Prueba de Levene multivariada para la comparación de matrices de covarianza en presencia de datos faltantes

Por:

Mario José Pacheco López

UNIVERSIDAD NACIONAL DE COLOMBIA

FACULTAD DE CIENCIAS

ESCUELA DE ESTADÍSTICA

MEDELLÍN - ANTIOQUIA

OCTUBRE DE 2009

Prueba de Levene multivariada para la comparación de matrices de covarianza en presencia de datos faltantes

Por:
Mario José Pacheco López

Presentado como requisito parcial para optar al título de
MAGISTER EN ESTADÍSTICA

Director:
Juan Carlos Correa, Ph.D.

UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE CIENCIAS
ESCUELA DE ESTADÍSTICA
MEDELLÍN - ANTIOQUIA
OCTUBRE DE 2009

UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE CIENCIAS
ESCUELA DE ESTADÍSTICA

Los jurados abajo firmantes certifican que han leido y recomiendan a la **Facultad de Ciencias** aprobar el trabajo de grado titulado "**Prueba de Levene multivariada para la comparación de matrices de covarianza en presencia de datos faltantes**" presentado por **Mario José Pacheco López** como requisito parcial para optar al título de **Magister en Estadística**.

Fecha: Octubre de 2009

Director: _____
Juan Carlos Correa, Ph.D.

Jurados: _____
Francisco Castrillón M., MSc.

_____
René Iral P., MSc.

A mi madre

Emira López B.

Índice general

Índice de tablas

Índice de figuras

Resumen

El método para comparar matrices de covarianzas más citado en la literatura es la prueba de Bartlett, la cual es una prueba de razón de verosimilitud modificada, pero esta prueba es sensible a violaciones del supuesto de normalidad multivariada. Otro procedimiento para la comparación de matrices de covarianza es la extensión de la prueba de Levene, la cual consiste en dos generalizaciones multivariadas para la homogeneidad de varianzas; estas son robustas a desviaciones del supuesto de normalidad, con la ventaja adicional de la simplicidad computacional inducida por el procedimiento univariado de prueba. En el siguiente trabajo se examina el comportamiento de la prueba de Levene multivariada en presencia de datos faltantes; un estudio de simulación es llevado a cabo para evaluar su comportamiento en comparación a la prueba de razón de verosimilitud modificada en términos de los niveles de significancia nominal y real. Se encuentra que la prueba de Levene tiene un buen comportamiento para los datos normales y no normales en muestras pequeñas y grandes, como en la presencia de datos faltantes.

Palabras clave: *Datos faltantes; Pruebas de homogeneidad de matrices de covarianza; Prueba de Levene; Prueba de razón de verosimilitud.*

Abstract

The method for comparing covariance matrices most cited in the literature is the Bartlett test, which is a likelihood ratio test changed, but this test is sensitive to violations of multivariate normality assumption. Another procedure for the comparison of covariance matrices is the extension of the Levene test, which consists of two multivariate generalizations for homogeneity of variances, and these are robust to deviations from the assumption normality, with the added advantage of computational simplicity induced by univariate test. In the following work examines the behavior of multivariate Levene's test in the presence of missing data; a simulation study is conducted to evaluate its performance compared to the likelihood ratio test changed in terms of nominal significance levels and real. It is found that the Levene test has a good behavior for normal and nonnormal data in small samples and large, as in the presence of missing data.

Key words: *Likelihood ratio test; Levene's test; Missing data; Test of homogeneity of covariance matrices.*

Agradecimientos

El autor agradece a todos los profesores, compañeros de maestría y demás integrantes de la Escuela de Estadística. En especial agradece la asesoría del profesor Juan Carlos Correa en el desarrollo de este trabajo de grado.

Medellín, Antioquia — Mario Pacheco López
Septiembre de 2009

Introducción

La importancia de la comparación de matrices de varianzas y covarianzas en el análisis estadístico de datos, descansa en el hecho que muchos de los análisis estadísticos estándar requieren comparar las matrices de varianzas y covarianzas, como es el caso del análisis discriminante y el análisis de varianza multivariado, o de una manera indirecta como en los análisis preliminares del análisis de conglomerados y del análisis en componentes principales.

Existen diferentes pruebas para la comparación de matrices de varianzas y covarianzas. La más usada es la prueba basada en la estadística de razón de verosimilitud (LRT), siendo esta muy sensible a violaciones del supuesto de normalidad (Tiku y Balakrishnan (1985), O'Brien (1992), Aslam y Rocke (2005)). Entre los procedimientos de prueba robustos que se encuentran en la literatura especializada están algunas modificaciones a la prueba LRT, como es la obtención de los valores críticos de la prueba mediante remuestreo bootstrap (Zhang y Boos (1992)) y el empleo de S estimadores de la matriz de varianzas y covarianzas (Aslam y Rocke (2005)). Otras estadísticas propuestas se basan en el estadístico de Wald bajo distribuciones normales multivariadas y distribuciones elípticas multivariadas (Schott (2001)).

Un procedimiento robusto, que se comporta mejor que la LRT en presencia de desviaciones del supuesto de normalidad multivariada es el propuesto por O'Brien (1992). Esta es una extensión de la prueba de Levene que ofrece un procedimiento que, además de ser robusto, es de muy fácil aplicación y conserva las propiedades del procedimiento de prueba univariado.

En el caso datos faltantes, existen diferentes procedimientos de prueba como son la LRT con datos faltantes, la cual basa la obtención de las verosimilitudes en, por ejemplo, el algoritmo EM. También están los propuestos por Jamshidian y Schott (2007), que emplean modificaciones de las pruebas de Wald propuestas por Schott (2001) junto con una estrategia de eliminación de observaciones faltantes.

En el presente trabajo de investigación se estudia -vía simulación-,el comportamiento de la prueba de Levene multivariada para la comparación de matrices de varianzas y covarianzas y se analiza la robustez ésta frente a la prueba LRT bajo normalidad, no normalidad y con la presencia de datos faltantes. La desviación del supuesto de normalidad multivariada se logra empleando distribuciones normales sesgadas con diferentes valores de sus parámetros de sesgo.

El objetivo principal de este trabajo fue examinar el comportamiento de la prueba de Levene multivariada para probar homogeneidad de matrices de varianzas y covarianzas bajo desviaciones del supuesto de normalidad y en presencia de datos faltantes. Para esto se establecieron escenarios de simulación para la prueba de Levene bajo desviaciones del supuesto de normalidad multivariada y se analizó la robustez de la prueba de Levene bajo estos escenarios. La desviación del supuesto de normalidad multivariada se logró empleando distribuciones normales sesgadas con diferentes valores de sus parámetros de sesgo. Se establecieron además escenarios de simulación para la prueba de Levene y la prueba LRT en presencia de datos faltantes y se analizó la robustez de las pruebas en presencia de los datos faltantes. Para la simulación de los datos faltantes se usó la metodología de datos faltantes completamente aleatorios con diferentes porcentajes de dichos datos faltantes. Finalmente se comparó mediante simulación estadística la prueba de Levene con la prueba LRT en términos del nivel de significancia de la prueba.

Capítulo 1

Pruebas de igualdad de matrices de covarianza

Existen contrastes para una gran variedad de hipótesis sobre la matriz de covarianzas de una población, o sobre las matrices de covarianzas de más de una población. Se considera aquí de una manera general el contraste de hipótesis para la igualdad de matrices de covarianza de g poblaciones normales.

1.1. Contraste de interés

Suponga que se tiene un conjunto de g poblaciones p variadas. Estamos interesados en contrastar el sistema de hipótesis:

$$\begin{aligned} H_0 &: \Sigma_1 = \Sigma_2 = \cdots = \Sigma_g \\ H_1 &: \Sigma_i \neq \Sigma_s \quad \text{para algún, } i \neq s,\ i, s = 1, 2, ..., g \end{aligned} \tag{1.1}$$

donde Σ_i corresponde a la matriz de varianzas y covarianzas de la población i-ésima.

1.2. Prueba de razón de verosimilitud, LRT

Si la distribución de la muestra aleatoria $X_1, X_2, ..., X_n$ depende del vector de parámetros θ y si $H_0 : \theta \in \Omega_0$ y $H_1 : \theta \in \Omega_1$, entonces la estadística de razón de verosimilitud para probar H_0 contra H_1 se define como (Zhang y Boos (1992))

$$\ell = \frac{\sup_{\theta \in \Omega_0} L(\theta)}{\sup_{\theta \in \Omega_1 \cup \Omega_0} L(\theta)}$$

En el caso que se quiera probar $H_0 : \Sigma_1 = \Sigma_2 = \cdots = \Sigma_g$ contra $H_1 : \Sigma_i \neq \Sigma_s$ para algún $i \neq s$, $i, s = 1, 2, ..., g$, basados en g muestras de vectores aleatorios independientes $p \times 1$ $\{\mathbf{X}_{i1}, ..., \mathbf{X}_{in_i}\}$, $i = 1, ..., g$, la estadística de razón de verosimilitud toma la forma

$$\ell = \frac{|S|^{-(N-g)/2}}{\prod_{i=1}^{g} |S_i|^{-(n_i-1)/2}}$$

donde

$$S_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i\right) \left(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i\right)' \tag{1.2}$$

$$\bar{\mathbf{X}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{X}_{ij}$$

$$S = \sum_{i=1}^{g} \frac{n_i - 1}{N - g} S_i \tag{1.3}$$

y

$$N = \sum_{i=1}^{g} n_i \tag{1.4}$$

Bajo la hipótesis nula de igualdad de matrices de varianzas covarianzas y el supuesto de normalidad multivariada

$$R = -2 \log \ell = (N - g) \log |S| - \sum_{i=1}^{g} (n_i - 1) \log |S_i| \tag{1.5}$$

sigue aproximadamente una distribución χ^2 con $v = \frac{1}{2} p (p + 1) (g - 1)$ grados de

libertad. Esto es $-2\log\ell \xrightarrow{d} \chi^2_v$ conforme mín $(n_1, ..., n_g) \to \infty$ con $n_i/N \to \lambda_i \in (0, 1)$ para $i = 1, ..., g$ (Zhang y Boos (1992)).

1.3. Prueba de Levene multivariada

En el caso univariado, sea x_{ij} un conjunto de $j = 1, ..., n_i$ observaciones en cada uno de g grupos, $i = 1, ..., g$. La estadística de prueba de Levene, es la razón F de un ANOVA, comparando los g grupos, calculada sobre las desviaciones absolutas $z_{ij} =\mid x_{ij} - \bar{x}_i \mid$ a la media del grupo $\bar{x}_i$ (Anderson (2006)).

