Anales Científicos, 77 (2): 141-146 (2016)
ISSN 0255-0407 (Versión impresa)
ISSN 2519-7398 (Versión electrónica)

DOI: http://dx.doi.org/10.21704/ac.v77i2.483

Presentado:16/09//2015
Aceptado:24/09/2015

COMPARACIÓN DE PRUEBAS DE NORMALIDAD MULTIVARIADA

COMPARISON TESTS MULTIVARIATE NORMAL

[1]Jaime Carlos Porras Cerron

Resumen

La distribución Normal Multivariada es utilizada como supuesto de muchos análisis estadísticos paramétricos tales como: MANOVA, Análisis Discriminante Lineal, Análisis de Componentes Principales, Correlación Canónica, entre otras. Es importante verificar el cumplimiento de este supuesto para que los resultados obtenidos con los métodos mencionados tengan validez. En la literatura estadística, existen diferentes pruebas que permiten verificar la normalidad multivariada. Sin embargo, aun no se ha estudiado lo suficiente, los criterios para determinar cuál es la prueba más adecuada que se debe utilizar bajo ciertas condiciones como: tamaño de muestra, número de variables, variabilidad conjunta. En la investigación se utilizó simulación de Monte Carlo para la comparación de cuatro pruebas de normalidad multivariada que son: Mardia, Henze-Zinkler, Shapiro-Wilk Generalizada y Royston. Se determinó que no existen diferencias significativas en la potencia de las pruebas consideradas en la presente investigación. Por otro lado, la implementación de las comparaciones se hizo con ayuda del programa estadístico R.

Palabras clave: Normalidad Multivariada, Simulación de Monte Carlo, Variabilidad total, Prueba de Mardia, Prueba de Henze-Zirkler, Prueba de Royston, Prueba de Shapiro Wilk Generalizada.

Abstract

The Multivariate Normal distribution is used as a course of many parametric statistical analyzes such as: MANOVA, Linear Discriminant Analysis, Principal Components Analysis, Canonical Correlation, among others. It is important to verify compliance with this course for the results obtained with the above methods are valid. In the statistical literature, there are different tests to verify multivariate normality. However, it has not yet been studied enough, the criteria to determine the most appropriate test to be used under certain conditions such as sample size, number of variables, joint variability. Mardia, Henze-Zinkler, Shapiro-Wilk Pervasive and Royston: Monte Carlo simulation for comparison of four tests of multivariate normality that are used in research. It was determined that there are no significant differences in the strength of the evidence considered in this investigation. Furthermore, the implementation of comparisons made using the statistical program R.

Keywords: Multivariate Normality, Monte Carlo Simulation, Total Variability, Mardia test, test-Zirkler Henze, Royston test, Shapiro Wilk test Generalized.

1. Introducción

Muchos métodos de análisis estadísticos como: el Análisis Multivariado de la Varianza (MANOVA), el Análisis Discriminante Lineal (ADL), el Análisis de Componentes Principales (ACP), Correlación Canónica (CC), entre otros, requieren el cumplimiento del supuesto de normalidad multivariada. Si los datos provienen de una distribución normal multivariada (exacta o aproximadamente), los métodos antes mencionados podrían brindar resultados confiables. Caso contrario, el rendimiento de los métodos podría disminuir dramáticamente, es decir sus resultados no serian confiables.

Para verificar si un conjunto de datos proviene de una distribución normal multivariada se puede hacer uso de gráficos (procedimientos descriptivos) o de pruebas estadísticas (procedimientos inferenciales).Si bien es cierto que los métodos gráficos son más fáciles de interpretar, las pruebas estadísticas nos permiten una mejor generalización de los resultados.

La presente investigación tiene como principal objetivo comparar cuatro pruebas estadísticas que permiten evaluar si un conjunto de datos se ajusta a una distribución normal multivariada. Las pruebas a utilizar son: Mardia,

[1] Departamento de Estadística e Informática, Facultad de Economia y Planificacion. UNALM. E-mail: jaimepc@lamolina.edu.pe

Henze-Zinkler, Shapiro-Wilk Generalizada y Royston.

