unitario. En dimensión superior sus combinaciones lineales son densas (con respecto a la norma sup) en el espacio de funciones continuas en Ω_k.

Se denota a $\mathcal{H}_j$ como el conjunto de esféricos armónicos de grado j en la base ortonormal y $\mathcal{G}_j = \displaystyle\bigcup_{0 \le i \le j} \mathcal{H}_i$. El número de esféricos armónicos linealmente independientes de grado j, en dimensión k, está dado por $LI(k,j) = \begin{pmatrix} k+j-1 \\ j \end{pmatrix} - \begin{pmatrix} k+j-3 \\ j-2 \end{pmatrix}$ con $LI(k,0) = 1$ y $LI(k,1) = k$, para todo k.

Para $x \neq 0$ en $\mathbb{R}^k$ y un número positivo j, se definen las funciones

$$r_j(x) = \| x \|^j \quad y \quad u(x) = \frac{x}{\| x \|}$$

r_1 y u son las coordenadas polares de x.

El estadístico basado en promedios sobre la muestra estandarizada de esféricos armónicos y funciones radiales multivariados, es dado en (Manzotti y Quiroz (2001)) como $Z^2_{2,n}$ y se obtiene reemplazando las funciones f_j dadas en (2.22) por: r_1 y $r_3(h \circ u)$, para cada $h \in \mathcal{G}_2$. El número total de funciones usadas en dimensión k es, $q = \begin{pmatrix} k+1 \\ 2 \end{pmatrix} + k + 1$ y el número de parámetros a estimar es $s = \begin{pmatrix} k+1 \\ 2 \end{pmatrix} + k$.

Sean $h_1, h_2, h_3, \ldots h_{q-2}$ funciones en $\mathcal{H}_1$ y $\mathcal{H}_2$, $f_i = r^3(h_i \circ u)$, para $1 \leq i \leq q-2$, $f_{q-1} = r$, $f_q = r^3$ y $\beta = \Gamma(\frac{k+1}{2})/\Gamma(\frac{k}{2})$. Con esa elección de las funciones la matriz de covarianza $\mathbf{V}$ ahora es dada por

$$\mathbf{V} = \begin{pmatrix} \alpha I_{q-2} & 0 \\ 0 & W \end{pmatrix}$$

donde $\alpha = k(k+2)(k+4)$ y W es una matriz por bloque 2×2 dada por $W = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix}$ con $w_{11} = k - 2\beta^2$, $w_{12} = k(k+2) - 2(k+1)\beta^2$ y $w_{22} = \alpha - 2(k+1)^2\beta^2$

Invariancia afín y convergencia asintótica:

Manzotti y Quiroz (2001) probaron que el estadístico de esféricos armónicos y funciones radiales $Z^2_{2,n}$ es invariante bajo transformaciones lineales de los datos, es decir, si $\mathbf{Y} = \mathbf{AX} + b$, entonces el estadístico de esféricos armónicos y funciones radiales de $\mathbf{Y}$ y de $\mathbf{X}$ son iguales, para cada matriz $\mathbf{A} \in \mathbb{R}^{k\times k}$ no singular y cada vector $b \in \mathbb{R}^k$.

Bajo la hipótesis nula de normalidad multivariada, Manzotti y Quiroz (2001) demostraron

$$Z^2_{2,n} \longrightarrow \sum_{i=1}^{s} \lambda_i G_i^2 + G_q^2 \quad \text{cuando } n \to \infty \tag{2.23}$$

es decir, el estadístico tiene como distribución límite, una combinación lineal de distribuciones Chi-cuadrado, los G_i con $1 \leq i \leq q$ son variables $i.i.d\ \mathcal{N}(0,1)$ y los λ_i son los

autovalores diferentes de cero de la matriz $I_k - \Pi - \mathbf{V}^{-1/2} D J^{-1} D^T \mathbf{V}^{-1/2}$, donde

$$\Pi = \begin{pmatrix} 0 & 0 \\ 0 & \Pi_{2,2} \end{pmatrix} \text{ y } \mathbf{V}^{-1/2} D J^{-1} D^T \mathbf{V}^{-1/2} = \begin{pmatrix} \frac{d_0^2}{\alpha} I_k & 0 & 0 \\ 0 & \frac{d_1^2}{\alpha} I_{q-k-2} & 0 \\ 0 & 0 & 2kRER \end{pmatrix}$$

con $d_0 = \sqrt{k}(k+2)$, $d_1 = \sqrt{2}(k+1)(k+3)\beta/\sqrt{k(k+2)}$. $R = W^{-1/2}$ es una matriz real simétrica 2×2, $E = \begin{pmatrix} e_1^2 & e_1 e_2 \\ e_1 e_2 & e_2^2 \end{pmatrix}$, $\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = R \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$ y $\Pi_{2,2}$ es la matriz de proyección 2×2 sobre el subespacio de $\mathbb{R}^2$ ortogonal a $\mathbf{z}$, es decir,

$$\Pi_{2,2} = \frac{\mathbf{v}\mathbf{v}^\mathbf{t}}{\|\mathbf{v}\|^2} \text{ con } \mathbf{v} = \begin{pmatrix} z_2 \\ -z_1 \end{pmatrix}.$$

Los λ_i vienen dados por

$$\begin{aligned} \lambda_i &= \frac{2}{k+4} && \text{repetido } k \text{ veces} \\ \lambda_i &= 1 - \frac{2(k+1)^2(k+3)^2\beta^2}{k^2(k+2)^2(k+4)} && \text{repetido } \begin{pmatrix} k+1 \\ 2 \end{pmatrix} - 1 \text{ veces.} \end{aligned} \tag{2.24}$$

Hay un último autovalor diferente de cero de la matriz $I_2 - \Pi_{2,2} - 2qRER$, con multiplicidad uno. Este autovalor no tiene una fórmula cerrada y aunque se podría calcular, la expresión que se obtendría resulta ser muy compleja. Por ser el más pequeño (es un valor casi despreciable) de todos los autovalores que aparecen en (2.23) no es tomado en cuenta en este estudio.

Capítulo 3

CUANTILES MONTE CARLO

En la literatura, existe teoría asintótica para la mayoría de los estadísticos estudiados, lo cual permitió establecer la comparación entre los cuantiles Monte Carlo y los cuantiles asintóticos (los que aparecen en la tabla con $n = \infty$), a excepción del estadístico de Bowman y Foster para el cual no se tenia una distribución asintótica al momento de realizar la simulación y del estadístico de Henze y Wagner que a pesar de que se tiene teoría asintótica los cuantiles resultan difíciles de simular, ya que la distribución límite corresponde a una combinación lineal de variables aleatorias normal estándar y de autovalores de un operador integral, para los cuales, no es tan sencillo encontrar una forma explícita.

Bajo la hipótesis nula de normalidad multivariada se estimaron cuantiles de la distribución asintótica de cada estadístico evaluando, mediante simulaciones Monte Carlo, la convergencia de los cuantiles simulados a los cuantiles límites (teóricos). Para ello se genero en cada caso $m = 10000$ muestras k-dimensionales (k= 2, 5, 8 y 10) normal estándar de tamaño n= 20, 50, 100 y 200, calculando el estadístico de interés sobre cada muestra y extrayendo los cuantiles de los 10.000 valores observados del estadístico.

Los cuantiles Monte Carlo obtenidos se compararon, cuando fue posible, con los cuantiles asintóticos proporcionados por la teoría existente en la literatura. En este sentido, se verificó que existe teoría asintótica que permite establecer cuantiles para la mayoría de los estadísticos considerados, con la excepción de los estadísticos de Bowman y Foster (1993) y el de Henze y Wagner (1997). Para este último, aunque existe una teoría asintótica, la distribución límite corresponde a un funcional complejo de un proceso estocástico multivariado, cuyos cuantiles resultan muy difíciles de evaluar por simulación.

3.1. Resultados y análisis del estudio de simulación

A continuación se muestran los resultados obtenidos de los cuantiles Monte Carlo para los distintos estadísticos de sesgo y curtosis. En las Tablas 3.1 hasta la 3.5 se dan los resultados de los estadísticos de sesgo y curtosis, en las Tablas 3.6 hasta la 3.9 se tienen los resultados del estadístico basado en la función característica empírica (con parámetro β= 0.1, 0.5, 1 y 3), en la Tabla 3.10 se tienen los resultados del estadístico de estimación de densidad, y en la Tabla 3.11 los resultados del estadístico de esfericos armónicos y de funciones radiales. En estas tablas, la fila $n = \infty$ corresponde a los cuantiles para la distribución límite.

Tabla 3.1: Cuantiles Monte Carlo para el estadístico de sesgo de Mardia ($\widetilde{\mathbf{b_{1,k}}}$), en dimensión k y tamaño muestral n.

		Cuantiles Simulados					
k	n	50 %	90 %	92.5 %	95 %	97.5 %	99 %
2	20	2.134	5.411	5.931	6.789	8.194	10.188
	50	2.726	6.908	7.578	8.651	10.418	12.975
	100	2.999	7.282	8.112	9.127	10.915	13.372
	200	3.176	7.664	8.383	9.515	11.356	13.907
	∞	3.357	7.779	8.496	9.488	11.143	13.277
5	20	24.29	32.96	34.36	36.27	38.97	42.25
	50	29.62	41.33	42.85	45.46	49.12	53.73
	100	31.88	43.76	45.43	47.87	51.41	57.00
	200	32.86	45.13	46.83	49.26	52.60	58.22
	∞	34.34	46.06	47.66	49.80	53.20	57.34
8	20	86.11	100.1	102.0	104.7	108.7	114.5
	50	103.9	125.3	128.8	132.8	138.5	145.2
	100	111.4	133.6	136.6	140.8	148.1	156.7
	200	115.1	137.5	140.3	144.4	150.5	158.8
	∞	119.3	140.2	142.0	146.6	152.2	158.0
10	20	158.9	175.9	178.2	181.0	186.1	192.2
	50	192.1	220.4	224.1	229.1	238.2	249.4
	100	204.8	235.0	238.9	244.1	251.8	260.8
	200	212.3	241.3	245.3	250.2	257.9	267.9
	∞	219.3	247.3	250.9	255.6	262.0	271.7

En la Tabla 3.1 se muestran los cuantiles Monte Carlo del estadístico $\widetilde{\mathbf{b_{1,k}}}$, en dimensión k y tamaño muestral n. La convergencia es algo lenta y monótona creciente con n, es decir, los cuantiles Monte Carlo convergen desde abajo a los valores límites. Los cuantiles límites proporcionan una buena aproximación a los cuantiles de la muestra finita para $n \geq 200$. El estadístico simulado converge asintóticamente a una distribución chi-Cuadrado con $\frac{k(k+1)(k+2)}{6}$ grados de libertad (ver (2.3)).

Tabla 3.2: Cuantiles Monte Carlo para el estadístico de curtosis de Mardia ($\widetilde{\mathbf{b_{2,k}}}$), en dimensión k y tamaño muestral n.

		Cuantiles Simulados						
k	n	2.5 %	50 %	90 %	92.5 %	95 %	97.5 %	99 %
2	20	-1.254	-0.503	0.379	0.522	0.712	1.021	1.466
	50	-1.529	-0.383	0.777	0.944	1.194	1.646	2.157
	100	-1.632	-0.303	0.979	1.151	1.414	1.843	2.431
	200	-1.713	-0.240	1.093	1.267	1.526	1.960	2.510
5	20	-1.681	-0.871	-0.177	-0.069	0.057	0.303	0.554
	50	-1.822	-0.639	0.429	0.576	0.772	1.120	1.529
	100	-1.872	-0.459	0.716	0.891	1.122	1.449	1.891
	200	-1.958	-0.356	0.917	1.120	1.350	1.742	2.121
8	20	-1.925	-1.280	-0.728	-0.646	-0.541	-0.388	-0.201
	50	-2.086	-0.905	0.0821	0.215	0.394	0.686	0.999
	100	-2.111	-0.664	0.478	0.635	0.839	1.188	1.598
	200	-2.132	-0.469	0.759	0.928	1.164	1.511	1.988
10	20	-2.047	-1.561	-1.094	-1.029	-0.954	-0.834	-0.658
	50	-2.174	-1.092	-0.165	-0.033	0.124	0.360	0.702
	100	-2.291	-0.831	0.306	0.451	0.661	0.996	1.371
	200	-2.164	-0.562	0.679	0.831	1.041	1.391	1.816
	∞	-1.960	0.000	1.282	1.440	1.645	1.960	2.326

En la Tabla 3.2 se muestran los cuantiles simulados del estadístico $\widetilde{\mathbf{b_{2,k}}}$. La convergencia a los cuantiles límites es, en este caso, sumamente lenta, y este problema se agrava al aumentar la dimensión, por lo que resulta imprescindible el empleo de tablas. Los cuantiles obtenidos están por debajo de los cuantiles límites. Se obtiene una buena aproximación a los cuantiles límites para dimensión $k = 2$ y $n \geq 200$. El estadístico simulado converge asintóticamente a una distribución normal estándar (ver (2.4)).

Tabla 3.3: Cuantiles Monte Carlo para el estadístico de sesgo de Srivastava ($\widetilde{\mathbf{\bar{b}^2_{1,k}}}$), en dimensión k y tamaño muestral n.

		Cuantiles Simulados					
k	n	50 %	90 %	92.5 %	95 %	97.5 %	99 %
2	20	0.8152	3.007	3.458	4.118	5.296	7.018
	50	1.053	3.842	4.346	5.058	6.474	8.852
	100	1.229	4.253	4.796	5.597	6.957	8.918
	200	1.251	4.519	5.114	5.930	7.544	9.458
	∞	1.386	4.605	5.181	5.991	7.378	9.210
5	20	2.637	6.091	6.663	7.652	9.285	11.32
	50	3.513	8.083	8.932	10.00	11.79	14.32
	100	3.885	8.599	9.421	10.57	12.27	14.28
	200	4.071	8.891	9.700	10.82	12.50	15.20
	∞	4.351	9.236	10.01	11.07	12.83	15.07
8	20	4.424	8.768	9.553	10.54	12.23	14.47
	50	6.013	11.40	12.48	13.64	15.78	18.56
	100	6.607	12.361	13.41	14.63	16.75	19.40
	200	7.011	13.01	13.97	15.21	17.57	20.49
	∞	7.344	13.36	14.27	15.51	17.53	20.09
10	20	5.728	10.72	11.56	12.69	14.77	17.33
	50	7.567	13.89	14.89	16.31	18.79	22.04
	100	8.364	14.74	15.82	17.26	19.82	22.62
	200	8.863	15.40	16.45	17.76	19.95	22.58
	∞	9.342	15.99	16.97	18.31	20.48	23.21

En la Tabla 3.3 se muestran los cuantiles simulados del estadístico $\widetilde{\mathbf{\bar{b}^2_{1,k}}}$. Hay convergencia a los cuantiles límites, la convergencia es monótona creciente con n y es relativamente más rápida que la del estadístico de sesgo de Mardia (Tabla 3.1), salvo para ciertas fluctuaciones atribuibles al procedimiento de simulación. Se obtiene buena aproximación a los cuantiles límites a partir de $n = 200$. El estadístico simulado converge asintóticamente a una distribución chi-cuadrado con k grados de libertad (ver (2.7)).

Tabla 3.4: Cuantiles Monte Carlo para el estadístico de curtosis de Srivastava ($\widetilde{\mathbf{b_{2,k}}}$), en dimensión k y tamaño muestral n.