Una prueba multivariada para (1.1), análoga a la prueba univariada de Levene, propuesta por O'Brien (1992) se basa en la razón F del ANOVA calculado sobre las distancias euclídeas de puntos individuales de cada grupo a su centroide $\mathbf{c}_i$

$$z^c_{ij} = \Delta\left(\mathbf{x}_{ij}, \mathbf{c}_i\right)$$

donde el vector centroide se define como el punto que minimiza la suma de cuadrados de las distancias a cada punto

$$\sum_{j=1}^{n_i} \Delta^2\left(\mathbf{x}_{ij}, \mathbf{c}_i\right)$$

y

$$\Delta\left(\mathbf{x}_{ij}, \mathbf{c}_i\right) = \sqrt{\sum_{k=1}^{p} \left(x_{ikj} - c_{ik}\right)^2}$$

Este vector centroide usualmente se asume como el vector de medias muestrales, definido como $\bar{\mathbf{x}} = (\bar{x}_1, ..., \bar{x}_p)$.

Luego, la estadística F del ANOVA que se utiliza para probar H_0 es de la forma:

$$F = \frac{(N-g)\sum_{i=1}^{g} n_i \left(\bar{z}^c_{i\cdot} - \bar{z}^c_{\cdot\cdot}\right)^2}{(g-1)\sum_{i=1}^{g}\sum_{j=1}^{n_i} \left(z^c_{ij} - \bar{z}^c_{i\cdot}\right)^2}$$

con $N = \sum_{i=1}^{g} n_i$, $\bar{z}_{i\cdot}^{c} = \sum_{j=1}^{n_i} z_{ij}$ y $\bar{z}_{\cdot\cdot} = \sum_{i=1}^{g} z_{i\cdot}^{c}$.

Bajo H_0 la estadística F sigue aproximadamente una distribución F con $g-1$ y $N-g$ grados de libertad, Anderson (2006).

Brown y Forsythe (1974) y Anderson (2006) sugieren el uso de medianas para la prueba de Levene como una opción robusta para esta prueba:

$$z_{ij}^{m} = \Delta\left(\mathbf{x}_{ij}, \mathbf{m}_i\right)$$

donde $\mathbf{m}_i$ se define como el punto que minimiza la suma de las distancias a los puntos individuales del grupo i-ésimo

$$\sum_{j=1}^{n_i} \Delta\left(\mathbf{x}_{ij}, \mathbf{m}_i\right)$$

En este caso la mediana espacial no necesariamente es definida como el vector de medianas individuales para cada variable. En algunos paquetes estadísticos como el R se encuentran rutinas implementadas para encontrar medianas espaciales de este tipo.

1.4. Otros procedimientos de prueba

En la literatura especializada se encuentra un conjunto de procedimientos que pretenden de alguna manera resolver el problema que tiene la estadística de razón de verosimilitud usual frente a la no normalidad y que afecta directamente la distribución de la estadística LRT. A continuación se muestra en resumen algunos procedimientos de prueba que, aunque no serán usados para efectos de simulaciones ni contrastes, sí serán útiles al momento de realizar la discusión del trabajo.

1.4.1. Modificaciones a la prueba LRT

Entre las modificaciones propuestas en la literatura para volver robusta la prueba LRT se encuentra la propuesta por Zhang y Boos (1992) de reemplazar los valores

críticos basados en la aproximación χ^2 por valores críticos bootstrap adecuados para la estadística de razón de verosimilitud. Otra propuesta consiste en una modificación de la prueba de Bartlett a través de la estimación de la matriz de varianzas y covarianzas mediante un S estimador (Aslam y Rocke (2005)). Estas propuestas convierten la LRT clásica en una LRT robusta frente a desviaciones del supuesto de normalidad.

1.4.1.1. Valores críticos bootstrap

Para conseguir valores críticos adecuados en la prueba LRT se pueden calcular dichos valores vía metodología bootstrap, como lo describen Zhang y Boos (1992).

Considérese el espacio de remuestreo

$$R_s = \left\{ \mathbf{X}_{ij} - \bar{\mathbf{X}}, \quad j = 1, 2, ..., n_i, \quad i = 1, ..., g \right\}$$

La idea es extraer B muestras aleatorias independientes con remplazo de R_s y calcular el estadístico LRT para cada conjunto o muestra, escribiendo a cada estadística de razón de verosimilitud como $L_1^*, ..., L_B^*$ y notando que su distribución empírica es la estimación bootstrap de la distribución nula de L en (1.5).

Dado que H_0 en (1.1) es rechazada para valores grandes de L, el valor crítico α para L es el percentil $(1-\alpha)$ de la distribución de los L^* y el valor p bootstrap $\hat{p}_B$ es la proporción de los L_i^* que son menores o iguales que el valor observado de L en la muestra original.

1.4.1.2. S estimador de Σ

Una modificación de la prueba LRT clásica mediante la estimación de la matriz de varianzas y covarianzas mediante un S estimador es propuesto por Aslam y Rocke (2005). Esta propuesta convierte la LRT clásica en una LRT robusta.

Sea $\rho : \Re^+ \to \Re^+$ una función dos veces diferenciable, continua, simétrica y no decreciente con $\rho(0) = 0$ y $\rho(x) = \rho(c)$ para todo $x \geq c$. Dado un conjunto de n

puntos en $\Re^p$, el S estimador, $\left(\tilde{\mu}, \tilde{\Sigma}\right)$, se define como el mínimo $\mid \Sigma \mid$ sujeto a

$$n^{-1} \sum_i \rho\left(d_i\right) = b_0$$

donde

$$d_i^2 = \left(x_i - \mu\right)' \Sigma^{-1} \left(x_i - \mu\right)$$

y $b_0 = E\left(\rho\left(d\right)\right)$ con $d \sim N\left(0, 1\right)$.

La función $\rho\left(d\right)$ que escogen Aslam y Rocke (2005) es la denominada función ρ_{tb} por que minimiza la sensibilidad a observaciones atípicas:

$$\rho_{tb} = \begin{cases} \frac{d^2}{2} & d \leq a \\ \left(\frac{a^2}{2} - \frac{a^2\left(a^4 - 5a^2b^2 + 15b^4\right)}{30b^4}\right. & \\ +d^2\left(\frac{1}{2} + \frac{a^4}{2b^4} - \frac{a^2}{b^2}\right) + d^3\left(\frac{4a}{3b^2} - \frac{4a^3}{3b^4}\right) & \\ +d^4\left(\frac{3a^2}{2b^4} - \frac{1}{2b^2}\right) - \frac{d^5 4a}{5b^4} + \frac{d^6}{6b^4}\Big) & a \leq d \leq a + b \\ \frac{a^2}{2} + \frac{b(5b + 16a)}{30} & d \geq a + b \end{cases}$$

Con a y b constantes positivas.

Luego, para una muestra $X_{n \times p}$ de una $N\left(0, \Sigma\right)$ y S_{tb} el S estimador de Σ usando ρ_{tb}, mS_{tb} sigue aproximadamente una distribución Wishart $cW_p\left(\Sigma, m\right)$, donde c es una constante que satisface

$$E\left(S_{tb}\right) = c\Sigma$$

y m son los grados de libertad.

Los valores m y c pueden ser estimados como

$$\hat{m} = 2/\widehat{CV}$$

y

$$\hat{c} = \frac{1}{p} \sum_{i=1}^{p} s_{ii}$$

donde $\widehat{CV}$ es el estimador del coeficiente de variación de los elementos de la diagonal del S estimador de Σ y s_{ii} son los elementos de la diagonal de S_{tb}.

De esta forma la LRT robusta propuesta para el caso de g muestras es

$$\ell_R = \frac{\prod_{i=1}^{g} \left|S_{tb}^{(i)}\right|^{v_i/2}}{\left|\frac{\sum_{i=1}^{g} v_i S_{tb}^{(i)}}{\sum_{i=1}^{g} v_i}\right|^{\sum_{i=1}^{g} v_i/2}}$$

donde v_i son los grados de libertad asociados a $S_{tb}^{(i)}$.

La distribución de $-2\rho \log \ell_R$ se encuentra como

$$P\left(-2\rho \log \ell_R \leq x\right) = P\left(\chi_f^2 \leq x\right) + \frac{\gamma}{M^2}\left[P\left(\chi_{f+4}^2 \leq x\right) - P\left(\chi_f^2 \leq x\right)\right] + O\left(M^{-3}\right)$$

donde $M = \rho n$,

$$n = \sum_{i=1}^{g} n_i$$

$$f = p(p+1)(g-1)/2$$

$$\rho = 1 - \frac{2p^2+3p-1}{6(p+1)(g-1)n}\left(\left(\sum_{i=1}^{g} 1/k_i\right) - 1\right)$$

$$k_i = \frac{n_i}{\sum_{i=1}^{g} n_i}$$

$$\gamma = M^2 \frac{p(p+1)}{48\left(n\rho\right)^2}\left[(p-1)(p-2)\left(\left(\sum_{i=1}^{g} 1/k_i\right) - 1\right) - 6(g-1)\left(n\left(1-\rho\right)\right)^2\right]$$

1.4.2. Pruebas de Wald

En el caso de poblaciones normales p variadas Schott (2001) desarrolla un procedimiento de prueba para (1.1) basada en la estadística de Wald. Este mismo autor

propone, para cuando se pueda asumir poblaciones elípticas[1] con parámetros de curtosis común es una prueba basada en una estadística de Wald que aproveche el supuesto de distribuciones elípticas.

La propuesta de Schott (2001) para probar (1.1) consiste en el cálculo de la estadística

$$T_1 = \frac{N-g}{2}\left\{\sum_{i=1}^{g}\gamma_i \mathrm{tr}\left(S_i S^{-1} S_i S^{-1}\right) - \sum_{i=1}^{g}\sum_{j=1}^{g}\gamma_i\gamma_j \mathrm{tr}\left(S_i S^{-1} S_j S^{-1}\right)\right\}$$

donde $\gamma_i = \frac{n_i - 1}{N-g}$ y S_i, S y N definidos como en (1.2), (1.3) y (1.4), respectivamente.

Bajo H_0 y el supuesto de normalidad multivariada la estadística T_1 tiene una distribución χ^2 con $v = \frac{1}{2}(g-1)\,p\,(p+1)$.

