Comparación entre análisis discriminante no métrico y regresión logística multinomial

Freddy Hernández Barajas

Director: Juan Carlos Correa Morales, Ph.D

Trabajo presentado como requisito para optar
al título de Magister en Estadística

Posgrado en Estadística

Universidad Nacional de Colombia - Sede Medellín

2007

Dedicatoria

A Dios por todo lo que me ha regalado.

Agradecimientos

Gracias a mis compañeros y profesores, en especial al profesor Juan Carlos Correa por la oportunidad de trabajar con él.

Resumen

En este trabajo se presenta un estudio de comparación entre las técnicas de clasificación análisis discriminante no métrico, regresión logística multinomial; adicionalmente, se considera la técnica tradicional análisis discriminante lineal ya que la función de clasificación de ésta sirve como punto de partida para análisis discriminante no métrico. Los escenarios bajo los cuales se llevó a cabo el estudio fueron los siguientes: tres grupos distribuidos normal bivariado con matrices de varianza y covarianza iguales y diferentes, siete grupos distribuidos normal con tres variables y matrices de varianza y covarianza diferentes, tres grupos distribuidos Logit-normal, Lognormal y $Sinh^{-1}$-normal con dos variables. Los desempeños de las técnicas fueron medidos con la tasa de clasificación errónea. Las técnicas de Regresión Logística Multinomial y Análisis Discriminante Lineal obtuvieron tasas de clasificación errónea muy similares en todo el estudio e inferiores que Análisis Discriminante no Métrico. Se construyeron funciones en el programa R para implementar el algoritmo propuesto por Choulakian y Almhana (2001) para técnica Análisis Discriminante no Métrico. Se llevó a cabo una aplicación de las tres técnicas utilizando una base de datos real sobre datos antropométricos de trabajadores colombianos. En esta aplicación se encontró que las mejores clasificaciones fueron obtenidas por Regresión Logística Multinomial y Análisis Discriminante Lineal al clasificar personas en tres grupos predeterminados por el índice de masa corporal. Finalmente, del estudio se recomienda la utilización de las técnicas regresión logística multinomial y análisis discriminante lineal para realizar clasificaciones en situaciones donde la distribución de los datos sea próxima a la distribución normal multivariada con variables explicativas cuantitativas.

Summary

In this paper we show the results of a comparison simulation study for three classification techniques: Multinomial Logistic Regression (MLR), No Metric Discriminant Analysis (NDA) and Linear Discriminant Analysis (LDA). The measure used to compare the performance of the three techniques was the Error Classification Rate (ECR). We found that MLR and LDA techniques have similar performance and that they are better that NDA when the population multivariate distribution is Normal or Logit-Normal. We illustrated the application of the three techniques using a Colombian workers anthropometric data base.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Uno de los campos importantes de la estadística es la clasificiación de nuevas observaciones en una de varias poblaciones dada una discriminación por medio de las características consideradas como relevantes en las observaciones. Fisher propuso una herramienta para lograr esta tarea llamada Análisis Discriminante Lineal (LDA) el cual divide el espacio muestral en subespacios mediante hiperplanos para separar lo mejor posible las poblaciones en cuestión. Efron (1975) mostró que la regresión logística tiene un mejor desempeño que el Análisis Discriminante Lineal cuando se violan los supuestos de multinormalidad y homocedasticidad. Raveh (1989) propuso el Análisis Discriminante no Métrico (NDA) el cual se basa en una regla de separación diferente al Análisis de Discriminante Lineal con resultados muy similares a la metodología de Fisher y cuya ventaja radica en que no requiere supuestos distribucionales. Adicionalmente, se cuenta con Regresión Logística Multinomial (MLR) que no requiere supuestos distribucionales y que permite realizar clasificaciones de observaciones en una de varias categorías. Hasta el momento sólo se ha publicado una comparación de Análisis Discriminante no Métrico contra Regresión Logística para el caso de dos poblaciones, Usuga (2006). En el presente documento se realiza una comparación entre Análisis Discriminante no Métrico y Regresión Logística multinomial para el caso de más de dos grupos y bajo diferentes escenarios de simulación.

En el capítulo 2 del presente trabajo se puede encontrar una descripción de las técnicas, sus reglas de clasificación y los trabajos anteriores sobre comparaciones entre ellas. En el capítulo 3 se pueden encontrar las características de los cuatro escenarios de simulación, sus resultados y las conclusiones a las cuales se llegó. En el capítulo 4 se encuentra una aplicación de a una base de datos reales llamada Acopla95 que contiene mediciones antropométricas de 2100 trabajadores colombianos en 1995. En la parte final del documento se puede encontrar el anexo que corresponde a la documentación de las funciones creadas en el programa R para implementar la técnica análisis discriminante lineal.

Capítulo 2

Marco teórico

2.1. El problema de clasificación

Supóngase que se tienen G grupos o clases de elementos. Adicionalmente, supóngase que se tiene una nueva observación que posee p características y cuyo interés es *asignarla* o *clasificarla* en uno de los G grupos. El problema de clasificación es encontrar la relación entre el vector de características de la nueva observación y la pertenencia a uno de los grupos. Para asignar la nueva observación a uno de los grupos se debe construir una Función Discriminante que sirva como regla para determinar a qué grupo pertenece una nueva observación basados en unas características relevantes. La clasificación tiene utilidad en diversas áreas de la ciencia, la tecnología y los negocios, entre las cuales se pueden destacar el diagnóstico de enfermedades, la asignación de créditos, identificación de criminales mediante el uso de huellas digitales, identificación de compradores potenciales de un producto entre otras.

2.2. Análisis Discriminante Lineal

Supongase que se tienen tienen G grupos cada uno con $n_1, n_2, \ldots, n_g$ observaciones p variadas, además cada grupo posee vector de medias $\bar{\mathbf{x}}_\mathbf{i}$, matriz de covarianzas $\mathbf{S_i}$ para $i = 1, 2, \ldots, G$

y vector de medias general $\bar{\mathbf{x}}$.

La técnica LDA se basa en los supuestos de normalidad multivariada y de igualdad entre las matrices de varianzas y covarianzas de cada uno de los grupos. El objetivo de LDA es encontrar un vector $\mathbf{a}'$ de p variables de tal manera que se maximice λ definido por:

$$\lambda = \frac{a'Ba}{a'Wa}, \tag{2.1}$$

donde B y W son dos matrices que representan cada una la variabilidad entre los grupos y la variabilidad dentro de los grupos respectivamente, dadas por;

$$B = \sum_{i=1}^{g} n_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})', \tag{2.2}$$

$$W = \sum_{i=1}^{g} (n_i - 1)S_i. \tag{2.3}$$

Los valores de a que maximizan λ se pueden encontrar por medio de los vectores propios $\hat{e}_1, \hat{e}_2, \cdots, \hat{e}_s$ asociados con los valores propios positivos[1] $\hat{\lambda}_1, \hat{\lambda}_2, \cdots, \hat{\lambda}_s$ de $W^{-1}B$.

De esta manera si $\hat{a} = \hat{e_1}$ entonces $\hat{a}$ se le denomina primer discriminante lineal (LD_1), si $\hat{a} = \hat{e_2}$ entonces $\hat{a}$ se le denomina segundo discriminante lineal (LD_2) y así hasta $\hat{a} = \hat{e_s}$, en cuyo caso $\hat{a}$ se denomina s-ésimo discriminante lineal (LD_s).

La regla para clasificar una nueva observación x es asignar a la población o grupo i si se cumple que: $\sum_{j=1}^{r} \left[\hat{a}_j'(x - \bar{x}_i)\right]^2$ es mínimo[2].

[1] Donde $s = min\{p, g-1\}$

[2] Donde $r \leq s$.

2.3. Análisis Discriminante no Métrico

Raveh (1989) propuso un procedimiento de discriminante no métrico (NDA) basado en la maximización de un índice de separación entre dos grupos. Guttman (1988) generalizó el índice propuesto por Raveh para múltiples grupos y lo llamó *disco* (discrimination coefficient).

El NDA tiene como objetivo determinar una función discriminante de tal manera que se maximice el cociente entre la variabilidad entre grupos con la variabilidad dentro de grupos.

Supóngase que se tienen G grupos p-variados $X(1), X(2), \ldots, X(G)$ cada uno con $n_1, n_2, \ldots, n_G$ observaciones[3], el elemento denotado por $X_i(j)$ corresponde a la observación i del grupo j. Al conjunto formado por las observaciones de todos los grupos se denomina *Conjunto de Entrenamiento*.

Sea η un vector p-dimensional, y $z_i(g)$ la variable aleatoria definida así: $z_i(g) = \eta' X_i(g)$ que representa el *Score* de la i-ésima observación del grupo g-ésimo dado por η. El índice *disco* entre G grupos está dado por:

$$disco = \frac{\sum_{g=1}^{G}\sum_{h=1}^{G} n_g n_h \left|\bar{z}(g) - \bar{z}(h)\right|}{\sum_{g=1}^{G}\sum_{h=1}^{G}\sum_{i=1}^{n_g}\sum_{j=1}^{n_h} \left|z_i(g) - z_j(h)\right|}, \tag{2.4}$$

donde $\bar{z}(g)$ representa el promedio de los scores para las observaciones del grupo g.

El numerador en (2.4) se puede escribir como

$$\sum_{g=1}^{G}\sum_{h=1}^{G}\left|\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\left[z_i(g) - z_j(h)\right]\right|,$$

[3]En total n observaciones, $n = \sum_{i=1}^{G} n_i$

donde el valor absoluto corresponde a una medida de la separación entre los grupos h y g.

El denominador en (2.4) contiene el elemento

$$\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}|z_i(g)-z_j(h)|,$$

que representa la variación total entre los grupos h y g, cuantificado mediante desviaciones absolutas.

En virtud de la siguiente desigualdad

$$\sum_{g=1}^{G}\sum_{h=1}^{G}\left|\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}[z_i(g)-z_j(h)]\right| \leq \sum_{g=1}^{G}\sum_{h=1}^{G}\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}|z_i(g)-z_j(h)|,$$

se obtiene que la ecuación (2.4) satisface $0 \leq disco \leq 1$.