		Cuantiles Simulados						
k	n	2.5 %	50 %	90 %	92.5 %	95 %	97.5 %	99 %
2	20	-1.579	-0.812	0.116	0.268	0.477	0.851	1.322
	50	-1.722	-0.598	0.565	0.751	0.984	1.413	2.068
	100	-1.751	-0.456	0.832	1.043	1.330	1.783	2.256
	200	-1.807	-0.333	1.013	1.218	1.479	1.890	2.468
5	20	-2.053	-1.211	-0.283	-0.150	0.035	0.370	0.799
	50	-2.149	-0.8368	0.329	0.517	0.779	1.147	1.598
	100	-2.067	-0.609	0.624	0.805	1.055	1.442	1.930
	200	-2.182	-0.4591	0.8376	1.0253	1.279	1.645	2.071
8	20	-2.471	-1.496	-0.556	-0.417	-0.239	0.046	0.411
	50	-2.397	-1.022	0.136	0.296	0.514	0.877	1.268
	100	-2.355	-0.745	0.495	0.662	0.915	1.247	1.713
	200	-2.215	-0.536	0.750	0.928	1.153	1.517	1.967
10	20	-2.670	-1.656	-0.723	-0.590	-0.423	-0.113	0.180
	50	-2.500	-1.120	0.037	0.209	0.412	0.789	1.23
	100	-2.412	-0.806	0.430	0.587	0.834	1.175	1.587
	200	-2.333	-0.594	0.706	0.889	1.134	1.495	1.889
	∞	-1.960	0.000	1.282	1.440	1.645	1.960	2.326

En la Tabla 3.4 se muestran los cuantiles simulados del estadístico $\widetilde{\mathbf{b_{2,k}}}$. El estadístico simulado converge lentamente a una distribución normal estándar (ver (2.8)), como en el caso del estadístico de curtosis de Mardia (Tabla 3.2).

Tabla 3.5: Cuantiles Monte Carlo para el estadístico de Brito, Balakrishnan y Quiros ($Q_{n,2}$), en dimensión k y tamaño muestral n.

		Cuantiles Simulados					
k	n	50 %	90 %	92.5 %	95 %	97.5 %	99 %
2	20	0.787	2.823	3.242	3.827	4.854	6.225
	50	1.071	3.814	4.333	5.056	6.438	8.542
	100	1.217	4.286	4.781	5.595	6.818	9.174
	200	1.305	4.287	4.859	5.704	7.150	9.100
	∞	1.386	4.605	5.180	5.991	7.378	9.210
5	20	2.150	4.663	5.071	5.671	6.490	7.882
	50	3.233	7.197	7.754	8.728	10.392	12.525
	100	3.719	8.037	8.771	9.814	11.59	14.17
	200	4.047	8.703	9.471	10.52	12.38	14.62
	∞	4.351	9.236	10.008	11.07	12.83	15.09
8	20	2.901	5.189	5.475	6.000	6.724	7.518
	50	5.123	9.660	10.37	11.38	12.96	15.25
	100	6.175	11.47	12.25	13.39	15.13	18.08
	200	6.767	12.48	13.32	14.58	16.65	18.81
	∞	7.344	13.36	14.27	15.51	17.53	20.09
10	20	3.035	4.990	5.267	5.595	6.172	7.062
	50	6.274	10.979	11.68	12.80	14.44	16.52
	100	7.640	13.41	14.26	15.35	17.27	19.67
	200	8.407	14.64	15.58	16.87	19.05	21.33
	∞	9.342	15.99	16.97	18.31	20.48	23.21

En la Tabla 3.5 se muestran los cuantiles Monte Carlo del estadístico de sesgo $Q_{n,2}$ dado en (2.11). Hay convergencia moderadamente rápida en dimensión $k = 2$, mientras que para el resto de las dimensiones la convergencia es moderadamente lenta (para $n = 200$ hay una aceptable aproximación a los cuantiles limites), la convergencia es monótona creciente con n y para cada dimensión k. El estadístico converge asintóticamente a una distribución chi-cuadrado con k grados de libertad (ver (2.13)).

Tabla 3.6: Cuantiles Monte Carlo para el estadístico de Henze y Wagner ($\mathbf{T}_{n,\beta}$), en dimensión $k = 2$ y tamaño muestral n.

	Parámetro	k=2					
n	β	50 %	90 %	92.5 %	95 %	97.5 %	99 %
	0.1	1.245e-05	1.983e-05	2.136e-05	2.322e-05	2.730e-05	3.155e-05
	0.5	0.025	0.0564	0.062	0.071	0.084	0.103
20	1	0.258	0.444	0.477	0.518	0.589	0.672
	3	0.8249	1.078	1.115	1.166	1.254	1.364
	0.1	8.333e-06	1.694e-05	1.858e-05	2.100e-05	2.533e-05	3.075e-05
	0.5	0.029	0.064	0.069	0.0776	0.093	0.112
50	1	0.266	0.465	0.496	0.544	0.617	0.6971
	3	0.8304	1.097	1.133	1.180	1.285	1.390
	0.1	7.016e-06	1.605e-05	1.755e-05	2.006e-05	2.453e-05	3.120e-05
	0.5	0.0306	0.065	0.071	0.079	0.094	0.113
100	1	0.268	0.471	0.506	0.554	0.634	0.735
	3	0.828	1.094	1.131	1.186	1.269	1.368
	0.1	6.582e-06	1.577e-05	1.746e-05	1.980e-05	2.396e-05	2.943e-05
	0.5	0.031	0.064	0.069	0.077	0.090	0.110
200	1	0.267	0.477	0.513	0.556	0.627	0.719
	3	0.828	1.093	1.127	1.178	1.263	1.366

En la Tabla 3.6 se tienen los cuantiles Monte Carlo del estadístico de Henze y Wagner en dimensión 2. Aunque no se dispone de los valores límites se observa que hay convergencia relativamente rápida para los diferentes valores de β y en particular para el parámetro $\beta = 3$.

Tabla 3.7: Cuantiles Monte Carlo para el estadístico de Henze y Wagner ($\mathbf{T}_{n,\beta}$), en dimensión $k = 5$ y tamaño muestral n.

	Parámetro	k=5					
n	β	50 %	90 %	92.5 %	95 %	97.5 %	99 %
	0.1	6.686e-05	8.338e-05	8.572e-05	8.914e-05	9.506e-05	1.023e-04
	0.5	0.133	0.174	0.179	0.186	0.202	0.216
20	1	0.694	0.787	0.800	0.818	0.849	0.884
	3	0.9923	1.009	1.013	1.010	1.031	1.048
	0.1	5.294e-05	7.471e-05	7.820e-05	8.290e-05	9.170e-05	1.016e-04
	0.5	0.145	0.191	0.198	0.207	0.223	0.243
50	1	0.698	0.806	0.822	0.843	0.878	0.917
	3	0.9933	1.015	1.018	1.023	1.030	1.039
	0.1	4.907e-05	7.157e-05	7.480e-05	8.025e-05	8.872e-05	9.874e-05
	0.5	0.149	0.195	0.202	0.212	0.228	0.247
100	1	0.695	0.809	0.824	0.843	0.878	0.918
	3	0.9941	1.015	1.017	1.021	1.027	1.035
	0.1	4.745e-05	7.011e-05	7.320e-05	7.756e-05	8.434e-05	9.517e-05
	0.5	0.153	0.198	0.205	0.213	0.230	0.247
200	1	0.698	0.811	0.826	0.845	0.883	0.932
	3	0.9945	1.015	1.018	1.021	1.027	1.034

En la Tabla 3.7 se tienen los cuantiles Monte Carlo del estadístico de Henze y Wagner en dimensión 5. Se observa que para $\beta = 0{,}1$ hay convergencia algo lenta y la convergencia es decreciente. Para valores del parámetro en el intervalo [0.5,3] hay convergencia rápida.

Tabla 3.8: Cuantiles Monte Carlo para el estadístico de Henze y Wagner ($\mathbf{T}_{n,\beta}$), en dimensión $k = 8$ y tamaño muestral n.

	Parámetro	k=8					
n	β	50 %	90 %	92.5 %	95 %	97.5 %	99 %
	0.1	1.732e-04	1.970e-04	2.00e-04	2.055e-04	2.1259	2.213e-04
	0.5	0.286	0.318	0.324	0.331	0.341	0.355
20	1	0.907	0.936	0.942	0.949	0.959	0.973
	3	0.999	0.999	0.999	0.999	1.000	1.001
	0.1	1.504e-04	1.866e-04	1.922e-04	1.996e-04	2.122e-04	2.270e-04
	0.5	0.306	0.346	0.352	0.361	0.375	0.392
50	1	0.901	0.942	0.947	0.955	0.966	0.979
	3	0.9997	1.000	1.000	1.001	1.002	1.004
	0.1	1.452e-04	1.823e-04	1.873e-04	1.959e-04	2.0976e-04	2.262e-04
	0.5	0.306	0.346	0.352	0.361	0.375	0.392
100	1	0.899	0.944	0.949	0.958	0.971	0.987
	3	0.9996	1.001	1.001	1.001	1.002	1.004
	0.1	1.432e-04	1.805e-04	1.861e-04	1.939e-04	2.064e-04	2.231e-04
	0.5	0.317	0.358	0.365	0.373	0.388	0.406
200	1	0.899	0.944	0.949	0.958	0.971	0.988
	3	0.9996	1.001	1.001	1.002	1.003	1.004

En la Tabla 3.8 se tienen los cuantiles Monte Carlo del estadístico basado en la función característica empírica de Henze y Wagner en dimensión 8. Aunque no disponemos de valores límites se observa convergencia rápida para valores de β>0.1 y en especial para $\beta = 3$.

Tabla 3.9: Cuantiles Monte Carlo para el estadístico de Henze y Wagner ($\mathbf{T}_{n,\beta}$), en dimensión $k = 10$ y tamaño muestral n.

	Parámetro	k=10					
n	β	50 %	90 %	92.5 %	95 %	97.5 %	99 %
	0.1	2.781e-04	3.059e-04	3.094e-04	3.147e-04	3.227e-04	3.321 e-04
	0.5	0.398	0.423	0.425	0.431	0.438	0.447
20	1	0.963	0.975	0.977	0.979	0.985	0.993
	3	0.9990	0.9990	0.9990	0.9990	1.000	1.000
	0.1	2.527e-04	2.992e-04	3.056e-04	3.147e-04	3.309e-04	3.509e-04
	0.5	0.416	0.452	0.457	0.464	0.474	0.488
50	1	0.957	0.976	0.979	0.983	0.989 0.996	
	3	0.9990	0.9990	0.9990	1.000	1.000	1.000
	0.1	2.484e-04	2.983e-04	3.065e-04	3.165e-04	3.317e-04	3.522e-04
	0.5	0.4219	0.460	0.465	0.472	0.483	0.496
100	1	0.956	0.977	0.980	0.983	0.989	0.997
	3	0.9990	1.000	1.000	1.000	1.000	1.000
	0.1	2.465e-04	2.952e-04	3.021e-04	3.125e-04	3.252e-04	3.445e-04
	0.5	0.425	0.464	0.469	0.475	0.488	0.502
200	1	0.955	0.978	0.981	0.9844	0.991	0.999
	3	0.9990	1.000	1.000	1.000	1.000	1.001

En la Tabla 3.9 se tienen los cuantiles Monte Carlo del estadístico basado en la función característica empírica de Henze y Wagner en dimensión 10. Los resultados son muy parecidos a los obtenidos en la Tabla 3.8.

De acuerdo a los resultados obtenidos en todas las dimensiones consideradas y para valores del parámetro $\beta = 1$ y $\beta = 3$, se tiene que hay convergencia rápida de los cuantiles Monte Carlo, la convergencia es monótona creciente con n en algunos casos y en otros casos hay convergencia oscilante, por ejemplo a partir de $n = 50$ los cuantiles parecen comenzar a fluctuar alrededor de sus valores límites (ver Tabla 3.8). Como se dijo arriba, la distribución asintótica del estadístico de Henze y Wagner no es sencilla evaluar por simulación (véase (Henze y Wagner (1997)), páginas: 5-13).

Tabla 3.10: Cuantiles Monte Carlo para el estadístico de Bowman y Foster ($\widetilde{\mathbf{J}}^2$), en dimensión k y tamaño muestral n

		Cuantiles Simulados					
k	n	50 %	90 %	92.5 %	95 %	97.5 %	99 %
2	20	1.716 e-02	2.766 e-02	2.912 e-02	3.179 e-02	3.687 e-02	4.303 e-02
	50	1.857 e-02	2.941 e-02	3.124 e-02	3.470 e-02	3.921 e-02	4.385 e-02
	100	1.950 e-02	2.971 e-02	3.131 e-02	3.360 e-02	3.702 e-02	4.418 e-02
	200	1.908 e-02	2.979 e-02	3.116 e-02	3.306 e-02	3.547 e-02	4.299 e-02
5	20	2.100e-03	2.372e-03	2.406e-03	2.469e-03	2.557e-03	2.666e-03
	50	2.429e-03	2.724e-03	2.772e-03	2.824e-03	2.911e-03	3.025e-03
	100	2.640e-03	2.962e-03	2.998e-03	3.063e-03	3.157e-03	3.278e-03
	200	2.867e-03	3.158e-03	3.215e-03	3.252e-03	3.344e-03	3.458e-03
8	20	1.091e-04	1.120e-04	1.137e-04	1.144e-04	1.157e-04	1.174e-04
	50	1.300e-04	1.362e-04	1.368e-04	1.379e-04	1.395e-04	1.414e-04
	100	1.497e-04	1.554e-04	1.562e-04	1.572e-04	1.587e-04	1.608e-04
	200	1.706e-04	1.763e-04	1.769e-04	1.781e-04	1.796e-04	1.814e-04
10	20	1.261e-05	1.279e-05	1.282e-05	1.287e-05	1.294e-05	1.306e-05
	50	1.553e-05	1.584e-05	1.588e-05	1.595e-05	1.605e-05	1.617e-05
	100	1.818e-05	1.852e-05	1.858e-05	1.862e-05	1.872e-05	1.884e-05
	200	2.126e-05	2.162e-05	2.166e-05	2.173e-05	2.184e-05	2.196e-05

En la Tabla 3.10, se muestran los cuantiles Monte Carlo del estadístico de estimación de densidad de Bowman y Foster estandarizado por $\sqrt{n}$, es decir, $\widetilde{\mathbf{J}}^2 = \sqrt{n}\mathbf{J}^2$. El estadístico $\mathbf{J}^2$ tiende a cero, pero al estandarizarlo con $\sqrt{n}$ parece converger aunque muy lentamente en las dimensiones más altas. Al momento de realizar este estudio no se encontró en la literatura resultados asintóticos sobre este estadístico (véase (Bowman y Foster (1993))).

Tabla 3.11: Cuantiles Monte Carlo para el estadístico de esféricos armónicos y funciones radiales de Manzotti y Quiroz ($Z^2_{2,n}$), en dimensión k y tamaño muestral n.

		Cuantiles Simulados					
k	n	50 %	90 %	92.5 %	95 %	97.5 %	99 %
2	20	1.056	2.658	2.938	3.340	4.116	5.164
	50	1.199	3.256	3.659	4.198	5.210	6.741
	100	1.288	3.498	3.963	4.557	5.805	7.348
	200	1.341	3.649	4.137	4.841	6.045	7.843
	∞	1.458	3.780	4.234	4.896	5.979	7.561
5	20	2.003	3.522	3.776	4.088	4.523	5.111
	50	2.223	4.243	4.564	5.029	5.867	7.041
	100	2.301	4.530	4.925	5.507	6.462	8.070
	200	2.401	4.627	5.080	5.696	6.683	8.408
	∞	2.458	4.759	5.167	5.907	6.928	8.770
8	20	3.074	4.675	4.898	5.187	5.592	6.046
	50	3.109	5.144	5.454	5.919	6.593	7.662
	100	3.215	5.331	5.710	6.205	7.119	8.763
	200	3.305	5.558	5.951	6.644	7.678	9.065
	∞	3.401	5.669	6.104	6.760	7.892	9.372
10	20	3.979	5.548	5.710	5.944	6.295	6.685
	50	3.705	5.817	6.141	6.609	7.414	8.534
	100	3.808	5.984	6.354	6.893	7.969	9.189
	200	3.893	6.141	6.556	7.162	8.132	9.562
	∞	3.949	6.167	6.639	7.276	8.383	10.02

En la Tabla 3.11 se muestran los cuantiles Monte Carlo del estadístico de esféricos armónicos y funciones radiales dado en (2.22). Hay convergencia moderadamente rápida, la convergencia es monótona creciente con n. Los cuantiles límites proporcionan una muy buena aproximación a la de la muestra finita para $n \geq 200$. Los valores obtenidos en todas las dimensiones son más pequeños que los cuantiles limites. La distribución asintótica del estadístico es una combinación lineal de chi-cuadrado, con coeficientes λ_i dados en (2.24).