En el caso que se pueda asumir poblaciones elípticas con parámetros de curtosis común, k, Schott (2001) desarrolla una prueba de Wald para probar (1.1) basada en la estadística

$$\begin{aligned} T_2 \;=\; & n\left(\sum_{i=1}^{g}\left\{\frac{1}{2}\hat{\delta}_1\gamma_i \mathrm{tr}\left(S_i S^{-1} S_i S^{-1}\right) - \hat{\delta}_2\gamma_i \mathrm{tr}\left(S_i S^{-1}\right)\right\}\right. \\ & \left. - \sum_{i=1}^{g}\sum_{j=1}^{g}\left\{\frac{1}{2}\hat{\delta}_1\gamma_i\gamma_j \mathrm{tr}\left(S_i S^{-1} S_j S^{-1}\right) - \hat{\delta}_2\gamma_i\gamma_j \mathrm{tr}\left(S_i S^{-1}\right)\mathrm{tr}\left(S_j S^{-1}\right)\right\}\right) \end{aligned}$$

la cual bajo H_0 sigue una distribución χ^2 con $v = (g-1)\,p\,(p+1)/2$ grados de libertad y

$$\hat{\delta}_1 = \left(1+\hat{k}\right)^{-1}$$

y

$$\hat{\delta}_2 = \frac{\hat{k}}{2\left(1+\hat{k}\right)\left[2\left(1+\hat{k}\right)+p\hat{k}\right]}$$

con $\hat{k}$ un estimador consistente del parámetro de curtosis k el cual se puede conseguir

[1] Un vector aleatorio $\mathbf{x}$, de tamaño $p \times 1$, tiene una distribución elíptica con vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas Σ si su función característica es de la forma: $\phi(\mathbf{t}) = e^{i\mathbf{t}'\boldsymbol{\mu}}\psi\left(\mathbf{t}'\Sigma\mathbf{t}\right)$, donde ψ es una función de escala tal que $\psi(0) = 1/2$.

como

$$\hat{k} = \sum_{l=1}^{g} \hat{k}^{(i)}/g$$

con $\hat{k}^{(i)}$ un estimador del parámetro de curtosis del i-ésimo grupo, el cual puede ser estimado con la expresión

$$\hat{k}^{(i)} = \frac{1}{3p} \sum_{j=1}^{p} \frac{z_j^{(i)}}{w_j^{(i)}} - 1\,,$$

Definiendo a las variables $z_j^{(i)}$ y $w_j^{(i)}$ en $\hat{k}^{(i)}$ como

$$z_j^{(i)} = \frac{1}{n_j^{(i)} - 4} \left\{ \sum_{l=1}^{n_j} \left(x_l^{(i)} - \bar{x}_j^{(i)} \right)^4 - 6 \left(s_{jj}^{(i)} \right)^2 \right\}$$

y

$$w_j^{(i)} = \frac{n_j^{(i)}}{n_j^{(i)} - 1} \left\{ \left(s_{jj}^{(i)} \right)^2 - \frac{z_j^{(i)}}{n_j^{(i)}} \right\}$$

1.4.3. Prueba de Tiku y Balakrishnan

Tiku y Balakrishnan (1985) proponen un procedimiento de prueba para igualdad de matrices de varianzas y covarianzas modificando la estadística T^2 para probar igualdad de vectores de media. El inconveniente con este procedimiento de prueba es que solo compara dos poblaciones normales bivariadas.

El procedimiento de prueba es como sigue. Considere dos muestras bivariadas provenientes de dos poblaciones normales bivariadas independientes,

$$\begin{pmatrix} x_{1j} \\ x_{2j} \end{pmatrix},\ j = 1, 2, ..., n_1 \quad \text{y} \quad \begin{pmatrix} y_{1j} \\ y_{2j} \end{pmatrix},\ j = 1, 2, ..., n_2$$

Definiendo

$$\begin{aligned} u_{1j} &= (x_{1j} - \bar{x}_1)^2 \\ u_{2j} &= (x_{2\cdot 1j} - \bar{x}_{2\cdot 1})^2 \end{aligned}$$

con $j = 1, 2, ..., n_1$ y

$$\begin{aligned} v_{1j} &= (y_{1j} - \bar{y}_1)^2 \\ v_{2j} &= (y_{2\cdot 1j} - \bar{y}_{2\cdot 1})^2 \end{aligned}$$

con $j = 1, 2, ..., n_2$. Donde

$$\begin{aligned} x_{2\cdot 1j} &= x_{2j} - \hat{b}x_{1j} \\ y_{2\cdot 1j} &= y_{2j} - \hat{b}y_{1j} \end{aligned}$$

y

$$\hat{b} = \frac{S_{x12} + S_{y12}}{S_{x11} + S_{y11}}$$

donde S_{zkl} corresponde a la covarianza muestral entre las variables k y l de una variable z.

Se define entonces la estadística T^2 de Hotelling como

$$T^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} (\bar{w}_1, \bar{w}_2) \begin{pmatrix} \phi_1^2 & 0 \\ 0 & \hat{\phi}_2^2 \end{pmatrix} \begin{pmatrix} \bar{w}_1 \\ \bar{w}_2 \end{pmatrix}$$

donde $w_1 = \bar{u}_1 - \bar{v}_1$ y $w_2 = \bar{u}_2 - \bar{v}_2$ y

$$\hat{\phi}_k^2 = \frac{S_{ukk} + S_{vkk}}{n_1 + n_2 - 2}, \quad k = 1, 2$$

Tiku y Balakrishnan (1985) demuestran que bajo el supuesto de normalidad bivariada y tamaños muestrales grandes

$$\frac{n_1 + n_2 - 3}{2\,(n_1 + n_2 - 2)} T^2$$

sigue aproximadamente una distribución F con 2 y $n_1 + n_2 - 3$ grados de libertad.

Capítulo 2

El problema de los datos faltantes

Los métodos estadísticos fueron desarrollados para analizar conjuntos rectangulares de datos que formen matrices donde las filas representan unidades, o también llamadas casos, observaciones o sujetos dependiendo del contexto, y las columnas representan variables o atributos asociados a esas unidades.

El problema que confrontan muchos estudios surge cuando algunas de las entradas de esa matriz bajo investigación no son observables, ya sea por problemas en la recolección de la información, problemas con el equipo de investigación o, en los casos de encuestas o estudios con personas, por la acción de éstas al no responder a las preguntas ya sea por falta de información, desconocimiento, vergüenza o temor a contestar, entre otras. Otro origen del problema de datos faltantes es la edición de datos.

2.1. Mecanismos que llevan a datos faltantes

Para manejar el problema de datos faltantes es necesario saber cuáles son las causas que llevan a la presencia de éstos y cuál es la naturaleza de los valores faltantes en las variables y su relación con los valores de otras variables dentro del conjunto de datos.

Definimos a $X = (x_{ij})$ como la matriz de datos con n filas y p columnas y $R = (r_{ij})$ la matriz indicadora de datos faltantes, la cual toma valor de 1 cuando el dato es faltante

y 0 cuando está disponible. Los mecanismos o causas que conllevan a la ausencia de valores en un conjunto de datos están sujetos a una probabilidad condicional entre los datos X y la matriz indicadora R dada por $f(R \mid X, \theta)$ donde θ denota los parámetros a estimarse.

Se describen tres tipos de mecanismos posibles para explicar la ausencia de valores en un conjunto de datos (Little y Rubin, 2002).

- Si la falta de datos no depende en sí de la matriz de datos X, es decir, que si tenemos;

  $$f(R \mid X, \theta) = f(R \mid \theta)$$

  para toda X, los datos se conocen como faltantes por completa aleatoriedad llamados por sus siglas en inglés por MCAR (*Missing Completely at Random*).

- Por otro lado, se define la siguiente partición; (X_{obs}, X_{miss}) en los datos de la matriz X donde X_{obs} representan los datos observados y X_{miss} representan los datos faltantes. Entonces un mecanismo menos restrictivo que el anterior es aquel donde la falta de datos depende sólo de X_{obs}, es decir que:

  $$f(R \mid X, \theta) = f(R \mid X_{obs}, \theta)$$

  para todo X_{miss}, los datos se conocen como faltantes por aleatoriedad y se denotan como MAR (*missing at random*), por sus siglas en inglés.

- En última instancia si la distribución de los datos faltantes depende de X_{miss}, y también puede que de X_{obs}, entonces decimos que los datos no son faltantes por aleatoriedad y se denotan por sus siglas en inglés NMAR (*not missing at random*). Este mecanismo obedece la siguiente distribución:

  $$f(R \mid X, \theta) = f(R \mid X_{obs}, X_{miss}, \theta)$$

Los mecanismos de datos faltantes se definen de acuerdo a cuan restrictivos son. El más restrictivo es el mecanismo MCAR, lo cual lo hace poco ocurrente en la vida real

pues la falta de datos es más usual que se deba a causas asociadas a un sector específico de la población. No obstante el mecanismo MAR no es el más común de todos pues todavía tiene la restricción de que la falta de datos depende sólo de los valores del conjunto de datos que son observados. Por lo que el mecanismo más frecuente en la realidad es el NMAR.

Por su naturaleza, los mecanismos MAR y NMAR son muy difíciles de implementar, pues, al ser los más comunes y menos restrictivos, sería prácticamente imposible crear simulaciones para todos los casos; no obstante el mecanismo MCAR, -el más restrictivo de todos-, permite escoger una simulación dentro de un marco limitado.

Para realizar el estudio de datos faltántes un método muy fácil de implementar consiste en eliminar las observaciones en la matriz de datos en donde aparezca, al menos, un valor faltante para una variable o atributo. Hacerlo así crea sesgos considerables dependiendo de la cantidad de datos eliminados; esto, además, incrementa la variabilidad de las estimaciones, es decir, si la proporción de valores faltantes es mínima, la estimación genera resultados satisfactorios, de aquí que sea de considerable importancia el mecanismo de generación de datos faltantes utilizado. Hay mecanismos que son un poco más flexibles que otros y permiten sólo la eliminación de datos correspondientes a una variable en particular, en otros casos es necesario eliminar todo el conjunto de observaciones del mismo individuo seleccionado de la muestra.

2.2. Solución al problema de datos faltantes: Métodos de imputación

Se describen a continuación algunas de las soluciones más comunes al problema de los datos faltantes. Aunque no serán usadas en este trabajo, estas servirán al momento de la discusión de los resultados del trabajo y sugerencias para futuros trabajos.

Los métodos de imputación se pueden definir simplemente como promedios o selecciones provenientes de una distribución de predicción de los valores faltantes que se basa en los valores observados.

1. Imputación por la media muestral (IMEAN): La imputación por la media muestral es uno de los métodos más antiguos de completar los datos faltantes en una matriz. El método consiste en completar los datos que faltan de la matriz X variable por variable, promediando los valores observados en cada variable y tomando ese promedio como la imputación o relleno para los datos faltantes. Este método de imputación es útil para variables numéricas continuas. Además es uno de los métodos más fáciles de aplicar y casi todos los programas estadísticos lo tienen. Sin embargo, por ser la media muestral una medida de tendencia central tiende a disminuir la variabilidad de los datos en la variable y esta disminución en la varibilidad puede afectar de manera significativa la comparación de matrices de covarianza.

2. Imputación por muestreo aleatorio (IRS): Este tipo de imputación es muy sencillo y funciona para cualquier tipo de variable. Consiste en tomar un muestreo aleatorio con remplazo de los valores observados en una variable y usarlos como reemplazos o sustituciones para los valores faltantes dentro de la misma variable.