El coeficiente *disco* es igual a cero si y solamente si todos los grupos tienen la misma media, y es igual a 1 si no existe superposición entre los scores de ningún par de grupos.

En virtud de $z_i(g) = \eta' X_i(g)$, la ecuación (2.4) puede escribirse de la siguiente manera:

$$disco = \frac{\sum_{g=1}^{G}\sum_{h=1}^{G} n_g n_h \left|\eta'\left[\bar{X}(g)-\bar{X}(h)\right]\right|}{\sum_{g=1}^{G}\sum_{h=1}^{G}\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\left|\eta'\left[X_i(g)-X_j(h)\right]\right|}, \tag{2.5}$$

El NDA propuesto por Raveh (1989) consiste en la búsqueda de η tal que maximice el coeficiente *disco* dado en la ecuación (2.5).

Disco en (2.5) puede ser escrito en forma matricial, para esto se definen dos matrices B_{gh} y $V_{gh}(i,j)$ ambas de orden $p \times p$ de la siguiente manera:

$$B_{gh} = \left[\bar{X}(g) - \bar{X}(h)\right]\left[\bar{X}(g) - \bar{X}(h)\right]', \tag{2.6}$$

$$V_{gh}(i,j) = \left[X_i(g) - X_j(h)\right]\left[X_i(g) - X_j(h)\right]'. \tag{2.7}$$

Usando la identidad $|a| = a^2/|a| \quad$ con $\quad a \neq 0$ se tiene que:

$$\left|\eta'\left[\bar{X}(g) - \bar{X}(h)\right]\right| = \frac{\eta' B_{gh}(i,j)\eta}{\sqrt{\eta' B_{gh}(i,j)\eta}} \quad \text{si} \quad B_{gh} \neq 0.$$

Además,

$$\left|\eta'\left[X_i(g) - X_i(h)\right]\right| = \frac{\eta' V_{gh}(i,j)\eta}{\sqrt{\eta' V_{gh}(i,j)\eta}} \quad \text{si} \quad V_{gh}(i,j) \neq 0.$$

De esta manera *disco* en (2.5) puede ser representado en notación matricial como una función del vector η así:

$$disco(\eta) = \frac{\eta' B(\eta)\eta}{\eta' V(\eta)\eta}, \tag{2.8}$$

donde $B(\eta)$ y $V(\eta)$ son matrices simétricas de orden $p \times p$ que sólo dependen del parámetro η de la siguiente manera:

$$B(\eta) = \sum_{g=1}^{G}\sum_{h=1}^{G} \frac{n_g n_h B_{gh}}{\sqrt{\eta' B_{gh}\eta}}$$

y

$$V(\eta) = \sum_{g=1}^{G}\sum_{h=1}^{G}\sum_{i=1}^{n_g}\sum_{j=1}^{n_h} \frac{V_{gh}(i,j)}{\sqrt{\eta' V_{gh}(i,j)\eta}}.$$

Para maximizar *disco* en (2.5) con respecto a η, Choulakian y Almhana (2001) propusieron el siguiente algoritmo:

1. Se comienza con $\eta_0 = \theta^*$ siendo θ^* el eigenvector[4] (LD_1) .

2. Calcular $\eta_{k+1} = \eta_k \left[1 - 2 \times disco(\eta_k)\right] + 2 \times V(\eta_k)^{-1} B(\eta_k)\eta_k$, para $k = 0, 1, 2, \ldots$

3. El proceso se detiene cuando $|disco(\eta_{k+1}) - disco(\eta_k)| \leq \epsilon$ donde ϵ es un valor real positivo definido con anterioridad, por ejemplo, $\epsilon = 10^{-5}$.

4. El valor óptimo de la función discriminante η se obtiene haciendo $\eta = \eta_k$.

La convergencia del algoritmo anterior fue demostrada por Choulakian y Almhana (2001).

Luego de encontrar el valor óptimo de η éste puede ser utilizado para clasificar nuevas observaciones en uno de los G grupos. Para realizar la clasificación se determinan $G - 1$ puntos de corte (CP) de la siguiente manera: se toman los n *disco scores* $z_i(g)$ con $i = 1, 2, \ldots, n_g$ y con $g = 1, 2, \ldots, G$; sin pérdida de generalidad, se puede suponer que los primeros n_1 scores son menores que los segundos n_2 y así sucesivamente. El punto de corte CP que separa los grupos 1 y 2 es igual al percentil $100(n_1/n)\,\%$ de los n scores ordenados, el segundo CP que separa los grupos 2 y 3 es igual al percentil $100((n_1 + n_2)/n)\,\%$ de los n scores ordenados, de manera similar se obtienen los $G - 3$ CP restantes.

Si los n scores de los G grupos no se superponen la anterior regla indica que el punto de corte CP entre el g-ésimo grupo y el $g + 1$-ésimo grupo sería:

$$\frac{max_{i=1,\ldots,n_g} z_i(g) + min_{i=1,\ldots,n_{g+1}} z(g+1)}{2},$$

lo cual asegura que la tasa de clasificaciones incorrectas de las observaciones en el conjunto de entrenamiento sea cero.

[4]Obtenido mediante Análisis Discriminante Lineal

2.4. Regresión Logística Multinomial

La técnica MLR consiste en la estimación de la probabilidad de que una observación $\mathbf{x}$ pertenezca a cada uno de los grupos, dados valores de las p variables que conforman la observación.

El modelo compara $G-1$ categorías contra una categoría de referencia. Dadas n observaciones $(y_i, \mathbf{x}_i)$ donde $\mathbf{x}_i$ es un vector con p variables y y_i es una v.a. independiente Multinomial con valores $1, 2, \ldots, G$, la cual indica el grupo al cual pertenece cada observación, la probabilidad condicional de pertenencia de $\mathbf{x}_i$ a cada grupo está dada por:

$$P(y = j|\mathbf{x}_i) = \frac{e^{\alpha_{1j}+\beta'_{1j}\mathbf{x}_i}}{1 + \sum_{k=2}^{G} e^{\alpha_{1j}+\beta'_{1k}\mathbf{x}_i}} \tag{2.9}$$

donde $\alpha_{11} = 0$ y $\beta_{11} = \mathbf{0}$.

La regla de clasificación consiste en que dada una nueva observación con p variables se calcula la probabilidad de que ésta observación pertenezca a cada uno de los G grupos y luego se asigna al grupo que presentó la mayor probabilidad.

La ventaja de MLR es que no requiere supuestos distribucionales como LDA y por lo tanto se puede aplicar a distribuciones multivariadas con variables cuantitativas y/o cualitativas.

2.5. Desarrollo histórico de las técnicas de clasificación y estudios de comparación.

2.5.1. Análisis Discriminante Lineal

LDA es una de las técnicas más populares para clasificar cuando se cumplen los supuestos de normalidad multivariada con igualdad de matrices de varianzas y covarianzas entre los grupos. Fisher (1936) propuso la metodología para discriminar en el caso de dos grupos. Welch (1939) mostró la optimalidad de LDA bajo condiciones de normalidad multivariada. Rao (1948) propuso el análisis discriminante canónico que es una generalización de análisis discriminante lineal para el caso de varios grupos.

Clunies y Riffenburgh (1960) y Anderson y Bahadur (1962) encontraron una función discriminante para el caso donde no se cumple la igualdad de matrices de varianzas y covarianzas. Chernoff (1972) sugirió medidas de qué tan bien se discrimina para el caso de dos grupos.

Recientemente Hawking (1997) y Croux (2001) presentaron un procedimiento llamado *high breakdown* para remover outliers que pueden afectar las estimaciones de los vectores de medias y las matrices de varianzas y covarianzas en LDA y que por lo tanto afectan las funciones de clasificación. Cheng et al. (2002) proponen dos formas alternativas de realizar las estimaciones del vector de medias y de la matriz de varianzas y covarianzas cuando faltan datos en las observaciones. Las propuestas consisten básicamente en combinar el algoritmo ER (expectation-robust) propuesto por Little and Smith (1987) con los estimadores *high breakdown*.

2.5.2. Regresión Logística Multinomial

Cornfield (1962), Cox (1966) y Day y Kerridge (1967) sugirieron el modelo de regresión logística para una variable respuesta binaria, Anderson (1972) propuso el modelo de regresión logística multinomial o policótomo.

Pregibon (1981) propuso un método para detectar posibles observaciones atípicas en la muestra de entrenamiento que se usa para encontrar los estimadores de máxima verosimilitud para un modelo de regresión logístico. Trevor y Ferry (1991) publicaron un artículo donde muestran un nuevo modelo de regresión logística robusto que mostró tener mejor desempeño en un estudio de simulación y en un estudio dental con ratas. Carroll and Pederson (1993) mostraron que existen otras dos formas de estimaciones resistentes y robustas que tienen sesgos iguales o menores en las estimaciones cuando se tienen observaciones atípicas en el conjunto de obsevaciones.

2.5.3. Análisis Discriminante no Métrico

Raveh (1983) y (1989) propuso el análisis discriminante no métrico que toma el primer discriminante lineal (LDA) y lo utiliza como punto de partida para construir su propio discriminante con el fin de realizar las clasificaciones, esto basado en un índice de separación llamado *DISCO*. Guttman (1988) generalizó el índice de separación propuesto por Raveh y lo generalizó para el caso de múltiples grupos. Por último, Choulakian y Almhana (2001) proponen un algoritmo para calcular los parámetros de la función discriminante (η) por medio de NDA y que asegura la maximización del índice *DISCO*.