3.2. Complejidad computacional y complejidad de los algoritmos

Una vez implementados cada uno de los estadísticos con el software estadístico R, se puede dar una breve descripción acerca de la complejidad de los algoritmos existentes y la complejidad computacional en función del tamaño muestral y las dimensiones.

A efectos de comparación, se da el tiempo de corrida para calcular cada estadístico m=10.000 veces en dimensión k =5, para $n = 100$. Para realizar los cálculos y simulaciones en este estudio se uso un procesador Intel Core2 Duo, 2.8 GHz, 2GB de memoria RAM.

Como todos los procedimientos requieren el cálculo de una matriz cuadrada simétrica real y definida positiva ($\mathbf{S}^{-1}$), no se incluyo este cálculo, que es de complejidad $O(k^2)$, aproximadamente, en las comparaciones que se dan a continuación. De los ocho estadísticos solo se dará el detalle del cálculo de la complejidad para el estadístico de sesgo de Mardia, indicando para los demás solamente el resultado.

1. **Estadístico de sesgo de Mardia:**

$$b_{1,k} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{(X_i - \overline{\mathbf{X}})^T\mathbf{S}^{-1}(X_j - \overline{\mathbf{X}})\}^3$$

El número de operaciones para producir $\mathbf{S}^{-1}(X_j - \overline{\mathbf{X}})$ es:
$k(k + k - 1)$
el número de operaciones para producir $(X_i - \overline{\mathbf{X}})^T\mathbf{S}^{-1}(X_j - \overline{\mathbf{X}})$ es:
$(k + k - 1) + k(k + k - 1)$
El número de operaciones para producir $\{(X_i - \overline{\mathbf{X}})^T\mathbf{S}^{-1}(X_j - \overline{\mathbf{X}})\}^3$ es:
$(k + k - 1) + k(k + k - 1) + 1$
Sumar en i y en j produce:
$[(k + k - 1) + k(k + k - 1) + 1]n^2$

De lo anterior se tiene que la complejidad del estadístico ($b_{1,k}$) es de orden n^2k^2. El algoritmo es un poco lento y el tiempo ejecución es de aproximadamente 0.2 segundos.

2. **Estadístico de curtosis de Mardia:**

   Para este estadístico se hace un análisis similar al del estadístico de sesgo de Mardia y se obtiene que la complejidad es de orden nk^2. El algoritmo es mucho más rápido que el de sesgo. El tiempo de ejecución es de aproximadamente 0.09 segundos.

3. **Estadístico de sesgo de Srivastava:**

   Este estadístico tiene complejidad de orden nk^2. El algoritmo es más rápido que el estadístico de sesgo de Mardia pero más lento que el de curtosis de Mardia. El tiempo de ejecución es de aproximadamente 0.13 segundos para el tamaño de muestra considerada en nuestro ejemplo.

4. **Estadístico de curtosis de Srivastava:**

   Este estadístico tiene complejidad de orden nk^2. Es un poco más lento que el estadístico de sesgo de Mardia. El tiempo de ejecución de es de aproximadamente 0.23 segundos.

5. **Estadístico de sesgo de Balakrishnan, Brito y Quiroz:**

   El estadístico tiene complejidad de orden n^2k^2. Es más lento que el estadístico de curtosis de Srivastava. El tiempo de ejecución del algoritmo es de aproximadamente 0.26 segundos.

6. **Estadístico de función característica empírica de Henze y Wagner:** Tiene complejidad de orden n^2k^2. A pesar de que tiene el mismo orden de complejidad que el estadístico de sesgo $Q_{n,2}$, es mucho más lento que este en la práctica, con un tiempo de ejecución de aproximadamente 12.8 segundos en el ejemplo considerado.

7. **Estadístico basado en la función de densidad de Bowman y Foster:**

   Tiene complejidad de orden n^2k^2. Es bastante lento, aunque menos lento que estadístico $\mathbf{T}_{n,\beta}$. El tiempo de ejecución es de aproximadamente 6.5 segundos.

8. **Estadístico de esféricos armónicos y funciones radiales :**

   Tiene complejidad de orden n^2k^3. Este estadístico es más lento que los estadísticos de sesgo y curtosis $b_{1,k}$, $(b_{2,k})$, $\bar{b}_{1,k}^2$, $\bar{b}_{2,k}$ y $Q_{n,2}$, pero es más rápido que los U-estadísticos $\mathbf{T}_{n,\beta}$ y $\mathbf{J}^2$. El tiempo de ejecución es de aproximadamente 0.39 segundos.

En conclusión, se tiene que, por su complejidad y tiempo de ejecución, los procedimientos considerados se pueden clasificar en tres grupos:

i) Curtosis de Mardia, sesgo y curtosis de Srivastava: tienen complejidad lineal en n y resultan los más rápidos en la comparación.

ii) Sesgo de Mardia, sesgo de Balakrishnan, Brito y Quiroz y esféricos armónicos y funciones radiales: con complejidad cuadrática en n y tiempo de ejecución intermedio en la comparación.

iii) El estadístico basado en la función característica empírica de Henze-Zirkle y el de estimación de densidad de Bowman y Foster: aunque su complejidad es $O(n^2k^2)$ los procedimientos resultan los más lentos en la comparación.

Capítulo 4

POTENCIA MONTE CARLO

El concepto de potencia se atribuye históricamente, a Neyman y Pearson (1928, 1933). A partir de entonces, aparecieron una serie de trabajos que consideran la potencia estadística (Cox (1948), McNemar (1960), Sterling (1959), Tukey (1960), Tullock (1959)).

La bondad de una prueba estadística se mide por el tamaño de dos indices de error: α, que es la probabilidad de rechazar la hipótesis nula, cuando es verdadera (error tipo I), y β, que es la probabilidad de aceptar la hipótesis nula, cuando es falsa (error tipo II).

La potencia de una prueba estadística viene dada por la capacidad de rechazar la hipótesis nula correctamente, de modo que está determinada por la probabilidad de cometer un error tipo II. Por lo tanto, la potencia es el complemento de la probabilidad de un error tipo II $(1-\beta)$ (probabilidad de rechazar la hipótesis nula, cuando esta es falsa y debería ser rechazada).

En este capítulo se realiza un análisis comparativo de la potencia a un nivel de significancia α=0.05, de los ocho estadísticos propuestos como pruebas de bondad de ajuste a normalidad multivariada. Cada uno de los estadísticos considerados se calcula en muestras de tamaño n =20, 50, 100 y 200 provenientes de distribuciones alternativas en dimensiones k=2 y 5. En cada caso, el experimento se repite $m = 1000$ veces a efectos del cálculo de la potencia.

Las diferentes distribuciones alternativas multivariadas consideradas en esta comparación fueron las siguientes:

1. **Distribución lognormal multivariada** :

La distribución lognormal multivariada pertenece al sistema de traslación de Johnson (1987), Capítulo 5. Las distribuciones en este sistema pueden obtenerse aplicando transformaciones básicas a las coordenadas de un vector con distribución normal multivariada.
Sea $\mathbf{X} \sim N_k(\mu, \mathbf{\Sigma})$, entonces para cada coordenada X_i del vector $\mathbf{X}$ se aplica la transformación lognornormal:

$$Y_i = \lambda_i \exp(X_i) + \xi_i$$

obteniéndose el vector $\mathbf{Y} = (Y_1, Y_2, \ldots Y_k)^T$ con distribución Lognormal de parámetros μ, $\mathbf{\Sigma}$, (λ_1, λ_2,..., λ_k) y (ξ_1, ξ_2,...,ξ_k).

Dentro de esta familia se pueden dar distribuciones insesgadas, pero vamos a usar solo las que tienen los siguientes valores $\lambda_i = 1$, $\xi_i = \mu_i = 0$, (para $i \leq k$) y $\mathbf{\Sigma}$ vendrá dada, para $k = 2$, por $\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}$, con $(\sigma_1, \sigma_2, \rho)$= (0.5,0.5,0), (0.05,0.5,0.8) y (0.25,0.25,-0.5). Mientras que para $k = 5$

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_1 & 0 & 0 & 0 \\ \sigma_1\sigma_2\rho_1 & \sigma_2^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_1^2 & \sigma_1\sigma_2\rho_2 & \sigma_1^2\rho_1\rho_2 \\ 0 & 0 & \sigma_1\sigma_2\rho_2 & \sigma_2^2 & \sigma_1\sigma_2\rho_1 \\ 0 & 0 & \sigma_1^2\rho_1\rho_2 & \sigma_1\sigma_2\rho_1 & \sigma_1^2 \end{pmatrix}$$

con $(\sigma_1, \sigma_2, \rho_1, \rho_2)$= (0.5,0.5,0,0), (0.05,0.5,0.5,-0.5) y (0.25,0.25,0.25,-0.5).

Con los parámetros descritos, esta alternativa presenta distintas formas de sesgo y curtosis, es una distribución fuertemente sesgada y de soporte no acotado, como se aprecia en la Figura 4.1.

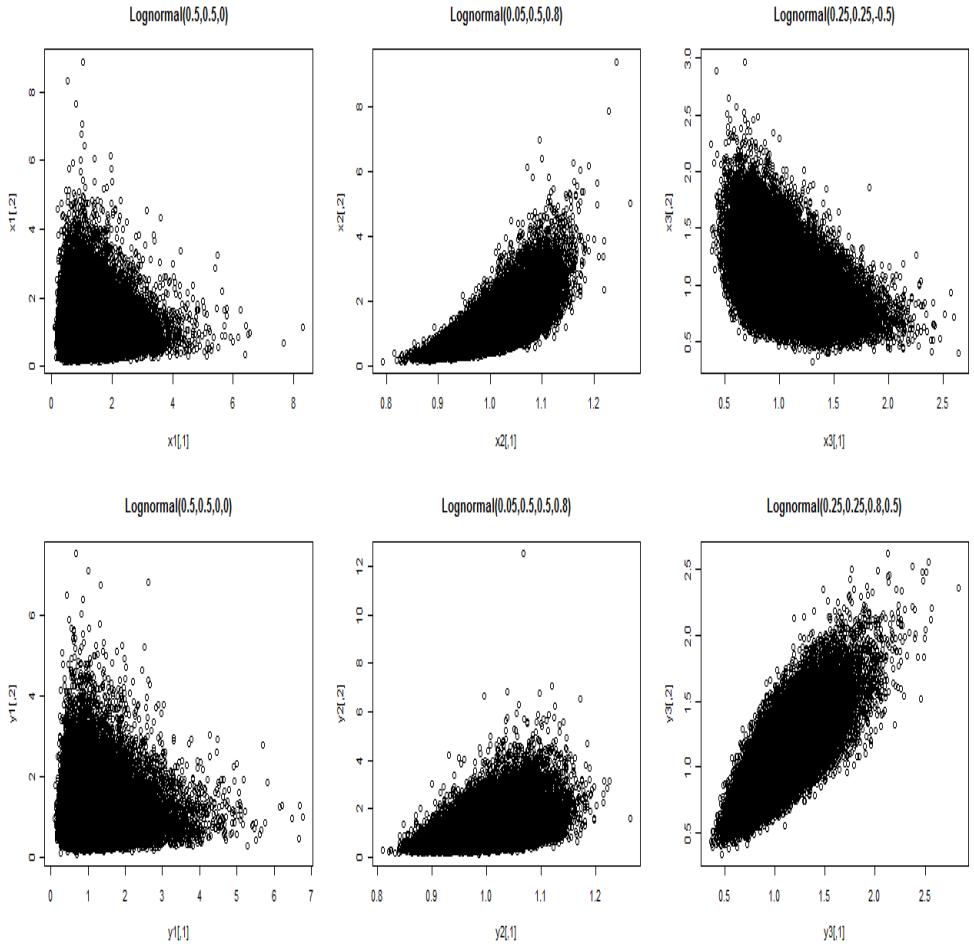

Figura 4.1: Gráfica de la Distribución lognormal multivariada en dimensión $k = 2$ con parámetros $(\sigma_1, \sigma_2, \rho)$= (0.5,0.5,0), (0.05,0.5,0.8) y (0.25,0.25,-0.5) y en dimensión $k = 5$ (la primera coordenada contra la segunda coordenada) con parámetros $(\sigma_1, \sigma_2, \rho_1, \rho_2)$=(0.5,0.5,0,0), (0.05,0.5,0.5,-0.5) y (0.25,0.25,0.25,-0.5).

2. **Distribución Logística Multivariada**:

   La distribución logística pertenece a la familia Burr-Pareto-Logística descrita por Johnson (1987), Capítulo 9. Esta distribución no está tan estrechamente relacionada con la distribución normal como lo están las que pertenecen al sistema de traslación de Johnson, las distribuciones de contornos elípticos o la distribución normal contaminada. Cada coordenada de la distribución logística es obtenida con la siguiente transformación (Johnson (1987), pág. 170, tabla 9.1):

$$Z_i = -log\left(\frac{Y_i}{X}\right)$$

donde el vector $(Y_1, Y_2, \ldots Y_k)^T$ tiene coordenadas i.i.d exp(1) y X se distribuye como una $\Gamma(\alpha, 1)$.

Los parámetros usados en nuestro estudio son α =0.5 y 2. Con 0.5 la distribución es sesgada a la izquierda, de colas pesadas, pero con 2 la distribución presenta distintas formas del sesgo y la curtosis, como se observa en la Figura 4.2.

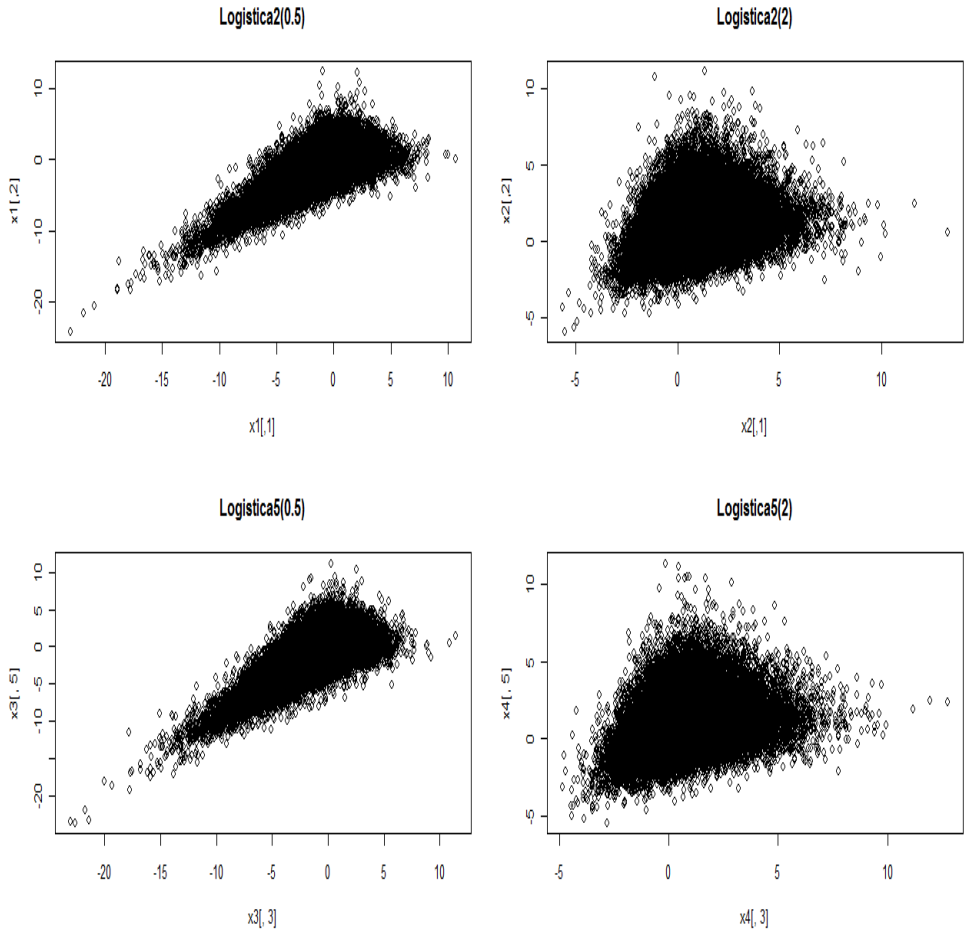

Figura 4.2: Gráfica de la Distribución logística multivariada en dimensión $k = 2$ y en dimensión $k = 5$ (la tercera coordenada contra la quinta coordenada) con parámetro α=0.5 y 2.