3. Imputación por vecinos más cercanos (KNN): En un intento por buscar un método general que fuera más certero en sus estimaciones se han creado métodos que hacen uso de métricas para medir distancias entre unidades basadas en los valores de las variables asociadas dentro del mismo conjunto de datos. Luego se calculan tales distancias y se procede a imputar los valores faltantes utilizando las unidades del conjunto de unidades completas más cercanas según la métrica. El algoritmo KNN imputa valores de esta forma utilizando como métrica la distancia euclidiana entre las unidades. Los pasos del algoritmo se explican a continuación:

    a) Particionar el conjunto de datos D en dos partes: Las unidades completamente observadas D_c y las unidades con valores faltantes D_m.

    b) Para cada unidad $\mathbf{x}_i$ en D_m, calcular las distancias entre las $\mathbf{x}_i$ y las

unidades completas $\mathbf{x}_c$ y escoger las k unidades más cercanas según la distancia euclideana. Este conjunto escogido para $\mathbf{x}_i$ se llama el conjunto D_k de los k vecinos más cercanos a $\mathbf{x}_i$, el cual contiene valores faltantes para una o varias variables o atributos.

c) Con los valores en D_k se imputarán los valores faltantes en $\mathbf{x}_i$ en cada variable j dependiendo del tipo de ésta. Si la variable j es del tipo continuo entonces se hace un promedio de los k vecinos en esta variable y ese promedio pasa a ser el valor de imputación para $x_{i,j}$. Por otro lado si la variable es de tipo binaria u ordinal se buscará el valor que más se repita dentro de los k vecinos y ese pasará a ser el valor de imputación. El proceso termina cuando las unidades en D_m se han imputado completamente.

Capítulo 3

Pruebas de igualdad de matrices de covarianza con datos faltantes

En muchas aplicaciones de análisis de datos multivariados se presenta el problema de observaciones faltantes para alguno de los individuos estudiados en algunas de las variables de interés (no necesariamente en todas las variables).

Ahora bien, la mayoría de los métodos que se exponen en la literatura para comparar matrices de covarianzas, asumen que los datos son observados completamente (Jamshidian y Schott (2006)). Por esta razón es necesario examinar el comportamiento de los procedimientos de prueba bajo este problema y analizar su efectividad frente a los ya propuestos.

3.1. La prueba LRT con datos faltantes

Para el caso de datos faltantes Jamshidian y Schott (2007) describen una prueba LRT modificada.

En el caso de que el vector de observaciones $\mathbf{x}_j$ $(j = 1, ..., n)$ sea observado de manera parcial, bajo el supuesto de normalidad multivariada la contribución de cada $\mathbf{x}_j$ a la log-verosimilitud es

$$\ell_j = -\frac{1}{2}\left\{p_j \log(2\pi) + \log|\Sigma_j^{oo}| + \left(\mathbf{x}_{\text{obs},j} - \boldsymbol{\mu}_j^{o}\right)^t \left(\Sigma_j^{oo}\right)^{-1}\left(\mathbf{x}_{\text{obs},j} - \boldsymbol{\mu}_j^{o}\right)\right\}$$

donde p_j es el número de componentes observados de $\mathbf{x}_j$ y $\boldsymbol{\mu}_j^{\mathrm{O}}$ y Σ_j^{OO} el subvector y submatriz de $\boldsymbol{\mu}$ y Σ correspondientes a los componentes observados de $\mathbf{x}_j$. De esta forma la log-verosimilitud observada es

$$\ell = \sum_{j=1}^{n} \ell_j$$

Luego la LRT requiere:

1. L_0: Valor máximo de la log-verosimilitud bajo la restricción impuesta por la hipótesis nula.

2. L_1: Valor máximo irrestricto de la log-verosimilitud.

Así la LRT es dada por

$$R = -2\left(L_0 - L_1\right)$$

donde R se demuestra sigue una distribución χ^2 con ν=$(g-1)\left(p + p\left(p+1\right)/2\right)$ grados de libertad, aproximadamente, para n grande. En el caso que se pueda asumir el vector de medias como $\mathbf{0}$, los grados de libertad son $(g-1)\,p\,(p+1)\,/2$.

3.2. Otros procedimientos de prueba

Otro procedimiento de prueba basado en una modificación de la LRT lo proponen Jamshidian y Schott (2007). Estos autores muestran una prueba LRT reescalando la hipótesis nula (1.1).

Las pruebas de Wald propuestas por Schott (2001) para datos completos también son modificadas al caso de datos incompletos. La idea de estos procedimientos de prueba, bajo normalidad y bajo el supuesto de poblaciones elípticas con parámetro de curtosis común, es eliminar recursivamente los valores faltantes en la estimación de medias y covarianzas necesarios para el cálculo de la estadística de Wald con una mínima pérdida de información, Jamshidian y Schott (2007).

3.2.1. Prueba LRT re-escalada

Jamshidian y Schott (2007) también muestran una prueba LRT re escalando la hipótesis nula (1.1). En esta prueba se desea probar la hipótesis

$$H_0 : \sigma^{(i)} - \sigma^{(1)} = 0 \qquad i = 2, ..., m$$

donde $\sigma^{(i)}$ denota el vector columna $v\left(\Sigma^{(i)}\right)$, de tamaño $\frac{1}{2}p(p+1) \times 1$, obtenido de $vec\left(\Sigma^{(i)}\right)$[1] al eliminar todos los elementos sobre la diagonal de $\Sigma^{(i)}$. Esta prueba se basa en la estadística

$$R^* = \left[(m-1)\, p\,(p+1)\,/2\right] R/traza\left(\hat{\Omega}V\right)$$

donde R corresponde a la LRT definida en la Sección 3.1 y

$$\hat{\Omega} = \hat{H}^{-1}\hat{G}\hat{H}^{-1}$$

con $\hat{H}$ y $\hat{G}$ los estimadores de

$$H(\boldsymbol{\sigma}) = -\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ell_i(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}^T}$$

$$G(\boldsymbol{\sigma}) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ell_i(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} \frac{\partial \ell_i(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}^T}$$

y

$$V = \hat{H} - \hat{H}\dot{\sigma}\left(\dot{\sigma}^T \hat{H} \dot{\sigma}\right)^{-1} \dot{\sigma}^T \hat{H} \qquad \dot{\sigma} = \partial\boldsymbol{\sigma}/\partial\boldsymbol{\sigma}^{(1)}$$

Bajo la hipótesis nula R^* sigue una distribución χ^2 con $v = (m-1)\, p\,(p+1)\,/2$ grados de libertad, aproximadamente.

[1] Sea $A = \{a_{ij}\}$, $i = 1, ..., m$, $j = 1, ..., n$, $vec\,(A) = (a_{11}, a_{12}, ..., a_{ij}, ..., a_{mn})$'

3.2.2. Pruebas de Wald

Jamshidian y Schott (2007) adaptan las pruebas de Wald expuestas Schott (2001) al caso de datos incompletos.

Haciendo uso de la notación *vec*, escribiendo la hipótesis de interés (1.1) como

$$H_0 : \theta = 0 \qquad \text{contra} \qquad H_1 : \theta \neq 0$$

donde

$$\theta = \left(\theta^{(1)T}, ..., \theta^{(g-1)T}\right)$$

y

$$\theta^{(i)} = v\left(\Sigma^{(i)} - \Sigma^{(m)}\right)$$

para $i = 1, 2, ..., g-1$.

Así, bajo el supuesto de normalidad multivariada, el procedimiento de esta prueba se basa en la estadística

$$T_1 = \sum_{i=1}^{g-1} \left(\hat{\theta}^{(i)} - \bar{\theta}_*\right)^T W_i \left(\hat{\theta}^{(i)} - \bar{\theta}_*\right)$$

donde

$$\bar{\theta}_* = W^{-1} \sum_{j=1}^{g} W_j \hat{\theta}^{(j)}$$

$$W = \sum_{i=1}^{g} W_i$$

$$W_i = \left(\hat{\Omega}^{(i)}\right)^{-1}$$

con $\hat{\Omega}^{(i)}$ la matriz de varianzas y covarianzas estimada de $v\left(\hat{\Sigma}^{(i)}\right)$.

Los elementos de $\hat{\Omega}^{(i)}$ son de la forma

$$\widehat{cov}\left(s_{hl}^{(i)}, s_{jk}^{(i)}\right) = \frac{n_{hljk}^{(i)}}{n_{hl}^{(i)} n_{jk}^{(i)}} \left(\hat{\sigma}_{hj}^{(i)} \hat{\sigma}_{lk}^{(i)} + \hat{\sigma}_{hk}^{(i)} \hat{\sigma}_{lj}^{(i)}\right) + \frac{n_{hljk}^{(i)} \left(n_{hljk}^{(i)} - 1\right) \left(\hat{\sigma}_{hj}^{(i)} \hat{\sigma}_{lk}^{(i)} + \hat{\sigma}_{hk}^{(i)} \hat{\sigma}_{lj}^{(i)}\right)}{n_{hl}^{(i)} n_{jk}^{(i)} \left(n_{hl}^{(i)} - 1\right) \left(n_{jk}^{(i)} - 1\right)}$$

y

$$\hat{\sigma}_{hl} = \sum_{i=1}^{g} \frac{n_{hl}^{(i)} - 1}{n_{hl} - g} s_{hl}^{(i)}$$

donde

$$n_{hl} = \sum_{i=1}^{g} n_{hl}^{(i)}$$

$$s_{hl}^{(i)} = \frac{1}{n_{hl}^{(i)} - 1} \sum_{t \in \mathcal{A}_{hl}^{(i)}} \left(x_{th}^{(i)} - \tilde{x}_{h}^{(i)} \right) \left(x_{tl}^{(i)} - \tilde{x}_{l}^{(i)} \right)$$

$$\tilde{x}_{h}^{(i)} = \frac{1}{n_{hl}^{(i)}} \sum_{t \in \mathcal{A}_{hl}^{(i)}} x_{th}^{(i)}$$

con $\mathcal{A}_{hl}^{(i)}$ indicando el conjunto de elementos observados para las variables h y l del grupo i-ésimo.

En el caso de poblaciones elípticas con parámetro de curtosis común, el estadístico de Wald toma la forma

$$T_2 = \sum_{i=1}^{g-1} \left(\hat{\theta}^{(i)} - \tilde{\theta}_* \right)^T V_i \left(\hat{\theta}^{(i)} - \tilde{\theta}_* \right)$$

donde

$$\tilde{\theta}_* = V^{-1} \sum_{j=1}^{g} V_j \hat{\theta}^{(j)}$$

$$V = \sum_{i=1}^{g} V_i$$

$$V_i = \left(\hat{\Omega}^{(i)} \right)^{-1}$$

con $\hat{\Omega}^{(i)}$ la matriz de varianzas y covarianzas estimada de $v\left(\hat{\Sigma}^{(i)}\right)$ cuyos elementos

toman la forma

$$\begin{aligned}\widehat{cov}\left(s_{hl}^{(i)}, s_{jr}^{(i)}\right) &= \frac{n_{hljr}^{(i)}}{n_{hl}^{(i)} n_{jr}^{(i)}}\left\{\hat{k}\hat{\sigma}_{hl}^{(i)}\hat{\sigma}_{jr}^{(i)} + \left(1+\hat{k}\right)\left(\hat{\sigma}_{hj}^{(i)}\hat{\sigma}_{lr}^{(i)} + \hat{\sigma}_{hr}^{(i)}\hat{\sigma}_{lj}^{(i)}\right)\right\} \\ &+ \frac{n_{hljr}^{(i)}\left(n_{hljr}^{(i)}-1\right)\left(\hat{\sigma}_{hj}^{(i)}\hat{\sigma}_{lr}^{(i)} + \hat{\sigma}_{hr}^{(i)}\hat{\sigma}_{lj}^{(i)}\right)}{n_{hl}^{(i)} n_{jr}^{(i)}\left(n_{hl}^{(i)}-1\right)\left(n_{jr}^{(i)}-1\right)}\end{aligned}$$

Un estimador de k se consigue como

$$\hat{k} = \sum_{l=1}^{g} \hat{k}^{(i)}/g$$

con

$$\hat{k}^{(i)} = \frac{1}{3p}\sum_{j=1}^{p}\frac{z_j^{(i)}}{w_j^{(i)}} - 1$$

$$z_j^{(i)} = \frac{1}{n_j^{(i)}-4}\left\{\sum_{t\in\mathcal{A}_j^{(i)}}\left(x_{tj}^{(i)} - \bar{x}_j^{(i)}\right)^4 - 6\left(s_{jj}^{(i)}\right)^2\right\}$$

$$w_j = \frac{n_j^{(i)}}{n_j^{(i)}-1}\left\{\left(s_{jj}^{(i)}\right)^2 - \frac{z_j^{(i)}}{n_j^{(i)}}\right\}$$

Se debe tener en cuenta que en la práctica el cáculo de $\widehat{cov}\left(s_{hl}^{(i)}, s_{jr}^{(i)}\right)$ puede volverse engorroso si se tiene en cuenta que este no se encuentra implementado en los paquetes estadísticos comunes y su implementación no es de fácil programación.