2.5.4. Comparaciones entre LDA y LR

Efron (1975) comparó las dos técnicas para el caso de dos grupos con igual matriz de varianzas y covarianzas, se encontró que la eficiencia relativa asintótica de LR con respecto a LDA está entre un medio y dos tercios. Crawley (1979) comparó LR con LDA para muestras pequeñas con dos grupos y encontró que para $\Sigma_1 = \Sigma_2$ LDA tiene un mejor desempeño que LR a la hora de clasificar, para el caso de $\Sigma_1 \neq \Sigma_2$ LR tuvo ligeramente un mejor desempeño y para el caso de dos poblaciones distribuidas no normal LR tuvo un desempeño muy superior a LDA. Harrell (1985) realizó una comparación entre las técnicas y consideró $\Sigma_1 = \Sigma_2$,

tamaños de muestra de 50 y 130 y distancias de Mahalanobis entre los vectores de medias de las dos poblaciones con valores que variaron desde 0.94 hasta 4.68, en el estudio se encontró que el desempeño de LDA fue mejor que LR pero que las diferencias no eran significativas.

Cherkaoui y Cleroux (1991) compararon para el caso de dos grupos LDA, LR, análisis discriminante cuadrático y tres métodos no paramétricos. En las conclusiones los autores plantean las condiciones en que cada una de las técnicas tienen un desempeño superior. Castrillón (1998) comparó LDA y LR para el caso de dos grupos y encontró que LR clasifica mejor cuando los supuestos de normalidad y matrices de varianzas y covarianzas se violan.

2.5.5. Comparaciones entre LDA y MLR

Shelley y Donner (1987) llevaron a cabo un estudio para medir la eficiencia relativa asintótica (ARE) de regresión logística multinomial comparada con análisis discriminante para el caso de poblaciones distribuidas normal multivariada con igual matriz de varianzas y covarianzas. Los casos que estudiaron fueron con dos, tres y cuatro grupos. Además, tuvieron en cuenta varias separaciones entre los grupos y estudiaron el efecto de la colinearidad entre los vectores del modelo de regresión logística sobre la ARE. Los autores encontraron que para el caso de vectores de clasificación colineales la ARE cambió de 50 % a 65 % para dos grupos y de 35 % a 95 % para el caso de cuatro grupos cuando la distancia entre el grupo de referencia y los demás estuvo en 3.0 a 3.5. Para el caso de vectores ortogonales se encontró que ARE decae rápidamente a medida que aparecen más grupos en el proceso de clasificación.

Capítulo 3

Estudio de simulación

En el presente trabajo se realizó la comparación entre Análisis Discriminante Lineal (LDA), Análisis Discriminante no Métrico (NDA) y Regresión Logística Multinomial (MLR) para el problema de clasificar observaciones nuevas a uno de tres o más poblaciones distribuidas normal multivariada y para poblaciones distribuidas Log-normal, $Sinh^{-1}$-normal y Logit-normal. Para el caso de poblaciones distribuidas normal multivariada, se trabajó con matrices de varianza y covarianza iguales y diferentes, además se consideraron varias combinaciones de tamaños muestrales.

3.1. Procedimiento de comparación

Las tres técnicas de clasificación fueron comparadas llevando a cabo el siguiente conjunto de pasos:

1. Se simula[1] una muestra por cada una de las G poblaciones con las cuales se forma al conjunto de entrenamiento. Se utilizaron diferentes poblaciones y diferentes parámetros. Adicionalmente, se consideraron varias combinaciones para los tamaños de muestra $n_1, n_2, \ldots, n_G$.

[1]Las simulaciones se llevarán a cabo usando el programa R.

2. Con el conjunto de entrenamiento se construyeron las funciones de discriminación para NDA y MLR y LDA.

3. Se generaron nuevas muestras como en el paso 1 y con ellas se formó un nuevo conjunto llamado de validación. A todas las observaciones de este conjunto se les clasificó con las funciones obtenidas en el paso 2.

4. Se calculó la tasa de clasificación errónea para el conjunto de validación del paso anterior.

5. Los pasos 3 y 4 se repetieron mil veces y se calculó la tasa promedio de clasificación errónea para cada uno de los procedimientos de clasificación.

3.2. Evaluación de las funciones de clasificación

La evaluación de los desempeños de las técnicas de clasificación se llevó a cabo utilizando la Tasa de Clasificación Errónea (**TCE**) que se define de la siguiente manera:

$$\mathbf{TCE} = \frac{\mathbf{NCE}}{\mathbf{NOBS}}$$

donde **NCE** corresponde al número de clasificaciones erradas por la técnica en el conjunto de validación y **NOBS** corresponde al número de observaciones en el conjunto de validación.

3.3. Escenarios de simulación

A continuación se muestran los diferentes escenarios en los cuales se realizaron las comparaciones entre las metodologías de clasificación.

- Escenario 1. Tres grupos normales bivariados con matrices de covarianza iguales.
  Aquí se consideraron tres grupos normales bivariados[2], la media del grupo 1 estuvo

[2]Los grupos fueron llamados grupo 1, grupo 2 y grupo 3.

ubicada siempre en el origen del plano cartesiano de $\mathbb{R}^2$ mientras que las medias de los otros dos grupos cambiaron de posición, pero siempre sobre los ejes cartesianos. Los tres grupos tuvieron la misma matriz de varianzas y covarianzas, se consideraron $\sigma_1 = \sigma_2 = 1$ con varios valores de correlación ρ de 0.1, 0.3, 0.5, 0.7 y 0.9. A continuación se muestran los diferentes casos para los vectores de medias.

- Situación A: $\mu_1 = (0, 0)$, $\mu_2 = (1, 0)$, $\mu_3 = (0, 1)$.
- Situación B: $\mu_1 = (0, 0)$, $\mu_2 = (1, 0)$, $\mu_3 = (0, 2)$.
- Situación C: $\mu_1 = (0, 0)$, $\mu_2 = (1, 0)$, $\mu_3 = (0, 3)$.
- Situación D: $\mu_1 = (0, 0)$, $\mu_2 = (2, 0)$, $\mu_3 = (0, 2)$.
- Situación E: $\mu_1 = (0, 0)$, $\mu_2 = (2, 0)$, $\mu_3 = (0, 3)$.

Otro aspecto importante que se tuvo en cuenta fue el tamaño de los grupos[3] que se denotan por n_1, n_2 y n_3 para los grupos 1, 2 y 3 respectivamente. Se utilizaron las siguientes combinaciones de tamaños muestrales.

- $n_1 = 20$, $n_2 = 20$, $n_3 = 20$.
- $n_1 = 50$, $n_2 = 50$, $n_3 = 50$.
- $n_1 = 100$, $n_2 = 100$, $n_3 = 100$.
- $n_1 = 20$, $n_2 = 50$, $n_3 = 100$.

Las comparaciones se llevaron a cabo utilizando las diferentes combinaciones anteriormente presentadas de ρ, medias y tamaños muestrales.

- Escenario 2. Tres grupos normales bivariados, matrices de covarianza diferentes.
  Para el caso de matrices de varianza y covarianza diferentes se trabajó de la siguiente manera: la matriz de covarianzas del grupo 1 se caracterizó por $\sigma_1 = \sigma_2 = 1$ con varios

[3]La cantidad de observaciones presentes del grupo 1 en los conjuntos de entrenamiento y validación fueron iguales, esto se aplicará para los demás grupos también.

valores de correlación ρ de 0.1, 0.3, 0.5, 0.7 y 0.9. La matriz de covarianzas de los grupos 1, 2 y 3 se denotan por Σ_1, Σ_2 y Σ_3. Los casos considerados fueron los siguientes:

- $\Sigma_2 = 2\Sigma_1$ y $\Sigma_3 = 2\Sigma_1$.
- $\Sigma_2 = 2\Sigma_1$ y $\Sigma_3 = 3\Sigma_1$.
- $\Sigma_2 = 3\Sigma_1$ y $\Sigma_3 = 3\Sigma_1$.
- $\Sigma_2 = 2\Sigma_1$ y $\Sigma_3 = 4\Sigma_1$.

Los vectores de medias fueron los siguientes:

- Situación A: $\mu_1 = (0, 0)$, $\mu_2 = (1, 0)$, $\mu_3 = (0, 1)$.
- Situación B: $\mu_1 = (0, 0)$, $\mu_2 = (1, 0)$, $\mu_3 = (0, 2)$.
- Situación C: $\mu_1 = (0, 0)$, $\mu_2 = (1, 0)$, $\mu_3 = (0, 3)$.
- Situación D: $\mu_1 = (0, 0)$, $\mu_2 = (2, 0)$, $\mu_3 = (0, 2)$.
- Situación E: $\mu_1 = (0, 0)$, $\mu_2 = (2, 0)$, $\mu_3 = (0, 3)$.

Los tamaños muestrales fueron los siguientes:

- $n_1 = 20$, $n_2 = 20$, $n_3 = 20$.
- $n_1 = 50$, $n_2 = 50$, $n_3 = 50$.

- Escenario 3. Siete grupos distribuidos normal multivariada.
  En este caso se consideraron siete grupos pertenecientes a diferentes poblaciones[4] distribuidas normal con tres variables. Los vectores de medias de cada una de las poblaciones se ubicaron sobre puntos positivos y negativos de cada uno de los tres ejes (en total seis grupos) y un último grupo sobre el origen para completar los siete grupos. Se consideraron tres situaciones (A, B y C) de alejamiento entre los grupos, los vectores de medias asociadas a cada una de las situaciones fueron los siguientes:

[4]Denotadas por $1, 2, \ldots, 7$

- Situación A: $\mu_1 = (0,0,0)$, $\mu_2 = (0,0,2)$, $\mu_3 = (0,0,-2)$, $\mu_4 = (0,2,0)$, $\mu_5 = (0,-2,0)$, $\mu_6 = (2,0,0)$, $\mu_7 = (-2,0,0)$.
- Situación B: $\mu_1 = (0,0,0)$, $\mu_2 = (0,0,2)$, $\mu_3 = (0,0,-2)$, $\mu_4 = (0,3,0)$, $\mu_5 = (0,-3,0)$, $\mu_6 = (3,0,0)$, $\mu_7 = (-3,0,0)$.
- Situación C: $\mu_1 = (0,0,0)$, $\mu_2 = (0,0,2)$, $\mu_3 = (0,0,-2)$, $\mu_4 = (0,4,0)$, $\mu_5 = (0,-4,0)$, $\mu_6 = (4,0,0)$, $\mu_7 = (-4,0,0)$.