3. **Distribución Normal Contaminada Multivariada**:
   Esta distribución es una mezcla de distribuciones normales y viene dada por

$$\rho\mathcal{N}_k(\mathbf{0},\mathbf{I}) + \varepsilon\mathcal{N}_k(\mu,\mathbf{I}).$$

   Donde $\mathbf{I}$ es ma matriz identidad, μ es el vector de medias, ε es la probabilidad de realizar el proceso $\mathcal{N}_k(\mu,\mathbf{I})$ y $\rho = (1-\varepsilon)$ es la probabilidad de realizar el proceso $\mathcal{N}_k(\mathbf{0},\mathbf{I})$.
   En nuestro estudio usaremos ε =0.05, 0.1 y $\mu = (3,3,\ldots 3)^T$ y $(4,4,\ldots 4)^T$.
   Esta distribución es sesgada, de soporte no acotado, como se aprecia en la Figura 4.3.

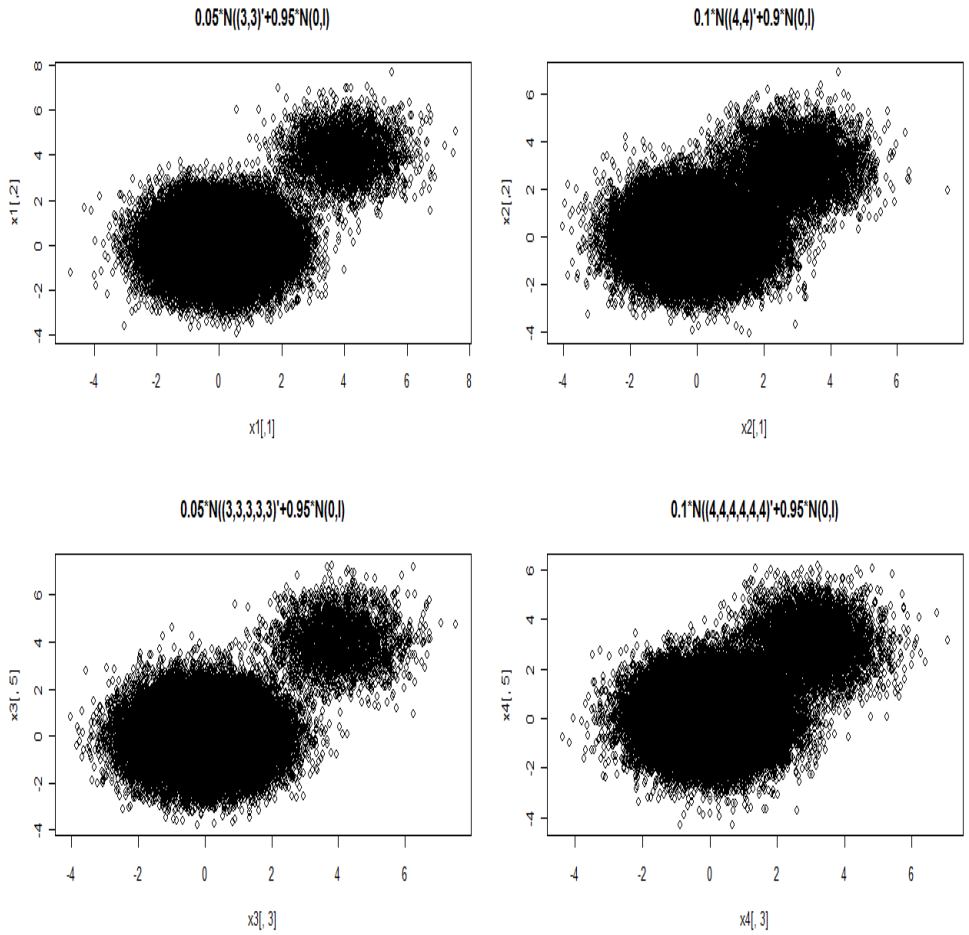

Figura 4.3: Gráfica de la Distribución Normal contaminada en dimensión $k = 2$ con ε =0.05, 0.1 y $\mu = (3,3,\ldots 3)^T$ y en dimensión $k = 5$ (la tercera coordenada contra la quinta coordenada) con ε =0.05, 0.1 y $\mu = (3,3,\ldots 4)^T$.

4. **Distribución Burr-Pareto-Logística**:

   Cada coordenada Z_i de la distribución Burr-Pareto-Logística es obtenida (Johnson (1987), pág. 167, densidad (9.10)) con la siguiente transformación:

$$Z_i = \left(1 + \frac{Y_i}{X}\right)^{-\alpha}$$

   donde el vector $(Y_1, Y_2, \ldots Y_k)^T$ tiene coordenas i.i.d exp(1) X se distribuye como una $\Gamma(\alpha, 1)$.

   Los valores del parámetro que se usan en este estudio son: $\alpha = 0.25$, 0.5, 1 y 2.

   Con los valores 0.25 y 0.5, la distribución es sesgada de colas livianas y soporte compacto. Mientras que para los valores 1 y 2 la distribución es simétrica de colas livianas y soporte compacto, muy parecida a la distribución uniforme en el cubo unitario, como se aprecia en la Figura 4.4.

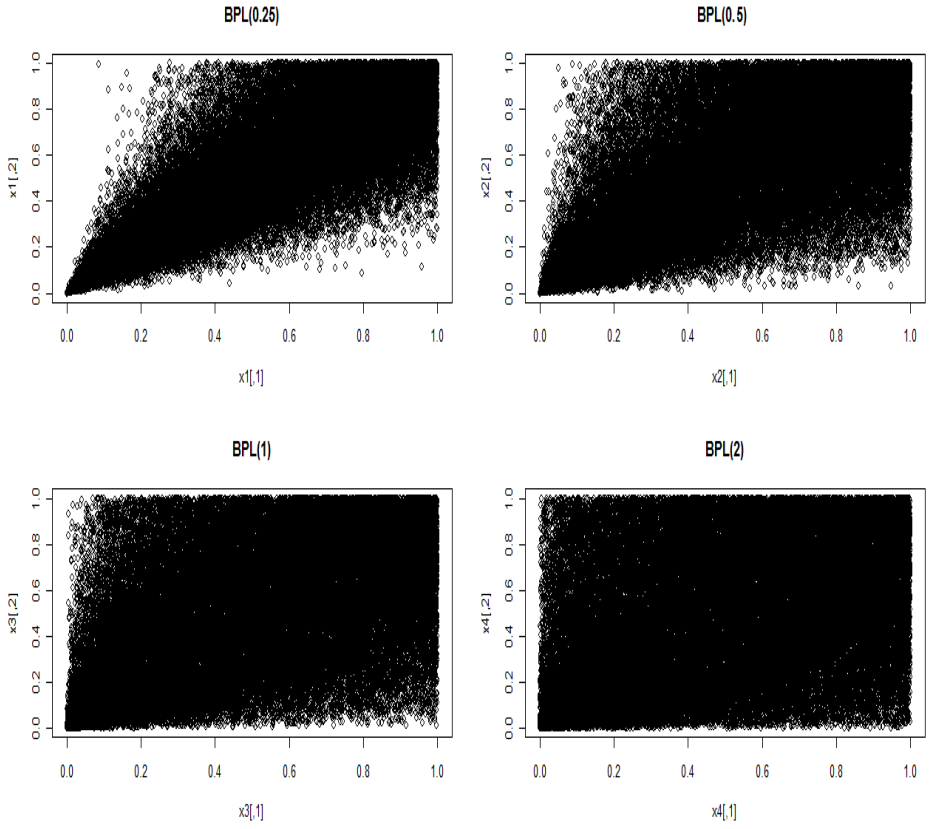

Figura 4.4: Gráfica de la Distribución Burr-Pareto-Logística en dimensión $k = 2$ y en dimensión $k = 5$ (la primera coordenada contra la segunda coordenada) con parámetro α=0.25 y 0.5.

5. **Distribución Weibull**:

   Se consideran muestras multivariadas, donde las coordenadas son i.i.d. Weibull(α, β), siendo α el parámetro de forma y β el parámetro de escala. Esta distribución coincide con la distribución exponencial cuando $\alpha = \beta = 1$. Los valores de los parámetros que se usan en este estudio son: (α,β)= (1,1), (1,2) y (1,2.5). Es una distribución asimétrica positiva, de soporte no acotado, como se observa en la Figura 4.5.

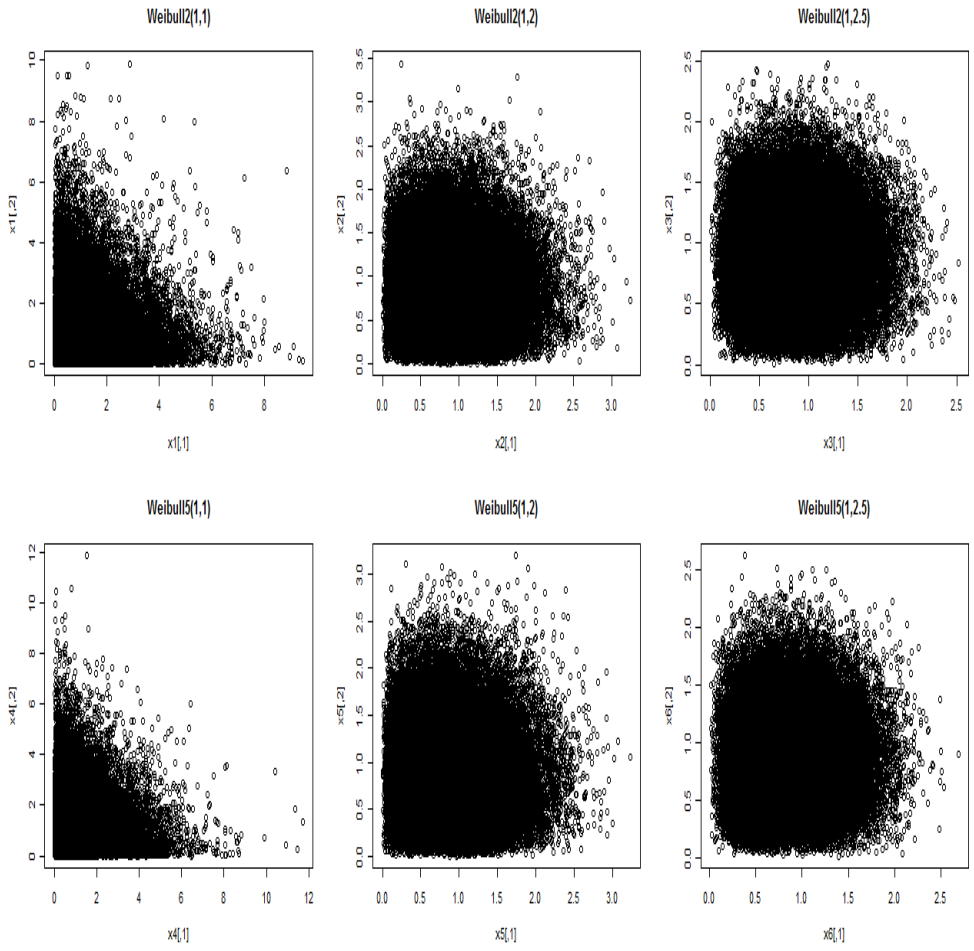

Figura 4.5: Gráfica de la Distribución Weibull en dimensión $k = 2$ y en dimensión $k = 5$ (la primera coordenada contra la segunda coordenada) para los parámetros (α,β)=(1,1), (1,2) y (1,2.5).

6. **Distribución seno hiperbólico inverso normal multivariada** :

   Es otra distribución dentro del sistema de traslación de Johnson (1987), Capítulo 5, pág. 63.

Para $\mathbf{X} \sim N_k(\mu, \mathbf{\Sigma})$, a cada componente X_i de $\mathbf{X}$ se aplica la transformación:

$$Y_i = \lambda_i \operatorname{sinh}(X_i) + \xi_i$$

obteniéndose el vector $\mathbf{Y} = (Y_1, Y_2, \ldots Y_k)^T$ con distribución seno hiperbólico inverso normal, con parámetros μ, $\mathbf{\Sigma}$, $(\lambda_1, \lambda_2, \ldots, \lambda_k)$ y $(\xi_1, \xi_2, \ldots, \xi_k)$. En nuestro caso usaremos los valores $\xi_i = 0$, $\lambda_i = 1$ (con $i \leq k$) y $\mathbf{\Sigma}$ es dado como en la distribución lognormal. Los parámetros usados en este estudio son los siguientes: Para $k = 2$, $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$= (0,2,0.1,1,0.8) y (0,2,0.25,0.25,0.5).
Mientras que para $k = 5$, $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho_1, \rho_2)$= (0,2,0.1,1,0,0) y (0,2,0.25,0.25,0,0). Dentro de esta familia se pueden dar distribuciones insesgadas, pero con estos parámetros, la distribución seno hiperbólico es sesgada y de soporte no acotado. Ver Figura 4.6.

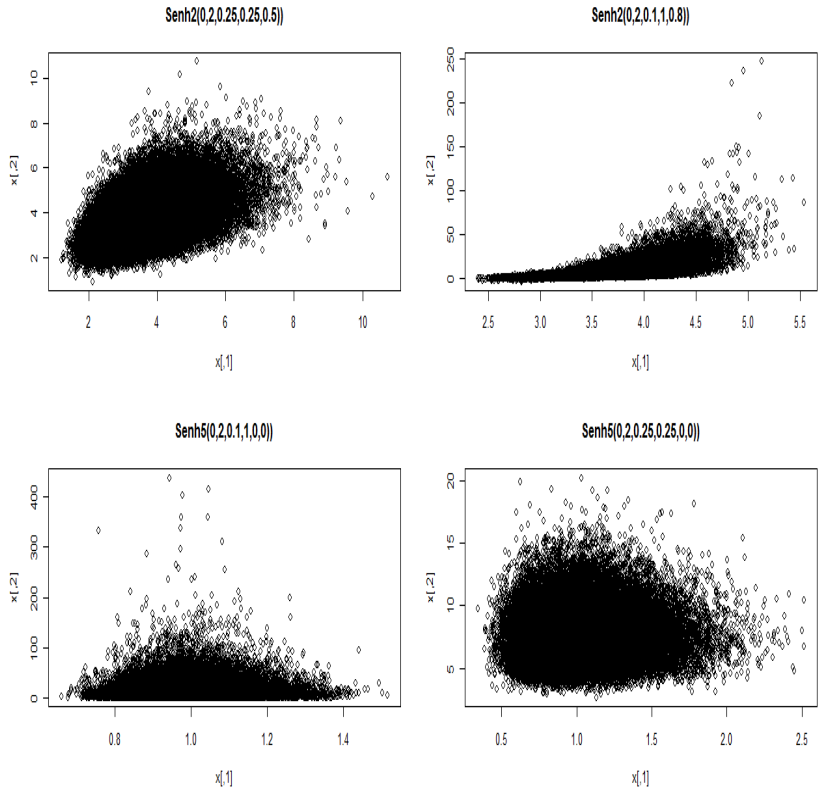

Figura 4.6: Gráfica de la Distribución Seno hiperbólico inverso normal mltivariado en dimensión $k = 2$ con parámetros $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$=(0,2,0.1,1,0.8) y (0,2,0.25,0.25,0.5) y en dimensión $k = 5$ (la primera coordenada contra la segunda coordenada) con parámetros $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho_1, \rho_2)$=(0,2,0.1,1,0,0) y (0,2,0.25,0.25,0,0).