Capítulo 4

Prueba de Levene con datos faltantes

Para el caso de datos faltantes no existe en la literatura una implementación de la prueba de Levene, aunque su implementación puede hacerse sin mayor complicación.

Como en la Sección (1.3) la estadística de Levene se basa en la razón F del ANOVA calculado sobre las distancias euclídeas de puntos individuales de cada grupo a su centroide

$$z_{ij}^{c} = \Delta\left(\mathbf{x}_{ij}, \mathbf{c}_{i}\right)$$

donde el vector centroide se define como el punto que minimiza la suma de cuadrados de las distancias a cada punto

$$\sum_{j=1}^{n_i} \Delta^{2}\left(\mathbf{x}_{ij}, \mathbf{c}_{i}\right)$$

y

$$\Delta\left(\mathbf{x}_{ij}, \mathbf{c}_{i}\right) = \sqrt{\sum_{k=1}^{p} \left(x_{ikj} - c_{ik}\right)^{2}}$$

Usando como en Jamshidian y Schott (2007) un método que utilice toda la información disponible cuando se estime el vector centroide necesario para llevar a cabo el procedimiento de prueba y si se escoge el vector de medias como el vector centroide, se procede a estimar el vector de medias estimando la media de la k-ésima variable

del i-ésimo grupo como:

$$\bar{x}_{ik} = \frac{1}{n_{ik}} \sum_{t \in A_{ik}} x_{itk}$$

donde A_{ik} representa el subconjunto de tamaño n_{ik}, observado de $\{1, 2, ..., n_i\}$, con $n_{ik} \leq n_i$.

Luego el estadístico F se calcula como:

$$F = \frac{(n_{obs} - g) \sum_{i=1}^{g} n_{i,obs} \left(\bar{z}_{i\cdot}^{c} - \bar{z}_{\cdot\cdot}^{c}\right)^2}{(g-1) \sum_{i=1}^{g} \sum_{t \in B_i} \left(z_{it}^{c} - \bar{z}_{i\cdot}^{c}\right)^2}$$

el cual sigue aproximadamente una distribución F con $g - 1$ y $n_{obs} - g$ grados de libertad y donde B_i representa el subconjunto, de tamaño $n_{i,obs}$, observado de los valores z_{it} de $\{1, 2, ..., n_i\}$.

Por su parte, el cálculo de $\bar{z}_{i\cdot}^{c}$ y $\bar{z}_{\cdot\cdot}^{c}$ se compone de la observaciones en la muestra que son observadas de manera completa

$$\begin{aligned} \bar{z}_{i\cdot}^{c} &= \frac{1}{n_{i,obs}} \sum_{t \in B_i} z_{it}^{c}, \\ \bar{z}_{\cdot\cdot}^{c} &= \frac{\sum_{i=1}^{g} n_{i,obs} \bar{z}_{i\cdot}^{c}}{n_{obs}} \end{aligned}$$

y

$$n_{obs} = \sum_{i=1}^{g} n_{i,obs}$$

Luego si se emplea la mediana espacial, la estadística de Levene se basará en la razón F del ANOVA calculado sobre las distancias euclídeas de cada punto a la mediana espacial

$$z_{ij}^{m} = \Delta\left(\mathbf{x}_{ij}, \mathbf{m}_i\right)$$

donde $\mathbf{m}_i$ se define como el punto que minimiza la suma de las distancias a los puntos

individuales del grupo i-ésimo

$$\sum_{j=1}^{n_i} \Delta\left(\mathbf{x}_{ij}, \mathbf{m}_i\right)$$

En el caso de la mediana espacial no es posible obtener una estadística que elimine únicamente el elemento de la variable faltante en cada individuo y por tanto es necesario eliminar el elemento completo dificultando así la efectividad de este procedimiento de prueba. De esta forma aunque la estadística F conserva la misma forma funcional que con el vector de medias estimado, el estimador de la mediana espacial no emplea toda la información disponible para su cálculo. Para el cálculo de la mediana espacial se emplea el algoritmo L_1-median descrito en Vardi y Zhang (1999) e implementado en R en la libreria SpatialNP.

En el Apéndice A2 se encuentra programada la función `mlevene` para R. En esta se calcula la estadística de Levene para el caso en el que se tienen observaciones faltantes, con la restricción de que cada individuo de la muestra tenga al menos una característica observada de las p características en estudio. La función calcula las estadísticas de Levene para datos faltantes descritas en este capítulo.

Para el caso de la estadística de Levene empleando el vector de medias la función `mlevene` emplea la metodología de utilización de toda la información disponible en la muestra para el cálculo de la estadística. Para el caso en el que se requiera la mediana espacial, la función invoca la función `spatial.median` de la librería SpatialNP. Con esta última función no es posible el cálculo de la mediana espacial utilizando toda la información disponible, así que se elimina el individuo completo. Como resultado de la función se tienen los valores de la estadística F con el vector de medias y con la mediana espacial (notados como Wc y Wm, respectivamente) y los valores p asociados a las estadísticas.

Capítulo 5

Estudio por simulación

El interés es examinar vía simulación el comportamiento de la prueba de Levene multivariada comparada con la prueba LRT bajo desviaciones del supuesto de normalidad multivariada y con la presencia de datos faltantes, comparando los niveles de significancia nominal y real de ambos procedimientos de prueba.

Se realizan las simulaciones de los niveles de significancia cuando los datos están incompletos. Los escenarios de simulación se dispusieron para analizar las diferencias entre matrices de covarianzas con diferentes números de variables ($p = 3, 5, 8$) y diferentes tamaños muestrales por grupo ($n = 30, 50, 100, 200, 500$). Por simplicidad se tomaron los mismos tamaños de muestra en cada grupo.

El número de poblaciones a comparar es dos. La hipótesis que fue contrastada fue $H_0 : \Sigma_1 = \Sigma_2$. Y se procedió a simular los niveles de significancia nominales del 5 % ($\alpha = 0{,}05$). Las muestras generadas en la simulación fueron simuladas con matríces de covarianza iguales a la matriz identidad, I_p.

5.1. Simulación de los datos faltantes

Para los propósitos de este estudio los resultados se basaron en el mecanismo de completa aleatoriedad MCAR. Todos los datos analizados en la simulación son completamente observados y, basados en este mecanismo, se remueven algunos de ellos

creando un conjunto de datos faltantes para cada conjunto de datos.

Para evitar la variabilidad en los resultados del experimento se usó un mecanismo MCAR completamente restrictivo donde cada valor dentro del conjunto de datos tiene exactamente la misma probabilidad de estar ausente. Los porcentajes de valores faltantes que se tomaron fueron 0, 10 y 30 por ciento.

5.2. Desviación de la normal multivariada

El mecanismo que se usó para simular las desviaciones del supuesto de normalidad multivariada se basó en el uso de poblaciones normales sesgadas multivariadas (Azzalini y Dalla Valle (1996)).

Una variable aleatoria $\mathbf{Z} = (Z_1, ..., Z_p)'$ se dice que tiene una distribución normal multivariada sesgada si su función de densidad es de la forma

$$f_p(\mathbf{z}) = 2\phi_p(\mathbf{z}; \Omega)\, \Phi(\boldsymbol{\lambda}'\mathbf{z}), \quad \mathbf{z} \in R^p$$

donde $\phi_p(z; \Omega)$ denota la función de densidad multivariada con marginales estandarizadas y matriz de correlación Ω, y $\Phi(\cdot)$ es la función de distribución acumulada normal.

En la normal multivariada sesgada el parámetro $\boldsymbol{\lambda}$ representa el parámetro de forma. Si $\boldsymbol{\lambda}$ es $\mathbf{0}$ la normal sesgada se reduce a la distribución normal multivariada con matriz de correlación Ω. En la Figura 1 se muestra el comportamiento de la densidad normal sesgada para diferentes valores de $\boldsymbol{\lambda}$.

5.3. Resultados de la simulación

En este apartado se presentan los resultados del proceso de simulación antes descrito con el fin de comparar las pruebas estadísticas de Levene y LRT.

Los programas de simulación se desarrollaron en el paquete estadístico R versión 2.9.0. En el apéndice se encuentran las funciones R para el calculo de los niveles