La potencia de prueba es el concepto que nos puede ayudar a elegir cuál es la mejor prueba en diferentes escenarios propuestos determinados según el tamaño de la muestra, el número de variables de la matriz de datos y la variabilidad total de los datos.

La obtención de la potencia de prueba que permite comparar estas pruebas se realizará a través de simulación de Monte Carlo mediante la elaboración de procedimientos obtenidos con ayuda del programa estadístico R.

2. Revisión de Literatura

Existen diversos procedimientos (gráficos y pruebas estadísticas) para verificar la normalidad multivariada.

Burdenski (2000) evaluó algunos procedimientos univariados como: gráfico Q-Q, diagrama de cajas, diagrama de tallos y hojas. También utilizó otros gráficos bivariados como: contorno, perspectiva y Chi cuadrado Q-Q. Asimismo, hizo uso de pruebas univariadas de Shapiro- Wilk y Kolmogorov-Smirnov.

Svantesson y Wallace (2003) aplicaron las pruebas de Royston y Henze-Zirkler a conjuntos de datos simulados con diferentes características.

De acuerdo a lo revisado por Mecklin y Mundfrom (2005), más de cincuenta métodos estadísticos están disponibles para verificar si un conjunto de datos proviene de una distribución normal multivariada.

Holgersson (2006) resaltó la importancia de los procedimientos gráficos, y presentó una herramienta gráfica simple basada en el diagrama de dispersión de dos variables correlacionadas que permite verificar si los datos provienen o no de una distribución normal multivariada.

Ramzan et al. (2013) aplicaron a datos reales los gráficos Chi- Cuadrado y beta Q-Q para verificar la normalidad univariada y multivariada.

Como se puede apreciar, existen muy pocos trabajos de investigación que permiten comparar diferentes pruebas de normalidad multivariada mediante su potencia de prueba.

3. Materiales y métodos

Materiales

Para realizar la aplicación del presente trabajo de investigación, se utilizaron diferentes funciones del programa estadístico R versión 3.2.2.

La función mvrnorm del paquete MASS permite la generación pseudoaleatoria de datos provenientes de una distribución normal multivariada.

La idea es generar diferentes escenarios. Estos escenarios implican que los conjuntos de datos presenten distintas características que incluyen: el tamaño de la muestra (n), el número de variables (p) y la variabilidad total de los datos (varianza generalizada). Para los datos simulados, se considerará un vector de medias igual a cero, dado que lo que se desea es priorizar la evaluación de la variabilidad de los datos.

Por ejemplo, si se fijan las siguientes características de un conjunto de datos: 50 observaciones, 8 variables y con una varianza generalizada de 5. Este tipo de conjunto de datos por simulación de Monte Carlo se repite r veces y en cada una de las repeticiones se evalúan las diferentes pruebas estadísticas de interés.

Los paquetes MVN y mvShapiroTest del R fueron utilizados para obtener los resultados de la evaluación a los datos generados de las diferentes pruebas de normalidad multivariada considerados en el presente estudio.

Finalmente, se elaboraron funciones (ver anexo) que permitan evaluar la potencia de las diferentes pruebas de normalidad multivariada consideradas.

En la tabla 1, se presenta una breve descripción de los conjuntos de datos que serán simulados. Para cada escenario propuesto, se indica: el tamaño de la muestra (n), el número de variables (p) y la variabilidad total de los datos (VT).

Tabla 1. Descripción de la estructura de los datos simulados.

Escenario	n	p	VT
1	30	3	1
2			144
3		7	1
4			144
5	100	3	1
6			144
7		7	1
8			144
9	500	3	1
10			144
11		7	1
12			144
13	1000	3	1
14			144
15		7	1
16			144

Fuente: Elaboración propia

En cada escenario, se estimará la potencia de prueba de las diferentes pruebas de normalidad multivariada.

Métodos

A continuación, se describen los aspectos teóricos de las diferentes pruebas de normalidad multivariada utilizadas en la presente investigación.