El tamaño de cada una de las muestras tomadas de cada población fue de diez, veinte y cincuenta. La matriz de covarianzas de los grupos $1, 2, \ldots, 7$ se denotan por Σ_1, Σ_2,..., Σ_7. Se consideraron matrices de covarianza diferentes y la matriz de covarianzas del grupo 1 (grupo base) se caracterizó porque las varianzas de las tres variables fueron $\sigma_1 = \sigma_2 = \sigma_3 = 1$. Se consideraron valores de correlación entre pares de variables ρ de 0.1, 0.3, 0.5, 0.7 y 0.9. Las estructuras generales de las matrices de varianzas y covarianzas fueron las siguientes dos:

- **Estructura 1**: $\Sigma_1 = \Sigma_2 = \Sigma_3$, $\Sigma_4 = \Sigma_5 = 2\Sigma_1$, $\Sigma_6 = \Sigma_7 = 4\Sigma_1$.
- **Estructura 2**: $\Sigma_1 = \Sigma_2 = \Sigma_3$, $\Sigma_4 = \Sigma_5 = 4\Sigma_1$, $\Sigma_6 = \Sigma_7 = 8\Sigma_1$.

- Escenario 4. Tres grupos con funciones de densidad especiales.
  En este escenario se llevó a cabo la comparación de las técnicas de clasificación usando tres funciones de densidad especiales como lo son: Lognormal, $Sinh^{-1}$-normal y Logit-normal. Para crear las muestras de entrenamiento y las de validación se generaron muestras de observaciones normales bivariadas con parámetros conocidos y luego se aplicó el sistema de transformación sugerido por Johnson (1987) para obtener observaciones distribuidas con las funciones de densidad especiales ya mencionadas.

  La estructura usada para este escenario fue la siguiente: se consideraron tres grupos normales bivariados[5], la media del grupo 1 estuvo ubicada siempre en el origen del plano cartesiano de $\mathbb{R}^2$ mientras que las medias de los otros dos grupos cambiaron de

[5]Los grupos fueron llamados grupo 1, grupo 2 y grupo 3.

posición, pero siempre sobre los ejes cartesianos. Los tres grupos tuvieron la misma matriz de varianzas y covarianzas, se consideraron $\sigma_1 = \sigma_2 = 1$ con varios valores de correlación ρ de 0.1, 0.3, 0.5, 0.7 y 0.9. A continuación se muestran los diferentes casos para los vectores de medias.

- Situación A: $\mu_1 = (0, 0)$, $\mu_2 = (1, 0)$, $\mu_3 = (0, 1)$.
- Situación B: $\mu_1 = (0, 0)$, $\mu_2 = (1, 0)$, $\mu_3 = (0, 2)$.
- Situación C: $\mu_1 = (0, 0)$, $\mu_2 = (1, 0)$, $\mu_3 = (0, 3)$.
- Situación D: $\mu_1 = (0, 0)$, $\mu_2 = (2, 0)$, $\mu_3 = (0, 2)$.
- Situación E: $\mu_1 = (0, 0)$, $\mu_2 = (2, 0)$, $\mu_3 = (0, 3)$.

Otro aspecto importante que se tuvo en cuenta fueron los tamaños de los grupos[6] que se denotan por n_1, n_2 y n_3 para los grupos 1, 2 y 3 respectivamente. Se utilizaron las siguientes combinaciones de tamaños muestrales.

- $n_1 = 20$, $n_2 = 20$, $n_3 = 20$.
- $n_1 = 50$, $n_2 = 50$, $n_3 = 50$.
- $n_1 = 100$, $n_2 = 100$, $n_3 = 100$.

Las combinaciones (5 situaciones, 5 correlaciones y 3 tamaños) de los parámetros de simulación arrojan 75 casos diferentes; para cada uno de los casos se consideraron adicionalmente las tres transformaciones para lograr que las muestras tuvieran las distribuciones definidas Lognormal, $Sinh^{-1}$-normal y Logit-normal.

[6]La cantidad de observaciones presentes del grupo 1 en los conjuntos de entrenamiento y validación fueron iguales.

3.4. Resultados obtenidos

3.4.1. Escenario 1.

En las figuras 3.1 a 3.4 se pueden observar los resultados para las diferentes situaciones de alejamiento (A, B, C, D y E) y todos los tamaños muestrales. Al analizar las figuras se encuentran patrones claros comunes en cada una de ellas, los resultados más sobresalientes son los siguientes:

- A medida que aumenta el alejamiento entre las poblaciones, las tasas de clasificación errónea disminuyen.
- A medida que aumenta el coeficiente de correlación en las poblaciones en estudio, las tasas de clasificación errónea disminuyen considerablemente.
- Cuando se aumenta el tamaño de muestra de 20 a 50 y luego a 100, se obtiene una disminución en las tasas de clasificación errónea.
- Cuando se consideran tamaños muestrales diferentes para cada población ($n_1 = 20$, $n_2 = 50$ y $n_2 = 100$) las tasas de clasificación errónea disminuyen considerablemente con respecto al caso de tamaños muestrales iguales ($n_1 = n_2 = n_2$) en cada población.
- La línea asociada con la tasa de clasificación de la técnica NDA siempre quedó por encima de las otras dos técnicas LDA y MLR.
- Los desempeños de LDA y MLR fueron muy similares, las líneas asociadas a estas dos técnicas se encuentran muy cerca.

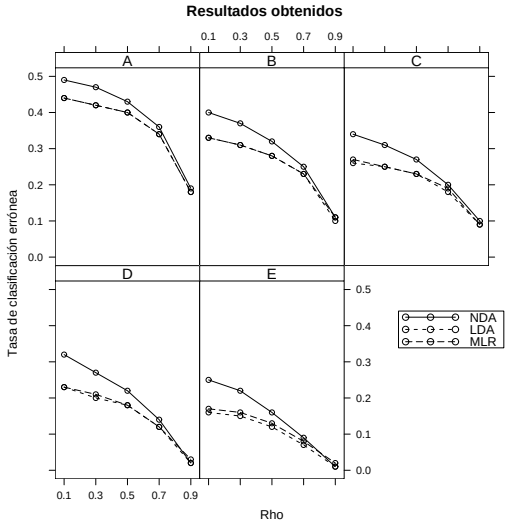

Figura 3.1: Escenario 1. Tamaños muestrales $n_1 = n_2 = n_3 = 20$

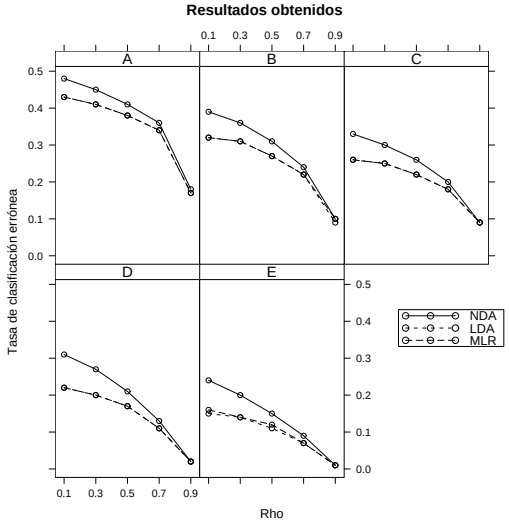

Figura 3.2: Escenario 1. Tamaños muestrales $n_1 = n_2 = n_3 = 50$

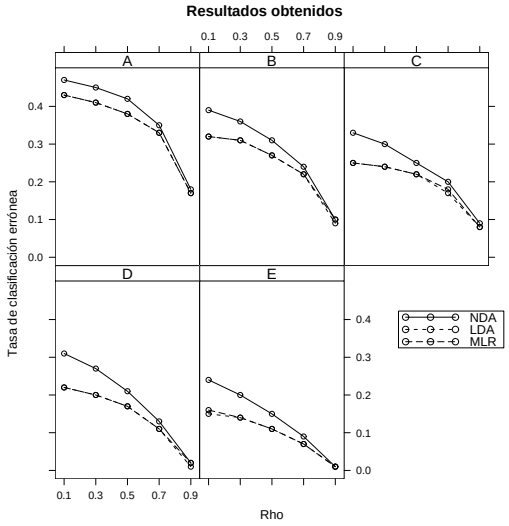

Figura 3.3: Escenario 1. Tamaños muestrales $n_1 = n_2 = n_3 = 100$

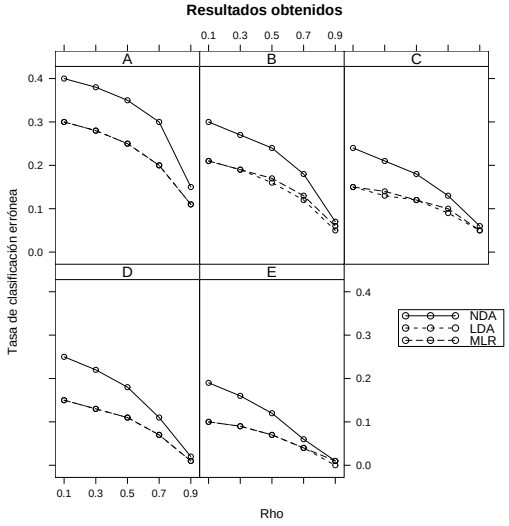

Figura 3.4: Escenario 1. Tamaños muestrales $n_1 = 20, n_2 = 50, n_3 = 100$

3.4.2. Escenario 2.

En esta parte de la presente sección se muestran los resultados obtenidos cuando se consideran diferentes estructuras para las matrices de varianzas y covarianzas de los grupos de comparación y diferentes tamaños muestrales.