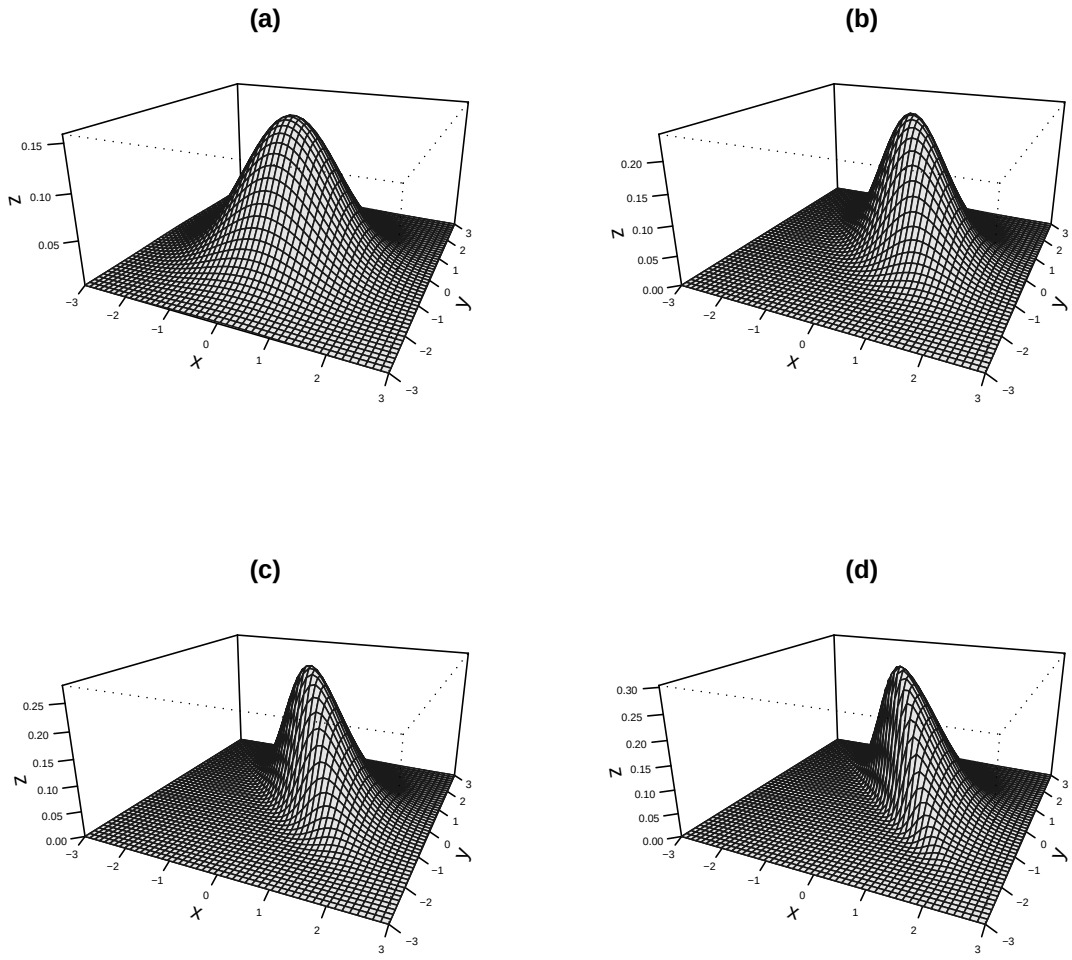

Figura 5.1: Densidades normal sesgada multivariada para diferentes valores del parámetro de forma, (a) $\boldsymbol{\lambda} = \mathbf{0}$, (b) $\boldsymbol{\lambda} = \mathbf{1,5}$, (c) $\boldsymbol{\lambda} = \mathbf{3}$, (d) $\boldsymbol{\lambda} = \mathbf{6}$

de significancia de las pruebas de Levene y LRT bajo los escenarios descritos en la metodología.

Las Tablas de la 5.1 a la 5.5 comparan los niveles de significancia para las pruebas de Levene y LRT. En estas el valor del parámetro $\boldsymbol{\lambda}$ es igual a $\mathbf{0}$, $\mathbf{0.7}$, $\mathbf{1.5}$, $\mathbf{3}$ y $\mathbf{6}$, respectivamente. En cada caso los datos para los dos grupos simulados tienen igual tamaño muestral por grupo $n_i = 30$, 50, 100, 200 y 500. También se considera un número de variables $p = 3$, 5 y 8. En cada caso, se simuló el nivel de significancia cuando la proporción de valores faltantes es 0, 10 y 30 por ciento. En todas las comparaciones el nivel de significancia nominal es de 5 %. Los resultados están basados en 5000 simulaciones. Por simplicidad, el valor promedio $\boldsymbol{\mu}$ de las muestras simuladas se asumió como $\mathbf{0}$.

La Tabla 5.1 muestra los resultados de la simulación para datos generados de una normal multivariada ($\boldsymbol{\lambda} = \mathbf{0}$). Es interesante notar que cuando los datos están completamente observados la LRT muestra niveles de significancia simulados mayores que los de la prueba de Levene y esta diferencia se incrementa conforme el número de variables incrementa también. Cuando el porcentaje de datos faltantes es diferente de 0 en cada grupo se nota que la prueba de Levene con $\bar{x}$ y Me tienen niveles de significancia menores que los de la LRT y a su vez están alrededor de 5 %.

En las Tablas 5.2 a la 5.5 se muestran los resultados para los datos simulados con un parámetro de forma $\boldsymbol{\lambda} \neq \mathbf{0}$. Aquí se nota como la LRT presenta niveles de significancia simulados superiores al nivel de significancia nominal, mientras la prueba de Levene muestra mejores resultados. Cuando se compara los niveles de significancia simulados con la prueba de Levene usando el vector de medias o la mediana espacial se observa que los resultados no difieren de manera significativa y además estos se mantienen al rededor del nivel de significancia nominal de 5 %.

En las figuras 5.2 y 5.3 se ilustra el comportamiento de los niveles de significancia cuando nos alejamos del supuesto de normalidad multivariada. En estas figuras el eje x contiene los valores de $\boldsymbol{\lambda}$ y el eje y los niveles de significancia simulados, manteniendo fijos a p (=3) y el tamaño de muestra n (=30 y 100, respectivamente) para cada

	$p=3$			$p=5$			$p=8$		
% Missing	0	10	30	0	10	30	0	10	30
$n_i = 30$									
LRT	0.053	0.146	0.514	0.090	0.293	0.933	0.175	0.774	0.933
Levene ($\bar{x}$)	0.056	0.041	0.050	0.042	0.046	0.058	0.038	0.054	0.078
Levene (Me)	0.054	0.048	0.062	0.044	0.043	0.060	0.034	0.056	0.070
$n_i = 50$									
LRT	0.056	0.131	0.416	0.074	0.215	0.797	0.098	0.487	0.797
Levene ($\bar{x}$)	0.050	0.042	0.047	0.040	0.056	0.063	0.058	0.049	0.046
Levene (Me)	0.052	0.040	0.047	0.045	0.064	0.050	0.060	0.037	0.037
$n_i = 100$									
LRT	0.046	0.106	0.370	0.053	0.199	0.696	0.068	0.354	0.696
Levene ($\bar{x}$)	0.047	0.054	0.047	0.047	0.060	0.056	0.042	0.070	0.048
Levene (Me)	0.053	0.057	0.055	0.050	0.054	0.044	0.051	0.066	0.044
$n_i = 200$									
LRT	0.045	0.106	0.328	0.063	0.141	0.605	0.068	0.272	0.605
Levene ($\bar{x}$)	0.049	0.049	0.051	0.042	0.044	0.051	0.055	0.046	0.054
Levene (Me)	0.055	0.056	0.040	0.041	0.046	0.047	0.048	0.052	0.047
$n_i = 500$									
LRT	0.052	0.114	0.323	0.045	0.135	0.598	0.064	0.243	0.598
Levene ($\bar{x}$)	0.053	0.047	0.047	0.049	0.054	0.055	0.043	0.050	0.037
Levene (Me)	0.055	0.046	0.044	0.056	0.056	0.054	0.042	0.050	0.038

Tabla 5.1: Niveles de significancia simulados, $\boldsymbol{\lambda} = \mathbf{0}$

	$p=3$			$p=5$			$p=8$		
% Missing	0	10	30	0	10	30	0	10	30
$n_i = 30$									
LRT	0.059	0.153	0.503	0.093	0.307	0.951	0.193	0.774	1.000
Levene ($\bar{x}$)	0.048	0.044	0.044	0.044	0.060	0.051	0.044	0.046	0.049
Levene (Me)	0.049	0.048	0.039	0.052	0.051	0.056	0.051	0.065	0.026
$n_i = 50$									
LRT	0.052	0.123	0.425	0.069	0.209	0.834	0.097	0.497	0.995
Levene ($\bar{x}$)	0.054	0.038	0.050	0.048	0.050	0.047	0.045	0.055	0.054
Levene (Me)	0.057	0.052	0.055	0.055	0.044	0.044	0.039	0.051	0.065
$n_i = 100$									
LRT	0.062	0.099	0.367	0.052	0.211	0.713	0.083	0.375	0.980
Levene ($\bar{x}$)	0.053	0.053	0.060	0.060	0.066	0.056	0.042	0.054	0.046
Levene (Me)	0.057	0.045	0.051	0.059	0.055	0.049	0.043	0.070	0.037
$n_i = 200$									
LRT	0.059	0.109	0.350	0.065	0.147	0.633	0.079	0.290	0.925
Levene ($\bar{x}$)	0.043	0.051	0.052	0.061	0.055	0.043	0.054	0.049	0.046
Levene (Me)	0.046	0.049	0.052	0.063	0.054	0.051	0.052	0.054	0.050
$n_i = 500$									
LRT	0.054	0.106	0.322	0.040	0.132	0.609	0.063	0.253	0.930
Levene ($\bar{x}$)	0.052	0.054	0.049	0.045	0.049	0.053	0.046	0.053	0.050
Levene (Me)	0.053	0.050	0.053	0.044	0.040	0.047	0.053	0.050	0.043

Tabla 5.2: Niveles de significancia simulados, $\boldsymbol{\lambda} = \mathbf{0,7}$

	$p=3$			$p=5$			$p=8$		
% Missing	0	10	30	0	10	30	0	10	30
$n_i = 30$									
LRT	0.060	0.170	0.531	0.105	0.317	0.950	0.192	0.787	1.000
Levene ($\bar{x}$)	0.043	0.051	0.049	0.038	0.050	0.061	0.037	0.050	0.038
Levene (Me)	0.040	0.068	0.055	0.044	0.052	0.067	0.056	0.051	0.043
$n_i = 50$									
LRT	0.073	0.142	0.443	0.074	0.234	0.830	0.115	0.508	1.000
Levene ($\bar{x}$)	0.054	0.053	0.045	0.061	0.041	0.053	0.052	0.048	0.049
Levene (Me)	0.053	0.057	0.057	0.063	0.033	0.041	0.052	0.050	0.056
$n_i = 100$									
LRT	0.063	0.130	0.399	0.061	0.216	0.710	0.084	0.367	0.974
Levene ($\bar{x}$)	0.057	0.051	0.046	0.055	0.045	0.045	0.059	0.042	0.042
Levene (Me)	0.053	0.043	0.046	0.058	0.057	0.044	0.065	0.053	0.039
$n_i = 200$									
LRT	0.070	0.117	0.366	0.069	0.178	0.659	0.080	0.298	0.968
Levene ($\bar{x}$)	0.047	0.043	0.041	0.047	0.046	0.058	0.056	0.043	0.054
Levene (Me)	0.047	0.040	0.041	0.047	0.038	0.054	0.061	0.043	0.056
$n_i = 500$									
LRT	0.063	0.123	0.334	0.047	0.149	0.624	0.075	0.270	0.880
Levene ($\bar{x}$)	0.041	0.049	0.045	0.048	0.051	0.055	0.047	0.053	0.039
Levene (Me)	0.038	0.053	0.049	0.048	0.052	0.048	0.049	0.055	0.044

Tabla 5.3: Niveles de significancia simulados, $\boldsymbol{\lambda} = \mathbf{1,5}$