7. **Distribución uniformemente distribuida sobre el cubo unitario** $[0,1]^k$:
   La distribución uniforme en el cubo unitario k-dimensional (con $k = 2$ y $k = 5$), es una distribución de simetría central, más no de simetría elipsoidal y soporte compacto, como se puede observar en la Figura 4.7a).

8. **Distribución Uniformemente distribuida sobre la bola unitaria** $[-1,1]^k$:
   La distribución uniforme en la bola unitaria k-dimensional (con $k = 2$ y $k = 5$), es una distribución esféricamente simétrica y soporte compacto, como se observa en la Figura 4.7b).

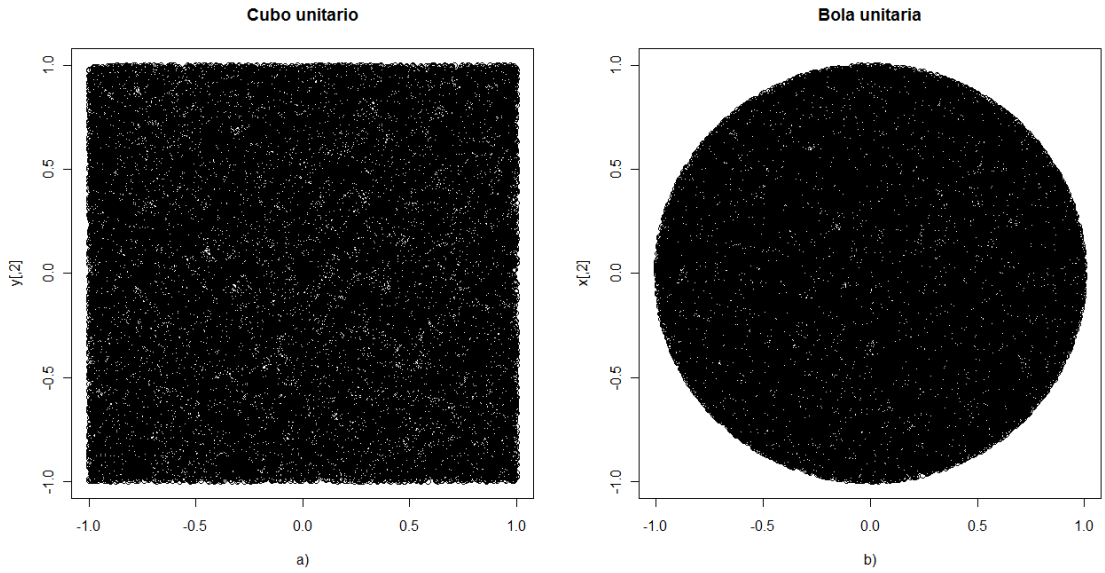

Figura 4.7: Gráfica de la Distribución uniformemente distribuida sobre el cubo unitario y sobre la bola unitaria en dimensión k =2

9. **Distribución Chi-Cuadrado con** d **grados de libertad**:
   Se consideran muestras multivariadas, donde cada componente es i.i.d como una chi-cuadrado con d grados de libertad.
   En este estudio se consideraron los siguientes valores para el parámetro d= 3 y 6. Esta distribución es fuertemente sesgada, de soporte no acotado, como se puede apreciar en la Figura 4.8.

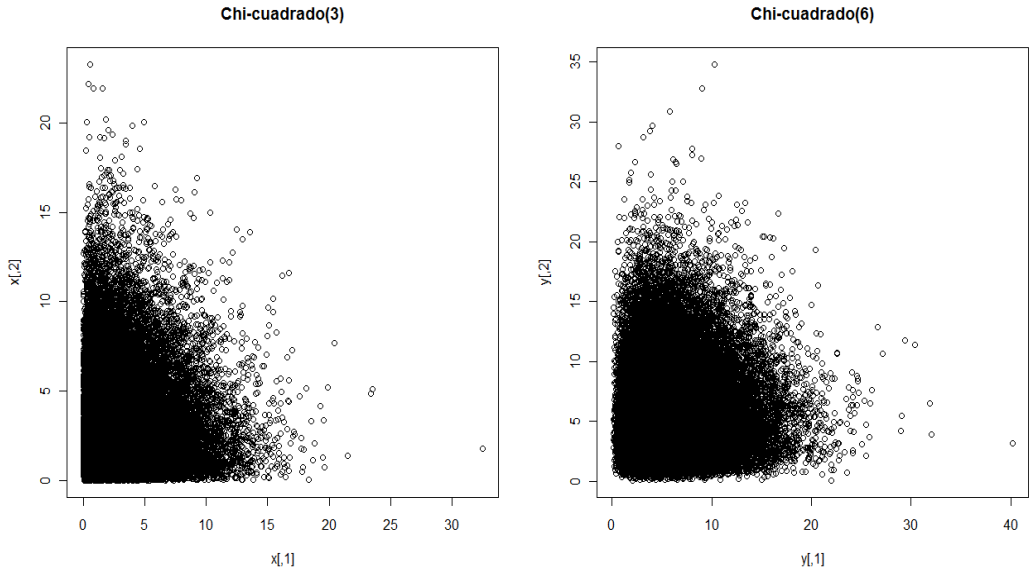

Figura 4.8: Gráfica de la Distribución chi-cuadrado con d= 3 y 6 grados de libertad, en dimensión k=2

10. **Distribución Normal-Sesgada de Arnold-Beaver**:
    Sea λ_0 un número real y un vector k-dimensional $\lambda_1 = (\lambda_{1,1}, \lambda_{1,2}, \ldots, \lambda_{1,k})$. Para obtener un vector normal-sesgado $X = (X_1, X_2, \ldots, X_k)^T$ se generan $k+1$ variables normal estándar i.i.d $Z_1, Z_2, \ldots, Z_k$ y W. Entonces, con $\kappa = \sqrt{1 + \|\lambda_1\|^2}$, $c = \dfrac{\lambda_0}{\kappa}$, se tiene que cada coordenada de X es de la forma:

$$X_i = \frac{Z_i - \lambda_{1,i} W(c)}{\kappa}$$

    donde $W(c)$ es W truncado por arriba en c. A grandes valores de $\lambda_{1,i}$ tiende a aumentar el sesgo en la coordenada X_i, mientras que para valores pequeños de la constante λ_0 aumenta el sesgo en todas las coordenadas.
    Los parámetros usados en este estudio para $k = 2$ son : λ_0= 1 , -1 y λ_1= (1,5), (2,2).
    Mientras que para $k = 5$: λ_0= 1 , -1 y λ_1=(1,2,3,4,5), (2,2,2,2,2).
    Como se observa en la Figura 4.9, la familia normal sesgada permite considerar datos con sesgo moderado.

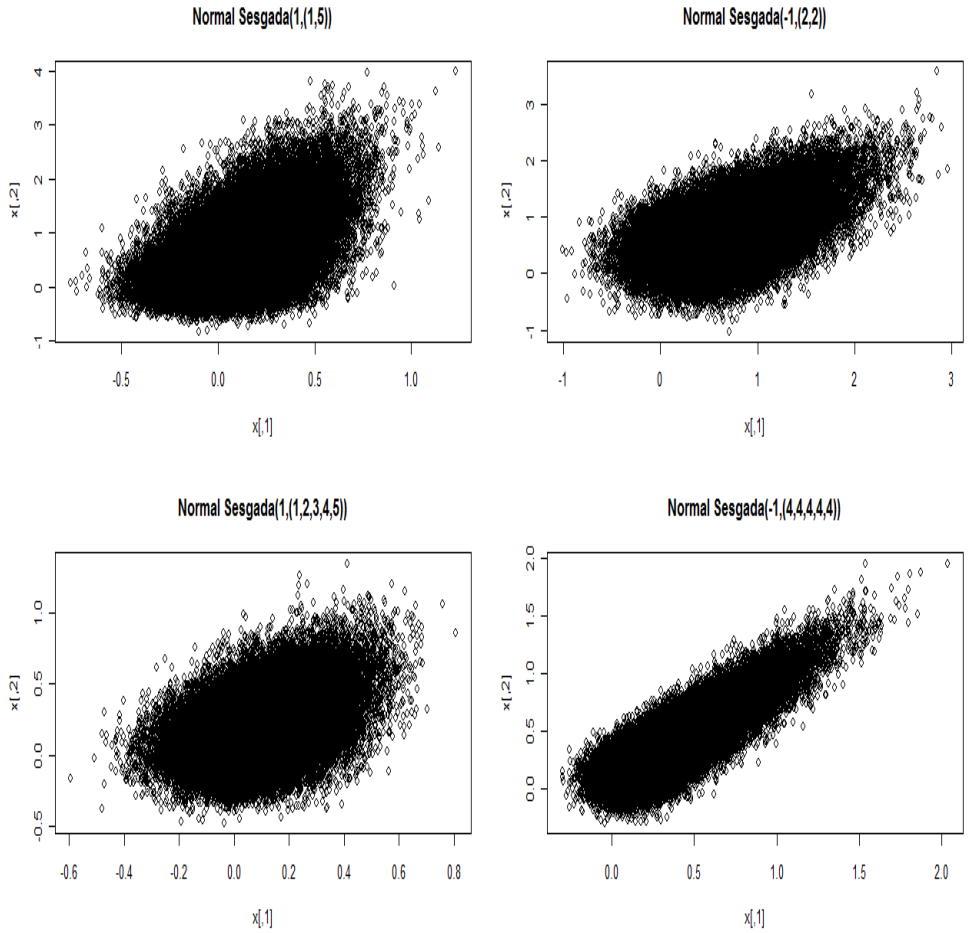

Figura 4.9: Gráfica de la Distribución normal sesgada en dimensión k=2 con parámetros λ_0= 1 , -1 y λ_1= (1,5), (2,2) y en dimensión k=5 con parámetros λ_0= 1 , -1 y λ_1=(1,2,3,4,5), (2,2,2,2,2).

11. **Distribución Logística Clásica**:

    Se considera una muestra multivariada con coordenadas i.i.d logísticas con parámetro de localización igual a 0 y parámetro de escala igual a 1.

    Las proyecciones bivariadas de esta distribución forman nubes aproximadamente elípticas con eje principal en la diagonal y presentan sesgo a la izquierda como puede apreciarse en la gráfica. Su forma es parecida a la distribución normal, pero tiene colas más pesadas que la normal, como se aprecia en la Figura 4.10.

    En la literatura, la distribución logística se ha considerado, históricamente, una de las alternativas más difíciles de detectar por las pruebas de normalidad.

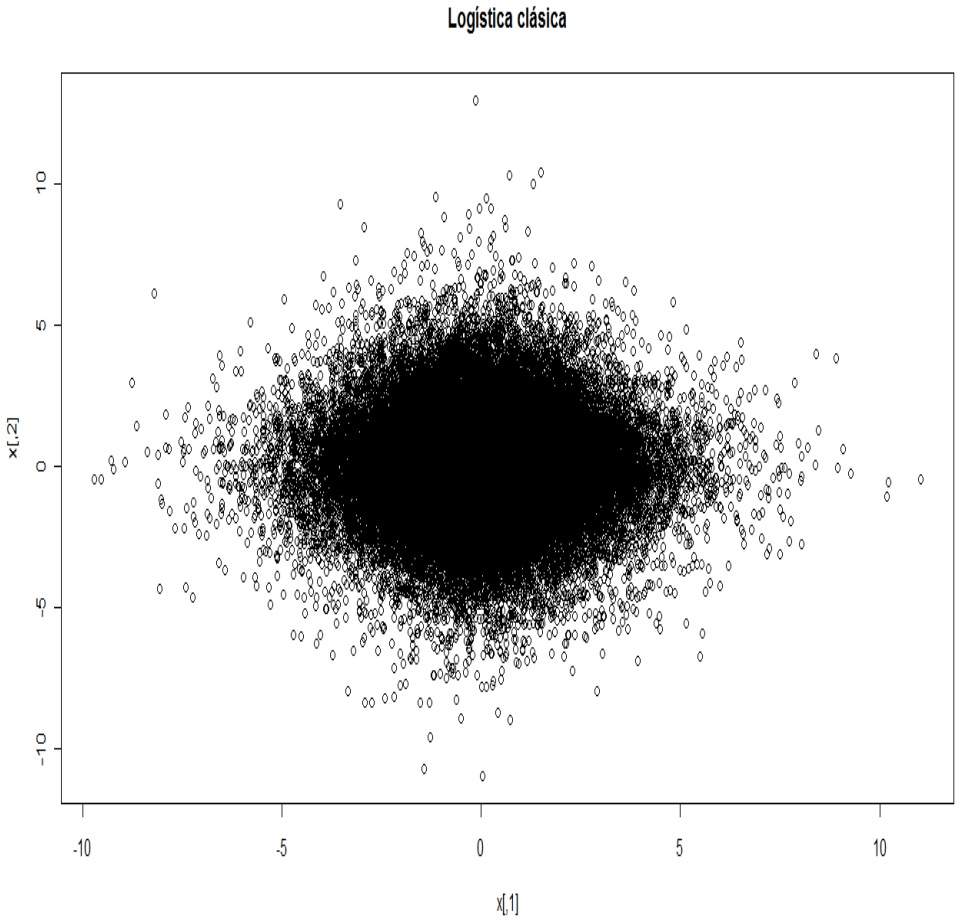

Figura 4.10: Gráfica de la Distribución logística clásica en dimensión k=2.

4.1. Resultados y análisis de la potencia Monte Carlo

A continuación se muestran los resultados obtenidos de las potencias Monte Carlo de los estadísticos considerados ($\widetilde{\mathbf{b_{1,k}}}$: estadístico de sesgo de Mardia, $\widetilde{\mathbf{b_{2,k}}}$: estadístico de curtosis de Mardia, $\widetilde{\mathbf{\bar{b}^2_{1,k}}}$: estadístico de sesgo de srivastava, $\widetilde{\mathbf{\bar{b}_{2,k}}}$: estadístico de curtosis de srivastava, $Q_{n,2}$: estadístico de sesgo de Balakrishnan, Brito y Quiroz, $\mathbf{T}_{n,\beta}$: estadístico basado en la función característica empírica de Henze y Wagner, $\widetilde{\mathbf{J}}^2$: estadístico de estimación de densidad de Bowman y Foster y $Z^2_{2,n}$: estadístico de esféricos armónicos y funciones radiales de Manzotti y quiroz), para los diferentes tamaños muestrales, en dimensiones k =2 y 5, y para el conjunto de alternativas no normales dadas.