Al observar las figuras 3.5 a 3.12 se notan nuevamente patrones claros de los cuales se pueden destacar los siguiente aspectos importantes:

- A medida que aumenta el alejamiento entre las poblaciones las tasas de clasificación errónea disminuyen.
- A medida que aumenta el coeficiente de correlación las tasas de clasificación errónea disminuyen considerablemente.
- Al considerar diferentes matrices de varianzas y covarianzas para los grupos, los desempeños de las técnicas se vieron afectados, el aumento promedio en la tasa de clasificación incorrecta fue de 6 %.
- Se observó también que cuando las matrices de varianzas y covarianzas eran diferentes con tamaños muestrales de 50 las tasas de clasificación incorrectas eran práctimente las mismas que cuando se tiene matrices de varianzas y covarianzas iguales, es decir, los desempeños de las técnicas fueron similares cuando los tamaños muestrales fueron de 50 sin importar si se cumplia el supuesto de igualdad de matriz de varianzas y covarianzas.
- La línea asociada con la tasa de clasificación de la técnica NDA siempre quedó por encima de las otras dos técnicas LDA y MLR.
- Los desempeños de LDA y MLR fueron muy similares, las líneas asociadas a estas dos técnicas se encuentran muy cerca.

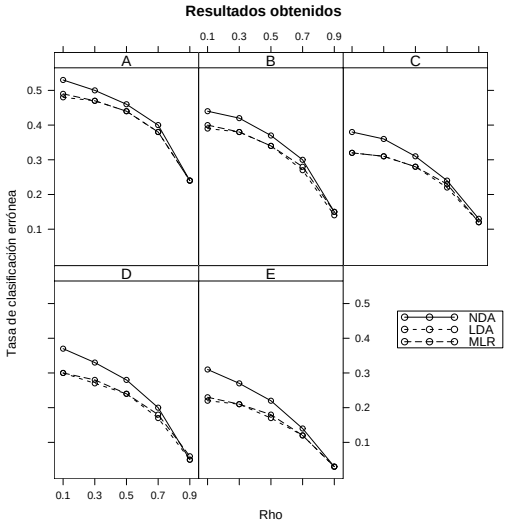

Figura 3.5: Tamaños muestrales $n_1 = n_2 = n_3 = 20$ y $\Sigma_2 = \Sigma_3 = 2\Sigma_1$

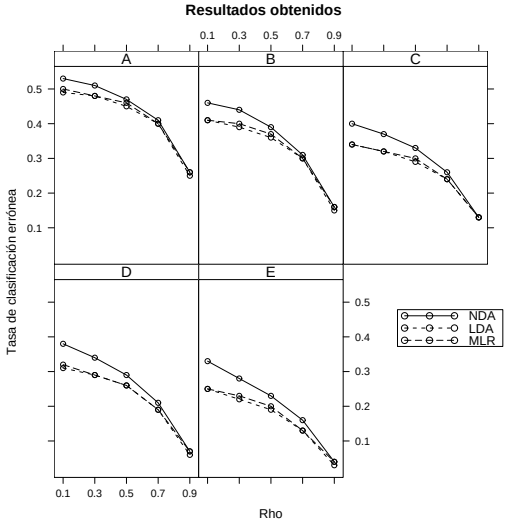

Figura 3.6: Tamaños muestrales $n_1 = n_2 = n_3 = 20$, $\Sigma_2 = 2\Sigma_1$ y $\Sigma_3 = 3\Sigma_1$

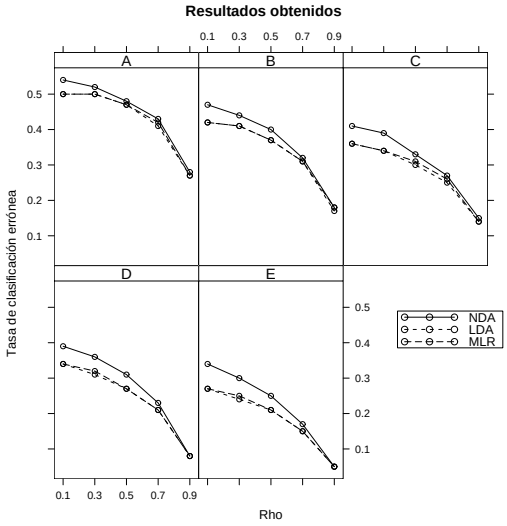

Figura 3.7: Tamaños muestrales $n_1 = n_2 = n_3 = 20$, $\Sigma_2 = 3\Sigma_1$ y $\Sigma_3 = 3\Sigma_1$

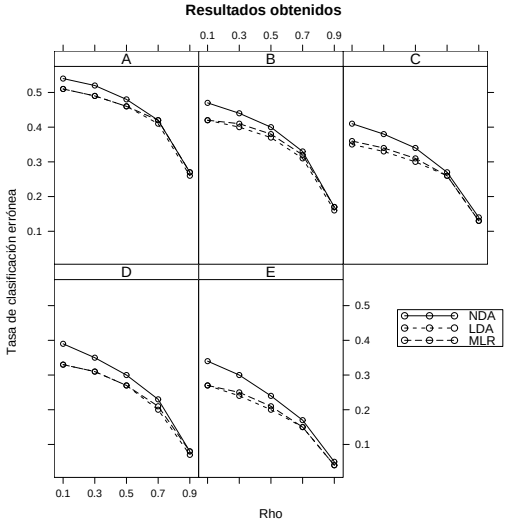

Figura 3.8: Tamaños muestrales $n_1 = n_2 = n_3 = 20$, $\Sigma_2 = 2\Sigma_1$ y $\Sigma_3 = 4\Sigma_1$

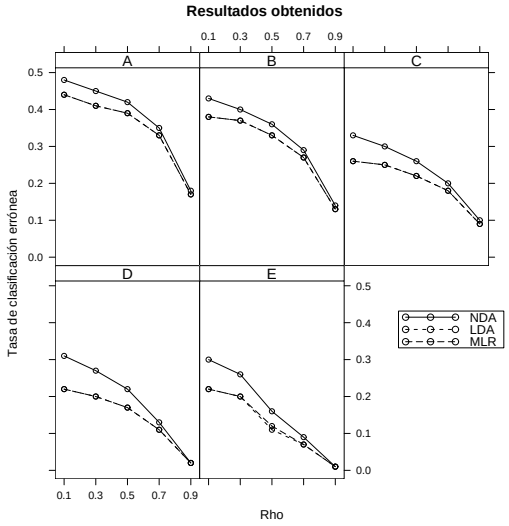

Figura 3.9: Tamaños muestrales $n_1 = n_2 = n_3 = 50$, $\Sigma_2 = 2\Sigma_1$ y $\Sigma_3 = 2\Sigma_1$

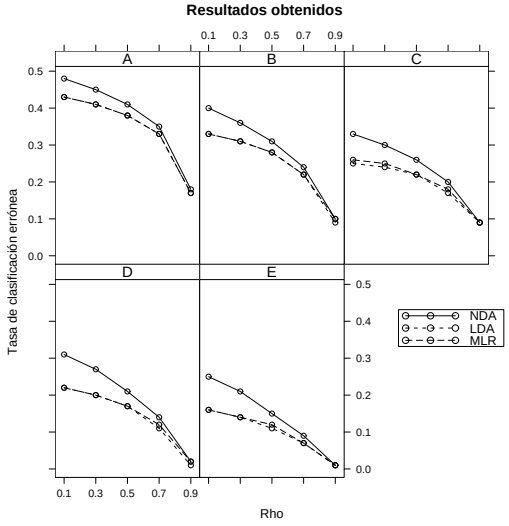

Figura 3.10: Tamaños muestrales $n_1 = n_2 = n_3 = 50$, $\Sigma_2 = 2\Sigma_1$ y $\Sigma_3 = 3\Sigma_1$

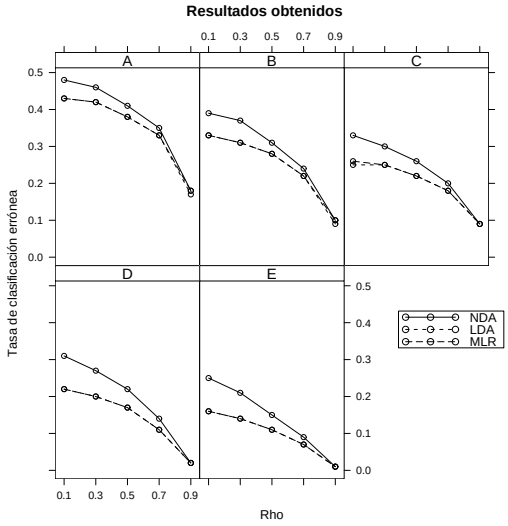

Figura 3.11: Tamaños muestrales $n_1 = n_2 = n_3 = 50$, $\Sigma_2 = 3\Sigma_1$ y $\Sigma_3 = 3\Sigma_1$

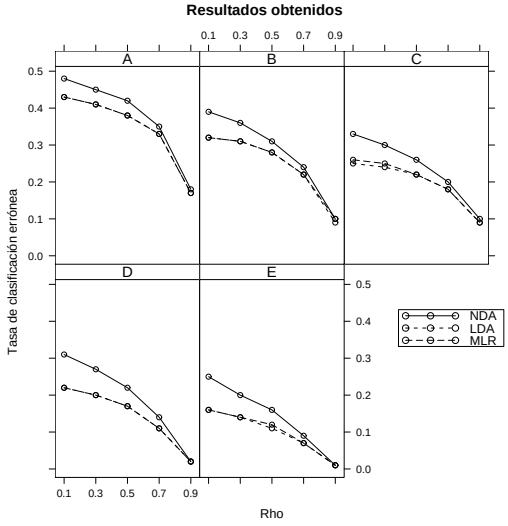

Figura 3.12: Tamaños muestrales $n_1 = n_2 = n_3 = 50$, $\Sigma_2 = 2\Sigma_1$ y $\Sigma_3 = 4\Sigma_1$

3.4.3. Escenario 3.

En el escenario 3 se compararon las tres técnicas de clasificación para el caso de siete grupos distribuidos normal con tres variables, los resultados obtenidos se muestran a continuación.