• Prueba de Mardia

Mardia (1970) propuso una prueba de normalidad multivariada la cual está basada en la extensión de la asimetría $\left(\hat{\gamma}_{1.p}\right)$ y curtosis $\left(\gamma_{2.p}\right)$

$$\hat{\gamma}_{1.p}=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}m_{ij}^3 \qquad \hat{\gamma}_{2.p}=\frac{1}{n}\sum_{i=1}^{n}m_{ii}^2 \tag{1}$$

Donde $m_{ij}=\left(x_i-\bar{x}\right)'S^{-1}\left(x_i-\bar{x}\right)$ es la distancia al cuadrado de Mahalanobis y p es el número de variables. La prueba estadística para la asimetría $(n/6)\hat{\gamma}_{1.p}$ se distribuye aproximadamente como una Chi Cuadrado con $p(p+1)(p+2)/6$ grados de libertad. Similarmente la prueba estadística para la curtosis $\gamma_{2.p}$ se distribuye aproximadamente normal con media $p(p+2)$ y varianza $8p(p+2)/n$.

• Prueba de Henze-Zirkler

La prueba de Henze-Zirkler está basada en la distancia funcional no negativa, la cual mide la distancia entre dos funciones de distribución. Si los datos presentan una distribución normal multivariada, la prueba estadística se distribuye aproximadamente como una lognormal. Primero, la media, varianza y el parámetro de suavización son calculados. Entonces, la media y la varianza son lognormalizados y el pvalor es estimado. La prueba estadística de normalidad multivariada de Henze-Zirkler es:

$$HZ=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}e^{-\frac{\beta^2}{2}D_{ij}}-2\left(1+\beta^2\right)^{-\frac{p}{2}}\sum_{i=1}^{n}e^{-\frac{\beta^2}{2\left(1+\beta^2\right)}D_i}+n\left(1+2\beta^2\right)^{-\frac{p}{2}} \tag{2}$$

Donde:

p : Número de variables.

$$\beta=\frac{1}{\sqrt{2}}\left(\frac{n(2p+1)}{4}\right)^{\frac{1}{p+4}}$$

$$D_{ij}=\left(x_i-x_j\right)'S^{-1}\left(x_i-x_j\right)$$

$$D_i=\left(x_i-\bar{x}\right)'S^{-1}\left(x_i-\bar{x}\right)=m_{ii}$$

En la expresión 2, D_i es la distancia al cuadrado de Mahalanobis de la i-ésima observación al centroide y D_{ij} es la distancia de Mahalanobis entre la i-ésima y j-ésima observación. Si los datos son normales multivariados el estadístico HZ es aproximadamente lognormal con media μ y varianza σ^2 dado por:

$$\mu=1-\frac{a^{-\frac{p}{2}}\left(1+p\beta^{\frac{2}{a}}+\left(p(p+2)\beta^4\right)\right)}{2a^2}$$

$$\sigma^2=2\left(1+4\beta^2\right)^{-\frac{p}{2}}+\frac{2a^{-p}\left(1+2p\beta^4\right)}{a^2}+\frac{3p(p+2)\beta^8}{4a^4}-4\varpi_\beta^{-\frac{p}{2}}\left(1+\frac{3p\beta^4}{2\varpi_\beta}+\frac{p(p+2)\beta^8}{2\varpi_\beta^2}\right)$$

Donde:

$a=1+2\beta^2$ y $\varpi_\beta=\left(1+\beta^2\right)\left(1+3\beta^2\right)$. La media y varianza log normalizada del estadístico HZ puede ser definido de la siguiente manera:

$$\log(\mu)=\log\left(\sqrt{\frac{\mu^4}{\sigma^2+\mu^2}}\right) \text{ y } \log\left(\sigma^2\right)=\log\left(\sqrt{\frac{\sigma^2+\mu^2}{\sigma^2}}\right) \tag{3}$$

Usando la distribución lognormal con parámetros μ y σ, se puede probar la significancia de la normalidad multivariada. La prueba de Wald para la normalidad multivariada es dada por:

$$z=\frac{\log(HZ)-\log(\mu)}{\log(\sigma)}$$

• Prueba de Royston

La prueba de Royston usa la estadística Shapiro-Wilk / Shapiro-Francia para probar la normalidad multivariada. Si la curtosis es mayor a 3, entonces se usa la prueba de Shapiro-Francia para distribuciones leptocurticas. Mientras que se usa la prueba de Shapiro-Wilk para distribuciones platicurticas.