El cálculo de la potencia del estadístico $\mathbf{T}_{n,\beta}$ se realizó con β =0.5, el cual es el parámetro dado en (Henze y Wagner (1997) y es diferente al error tipo II (β) :

Tabla 4.1: Potencia Monte Carlo contra la distribución Lognormal en dimensión 2 a un nivel de significancia α=0.05

Alternativa	n	Estadístico							
		$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{\bar{b}^2_{1,k}}}$	$\widetilde{\mathbf{\bar{b}_{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z^2_{2,n}$
Lognormal (0.5,0.5,0)	20	0.578	0.342	0.534	0.293	0.582	0.625	0.592	0.358
	50	0.958	0.727	0.912	0.66	0.966	0.982	0.937	0.825
	100	1	0.935	0.99	0.888	0.998	1	0.998	0.994
	200	1	0.996	1	0.991	1	1	1	1
Lognormal (0.05,0.5,0.8)	20	0.428	0.235	0.394	0.212	0.384	0.416	0.405	0.262
	50	0.861	0.519	0.856	0.454	0.801	0.908	0.796	0.615
	100	0.996	0.771	0.989	0.681	0.988	0.998	0.975	0.922
	200	1	0.959	1	0.91	1	1	1	0.998
Lognormal (0.25, 0.25,-0.5)	20	0.211	0.126	0.112	0.077	0.197	0.22	0.175	0.116
	50	0.579	0.233	0.363	0.126	0.512	0.589	0.356	0.301
	100	0.912	0.4	0.225	0.181	0.828	0.896	0.673	0.589
	200	0.999	0.636	0.609	0.263	0.986	0.999	0.924	0.898

Contra las tres alternativas dadas en la Tabla 4.1, se tiene que el estadístico de Henze y Wagner ofrece globalmente la mejor potencia, seguido por el estadístico de Bowman y Foster y por los estadísticos específicos de sesgo principalmente el de Mardia y el de Bahalakrisna, Brito y Quiroz. Los estadísticos de curtosis y el estadístico de esféricos

armónicos y funciones radiales presentan poca potencia frente a este tipo de alternativas. Para muestras pequeñas.

Tabla 4.2: Potencia Monte Carlo contra la distribución Lognormal en dimensión 5 a un nivel de significancia α=0.05

Alternativa		Estadístico							
	n	$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{\bar{b}^2_{1,k}}}$	$\widetilde{\mathbf{\bar{b}_{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z^2_{2,n}$
Lognormal (0.5,0.5,0,0)	20	0.604	0.485	0.535	0.371	0.626	0.687	0.537	0.25
	50	0.993	0.92	0.899	0.742	0.983	0.999	0.976	0.911
	100	1	0.999	0.997	0.954	1	1	1	1
	200	1	1	1	0.999	1	1	1	1
Lognormal (0.05,0.5,0.5,-0.5)	20	0.283	0.146	0.439	0.223	0.27	0.329	0.208	0.104
	50	0.813	0.495	0.875	0.528	0.746	0.848	0.576	0.551
	100	0.991	0.782	0.971	0.804	0.982	0.994	0.926	0.912
	200	1	0.963	0.999	0.962	1	1	0 .999	0.998
Lognormal (0.25,0.25,0.8,-0.5)	20	0.212	0.109	0.131	0.101	0.234	0.246	0.169	0.064
	50	0.654	0.377	0.343	0.216	0.658	0.736	0.427	0.387
	100	0.986	0.61	0.681	0.395	0.932	0.982	0.747	0.747
	200	1	0.883	0.959	0.679	1	1	0.982	0.986

De la Tabla 4.2, se tiene que al aumentar la dimensión no disminuye la potencia. En dos de las tres alternativas considerados el estadístico de Henze y Wagner presenta mejor potencia, mientras que para la segunda alternativa el estadístico de sesgo de Srivastava es claramente el mejor por tener potencia apreciable para los menores tamaños muestrales. También tienen buen desempeño los otros dos estadísticos de sesgo y el estadístico de Bowman y Foster. El estadístico de curtosis de Srivastava mejora notablemente. El estadístico que peor se comporta es el estadístico de curtosis de esféricos armónicos y funciones radiales.

Globalmente, considerando ambas dimensiones y distribuciones fuertemente sesgadas y de soporte no acotado, los que presentan mejor desempeño son los estadísticos de Henze y Wagner y el estadístico de Bowman y Foster.

Tabla 4.3: Potencia Monte Carlo contra la distribución Logística en dimensión 2 a un nivel de significancia α=0.05

Alternativa	n	Estadístico							
		$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{b_{1,k}^2}}$	$\widetilde{\mathbf{b_{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z_{2,n}^2$
Logística (0.5)	20	0.326	0.221	0.286	0.178	0.228	0.312	0.211	0.217
	50	0.694	0.461	0.62	0.477	0.478	0.634	0.453	0.538
	100	0.941	0.744	0.907	0.689	0.654	0.895	0.792	0.773
	200	0.998	0.927	0.994	0.925	0.846	0.998	0.966	0.966
Logística (2)	20	0.25	0.175	0.15	0.121	0.196	0.205	0.188	0.146
	50	0.515	0.303	0.278	0.255	0.331	0.466	0.3	0.381
	100	0.861	0.546	0.307	0.361	0.594	0.757	0.557	0.633
	200	0.983	0.804	0.333	0.617	0.782	0.971	0.796	0.876

Contra las dos alternativas consideradas en la Tabla 4.3, se obtiene que para la primera alternativa el estadístico de sesgo de Mardia fue quien presento mejor potencia, seguido por el estadístico de Henze y Wagner, el estadístico de sesgo de Srivastava, el estadístico de esféricos armónicos y el estadístico de sesgo de Balakrishnan, Brito y Quiroz. Para la segunda alternativa el estadístico de sesgo de Mardia sigue siendo el que presenta mejor potencia, seguido por el estadístico Henze y Wagner, el estadístico de esféricos armónicos, el estadístico de sesgo de Balakrishnan, Brito y Quiroz, curtosis de Mardia y el estadístico de Bowman y Foster, el que falla notablemente es el estadístico de curtosis de Srivastava.

Tabla 4.4: Potencia Monte Carlo contra la distribución Logística en dimensión 5 a un nivel de significancia α=0.05

Alternativa	n	Estadístico							
		$\widetilde{\mathbf{b}_{\mathbf{1,k}}}$	$\widetilde{\mathbf{b}_{\mathbf{2,k}}}$	$\widetilde{\mathbf{b}^{\mathbf{2}}_{\mathbf{1,k}}}$	$\widetilde{\mathbf{b}_{\mathbf{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z^2_{2,n}$
Logística (0.5)	20	0.356	0.231	0.373	0.232	0.195	0.3	0.18	0.094
	50	0.893	0.633	0.781	0.55	0.453	0.768	0.488	0.575
	100	0.995	0.916	0.976	0.873	0.618	0.993	0.814	0.898
	200	1	0.992	0.999	0.988	0.799	1	0.997	0.994
Logística (2)	20	0.271	0.179	0.158	0.132	0.182	0.27	0.154	0.072
	50	0.802	0.528	0.334	0.312	0.393	0.654	0.377	0.469
	100	0.987	0.82	0.588	0.621	0.59	0.962	0.7	0.822
	200	1	0.98	0.833	0.921	0.802	1	0.97	0.983

De la Tabla 4.4 se tiene que al aumentar la dimensión, la potencia mejora para la mayoría de los estadísticos, excepto para el estadístico de sesgo de Balakrishnan, Brito y Quiroz y el estadístico de esféricos armónicos y funciones radiales (cuando $n = 20$). En las dos alternativas consideradas el estadístico de sesgo de Mardia es el que tiene mejor desempeño, seguido por el estadístico de Henze y Wagner, mientras que el estadístico de esféricos armónicos es el que tiene peor desempeño para tamaños muestrales pequeños.

Globalmente, considerando ambas dimensiones y distribuciones con diferentes formas de sesgo y curtosis, se tiene que el estadístico de sesgo de Mardia y el estadístico de Henze y Wagner son los que tienen mejor desempeño.

Tabla 4.5: Potencia Monte Carlo contra la distribución Normal Contaminada en dimensión 2 a un nivel de significancia α=0.05

Alternativa	n	Estadístico							
		$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{\bar{b}^2_{1,k}}}$	$\widetilde{\mathbf{\bar{b}_{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z^2_{2,n}$
	20	0.316	0.211	0.336	0.229	0.318	0.323	0.284	0.237
$\varepsilon\mathcal{N}_2(\mu, I)+\rho\mathcal{N}_2(\mathbf{0}, I)$	50	0.71	0.474	0.736	0.571	0.69	0.705	0.45	0.635
ε= 0.05, ρ=0.95	100	0.916	0.778	0.913	0.824	0.931	0.919	0.713	0.898
$\mu = (3, 3)^T$	200	0.998	0.977	0.993	0.984	0.991	0.993	0.915	0.995
	20	0.596	0.324	0.748	0.405	0.603	0.704	0.651	0.429
$\varepsilon\mathcal{N}_2(\mu, I)+\rho\mathcal{N}_2(\mathbf{0}, I)$	50	0.985	0.576	0.988	0.673	0.967	0.985	0.953	0.901
ε= 0.1, ρ=0.9	100	1	0.778	1	0.862	1	1	0.999	1
$\mu = (4, 4)^T$	200	1	0.961	1	0.984	1	1	1	1

De la Tabla 4.5, se tiene que a menor contaminación (primera alternativa) el estadístico que presenta mejor desempeño es el estadístico de sesgo de Srivastava, seguido por el estadístico de Henze y Wagner y los otros dos estadísticos de sesgo. Para $n = 20$ los estadísticos de sesgo, el estadístico de Henze y Wagner tienen potencia aproximadamente igual a 3, el estadístico de esféricos armónicos y funciones radiales se comporta como un estadístico de curtosis. A mayor contaminación (segunda alternativa) el estadístico de sesgo de Srivastava y el estadístico de Henze y Wagner siguen siendo los que tienen mejor desempeño, el estadístico de Bowman y Foster mejora notablemente y los otros dos estadísticos de sesgo también presentan potencia apreciable. El estadístico de esféricos armónicos y funciones radiales sigue teniendo un comportamiento similar a los estadísticos de curtosis.

Tabla 4.6: Potencia Monte Carlo contra la distribución Normal Contaminada en dimensión 5 a un nivel de significancia α=0.05

Alternativa	n	Estadístico							
		$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{\bar{b}^2_{1,k}}}$	$\widetilde{\mathbf{\bar{b}_{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z^2_{2,n}$
	20	0.327	0.193	0.58	0.358	0.318	0.333	0.185	0.092
$\varepsilon\mathcal{N}_5(\mu, I)+\rho\mathcal{N}_5(\mathbf{0}, I)$	50	0.873	0.597	0.909	0.804	0.815	0.858	0.551	0.72
ε= 0.05, ρ=0.95	100	0.993	0.891	0.989	0.981	0.987	0.981	0.812	0.985
$\mu = (3,3,3,3,3)^T$	200	1	0.988	0.998	1	1	0.999	0.976	1
	20	0.442	0.247	0.808	0.431	0.44	0.5	0.381	0.177
$\varepsilon\mathcal{N}_5(\mu, I)+\rho\mathcal{N}_5(\mathbf{0}, I)$	50	0.888	0.945	0.997	0.667	0.887	0.99	0.945	0.654
ε= 0.1, ρ=0.9	100	0.999	0.598	1	0.83	0.997	1	0.999	0.951
$\mu = (4,4,4,4,4)^T$	200	1	0.782	1	0.967	1	1	1	1

De la Tabla 4.6, se tiene que para algunos estadísticos como curtosis de Mardia, Bowman y Foster y esféricos armónicos, la potencia disminuyo para $n = 20$, con relación al caso k=2. Los estadísticos de sesgo de Srivasta y el de Henze y Wagner siguen siendo los que presentan mejor desempeño, seguidos por el estadístico de sesgo de Balakrishnan, Brito y Quiroz y el estadístico de sesgo de Mardia. Para ambas alternativas el que tiene peor desempeño es el estadístico de esféricos armónicos.

Globalmente, en ambas dimensiones consideradas y frente a distribuciones sesgadas de soporte no acotado, los estadísticos que tienen mejor desempeño son el estadístico de sesgo de Mardia y el de Henze y Wagner, notándose que los estadísticos de curtosis y el estadístico de esféricos armónicos y funciones radiales presentan potencia más baja que los estadísticos de sesgo.

Tabla 4.7: Potencia Monte Carlo contra la distribución Burr-Pareto-Logística en dimensión 2 a un nivel de significancia α=0.05

Alternativa	n	Estadístico							
		$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{b^2_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z^2_{2,n}$
BPL (0.25)	20	0.136	0.126	0.098	0.083	0.098	0.115	0.42	0.143
	50	0.289	0.187	0.135	0.096	0.127	0.47	0.938	0.374
	100	0.675	0.265	0.161	0.117	0.174	0.948	1	0.769
	200	0.986	0.367	0.173	0.131	0.296	1	1	0.995
BPL (0.5)	20	0.047	0.215	0.048	0.051	0.037	0.056	0.275	0.224
	50	0.073	0.418	0.065	0.062	0.026	0.166	0.826	0.554
	100	0.193	0.716	0.05	0.112	0.024	0.765	0.999	0.906
	200	0.585	0.916	0.081	0.16	0.028	1	1	1
BPL (1)	20	0.012	0.251	0.03	0.096	0.017	0.02	0.217	0.246
	50	0.014	0.68	0.026	0.099	0.006	0.056	0.743	0.703
	100	0.027	0.955	0.022	0.234	0.001	0.386	0.995	0.987
	200	0.076	1	0.035	0.525	0.001	1	1	1
BPL (2)	20	0.004	0.336	0.013	0.143	0.013	0.009	0.171	0.288
	50	0.002	0.821	0.006	0.257	0.003	0.019	0.679	0.772
	100	0.004	0.995	0.012	0.454	0.002	0.223	0.986	0.989
	200	0.003	1	0.014	0.842	0	0.996	1	1

De la Tabla 4.7, se tiene que contra la primera alternativa el estadístico de Bowman y Foster es el que tiene mejor potencia, seguido por el estadístico de esféricos armónicos y funciones radiales, el estadístico de sesgo de Mardia, el estadístico de curtosis de Mardia y el estadístico de Henze y Wagner. Contra la segunda alternativa, el estadístico de Bowman y Foster sigue teniendo la mejor potencia, seguido por el estadístico de esféricos armónicos y el estadístico de curtosis de Mardia. Contra las dos últimas alternativas el estadístico de curtosis de Mardia es el que tiene mayor potencia, seguido por el estadístico de esféricos armónicos y el estadístico de Bowman y Foster. Frente a alternativas sesgadas de colas livianas y soporte compacto el que tiene mejor desempeño es el estadístico de Bowman y Foster y frente a alternativas simétricas de colas livianas y soporte compacto el estadístico de curtosis de Mardia es el que tiene mejor desempeño.

Tabla 4.8: Potencia Monte Carlo contra la distribución Burr-Pareto-Logística en dimensión 5 a un nivel de significancia α=0.05

Alternativa	n	Estadístico							
		$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{b^2_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z^2_{2,n}$
BPL (0.25)	20	0.392	0.232	0.233	0.208	0.362	0.545	0.766	0.145
	50	0.845	0.445	0.365	0.466	0.565	0.955	0.996	0.522
	100	0.995	0.702	0.536	0.782	0.766	1	1	0.907
	200	1	0.92	0.695	0.965	0.961	1	1	1
BPL (0.5)	20	0.155	0.088	0.119	0.073	0.121	0.224	0.504	0.095
	50	0.321	0.103	0.176	0.143	0.147	0.629	0.98	0.201
	100	0.7	0.09	0.296	0.236	0.158	0.989	1	0.459
	200	0.99	0.097	0.532	0.408	0.224	1	1	0.976
BPL (1)	20	0.031	0.059	0.067	0.072	0.04	0.065	0.247	0.169
	50	0.04	0.301	0.069	0.046	0.025	0.153	0.859	0.477
	100	0.11	0.543	0.119	0.038	0.017	0.727	0.999	0.849
	200	0.436	0.821	0.285	0.031	0.014	1	1	0.998
BPL (2)	20	0.012	0.118	0.031	0.084	0.015	0.015	0.106	0.256
	50	0.003	0.616	0.027	0.074	0.008	0.031	0.597	0.705
	100	0	0.955	0.034	0.171	0.005	0.349	0.975	0.985
	200	0.022	0.999	0.086	0.283	0.011	0.999	1	1

De la Tabla 4.8, se tiene que la potencia mejora con la dimensión y con el tamaño de la muestra. Contra las dos primeras alternativas (sesgadas de colas livianas y soporte compacto) el estadístico que tiene mejor potencia es el de Bowman y Foster, seguido por el estadístico de Henze y Wagner, el estadístico de sesgo de Mardia y el estadístico de sesgo de Balakrishnan, Brito y Quiroz. Contra la tercera alternativa (simétrica de colas livianas de soporte compacto), el estadístico que tiene mejor comportamiento es el de Bowman y Foster, seguido por el estadístico de esféricos armónicos y curtosis de Mardia. Contra la última alternativa (muy parecida a la distribución uniforme en el cubo unitario) el que tiene mejor potencia es el estadístico de esféricos armónicos, seguido por el estadístico de curtosis de Mardia y Bowman y Foster. Contra las dos últimas alternativas los estadísticos de sesgo no tuvieron potencia, como era de esperarse, ya que, con estos parámetros, las alternativas son simétricas.