	$p=3$			$p=5$			$p=8$		
% Missing	0	10	30	0	10	30	0	10	30
$n_i = 30$									
LRT	0.067	0.178	0.529	0.110	0.326	0.953	0.192	0.792	1.000
Levene ($\bar{x}$)	0.052	0.046	0.049	0.038	0.050	0.064	0.037	0.051	0.049
Levene (Me)	0.050	0.056	0.060	0.047	0.052	0.058	0.041	0.05	0.038
$n_i = 50$									
LRT	0.077	0.151	0.473	0.077	0.253	0.841	0.121	0.520	1.000
Levene ($\bar{x}$)	0.046	0.057	0.046	0.055	0.042	0.051	0.064	0.059	0.059
Levene (Me)	0.047	0.058	0.047	0.047	0.037	0.044	0.070	0.052	0.060
$n_i = 100$									
LRT	0.063	0.145	0.402	0.068	0.207	0.713	0.081	0.378	0.981
Levene ($\bar{x}$)	0.046	0.046	0.064	0.058	0.052	0.051	0.038	0.055	0.036
Levene (Me)	0.040	0.063	0.057	0.061	0.052	0.046	0.044	0.053	0.034
$n_i = 200$									
LRT	0.082	0.126	0.381	0.067	0.191	0.663	0.077	0.300	0.935
Levene ($\bar{x}$)	0.051	0.050	0.045	0.047	0.051	0.041	0.058	0.047	0.050
Levene (Me)	0.052	0.045	0.045	0.045	0.048	0.037	0.052	0.049	0.050
$n_i = 500$									
LRT	0.073	0.149	0.371	0.052	0.152	0.629	0.073	0.220	0.893
Levene ($\bar{x}$)	0.057	0.057	0.030	0.050	0.054	0.045	0.040	0.045	0.032
Levene (Me)	0.058	0.051	0.039	0.041	0.054	0.060	0.040	0.042	0.037

Tabla 5.4: Niveles de significancia simulados, $\boldsymbol{\lambda} = \mathbf{3}$

	$p=3$			$p=5$			$p=8$		
% Missing	0	10	30	0	10	30	0	10	30
$n_i = 30$									
LRT	0.070	0.178	0.531	0.110	0.333	0.955	0.194	0.790	1.000
Levene ($\bar{x}$)	0.042	0.053	0.059	0.046	0.061	0.046	0.045	0.045	0.029
Levene (Me)	0.049	0.050	0.047	0.047	0.045	0.059	0.058	0.053	0.049
$n_i = 50$									
LRT	0.073	0.147	0.475	0.085	0.258	0.846	0.130	0.527	1.000
Levene ($\bar{x}$)	0.049	0.041	0.057	0.054	0.055	0.045	0.048	0.054	0.050
Levene (Me)	0.048	0.048	0.049	0.060	0.043	0.050	0.045	0.058	0.056
$n_i = 100$									
LRT	0.070	0.147	0.409	0.073	0.206	0.726	0.085	0.380	0.984
Levene ($\bar{x}$)	0.047	0.039	0.045	0.041	0.048	0.060	0.055	0.058	0.038
Levene (Me)	0.043	0.053	0.042	0.042	0.051	0.052	0.056	0.056	0.036
$n_i = 200$									
LRT	0.089	0.138	0.397	0.070	0.188	0.665	0.079	0.297	0.947
Levene ($\bar{x}$)	0.048	0.046	0.058	0.059	0.045	0.041	0.052	0.049	0.047
Levene (Me)	0.042	0.043	0.056	0.057	0.045	0.050	0.058	0.049	0.045
$n_i = 500$									
LRT	0.074	0.150	0.378	0.051	0.154	0.621	0.075	0.213	0.896
Levene ($\bar{x}$)	0.057	0.043	0.032	0.052	0.053	0.044	0.048	0.057	0.030
Levene (Me)	0.065	0.050	0.041	0.053	0.058	0.047	0.052	0.054	0.038

Tabla 5.5: Niveles de significancia simulados, $\boldsymbol{\lambda} = \mathbf{6}$

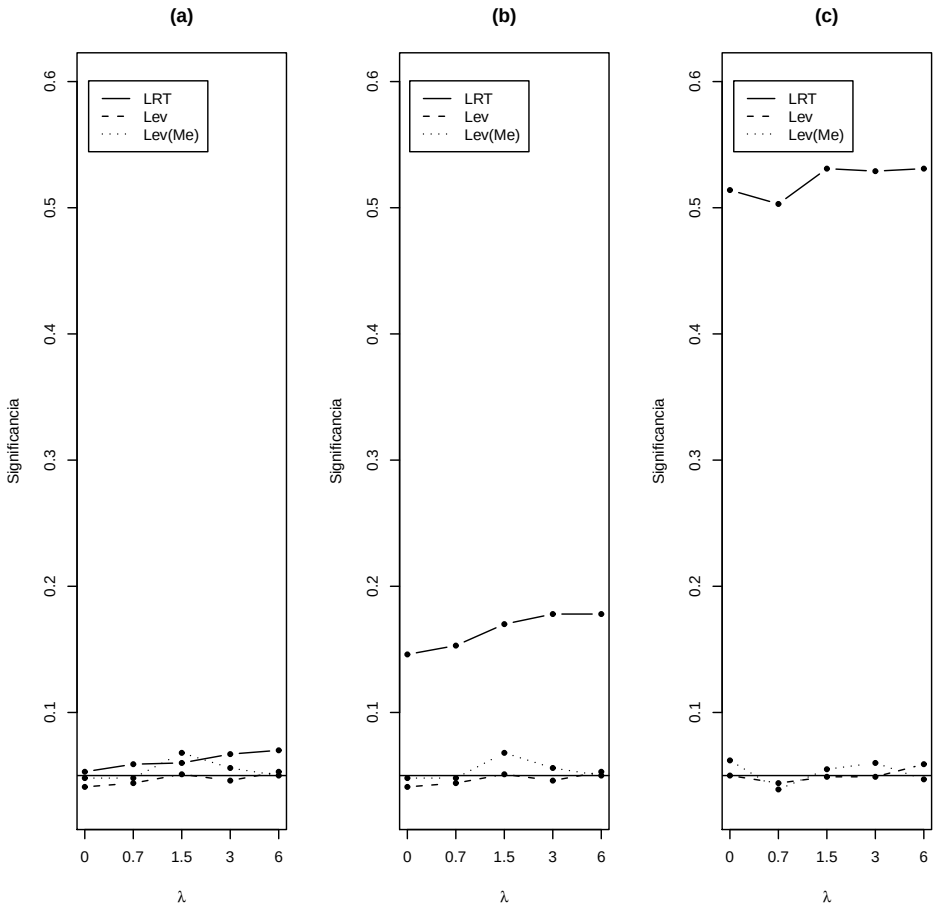

Figura 5.2: Niveles de significancia para $p = 3$, $n = 30$ y distintos valores de $\boldsymbol{\lambda}$. (a) % Missing = 0, (b) % Missing = 10, (c) % Missing = 30

porcentaje de valores missing. En estas se observa como los niveles de significancia de la prueba LRT crecen a medida que se aumenta el valor de $\boldsymbol{\lambda}$ alejándose del nivel de significancia nominal del 5 %. También se observa que el nivel de significancia simulado para las pruebas de Levene empleando el vector de medias o la mediana espacial se mantienen alrededor del nivel de significancia nominal. En los gráficos no se evidencia una gran diferencia entre las pruebas de Levene con el vector de medias y la mediana espacial.

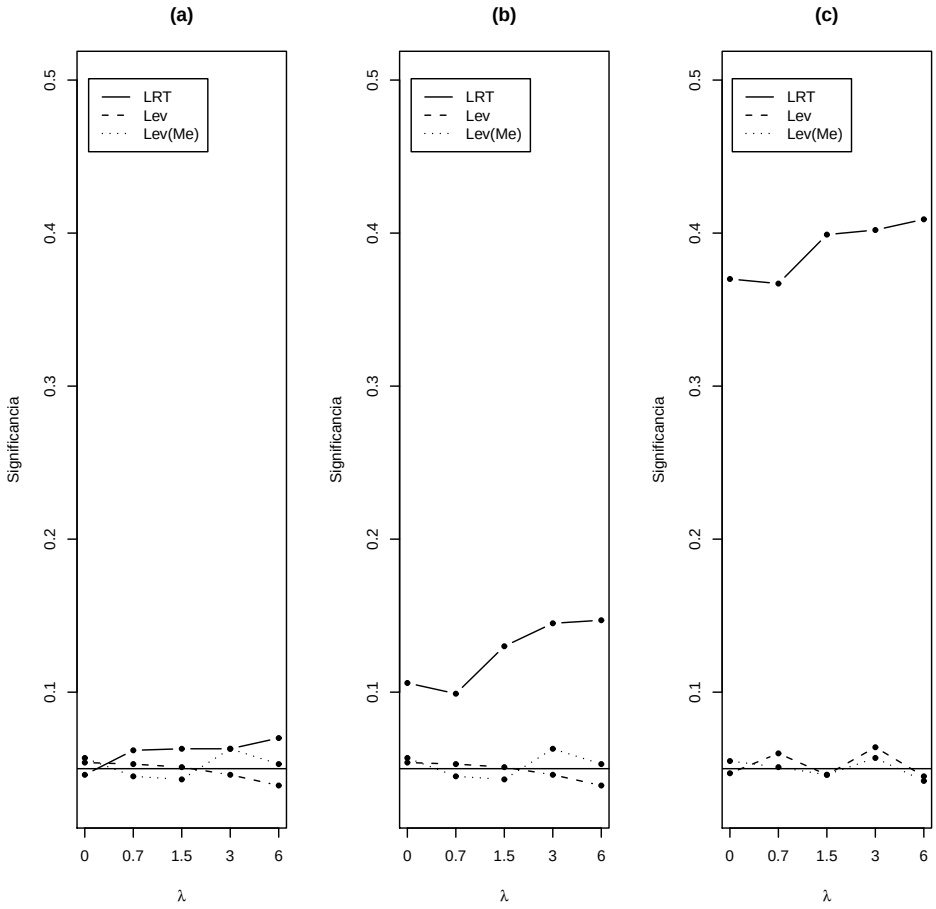

Figura 5.3: Niveles de significancia para $p = 3$, $n = 100$ y distintos valores de $\boldsymbol{\lambda}$. (a) % Missing = 0, (b) % Missing = 10, (c) % Missing = 30

Capítulo 6

Conclusiones y recomendaciones

Este trabajo discute el comportamiento de la extensión multivariada de la prueba de Levene comparada con la prueba LRT clásica en la comparación de matrices de covarianza. Los resultados son obtenidos vía simulación de los niveles de significancia nominal para ambas pruebas.

El procedimiento de prueba propuesto para la estadística de Levene en el caso de datos faltantes se basa en una estrategia simple de utilización de toda la información disponible a la hora de calcular la estadística. Se emplean el vector de medias y una definición de la mediana espacial para el cálculo de la estadística de Levene.

El principal hallazgo es el hecho de que la prueba de Levene supera a la LRT en la obtención de niveles de significancia predeterminados. Cuando los porcentajes de valores faltantes se incrementan se observa como la prueba de Levene en sus dos versiones simuladas tienen niveles de significancia nominal de alrededor del 5 %. También se confirma que la prueba LRT es sensible a violaciones del supuesto de normalidad multivariada, mientras que la prueba de Levene se mantiene robusta a las violaciones de ese supuesto. Lo mismo sucede cuando el número de variables se incrementa y el tamaño muestral también aumenta.