Si W_j es la prueba estadística de Shapiro-Wilk/Shapiro – Francia para la j-ésima variable (j=1,2,..,p) y Z_j son los valores obtenidos de la transformación para normalidad, entonces:

Si $4\leq n\leq 11$; $x=n$ y $\varpi_j=-\log[\gamma-\log(1-W_j)]$

Si $12\leq n\leq 2000$; $x=\log(n)$ y $\varpi_j=\log(1-W_j)$ (5)

Como se ha visto, x y ϖ_j cambian debido al tamaño de la muestra n. Usando ecuación 5 transforma los valores de cada variable aleatorio, obteniéndose la siguiente ecuación:

$$z=\frac{w_j-\mu}{\sigma}$$

Dondeγ, μ y σ son derivados de la siguiente aproximación polinomial:

$$\gamma=a_{0\gamma}+a_{1\gamma}x+a_{2\gamma}x^2+\cdots+a_{d\gamma}x^d$$

$$\mu=a_{0\mu}+a_{1\mu}x+a_{2\mu}x^2+\cdots+a_{d\mu}x^d$$

$$\log(\sigma)=a_{0\sigma}+a_{1\sigma}x+a_{2\sigma}x^2+\cdots+a_{d\sigma}x^d$$

La prueba estadística de Royston para normalidad univariada es dada por:

$$H=\frac{e^{\sum_{j=1}^{p}\psi_j}}{p}\sim\chi_e^2$$

Donde e es el equivalente grados de libertad y $\Phi(\cdot)$ es la función de distribución acumulada de la distribución normal estándar tal que:

$$e = p \Big/ \left[1 + (p-1)\bar{c}\right]$$

$$\psi_j = \left\{\Phi^{-1}\left[\Phi\left(-Z_j\right)/2\right]\right\}^2 \quad \text{j=1,2,…,p}$$

Como se puede apreciar en la última expresión, el término $\bar{c}$ debe ser calculado para poder calcular la significancia de la prueba estadística de Royston. El término $\bar{c}$ debe ser calculado de la siguiente manera:

$$\bar{c} = \sum_i \sum_j \frac{c_{ij}}{p(p-1)} \qquad \{c_{ij}\}_{i \neq j}$$

Donde

$$c_{ij} = \begin{cases} g(r_{ij}, n) & si\ i \neq j \\ 1 & si\ i = j \end{cases}$$

Con los límites de $g(\cdot)$ como $g(0,n) = 0$ y $g(1,n) = 1$. La función $g(\cdot)$ es definida de la siguiente manera:

$$g(r,n) = r^{\lambda}\left[1 - \frac{\mu}{\nu}(1-r)^{\mu}\right]$$

Los parámetros desconocidos, μ, λ y ν son estimados mediante simulación y se concluyó que se debe utilizar μ=0.715 y λ=5 para una muestra de tamaño 10≤n≤2000 y ν es una función cúbica la cual puede ser obtenida mediante:

$$\nu(n) = 0.21364 + 0.015124x^2 - 0.0018034x^3$$

Donde x=log(n)

- **Prueba de Shapiro Wilk Generalizada**

Si $X_1, \ldots X_n$ son vectores aleatorios idénticamente distribuidos en R^p, p≥1. Si $N^p(\mu, \Sigma)$ denota la densidad de una normal p-variada con vector de media μ y matriz de covarianza Σ.