Globalmente, en ambas dimensiones consideradas, el estadístico que presenta mejor desempeño es el estadístico de Bowman y Foster.

Tabla 4.9: Potencia Monte Carlo contra la distribución Weibull en dimensión 2 a un nivel de significancia α=0.05

Alternativa	n	Estadístico							
		$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{\bar{b}^2_{1,k}}}$	$\widetilde{\mathbf{\bar{b}_{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z^2_{2,n}$
Weibull (1,1)	20	0.79	0.479	0.728	0.401	0.777	0.837	0.875	0.484
	50	0.999	0.823	0.97	0.82	0.997	0.998	0.999	0.971
	100	1	0.986	0.998	0.962	1	1	1	1
	200	1	0.999	0.996	1	1	1	1	1
Weibull (1,2)	20	0.121	0.088	0.122	0.086	0.148	0.142	0.141	0.09
	50	0.39	0.133	0.361	0.115	0.397	0.466	0.306	0.204
	100	0.771	0.145	0.647	0.139	0.753	0.815	0.638	0.435
	200	0.99	0.165	0.879	0.151	0.993	0.994	0.933	0.885
Weibull (1,2.5)	20	0.038	0.073	0.052	0.065	0.067	0.053	0.057	0.063
	50	0.1	0.077	0.13	0.096	0.152	0.146	0.113	0.086
	100	0.219	0.112	0.202	0.086	0.258	0.318	0.21	0.148
	200	0.553	0.149	0.471	0.125	0.587	0.65	0.391	0.34

En la Tabla 4.9, se presentan tres alternativas (asimétricas positivas de soporte no acotado), obteniéndose que, contra la primera alternativa todos tienen buen desempeño, siendo los más destacados el estadístico de Bowman y Foster, el estadístico de Henze y Wagner y los estadísticos de sesgo. Contra la segunda alternativa el que tiene mejor potencia es el estadístico de Henze y Wagner, seguido por el estadístico de sesgo de Balakrishnan, Brito y Quiroz, el estadístico de sesgo de Mardia y el estadístico de sesgo de Srivastava. El estadístico de esféricos armónicos y funciones radiales tiene se comporta como un estadístico de curtosis

Tabla 4.10: Potencia Monte Carlo contra la distribución Weibull en dimensión 5 a un nivel de significancia α=0.05

Alternativa	n	Estadístico							
		$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{b^2_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z^2_{2,n}$
Weibull (1,1)	20	0.824	0.569	0.698	0.452	0.804	0.89	0.866	0.405
	50	1	0.965	0.971	0.854	1	1	0.999	0.984
	100	1	1	0.999	0.994	1	1	1	1
	200	1	1	1	1	1	1	1	1
Weibull (1,2)	20	0.082	0.051	0.084	0.072	0.098	0.13	0.087	0.056
	50	0.318	0.115	0.234	0.065	0.369	0.446	0.274	0.157
	100	0.8	0.137	0.498	0.111	0.784	0.886	0.605	0.397
	200	0.998	0.178	0.818	0.123	0.996	1	0.939	0.892
Weibull (1,2.5)	20	0.038	0.04	0.043	0.066	0.055	0.049	0.042	0.072
	50	0.072	0.064	0.075	0.052	0.095	0.1	0.105	0.082
	100	0.148	0.1	0.155	0.084	0.238	0.275	0.175	0.159
	200	0.497	0.103	0.324	0.061	0.565	0.697	0.348	0.384

En la Tabla 4.10, se observa que la potencia no cambia significativamente con la dimensión. Los resultados obtenidos en dimensión 5 son muy parecidos a los obtenidos en dimensión 2, ver Tabla 4.9. Los estadísticos de curtosis y el estadístico de esféricos armónicos presentan poca potencia frente a las dos últimas alternativas.

Globalmente, en ambas dimensiones consideradas y para distribuciones asimétricas positivas de soporte no acotado, el que tiene mejor desempeño es el estadístico de Henze y Wagner.

Tabla 4.11: Potencia Monte Carlo contra la distribución Seno Hiperbólico en dimensión 2 a un nivel de significancia α=0.05

Alternativa		Estadístico							
	n	$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{b_{1,k}^2}}$	$\widetilde{\mathbf{b_{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z_{2,n}^2$
	20	0.195	0.106	0.177	0.115	0.208	0.24	0.211	0.141
senh	50	0.473	0.203	0.462	0.201	0.524	0.585	0.416	0.354
(0,2,0.25,0.25,0.5)	100	0.872	0.362	0.754	0.306	0.85	0.903	0.732	0.661
	200	0.995	0.597	0.925	0.484	0.998	0.997	0.932	0.962
	20	0.69	0.445	0.819	0.504	0.727	0.811	0.8	0.572
senh	50	0.984	0.832	0.998	0.855	0.978	1	0.998	0.945
(0,2,0.1,1,0.8)	100	1	0.975	1	0.982	1	1	1	1
	200	1	1	1	1	1	1	1	1

Contra las dos alternativas presentadas en la Tabla 4.11, se tiene que para la primera alternativa el que tiene mejor potencia es el estadístico de Henze y Wagner, seguido por el estadístico de Bowman y Foster, el estadístico de sesgo de Balakrishnan, Brito y Quiroz y los estadísticos de sesgo de Mardia y Srivastava, para n=20 los estadísticos de sesgo tienen potencia cercana a 0.2. Los estadísticos de curtosis y el estadístico de esféricos armónicos presentan potencia baja frente a la primera alternativa. Contra la segunda alternativa todos tienen buen comportamiento, siendo los más destacados el estadístico de Henze y Wagner, el de Bowman y Foster y los estadísticos de sesgo.

Tabla 4.12: Potencia Monte Carlo contra la distribución Seno Hiperbólico en dimensión 5 a un nivel de significancia α=0.05

Alternativa		Estadístico							
	n	$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{b^2_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z^2_{2,n}$
	20	0.169	0.094	0.262	0.111	0.17	0.214	0.125	0.049
senh	50	0.598	0.297	0.617	0.302	0.577	0.643	0.368	0.336
(0,2,0.25,0.25,0,0)	100	0.936	0.525	0.92	0.54	0.901	0.95	0.683	0.684
	200	1	0.771	0.994	0.791	1	1	0.951	0.977
	20	0.682	0.467	0.844	0.622	0.661	0.76	0.605	0.362
senh	50	0.998	0.913	0.998	0.948	0.994	0.996	0.992	0.963
(0,2,0.1,1,0,0)	100	1	1	1	0.997	1	1	1	1
	200	1	1	1	1	1	1	1	1

De la Tabla 4.12, se puede observar que en general la potencia no parece cambiar con la dimensión. Contra la primera alternativa el estadístico de sesgo de Srivastava es quien tiene mejor potencia, seguido por el estadístico de Henze y Wagner y los estadísticos de sesgo de Mardia y Balakrishnan, Brito y Quiroz, mientras que los estadísticos de curtosis y el estadístico de esféricos armónicos tienen potencia muy baja frente a esta alternativa. Contra la segunda alternativa todos los estadísticos tienen buen comportamiento, siendo los más destacados el estadístico de sesgo de Srivastava, el estadístico de Henze y Wagner y los estadísticos de sesgo.

Globalmente, en ambas dimensiones consideradas y para distribuciones sesgadas de soporte no acotado, el estadístico que tiene mejor desempeño frente a este conjunto de alternativas es el estadístico de Henze y Wagner.

Tabla 4.13: Potencia Monte Carlo contra la distribución uniformemente distribuida sobre el cubo unitario en dimensión 2 a un nivel de significancia α=0.05

Alternativa		Estadístico							
	n	$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{\bar{b}^2_{1,k}}}$	$\widetilde{\mathbf{\bar{b}_{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z^2_{2,n}$
	20	0	0.343	0.011	0.239	0.005	0.015	0.125	0.316
Uniforme en el	50	0	0.891	0.001	0.576	0	0.009	0.64	0.828
cubo unitario	100	0	1	0.002	0.864	0	0.111	0.981	0.995
$[0,1]^2$	200	0	1	0	0.993	0	0.986	1	1

Contra alternativas de simetría central y de soporte compacto la Tabla 4.13, se observa que el estadístico que mejor desempeño tiene es el estadístico de curtosis de Mardia, seguido por el estadístico de esféricos armónicos, el estadístico de curtosis de Srivastava y el estadístico de Bowman y Foster, ya que son los que tienen potencia apreciable para tamaños muestrales pequeños. El estadístico de Bowman y Foster y el estadístico de esféricos armónicos y funciones radiales se comportan como estadísticos de curtosis. Los estadísticos de sesgo y el estadístico de Henze y Wagner no presentan potencia frente a este tipo de alternativas.

Tabla 4.14: Potencia Monte Carlo contra la distribución uniformemente distribuida sobre el cubo unitario en dimensión 5 a un nivel de significancia α=0.05

Alternativa		Estadístico							
	n	$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{\bar{b}^2_{1,k}}}$	$\widetilde{\mathbf{\bar{b}_{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z^2_{2,n}$
	20	0.002	0.212	0.021	0.146	0.002	0.004	0.066	0.377
Uniforme en el	50	0	0.896	0.001	0.332	0.001	0.004	0.455	0.907
cubo unitario	100	0	0.998	0.005	0.694	0	0.11	0.952	1
$[0,1]^5$	200	0	1	0	0.927	0	0.972	1	1

Contra la alternativa de simetría central central y soporte compacto en dimensión 5 Tabla 4.14, se tiene que la potencia no cambia significativamente con la dimensión. El estadístico que mejor desempeño tiene es el estadístico de esféricos armónicos y funciones radiales, seguido por el estadístico de curtosis de Mardia, el estadístico de Bowman y foster y el estadístico de curtosis de Srivastava. Los estadísticos de Henze y Wagner y Bowman y Foster tienen comportamiento similar a los estadísticos de sesgo.

Globalmente, en ambas dimensiones consideradas, los que tienen mejor desempeño son los estadísticos de esféricos armónicos y funciones radialesy el estadístico de curtosis de Mardia. Tal como era de esperarse, los estadísticos de sesgo no presentan potencia frente a esta alternativa y el estadístico de Henze y Wagner no presenta potencia para tamaños muestrales pequeños, pero para $n \geq 200$ el estadístico tiene un comportamiento consistente en el sentido que la potencia aumenta rápidamente.

Tabla 4.15: Potencia Monte Carlo contra la distribución uniformemente distribuida sobre la bola unitaria en dimensión 2 a un nivel de significancia α=0.05

Alternativa		Estadístico							
	n	$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{\bar{b}^2_{1,k}}}$	$\widetilde{\mathbf{\bar{b}_{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z^2_{2,n}$
Uniforme en la bola unitario	20	0.003	0.491	0.004	0.364	0.001	0.002	0.142	0.427
	50	0	0.98	0.002	0.89	0	0.013	0.65	0.928
	100	0	1	0.001	0.999	0.001	0.192	0.985	1
	200	0	1	0	1	0	0.997	1	1

Contra alternativas esféricamente simétricas y de soporte compacto, en dimensión 2 Tabla 4.13, se tiene que el estadístico que tiene mejor desempeño es el estadístico de curtosis de Mardia, seguido por el estadístico de esféricos armónicos y funciones radiales, el estadístico de curtosis de Srivastava y el estadístico de Bowman y Foster. Los estadísticos de sesgo y el estadístico de Henze y Wagner no presenta potencia frente a este tipo de alternativas.

Tabla 4.16: Potencia Monte Carlo contra la distribución uniformemente distribuida sobre la bola unitaria en dimensión 5 a un nivel de significancia α=0.05

Alternativa		Estadístico							
	n	$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{\bar{b}^2_{1,k}}}$	$\widetilde{\mathbf{\bar{b}_{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z^2_{2,n}$
Uniforme en la bola unitaria	20	0.001	0.456	0.009	0.249	0.002	0	0.044	0.625
	50	0	1	0.002	0.685	0	0.003	0.582	0.999
	100	0	1	0	0.986	0	0.303	0.986	1
	200	0	1	0	1	0	1	1	1

Contra alternativas esféricamente simétricas y de soporte compacto, en dimensión 5

Tabla 4.16, se observa que la potencia no cambia significativamente con la dimensión. El estadístico que tiene mejor potencia es el estadístico de esféricos armónicos y funciones radiales, seguido por el estadístico de curtosis de Mardia y curtosis de Srivastava.

Globalmente, en ambas dimensiones consideradas, los que tienen mejor desempeño son el estadístico de esféricos armónicos y funciones radiales y el estadístico de curtosis de Mardia. Los estadísticos de sesgo no presentan potencia frente a esta alternativa, el estadístico de Henze y Wagner no presenta potencia para tamaños muestrales pequeños, pero es consistente para $n \geq 200$.

Tabla 4.17: Potencia Monte Carlo contra la distribución Chi cuadrado en dimensión 2 a un nivel de significancia α=0.05

Alternativa		Estadístico							
	n	$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{b^2_{1,k}}}$	$\widetilde{\mathbf{b}_{\mathbf{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z^2_{2,n}$
χ^2_3	20	0.648	0.34	0.591	0.312	0.623	0.712	0.68	0.359
	50	0.985	0.703	0.919	0.658	0.985	0.995	0.988	0.867
	100	1	0.908	0.993	0.858	1	1	1	0.999
	200	1	0.996	1	0.986	1	1	1	1
χ^2_6	20	0.361	0.207	0.382	0.193	0.391	0.43	0.374	0.212
	50	0.849	0.382	0.776	0.406	0.872	0.914	0.796	0.568
	100	0.999	0.683	0.952	0.587	0.997	0.999	0.987	0.945
	200	1	0.893	0.994	0.827	1	1	1	1

Contra alternativas fuertemente sesgadas y de soporte no acotado, Tabla 4.17, se tiene que todos los estadísticos tienen un buen comportamiento, obteniéndose que contra la primera alternativa el que presenta mejor potencia es el estadístico de Henze y Wagner, seguido por el estadístico de Bowman y Foster, el estadístico de sesgo de Mardia, el estadístico de sesgo de Balakrishnan, Brito y Quiroz y el de sesgo de Srivastava, Mientras que contra la segunda alternativa el estadístico de Bowman y Foster sigue siendo el que tiene mejor potencia, seguido por el estadístico de sesgo de Balakrishnan, Brito y Quiroz, el estadístico de sesgo de Srivastava, el estadístico de Bowman y Foster y el estadístico de sesgo de Mardia.