En las figuras 3.13 a 3.15 se muestran los resultados para las diferentes situaciones de alejamiento, diferentes tamaños muestrales y diferentes estructuras de las matrices de varianzas y covarianzas. Los aspectos importantes a destacar son los siguientes:

- Si se observan las figuras 3.13, 3.14 y 3.15, la conclusión es que a medida que se alejan las poblaciones las tasas de clasificación errónea disminuyen en promedio 12 %.
- El efecto que tiene el aumento del coeficiente de correlación ρ entre pares de variables es el de disminuir las tasas de clasificación errónea.
- En el presente escenario de siete grupos se consideraron dos estructuras de matrices de varianzas y covarianzas, el efecto de pasar de la primera estructura a la segunda sobre las tasas de clasificación errónea fue de aumento.
- El aumento de los tamaños muestrales favoreció las tasas de clasificación errónea disminuyéndolas un promedio de 3 %.
- Los mejores desempeños correspondieron a las técnicas LDA y MLR las cuales tuvieron desempeños muy similares.
- La técnica NDA presentó las tasas de clasificación errónea más grandes.

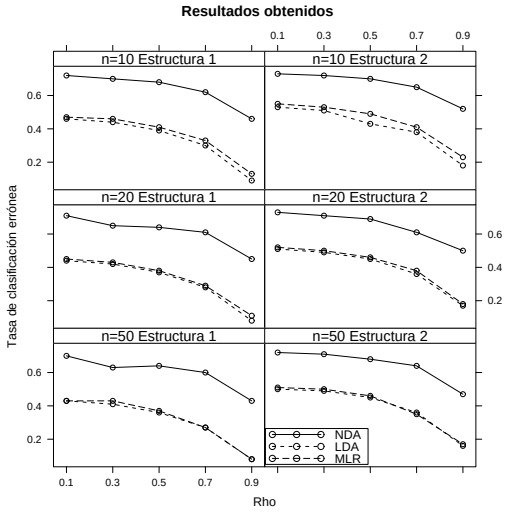

Figura 3.13: Situación A: $\mu_1 = (0,0,0)$ $\mu_2 = (0,0,2)$ $\mu_3 = (0,0,-2)$ $\mu_4 = (0,2,0)$ $\mu_5 = (0,-2,0)$ $\mu_6 = (2,0,0)$ $\mu_7 = (-2,0,0)$

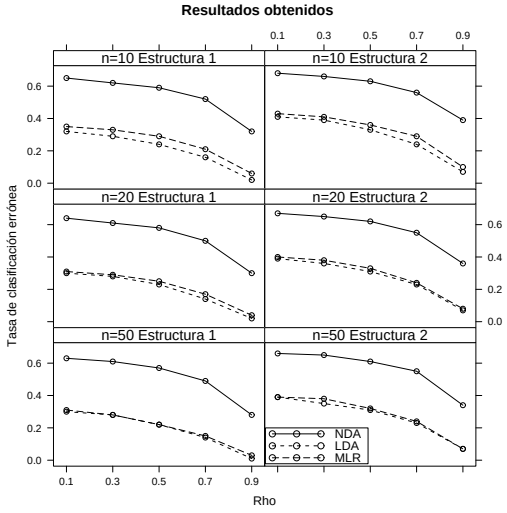

Figura 3.14: Situación B: $\mu_1 = (0,0,0)$ $\mu_2 = (0,0,2)$ $\mu_3 = (0,0,-2)$ $\mu_4 = (0,3,0)$ $\mu_5 = (0,-3,0)$ $\mu_6 = (3,0,0)$ $\mu_7 = (-3,0,0)$

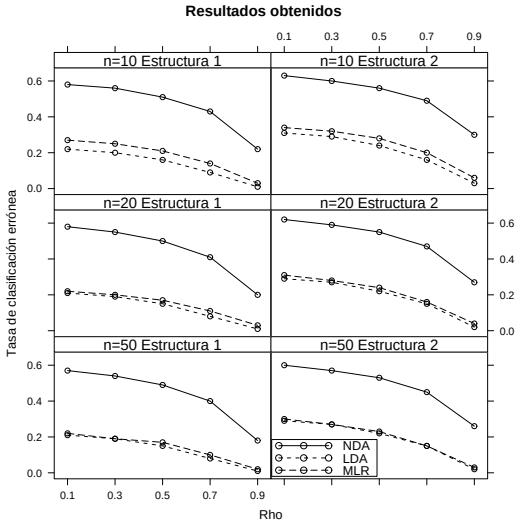

Figura 3.15: Situación C: $\mu_1 = (0,0,0)$ $\mu_2 = (0,0,2)$ $\mu_3 = (0,0,-2)$ $\mu_4 = (0,4,0)$ $\mu_5 = (0,-4,0)$ $\mu_6 = (4,0,0)$ $\mu_7 = (-4,0,0)$

3.4.4. Escenario 4.

Distribución Lognormal

En el presente apartado se muestran los resultados de las técnicas de clasificación para el caso de tres grupos distribuidos Lognormal con tres variables y diferentes tamaños muestrales.

En las figuras 3.16 a 3.18 se observan los desempeños de las técnicas, nuevamente aprecia un patrón en las curvas por lo cual se pueden destacar los siguientes aspectos:

- A medida que aumenta la separación entre los grupos (Situación A a E) se observa que las tasas de clasificación errónea disminuye.
- Por primera vez la técnica LDA tuvo un mal desempeño y fue superada por NDA y MLR considerablemente en la mayoría de los casos.

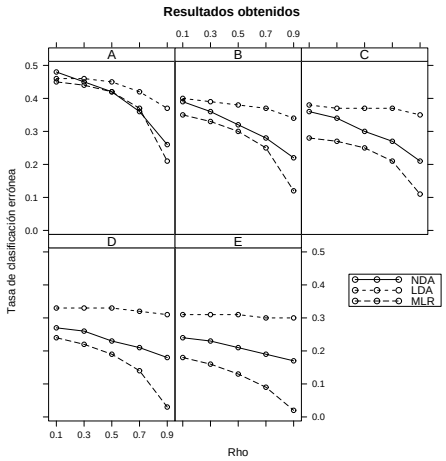

Figura 3.16: Distribución Lognormal con $n_1 = n_2 = n_3 = 20$

- Los mejores desempeños los obtuvo MLR y a medida que aumentaba la separación entre los grupos la diferencia con LDA y NDA fue mayor.
- A partir de la situación de alejamiento C la técnica LDA mantuvo su tasa de clasificación errónea independiente del coeficiente de correlación ρ.
- A medida que aumenta el coeficiente de correlación ρ los desempeños se hacen mejores en NDA y sobre todo en MLR.

Distribución $Sinh^{-1}$-normal

En las figuras 3.19 a 3.21 se pueden observar los resultados de las técnicas de clasificación cuando se consideraron tres grupos distribuidos $Sinh^{-1}$-normal con tres variables y diferentes tamaños muestrales. Nuevamente aquí se puede apreciar un patrón claro en los desempeños, es importante destacar los siguientes aspectos:

- Las tasas de clasificación errónea para NDA y MLR disminuyeron a medida que se

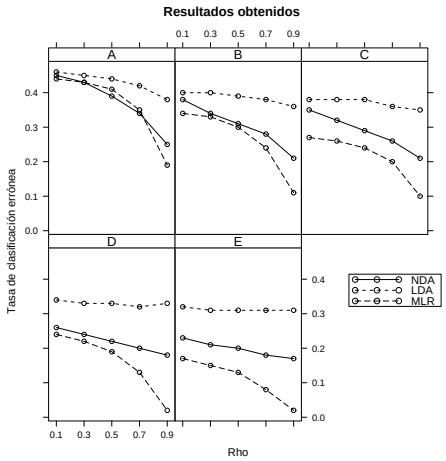

Figura 3.17: Distribución Lognormal con $n_1 = n_2 = n_3 = 50$

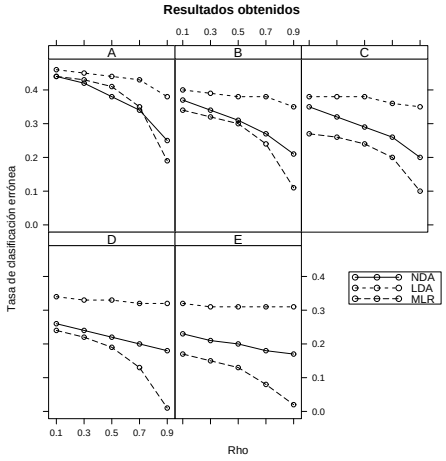

Figura 3.18: Distribución Lognormal con $n_1 = n_2 = n_3 = 100$

incrementó la situación de alejamiento entre los grupos, sin embargo, el desempeño para LDA no pareció mejorar a partir de la situación de alejamiento C.

- La técnica de clasificación con el mejor desempeño fue claramente MLR seguida por

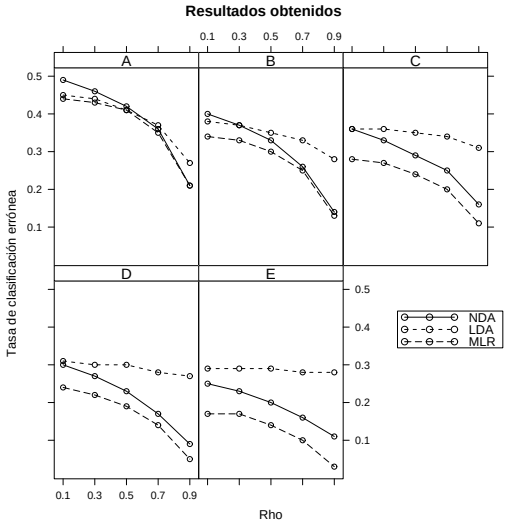

Figura 3.19: Distribución $Sinh^{-1}$-normal con $n_1 = n_2 = n_3 = 20$

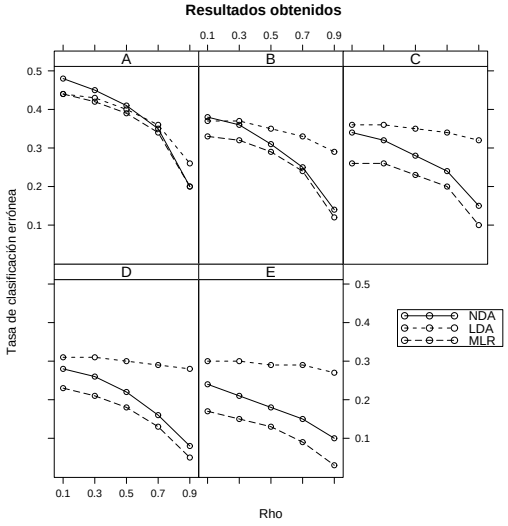

Figura 3.20: Distribución $Sinh^{-1}$-normal con $n_1 = n_2 = n_3 = 50$

NDA y por último con el peor desempeño LDA.