Si la hipótesis nula H_0: $X_1, \ldots X_n$ es una muestra de $N^p(\mu, \Sigma)$ donde μ y Σ son desconocidos, se propone la siguiente prueba estadística:

$$W^* = \frac{1}{p}\sum_{i=1}^{p} W_{Z_i}$$

Donde W_{Z_i} es el estadístico Shapiro-Wilk evaluado en la i-ésima coordenada de la observación transformada $Z_{i1},\ldots,Z_{in}$ i=1,…,p.

La prueba basada en W^* rechaza H_0 en una prueba de tamaño α si $W^* < c_{\alpha;n,p}$ donde $c_{\alpha;n,p}$ satisface la ecuación:

$$\alpha = P\left\{W^* < c_{\alpha;n,p} \,/\, H_0 \text{ es verdadero}\right\}$$

4. Resultados

Para la determinación de la mejor prueba, se consideró la potencia de prueba, y para la obtención de este valor se implementó una función en R (ver anexo).

Asimismo, para no favorecer a ninguna de las pruebas estadísticas se realizaron 100 repeticiones para cada uno de los escenarios propuestos.

Luego de realizar las corridas respectivas, se obtuvieron los resultados que se resumen en las tabla 2 (escenarios: 1, 5, 9, 13), 3 (escenarios: 2, 6, 10, 14), 4 (escenarios:3, 7, 11,15) y tabla 5 (escenarios: 4, 8, 12, 16); las cuales se presentan con un gráfico de líneas comparativo .

Tabla 2. Potencia de Prueba para p=3 y VT=1

Prueba de Normalidad	Tamaño de Muestra			
	30	100	500	1000
Mardia	0.96	0.98	0.96	0.94
Shapiro -Wilk	0.93	0.97	0.94	0.95
HZ	0.94	0.96	0.97	0.96
Royston	0.92	0.95	0.93	0.94

Fuente: Elaboración propia

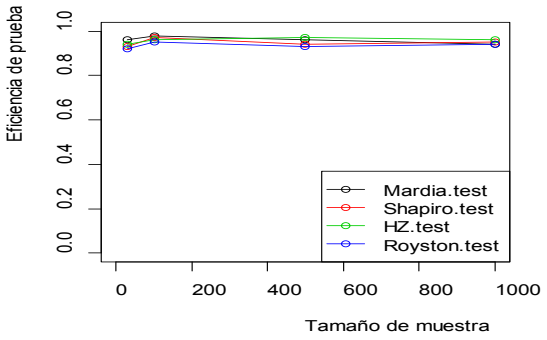


Gráfico 1. Potencia de Prueba para p=3 y VT=1.

Tabla 3. Potencia de Prueba para p=3 y VT=144

Prueba de Normalidad	Tamaño de Muestra			
	30	100	500	1000
Mardia	0.98	0.98	0.98	0.92
Shapiro -Wilk	0.95	0.98	0.96	0.94
HZ	0.97	0.93	0.94	0.93
Royston	0.93	0.98	0.96	0.92

Fuente: Elaboración propia

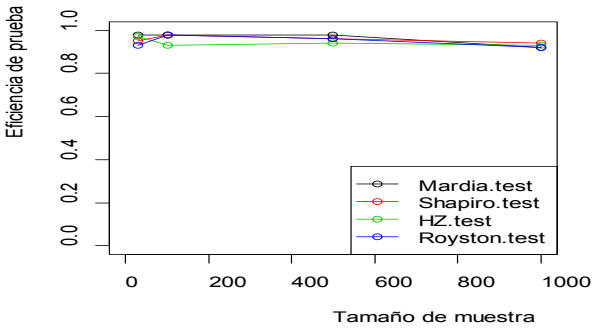


Gráfico 2. Potencia de Prueba para p=3 y VT=144.

Para la obtención de estos dos primeros resultados se ha considerado en la simulación pocas variables (p=3). La diferencia fundamental de los escenarios radica en la variabilidad generalizada que se ha tomado en cuenta. Esta variabilidad generalizada está definida como el determinante de la matriz de covarianza.