Tabla 4.18: Potencia Monte Carlo contra la distribución Chi cuadrado en dimensión 5 a un nivel de significancia α=0.05

Alternativa	n	Estadístico							
		$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{b^2_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z^2_{2,n}$
χ^2_3	20	0.632	0.39	0.508	0.313	0.613	0.739	0.665	0.227
	50	0.997	0.841	0.916	0.691	0.989	0.999	0.994	0.914
	100	1	0.991	0.993	0.93	1	1	1	1
	200	1	1	1	0.998	1	1	1	1
χ^2_6	20	0.318	0.183	0.301	0.17	0.378	0.465	0.334	0.086
	50	0.92	0.505	0.695	0.375	0.888	0.965	0.828	0.587
	100	1	0.828	0.946	0.653	0.999	1	0.994	0.978
	200	1	0.981	0.997	0.909	1	1	1	1

Contra la alternativa fuertemente sesgada en dimensión 5, Tabla 4.18, se tiene que la potencia no parece mejorar con la dimensión. Contra la primera alternativa todos tienen buen desempeño, siendo el estadístico de Henze y Wagner el que tiene mejor potencia, seguido por el estadístico de Bowman y Foster y los estadísticos de sesgo. Contra la segunda alternativa los estadísticos que tienen mejor desempeño son el de Henze y Wagner, el de sesgo de Balakrishnan, Brito y Quiroz, el de Bowman y Foster y los estadísticos de sesgo de Mardia y Srivastava. Mientras que los estadísticos que tienen potencia significativamente más baja son los de curtosis y el estadístico de esféricos armónicos y funciones radiales, este último es el que tiene peor desempeño.

Globalmente, en ambas dimensiones consideradas, los que tienen mejor desempeño son los estadísticos de Henze y Wagner, el de sesgo de Balakrishnan, Brito y Quiroz y el de sesgo de Mardia.

Tabla 4.19: Potencia Monte Carlo contra la distribución Normal sesgada en dimensión 2 a un nivel de significancia α=0.05

Alternativa	n	Estadístico							
		$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{\bar{b}^2_{1,k}}}$	$\widetilde{\mathbf{\bar{b}_{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z^2_{2,n}$
	20	0.113	0.08	0.164	0.093	0.138	0.153	0.146	0.087
λ_0= 1	50	0.308	0.104	0.462	0.134	0.334	0.407	0.296	0.16
λ_1= (1,5)	100	0.687	0.123	0.804	0.154	0.669	0.757	0.593	0.305
	200	0.968	0.169	0.991	0.207	0.956	0.989	0.886	0.707
	20	0.095	0.071	0.134	0.094	0.118	0.106	0.106	0.07
λ_0= -1	50	0.218	0.094	0.329	0.114	0.264	0.255	0.208	0.133
λ_1= (2,2)	100	0.438	0.14	0.601	0.153	0.494	0.489	0.322	0.244
	200	0.815	0.174	0.913	0.226	0.813	0.853	0.563	0.504

Contra la alternativa normal sesgada y en dimensión 2, Tabla 4.19, el estadístico que tiene mejor desempeño es el estadístico de sesgo de Srivastava, seguido por el estadístico de sesgo de Balakrishnan, Brito y Quiroz, el estadístico de Henze y Wagner y el estadístico de sesgo de Mardia. Los estadísticos de curtosis y el estadístico de esféricos armónicos y funciones radiales son los que tienen peor desempeño.

Tabla 4.20: Potencia Monte Carlo contra la distribución Normal sesgada en dimensión 5 a un nivel de significancia α=0.05

Alternativa	n	Estadístico							
		$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{\bar{b}^2_{1,k}}}$	$\widetilde{\mathbf{\bar{b}_{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z^2_{2,n}$
	20	0.071	0.046	0.141	0.095	0.063	0.08	0.066	0.062
λ_0= 1	50	0.147	0.067	0.356	0.077	0.156	0.186	0.139	0.085
λ_1=(1,2,3,4,5)	100	0.314	0.085	0.754	0.125	0.34	0.402	0.232	0.149
	200	0.697	0.095	0.985	0.158	0.69	0.833	0.511	0.329
	20	0.087	0.054	0.168	0.086	0.085	0.098	0.089	0.059
λ_0= -1	50	0.159	0.086	0.437	0.114	0.203	0.232	0.172	0.094
λ_1=(4,4,4,4,4)	100	0.414	0.104	0.822	0.173	0.451	0.51	0.307	0.165
	200	0.858	0.124	0.993	0.227	0.772	0.918	0.643	0.424

Contra la alternativa normal sesgada y en dimensión 5, Tabla 4.20, la potencia no cam-

bia significativamente con la dimensión. El estadístico que tiene mejor comportamiento es el de sesgo de Srivastava, seguido por el estadístico de Henze y Wagner y los estadísticos de sesgo de Balakrishnan, Brito y Quiroz y de Mardia. Los estadísticos de curtosis, el estadístico de esféricos armónicos y el estadístico de Bowman y Foster no se comportan bien frente a este tipo de alternativas, ya que presentan potencias muy pequeñas.

Globalmente, en ambas dimensiones consideradas, los estadísticos que tienen mejor desempeño son el estadístico de sesgo de Srivastava y el estadístico de Henze y Wagner.

Tabla 4.21: Potencia Monte Carlo contra la distribución Logística clásica en dimensión 2 a un nivel de significancia α=0.05

Alternativa		Estadístico							
	n	$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{\bar{b}^2_{1,k}}}$	$\widetilde{\mathbf{\bar{b}_{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z^2_{2,n}$
Logística clásica	20	0.156	0.113	0.141	0.124	0.137	0.128	0.12	0.151
	50	0.253	0.253	0.198	0.239	0.233	0.267	0.162	0.275
	100	0.294	0.469	0.238	0.401	0.292	0.38	0.28	0.511
	200	0.359	0.752	0.275	0.66	0.335	0.542	0.446	0.76

Contra la alternativa presentada en la Tabla 4.21, en dimensión 2, se tiene que el estadístico que tiene mejor desempeño es el estadístico de esféricos armónicos, seguido por el estadístico de curtosis de Mardia, curtosis de Srivastava y el estadístico de Henze y Wagner.

Tabla 4.22: Potencia Monte Carlo contra la alternativa Logística clásica en dimensión 5 a un nivel de significancia α=0.05

Alternativa		Estadístico							
	n	$\widetilde{\mathbf{b_{1,k}}}$	$\widetilde{\mathbf{b_{2,k}}}$	$\widetilde{\mathbf{\bar{b}^2_{1,k}}}$	$\widetilde{\mathbf{\bar{b}_{2,k}}}$	$Q_{n,2}$	$\mathbf{T}_{n,\beta}$	$\widetilde{\mathbf{J}}^2$	$Z^2_{2,n}$
Logística clásica	20	0.142	0.126	0.118	0.112	0.132	0.167	0.091	0.056
	50	0.318	0.362	0.19	0.22	0.27	0.323	0.147	0.313
	100	0.461	0.649	0.258	0.413	0.324	0.476	0.252	0.606
	200	0.546	0.909	0.308	0.706	0.401	0.701	0.477	0.903

Contra la alternativa presentada en la Tabla 4.22, en dimensión 5, se tiene que la potencia mejora con la dimensión y con n. A pesar de que esta alternativa es una distribución

difícil de detectar por las pruebas de normalidad, el estadístico de Henze y Wagner, los estadísticos de sesgo y los estadísticos de curtosis la detectan. Los estadísticos que tiene peor desempeño son los estadísticos de Bowman y Foster y el estadístico de esféricos armónicos y funciones radiales.

Globalmente, en ambas dimensiones consideradas, los estadísticos que tienen mejor desempeño son el de curtosis de Mardia y el de esféricos armónicos. Los estadísticos de sesgo y el estadístico de Bowman y Foster no se comportan bien frente a este tipo de alternativa, ya que presentan potencias muy pequeñas.

Capítulo 5

CONCLUSIONES Y RECOMENDACIONES

Una vez realizado el estudio comparativo entre los diferentes estadísticos de sesgo, curtosis, estimación de densidad, función característica empírica y esféricos armónicos, como métodos de bondad de ajuste basados en la hipótesis nula de normalidad multivariada se llegó a las siguientes conclusiones y recomendaciones:

5.1. Conclusiones

Los ocho estadísticos propuestos fueron implementados con el software estadístico R, obteniéndose que, de acuerdo a la complejidad de los algoritmos, a la complejidad computacional y al tiempo de ejecución los estadísticos se pueden clasificar en rápidos, lentos e intermedios. Los estadísticos de curtosis de Mardia y Srivastava y el estadístico de sesgo de Srivastava son los que resultan más rápidos en la comparación, los más lentos resultaron ser el estadístico de estimación de densidad de Bowman y Foster y el estadístico basado en la función característica empírica de Henze y Wagner, mientras que los que resultaron intermedios en la comparación fueron los estadísticos de sesgo de Mardia, sesgo de Brito, Balakrishnan y Quiroz y el estadístico de esféricos armónicos y funciones radiales de Manzotti y Quiroz .

Bajo la hipótesis nula de normalidad multivariada, se usó el método Monte Carlo para obtener cuantiles aproximados para muestras de tamaño finito y en diferentes dimensio-

nes y al compararlos con los cuantiles límites, en los casos donde fue posible, se obtuvo lo siguiente:

El estadístico de sesgo de Mardia presenta convergencia lenta a los cuantiles límites, la convergencia es monótona creciente.

El estadístico de curtosis de Mardia tiene convergencia sumamente lenta al aumentar la dimensión.

El estadístico de sesgo de Srivastava, presenta convergencia relativamente más rápido que el estadístico de sesgo de Mardia, mientras que el estadístico de curtosis converge lentamente.

El estadístico de sesgo de Brito, Balakrishnan y Quiroz presenta convergencia moderadamente rápida para dimensiones pequeñas, mientras que para dimensiones altas la convergencia es moderadamente lenta.

El estadístico basado en la de función característica empírica de Henze y Wagner, tiene una distribución muy concentrada. Presenta muy poca varianza alrededor de su centro. Esto lleva a la necesidad de tener cuidado con errores numéricos en el cálculo del estadístico. Este fenómeno se presenta en dimensiones altas. Este estadístico converge rápidamente para todas las dimensiones, para los diferentes valores del parámetro β y especialmente para $\beta = 3$, en algunos casos la convergencia es monótona creciente con n y a partir de $n = 50$ los cuantiles comienzan a oscilar alrededor de sus valores límites.

El estadístico de estimación de densidad (estandarizado) de Bowman y Foster converge más lentamente al aumentar la dimensión.

El estadístico de esféricos armónicos y funciones radiales de Manzotti y Quiroz convergen moderadamente rápido a los cuantiles límites y la convergencia es monótona creciente.

En la mayoría de los casos los cuantiles obtenidos están por debajo de los cuantiles límites, pero a pesar de esto, se obtiene una buena aproximación a los cuantiles límites para muestras finitas.

Al evaluar el desempeño de cada estadístico en términos de la potencia, para la hipótesis nula de normalidad multivariada, con datos en distintas dimensiones, distintos tamaños muestrales, distantas alternativas y a un nivel de significancia α=0.05 se obtuvo lo siguiente:

Los estadísticos de sesgo resultaron ser muy buenos competidores contra alternativas sesgadas, como era de esperarse.

El estadístico basado en la función característica empírica de Henze y Wagner tiende a comportarse como un estadístico de sesgo para tamaños muestrales pequeños. Al aumentar el tamaño de la muestra, puede también detectar alternativas insesgadas.

Los estadísticos de curtosis son buenos competidores contra alternativas simétricas.

El estadístico de esféricos armónicos y funciones radiales de Manzotti y Quiroz tiende a ser más efectivo, para muestras pequeñas contra alternativas simétricas que se alejan de la normalidad debido a la curtosis, aunque para tamaño de muestra grande, puede detectar alternativas sesgadas, este estadístico se comporta como un estadístico de curtosis porque mide la curtosis del radio.

El estadístico de estimación de densidad de Bowman y Foster tiene un comportamiento similar al estadístico de Henze y Wagner. Es buen competidor para algunas alternativas simétricas y en otros casos para algunas distribuciones sesgadas.

En general ningún método resulto ser uniformente superior a otro, ya que los resultados dependen de las características que tengan las distribuciones alternativas no normales consideradas.

5.2. Recomendaciones

De acuerdo a lo expuesto anteriormente se dan a continuación algunas recomendaciones para el analista de datos, en cuanto a cual o cuales estadísticos resultaron preferible tomando en cuenta la dimensión, el tamaño muestral y el tipo de distribución alternativa:

- Para distribuciones simétricas de soporte no acotado de colas livianas el estadístico de curtosis de Mardia y el estadístico de esféricos armónicos de Manzotti y Quiroz.
- Para distribuciones simétricas de soporte compacto los estadísticos estimación de densidad de Bowman y Foster el estadístico de esféricos armónicos de Manzotti y los estadísticos de curtosis.
- Para distribuciones asimétricas de soporte acotado, el estadístico basado en la función característica empírica de Henze y Wagner (con $\beta \geq 0.5$) y el estadístico de estimación de densidad de Bowman y Foster.
- Para distribuciones sesgadas, el estadístico de estimación de densidad de Henze y Wagner (con $\beta \geq 0.5$), Bowman y Foster y los estadísticos de sesgo.

Para investigaciones futuras, una estrategia que parece ser suficiente para detectar con alta probabilidad cualquiera de las alternativas consideradas en este estudio, sería combinar (a un nivel de $\alpha/2$ cada una) los estadísticos de Henze y Wagner y el estadístico de armónicos esférico y funciones radiales.

REFERENCIAS

Balakrishnan, Brito M. R., N. y Quiroz, A. (2004). «A vectorial notion of skewness and its use testing for multivariate symmetry». *Preprint*, pp. 1–13.

Baringhaus, L. y Henze, N. (1988). «A consistent test for multivariate normality based on the empirical characteristic function». *Metrika*, **V.35**, pp. 339–348.

Bowman, A. W. y Foster, P. J. (1993). «Adaptative Smoothing and density-Based tests of Multivariate Normality». *Journal of the American Statistical Association*, **V.88(422)**, pp. 529–537.

Farrel, P.; Warrera, M. y Naczk, K. (2006). «On tests for multivariate normality and associated simulation studies». *Journal of Stastistical Computation and Simulation*, **V.00(00)**, pp. 1–14.

Henze, N (1994). «On Mardia's kurtosis test for multivariate normality». *Communications in Statist. Theory and Methods*, **V.23**, pp. 1031–1045.

Henze, N. y Wagner, T. (1997). «A New Approach to the BHEP tests for Multivariate Normality.» *Journal of Multivariate Analysis*, **V.62**, pp. 1–23.

Henze, N. y Zirkler, B. (1990). «A class of invariant and consistent test for multivariate normality». *Communications in Statist. Theory and Methods*, **V.19**, pp. 3595–3617.

Johnson, M. (1987). *Multivariate statistical simulation.* Wiley, New York..

Johnson, Richard A. y Wichern., Dean W. (200). *Applied multivariate statistical analysis.* Prentice Hall.

Kotz, S.; Walakrishnan, N. y Jhonson, N. L. (2000). «Continuous Multivariate Distributions.» *Methods and Applications*, **V.1**.

Malkovich, J. F. y Afifi, A. A. (1973). «On test for multivariate Normality». *Journal of the American Statistical Association*, **V.68**, pp. 176–179.

MANZOTTI, A. y QUIROZ, A. (2001). «Spherical harmonics in quadratic forms for testing multivariate normality». *Journal of the American Statistical Association*, **V.10(1)**, pp. 87–104.

MARDIA, K.V (1970). «Measures of Multivariate skewness and kurtosis with applications». *Biometrika*, **V.57**, pp. 519–530.

MEKLIN, C. y MUNDFROM, D. (2005)). «A Monte Carlo comparison of the type I and type II error rates of tests of multivariate normality.» *Journal of Stastistical Computation and Simulation*, **V.75(2)**, pp. 93–107.

MOORE, D. y STUBBLBINE, J. (1981). «Chi-Square tests for multivariate normality whith application to commom stock prices». *Communications in Statistics. Theory and Methods*, **V.10(8)**, pp. 713–738.

RENCHER, ALVIN C. (2002). *Methods of multivariate analysis.* John Wiley, sons, INC, New York..

ROMEU, J. L. y OZTURK, A. (1993). «A comparative study of goodness of fit tests for multivariate normality.» *Journal of Multivariate Analysis*, **V.46**, pp. 309–334.

WARINGHAUS, L. y HENZE, N. (1992). «Limit distribution for Mardia's measure of multivariate skewness». *Annals of Statistics*, **V.20**, pp. 1889–1902.