- El desempeño de LDA se mantuvo constante alrededor del 29 % a partir de la situación de alejamiento D e independiente del coeficiente de correlación ρ, mientras que los desempeños de las otras dos técnicas mejoraron a medida que aumentaba el coeficiente de correlación.

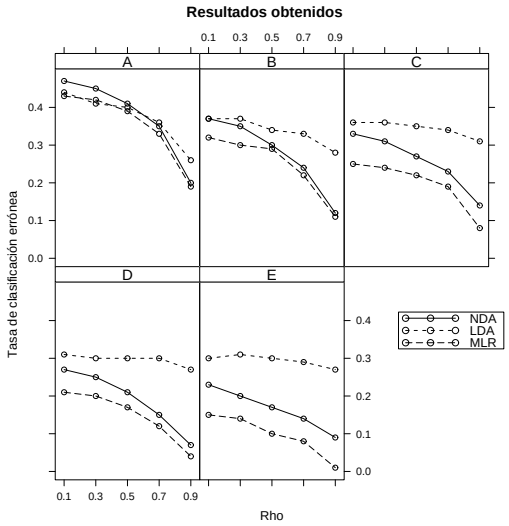

Figura 3.21: Distribución $Sinh^{-1}$-normal con $n_1 = n_2 = n_3 = 100$

Distribución Logit-normal

En las figuras 3.22 a 3.24 se encuentran los resultados para el caso de grupos distribuidos Logit-normal con tres variables y diferentes tamaños muestrales. Los aspectos a resaltar son los siguientes:

- A medida que aumenta la situación de alejamiento los desempeños de las técnicas mejoran considerablemente.

- A medida que aumenta el coeficiente de correlación ρ los desempeños de las técnicas mejoran.
- El aumento en los tamaños muestrales de los grupos de entrenamiento tuvo como consecuencia una ligera mejora en los desempeños de las técnicas.
- Los desempeños de LDA y MLR son muy similares y superiores a NDA para todos los casos.
- La técnica NDA presentó las tasas de clasificación errónea más altas.

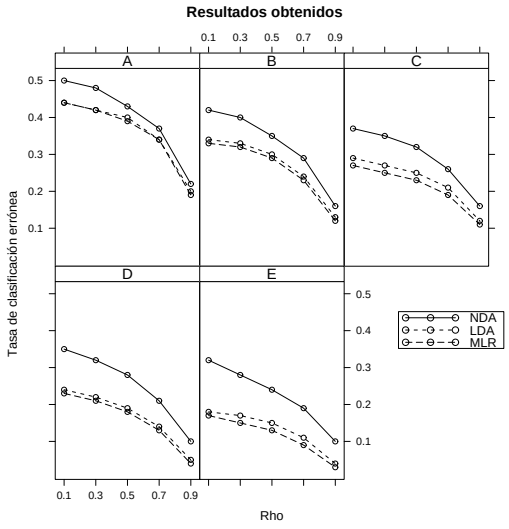

Figura 3.22: Distribución Logit-normal con $n_1 = n_2 = n_3 = 20$

Del estudio de simulación del escenario 4 se destacan los siguientes aspectos:

- Las tasas de clasificación errónea disminuyen a medida que aumenta el alejamiento entre los grupos y a medida que aumenta el coeficiente de correlación ρ.
- La técnica que mejor desempeño tuvo fue MLR.

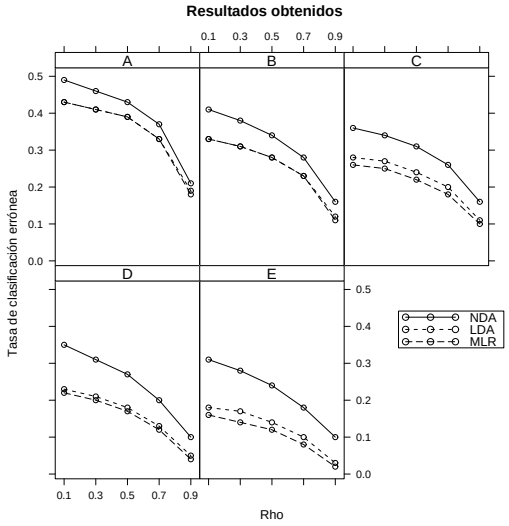

Figura 3.23: Distribución Logit-normal con $n_1 = n_2 = n_3 = 50$

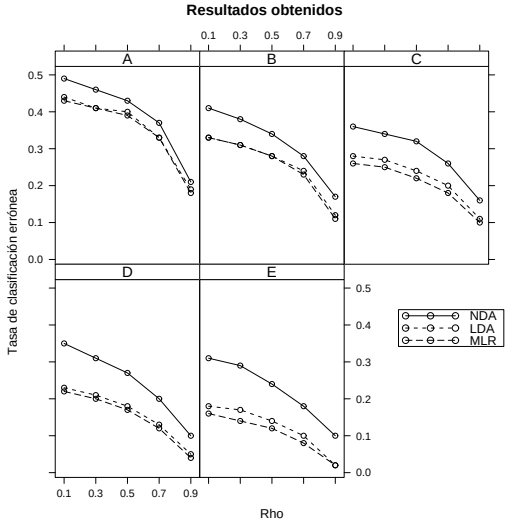

Figura 3.24: Distribución Logit-normal con $n_1 = n_2 = n_3 = 100$

- Para las distribuciones Log-normal y $Sinh^{-1}$-normal NDA tuvo un mejor desempeño que LDA.
- El aumento de tamaños de muestra generó una mejora en la clasificación cuando se pasó de veinte a cincuenta, sin embargo, cuando se pasó de cincuenta a cien el impacto en las tasas de clasificación errónea fue mínimo.

Capítulo 4

Aplicación

En la presente sección se mostrará una aplicación de las técnicas de clasificación NDA, LDA y MLR estudiadas usando para esto la base de datos de Parámetros antropométricos de la población laboral colombiana 1995 del estudio realizado por Estrada y et al (1998). La base de datos consta de las mediciones de 69 variables antropométricas tomadas a 2100 trabajadores activos, 785 de sexo femenino y 1315 de sexo masculino, de las ciudades Pereira, Santa Marta, Barranquilla, Cartagena, Bogotá, Bucaramanga, Barrancabermeja, Manizales, Medellín, Riohacha, Cali y Pasto cuyas edades estaban comprendidas entre 20 y 60 años.

La variable elegida para la formación de los grupos fue el Índice de Masa Corporal (IMC) que se define como $IMC = \frac{Peso(kg)}{Altura(cm)^2}$. Estándares médicos internacionales indican que existen tres grupos definidos a partir del IMC, los cuales servirán para la clasificación.

$$Grupo = \begin{cases} 1 & \text{Delgadez} \quad \text{si} \quad IMC < 18 \\ 2 & \text{Normal} \quad \text{si} \quad 18 \leq IMC \leq 25 \\ 3 & \text{Obesidad} \quad \text{si} \quad IMC > 25 \end{cases}$$

En la figura 4.1 se pueden observar claramente los tres grupos denotados por los números 1,

2 y 3. Los grupos se muestran en la siguiente figura.

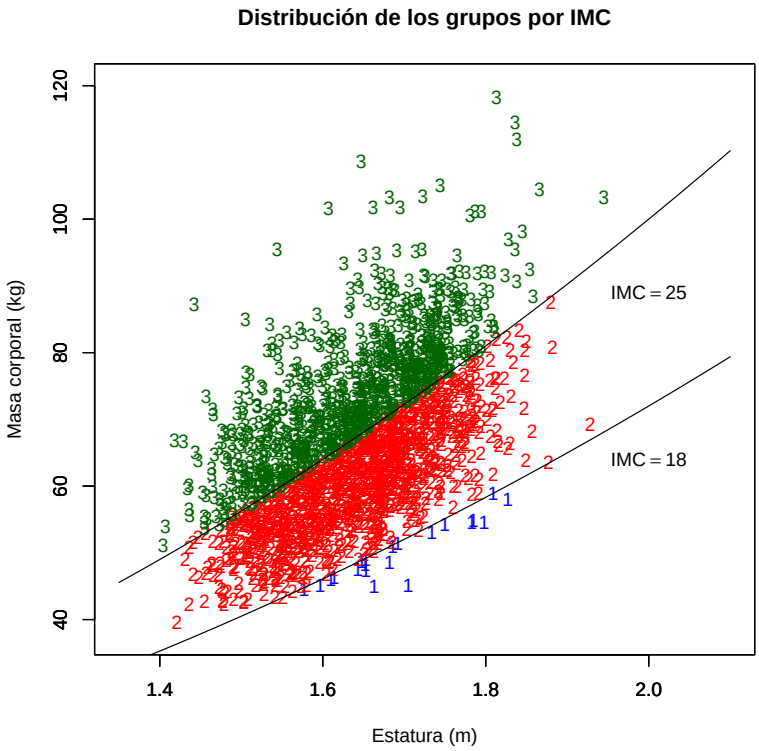

Figura 4.1: Distribución de los grupos

En el cuadro 4.1 se observan el número de observaciones que conforman cada uno de los grupos, mientras que de los grupos 2 y 3 se tienen más de 800 observaciones del grupo 1 sólo se dispone de 21 lo cual se tendrá en cuenta al momento de formar el conjunto de entrenamiento de las técnicas.