Aunque no se hicieron comparaciones de la prueba de Levene con el resto de pruebas que se encuentran en la literatura, cabe resaltar que la simplicidad del procedimiento

de prueba de la estadística de Levene hace a esta prueba mucho más atractiva que el resto de alternativas, si se tiene en cuenta por ejemplo que la forma funcional sencilla de la prueba de Wald cuando no se tienen datos faltantes no se conserva en su extensión al caso en el que se tengan datos faltantes. Además otros procedimientos de prueba no tienen adaptaciones al caso en el que se tengan datos faltantes, como es el caso de los métodos de Zhang y Boos (1992), Tiku y Balakrishnan (1985), Aslam y Rocke (2005), entre otros.

Por otra parte, la metodología de imputación de observaciones faltantes es una alternativa a la hora de solucionar el problema su existencia, pero no se encuentra documentado el efecto que pueda tener este procedimiento en ninguna de las pruebas de homogeneidad multivariadas referenciadas en este trabajo, además de que la escogencia de un método de imputación amplía el rango de posibilidades a la hora de evaluar los procedimientos de prueba.

Para la prueba de Levene no fue posible adaptar la metodología de utilización de toda la información disponible empleando la mediana espacial y por tanto fue necesario llevar a cabo el procedimiento de prueba de Levene con los elementos observados de manera completa. Aún así, en el estudio por simulación del Capítulo 5 no se observan diferencias relevantes entre el procedimiento de prueba de Levene empleando el vector de medias o la mediana espacial. Una salida a este inconveniente podría ser la imputación de los datos faltantes con algún mecanismo como los definidos en la Sección 2.2.

Se construyó una función en R llamada `mlevene` que calcula la estadística de Levene para el caso de datos faltantes empleando el vector de medias y la mediana espacial. Para el caso del vector de medias, la función emplea toda la información disponible en la muestra. Para el caso de la mediana espacial la función elimina todo el individuo que tiene al menos una característica faltante. Finalmente la función entrega las estadísticas F con sus respectivos valores p.

En la literatura especializada existen diferentes procedimientos univariados de prueba

para homogeneidad varianzas. Correa, Iral y Rojas (2006) analizan el comportamiento de algunas de estas pruebas vía siulación. La prueba de Levene multivariada aquí descrita consiste en una extensión multivariada de uno de estos procedimientos univariados de prueba. Un paso a seguir a este trabajo consistiría en adaptar algunos de estos procedimientos de prueba al caso multivariado y examinar su comportamiento cuando se tenga el problema de datos faltantes y frente a desviaciones del supuesto de normalidad multivariada o en el caso que se tengan observaciones atípicas.

Bibliografía

[1] Anderson, M., 2006. Distance-based tests for homogeneity of multivariate dispersion. Biometrics.

[2] Aslam, S. y Rocke D., 2005. A robust testing procedure for the equality of covariance matrices. Computational Statistics & Data Analysis 49, 863–874.

[3] Azzalini A. y Dalla Valle, A., 1996. The multivariate skew-normal distribution. Biometrika 83, 715–726.

[4] Brown, M. y Forsythe, A. 1974. Robust tests for the equality of variances. Journal of the American Statistical Association 69, 364–376.

[5] Correa, J., Iral, R. y Rojas, L. 2006. Estudio de potencia de pruebas de homogeneidad de varianza. *Rev. Col. de Est.* 29, 57–76.

[6] Jamshidian, M. y Schott, J., 2007. Testing equality of covariance matrices when data are incomplete. *Computational Statistics & Data Analysis* **51**, 4227–4239.

[7] Little, R. y Rubin, D., 2002. Statistical Analysis With Missing Data, second ed. Wiley, New York.

[8] O'brien, P., 1992. Robust procedures for testing equality of covariance matrices. Biometrics 48, 819–827.

[9] Schott, J., 2001. Some tests for the equality of covariance matrices. J. Statatistical planning and inference 94, 25–36.

[10] Tiku, M. y N. Balakrishnan, N., 1985. Testing the equality of variance-covariance matrices the robust way. Communications in Statistics - Theory and Methods 12, 3033–3051

[11] Vardi, Y. y Zhang, C. (1999), The multivariate L1-median and associated data depth, PNAS, 97, 1423–1426.

[12] Zhang, J. y Boos, D., 1992. Bootstrap critical values for testing homogeneity of covariance matrices. J. of the American Statistical Association 87, 425-429.

Apéndice

A1 Generación de los datos faltantes

```
################### Generación de datos faltantes ###################
## Mecanismo de generación de missing MCAR.                        ##
## Se aplca sobre una matrix de datos n x p.                       ##
#####################################################################

gen.mis = function(X, na.s){
n = dim(X)[1]
p = dim(X)[2]
gmv = matrix(0, n, p)
gmv=apply(gmv,2,function(x) rbinom(n, 1, na.s))
X[gmv==1]=NA
list("X"=X)
}

############ Generación de muestras con missing ######################
## De acuerdo al tamaño de mu, omega y alpha se generan muestras   ##
## skew normal bivariadas y datos faltantes aleatorios MCAR.       ##
#####################################################################

muestras.sim = function(n1,n2,mu,omega1,omega2,alpha,na.s){
require(sn)
G1 = rmsn(n1, mu, omega1, alpha)
```

```
G2 = rmsn(n2, mu, omega2, alpha)
G=c(rep(1,n1),rep(2,n2))
G1 = gen.mis(G1, na.s)$X
G2 = gen.mis(G2, na.s)$X
muestra = cbind(G,rbind(G1,G2))
list("muestra"=muestra)
}
```

A2 Simulación de la prueba de Levene

```
######################## Estadística de levene ########################
## Se necesita que la base esté ordenada de acuerdo a las muestras ##
## con una primera columna conteniendo las etiquetas de los grupos.##
## Entrega la estadítica W calculada usando el vector de medias Wc ##
## y la estadística W usando la mediana espacial Wm.               ##
## Nota: Esta función sirve para cualquier número de variables     ##
#####################################################################

mlevene = function(muestras){
require(SpatialNP)
n.grupos = length(table(muestras[,1]))
ni = nimis = Zic = Zim = Sic = Sim = double(0)
for(i in 1:n.grupos){
muestrai = muestras[(muestras[,1]==i),-1]
ni[i] = length(muestrai[,1])
v.medias = apply(muestrai,2,mean,na.rm =T)
v.medianas = apply(muestrai,2,spatial.median,na.action=na.omit)
Zc = sqrt(apply((muestrai-rep(1,ni[i])%x%t(v.medias))^2,1,sum))
Zm = sqrt(apply((muestrai-rep(1,ni[i])%x%t(v.medianas))^2,1,sum))
Zic[i] = mean(Zc,na.rm =T)
```

```
Zim[i] = mean(Zm,na.rm =T)
Sic[i] = sum((Zc-Zic[i])^2,na.rm =T)
Sim[i] = sum((Zm-Zim[i])^2,na.rm =T)
nimis[i] = length(Zc[!is.na(Zc)])
}
Zc.. = sum(Zic*nimis)/sum(nimis)
Zm.. = sum(Zim*nimis)/sum(nimis)
Wc = (sum(nimis)-n.grupos)*sum(nimis*(Zic-Zc..)^2)/((n.grupos-1)*sum(Sic))
Wm = (sum(nimis)-n.grupos)*sum(nimis*(Zim-Zm..)^2)/((n.grupos-1)*sum(Sim))
pvalorc = 1-pf(Wc,(n.grupos-1),(sum(nimis)-n.grupos))
pvalorm = 1-pf(Wm,(n.grupos-1),(sum(nimis)-n.grupos))
list("Wc"=Wc, "Wm"=Wm, "pvalorc"=pvalorc, "pvalorm"=pvalorm)
}

#######################################################################
################## Aplicación de la prueba de Levene ##################
## Para comparar el nivel de significancia real y el simulado bajo ##
## la hipótesis nula. 2 grupos.                                      ##
#######################################################################

sim = function(p,omega1,omega2,alpha,n1,n2,nitera,na.s){
sigc = sigm = sigWc = sigWm = 0
res1 = res2 = matrix(0,ncol=3,nrow=5)
cuant = c(.01,.05,.1)
for(k in 1:3){
for(j in 1:5){
for(i in 1:nitera){
muestra = muestras.sim(n1[j],n2[j],mu,omega1,omega2,alpha,na.s)$muestra
sigm[i] = ifelse(mlevene(muestra)$pvalorm <= cuant[k],1,0)
sigc[i] = ifelse(mlevene(muestra)$pvalorc <= cuant[k],1,0)
}
```

```
sigWm[j] = mean(sigm,na.rm =T)
sigWc[j] = mean(sigc,na.rm =T)
}
res1[,k] = sigWc
res2[,k] = sigWm
}
list("resWc"=res1,"resWm"=res2)
}
```

A3 Simulación de la LRT

```
############################## Prueba LRT ##############################
## Se necesita que la base esté ordenada de acuerdo a las muestras ##
## con una primera columna conteniendo las etiquetas de los grupos.##
## Entrega la estadítica LRT calculada.                            ##
## Nota: Esta función sirve para cualquier número de variables     ##
#######################################################################
Test.LRT <- function(A){
require(norm)
k <- length(table(A[,1]))
L0 <- 0
L1 <- 1
for(i in 1:k){
pre = prelim.norm(A[A[,1]==i,2:ncol(A)])
sig = getparam.norm(pre,em.norm(pre))$sigma
Si <- (nrow(A[A[,1]==i,])-1)*sig
L1 <- L1* ( ( det(Si/(nrow(A[A[,1]==i,])-1)) )^((nrow(A[A[,1]==i,])-1)/2)
)
L0 <- L0+Si
```

```
}
p = ncol(A)-1
g = 1-(2*p^2+3*p-1)/(6*(p+1)*(k-1))*(sum(1/(table(A[,1])-1))-1/(nrow(A)-k))
l <- L1/( ( det(L0/(nrow(A)-k)) )^((nrow(A)-k)/2) )
est <- -2*log(l)
list("L"=est)
}

#########################################################################

######################## Aplicación de la LRT ############################
## Para comparar el nivel de significancia real y el simulado bajo ##
## la hipótesis nula. 2 grupos.                                        ##
#########################################################################

simL = function(p,omega1,omega2,alpha,n1,n2,nitera,na.s){
sig = sigL = 0
res1 = matrix(0,ncol=3,nrow=5)
cuant = c(.01,.05,.1)
for(k in 1:3){
for(j in 1:5){
for(i in 1:nitera){
muestra = muestras.sim(n1[j],n2[j],mu,omega1,omega2,alpha,na.s)$muestra
sig[i] = ifelse(Test.LRT(muestra)$L >
qchisq((1-cuant[k]),((1/2)*p*(p+1)*(2-1))),1,0)
}
sigL[j] = mean(sig,na.rm =T)
}
res1[,k] = sigL
}
list("resL"=res1)
}
```