Se puede observar que no existen diferencias significativas entre las pruebas de normalidad multivariada analizadas para los distintos tamaños de muestra.

Se podría resaltar que si la variabilidad generalizada es alta y el tamaño de muestra aumenta, la potencia de prueba disminuye ligeramente.

Tabla 4. Potencia de Prueba para p=7 y VT=1.

Prueba de Normalidad	Tamaño de Muestra			
	30	100	500	1000
Mardia	1	0.94	0.91	0.9
Shapiro -Wilk	0.94	0.97	0.97	0.96
HZ	0.95	0.92	0.91	0.95
Royston	0.92	0.92	0.93	0.95

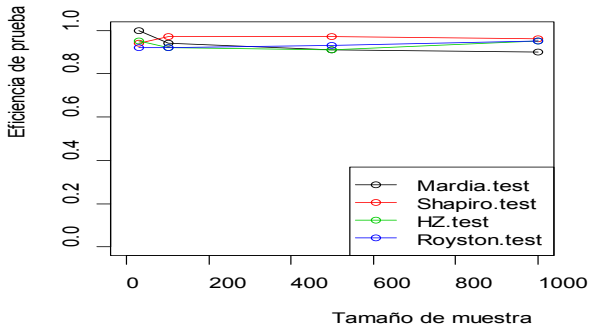


Gráfico 3. Potencia de Prueba para p=7 y VT=1.

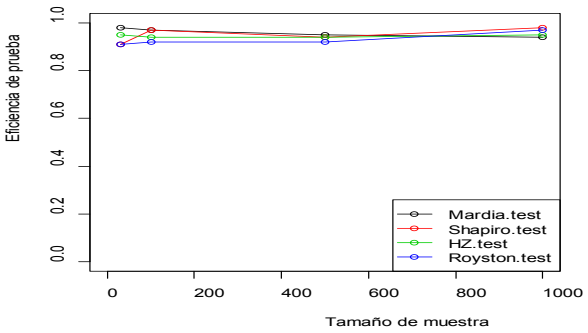
Fuente: Elaboración propia

Tabla 5. Potencia de Prueba para p=7 y VT=144.

Prueba de Normalidad	Tamaño de Muestra			
	30	100	500	1000
Mardia	0.98	0.97	0.95	0.94
Shapiro -Wilk	0.91	0.97	0.94	0.98
HZ	0.95	0.94	0.94	0.95
Royston	0.91	0.92	0.92	0.97

Fuente: Elaboración propia

Gráfico 4: Potencia de Prueba para p=7 y VT=144.



Estos dos últimos resultados se caracterizan, porque se utilizaron una mayor cantidad de variables en la simulación (p=7).

Se puede observar en los cuadros y gráficos para ambos niveles de variabilidad generalizada (baja y alta) que a medida que el tamaño de muestra se incrementa la potencia de la Prueba de Mardia disminuye ligeramente; mientras que en la prueba de Royston sucede todo lo contrario.

5. Conclusiones

El presente trabajo constituye un aporte en la investigación sobre la comparación de cuatro pruebas de normalidad multivariada. Sin embargo, realizando las respectivas modificaciones al programa de comparación elaborado en R se puede extender las comparaciones a más pruebas de normalidad multivariada que ofrece la literatura estadística.

Para investigaciones de este tipo, es necesario generar diferentes escenarios que involucren diferentes criterios como: tamaño de muestra, el número de variables consideradas en el estudio y la variabilidad generalizada. En esta investigación, se utilizaron 16 escenarios.

Se consideraron dos niveles de variabilidad, los cuales numéricamente fueron establecidos por el determinante de la matriz de covarianza. Sin embargo, no se encontraron diferencias significativas entre las pruebas en estudio.

Para generar las matrices de covarianza, se asumió independencia entre las variables. Lo que implica que

en futuras investigaciones se podría establecer un cierto grado de correlación entre las variables con el fin de determinar si los resultados obtenidos se mantienen o cambian.

Se ha demostrado que las cuatro pruebas consideradas no presentan diferencias significativas en su potencia de prueba. Sin embargo, cabe resaltar que cuando se trabaja con una mayor cantidad de variables la potencia de la prueba de Mardía disminuye ligeramente mientras que la potencia de prueba de la prueba de Royston aumenta a medida que el tamaño de muestra aumenta.

Finalmente, es necesario mencionar que el programa estadístico R es muy útil para implementar los procedimientos de comparación, debido a que los programas estadísticos comerciales no cuentan con las pruebas que se consideraron en la presente investigación.

6. Literatura citada

[1] **Burdenski, T (2000).** Evaluating univariate, bivariate, and multivariate normality using graphical and statistical procedures. Multiple Linear Regression.

[2] **Holgersson, H. (2006):** A graphical method for assessing multivariate normality. Computational Statistics.

[3] **Korkmaz, S., Goksuluk, D., Zararsiz, G. (2015).** MVN: An R Package for Assessing Multivariate Normality. Hacettepe University, Faculty of Medicine, Departament of Bioestatistics, Ankara, Turkey.

[4] **Ramzan, S., Maqbool, F., y Ramzan, S. (2013):** Evaluating multivariate normality: A graphical approach. Middle East Journal of Scientific Research.

[5] **Romeu, J., Ozturk, A. (1993):** A Comparative Study of Googness of Fit Tests for Multivariate Normality. CASE Center of Syracuse University. Academic Press Inc.

[6] **R version 3.2.2 Copyright © 2015**. The R Foundation for Statistical Computing.

[7] **Svantesson, T. y Wallace, J. (2003).** Tests for assessing multivariate normality and the covariance structure of mimo data. In Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). 2003 IEEE International Conference on, volume 4, pages 656–659. IEEE,

[8] **Timm, N. (2002)** Applied Multivariate Analysis. Springer texts in statistics.

[9] **Villasenor, J., González, E. (2009)** A Generalization of Shapiro-Wilk´s Test for Multivariate Normality.

Anexo

Función compara

Descripción: La función a la que se denominó "compara" permite obtener la potencia de prueba de las pruebas de normalidad multivariada de: Mardia, Henze-Zinkler, Shapiro-Wilk Generalizada y Royston.

Argumentos:

Esta función necesita como argumentos:

r: Número de veces que se desea realizar la simulación.

n: Vector de tamaños de muestra que se desean evaluar.

alfa: Nivel de significación con el cual se desea evaluar las pruebas estadísticas.

mu: Vector de medias de longitud p.

sigma: Matriz de covarianzas de dimensión pxp.

Valor:

La función brinda como resultado un cuadro de doble entrada donde se presentan las potencias de prueba para cada prueba de normalidad en cada tamaño de muestra considerado.

```
library(MVN)
library(mvShapiroTest)
compara<-function(r,n,alfa,mu,sigma)
{
resul2<-matrix(0,4,length(n))
rownames(resul2) <- c("Mardia.test","Shapiro.test",
"HZ.test","Royston.test" )
colnames(resul2) <- n
for (i in 1:length(n))
  {
resul1<-matrix(0,r,4)
for (j in 1:r)
  {
   datos<-as.matrix(mvrnorm(n[i],mu,sigma))
pvalores<-c(mardiaTest(datos, qqplot = FALSE)@p.
value.skew,mvShapiro.Test(datos)$p.value,hzTest(datos,
qqplot = FALSE)@ p.value, roystonTest(datos,qqplot=F
ALSE)@p.value)
resul1[j,]<-ifelse(pvalores>=alfa, 1, 0)
}
resul2[,i]<-apply(resul1,2,mean)
}
plot(c(0,n),seq(0,1,1/length(n)) ,type =
"n",xlab="Tamaño de muestra" ,ylab="Eficiencia de
prueba")
for(k in 1:4)
{
x<-as.vector(resul2[k,])
points(n,x,col=k)
lines(n,x,col=k)
}
legend("bottomright",rownames(resul2),lty=1,col=1:4
,pch=1)
return(resul2)
```