Se eligieron seis variables con las cuales se realizó el proceso de clasificación.Las dos primeras variables están asociadas con la talla de los trabajadores mientras que las siguientes cuatro están asociadas con la masa muscular y/o acumulamiento de grasa, las variables son las siguientes:

Grupo	Nºobs.
1	21
2	1181
3	897

Cuadro 4.1: Distribución de los grupos

$$\text{Variables} = \begin{cases} V1 & \text{Longitud del pie (cm)} \\ V2 & \text{Alcance vertical máximo (cm)} \\ V3 & \text{Perímetro del brazo relajado (cm)} \\ V4 & \text{Perímetro del muslo (cm)} \\ V5 & \text{Ancho transversal del tórax (cm)} \\ V6 & \text{Perímetro abdominal umbilical (cm)} \end{cases}$$

En la figura 4.2 se pueden observar los boxplot para cada una de las variables seleccionadas. Las cajas de los boxplot para las variables V1, V2 y V4 se traslapan mientras que para las demás variables parece existir una diferencia entre los grupos, es decir, las variables V3, V5 y V6 parecen discriminar mejor entre los tres grupos. Adicionalmente se puede ver que existe una relación directa entre el grupo y los boxplot de las variables V3, V5 y V6, es decir, se ve que las personas pertenecientes al grupo 3 (obesidad) tienden a tener valores más grandes en estas variables seguidos de los del grupo 2 y grupo 1. Lo anterior se puede ver como una situación lógica ya que a medida que una persona sea obesa ésta tenderá a tener mayor perímetro del brazo, del tórax y del abdomen. En la variable V4 que corresponde al perímetro del muslo se puede apreciar el mismo patrón aunque no tan marcado como en las variables anteriores. Por otra parte, se observa que que la mayor parte de los boxplot muestran distribuciones marginales simétricas con varianzas similares.

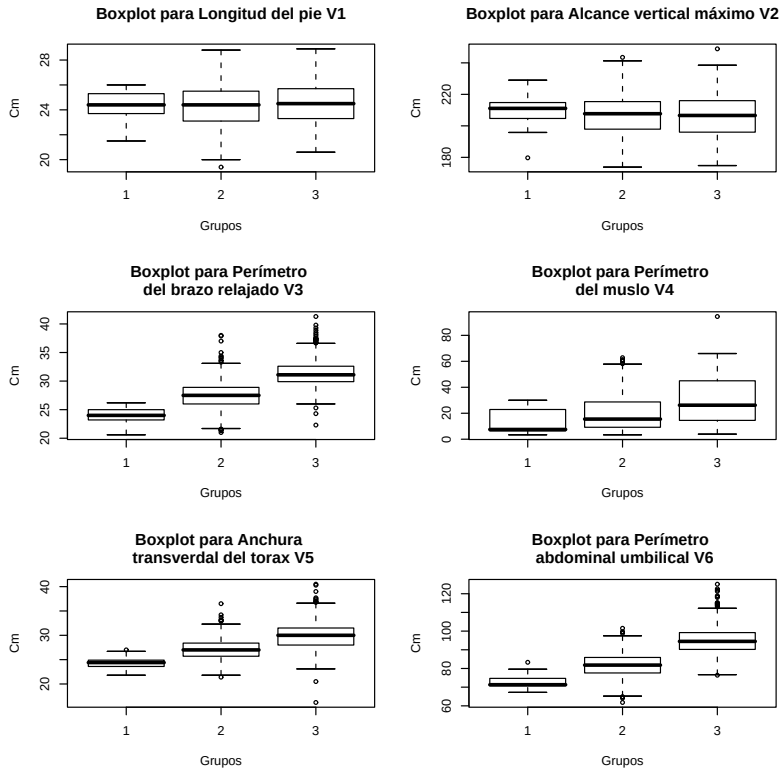

Figura 4.2: Boxplot para las variables de clasificación

En el cuadro 4.2 se presentan los resultados de las pruebas de simetria y kurtosis multivariada. Se puede observar que la hipótesis nula de simetría igual a cero se acepta[1] tanto a nivel general como grupal, por otra parte, la hipótesis nula de kurtosis igual a $p(p+2) = 24$ se rechaza en todos los casos, en conclusión, los datos para la aplicación cumplen el supuesto de sumetría multivariada pero no el supuesto de kurtosis multivariada.

En los cuadros 4.3, 4.4 y 4.5 se muestran las matrices de varianzas y covarianzas para los tres grupos definidos. Al utilizar el procedimiento para probar la hipótesis nula de igualdad de

[1] A un nivel de significancia del 5 %

Grupo	Simetría	Valor P	Kurtosis	Valor P
General	0.01	1	3418.47	0
1	16.7	0.38	168.57	0
2	0.01	1	1937.35	0
3	0.01	1	1723.73	0

Cuadro 4.2: Pruebas de simetría y kurtosis multivariada

varianzas mostrado en Díaz (2002), se encontró que el valor P de la prueba es menor que un nivel de significancia del 5 %, por lo tanto, se rechaza la hipótesis nula.

	v1	v2	v3	v4	v5	v6
v1	1.62	13.30	0.25	−3.76	1.17	−0.31
v2	13.30	141.03	0.61	−48.15	10.95	−10.18
v3	0.25	0.61	2.08	−0.85	0.22	0.87
v4	−3.76	−48.15	−0.85	79.05	−4.22	17.24
v5	1.17	10.95	0.22	−4.22	1.75	−0.56
v6	−0.31	−10.18	0.87	17.24	−0.56	14.46

Cuadro 4.3: Matriz de varianzas y covarianzas Grupo 1

Para llevar a cabo la aplicación de cada una de las técnicas se construyeron dos conjuntos a partir de la base original de 2100 observaciones y 6 variables. El conjunto de entrenamiento y el de validación estuvieron conformados por 10, 50 y 50 observaciones del grupo 1, 2 y 3 respectivamente. El tamaño de 10 tomado del grupo 1 se debió al número de observaciones totales en ese grupo que era de 21 lo que obligó a tomar un número pequeño de obsevaciones. El conjunto de validación estuvo formado por el restante de observaciones que no hicieron parte del conjunto de entrenamiento. Al conjunto de entrenamiento fueron aplicadas las técnicas para crear las funciones discrimiantes y luego con éstas funciones discriminantes se llevó a cabo la clasificación de las observaciones del conjunto de validación calculando la tasa de clasificación incorrecta para cada caso. Los resultados obtenidos se muestran en el cuadro 4.6.

	v1	v2	v3	v4	v5	v6
v1	2.70	17.18	1.58	−9.86	1.77	2.09
v2	17.18	151.89	11.13	−71.18	14.83	16.24
v3	1.58	11.13	4.99	−3.08	2.59	7.01
v4	−9.86	−71.18	−3.08	186.78	−7.08	10.60
v5	1.77	14.83	2.59	−7.08	4.18	6.29
v6	2.09	16.24	7.01	10.60	6.29	37.18

Cuadro 4.4: Matriz de varianzas y covarianzas Grupo 2

	v1	v2	v3	v4	v5	v6
v1	2.47	17.28	1.01	−10.03	2.16	2.09
v2	17.28	167.55	8.42	−65.69	20.43	18.17
v3	1.01	8.42	5.02	4.64	2.89	7.92
v4	−10.03	−65.69	4.64	288.97	−7.24	20.22
v5	2.16	20.43	2.89	−7.24	7.27	9.31
v6	2.09	18.17	7.92	20.22	9.31	52.23

Cuadro 4.5: Matriz de varianzas y covarianzas Grupo 3

En el cuadro 4.6 se pueden observar los resultados de la aplicación de las técnicas LDA, MLR y NDA para la base de datos con las variables seleccionadas. El caso 1 mostrado en el cuadro corresponde a las tasas de clasificación errónea cuando se consideran las seis variables. La técnica LDA generó la menor tasa de clasificación errónea con un 18 % seguida de MLR y NDA con diferencias de 1 % entre ellas.

Los casos 2 al 7 corresponden a los desempeños de las técnicas cuando se elimina a una de las variables y el proceso se lleva a cabo con las restantes variables. Se puede ver que cuando se excluyeron las variables V3, V5 y V6 las tasas tuvieron un aumento mayor que en los otros casos. En los casos 8 y 9 se muestran los resultados cuando se consideran sólo las variables V1, V2 y V4 las cuales en el boxplot de la figura 4.2 mostraron traslapes entre las cajas de los grupos en consideración. Se observa que cuando sólo se utilizan estas variables las tasas

TASA DE CLASIFICACIÓN ERRÓNEA					
Caso	Variables utilizadas	Variables excluidas	LDA	MLR	NDA
1	V1, V2, V3, V4, V5, V6		0.18	0.19	0.20
2	V2, V3, V4, V5, V6	V1	0.18	0.19	0.20
3	V1, V3, V4, V5, V6	V2	0.18	0.18	0.23
4	V1, V2, V4, V5, V6	V3	0.20	0.22	0.21
5	V1, V2, V3, V5, V6	V4	0.18	0.21	0.20
6	V1, V2, V3, V4, V6	V5	0.19	0.22	0.21
7	V1, V2, V3, V4, V5	V6	0.21	0.25	0.23
8	V1, V2	V3, V4, V5, V6	0.56	0.57	0.59
9	V1, V2, V4	V3, V5, V6	0.45	0.46	0.50
10	V3, V5, V6	V1, V2, V4	0.17	0.20	0.25

Cuadro 4.6: Resultados de la aplicación

de clasificación errónea aumentan por encima del doble lo cual indica que estas variables no diferencian claramente los grupos.

En el caso 10 solo se tienen en cuenta las variables V3, V5 y V6 que parecen ser las que mejor discriminan los grupos, LDA obtiene en este caso la mejor tasa hasta ahora del 17 % mientras que las otras dos técnicas no mejoraron los resultados obtenidos anteriormente.

El mejor desempeño de LDA 17 % se obtuvo en el caso 10, el de MLR 18 % se obtuvo en el caso 3 y el mejor de NDA 20 % se obtuvo en el caso 1, 2 y 5. Las diferencias entre los mejores desempeños de las técnicas son del 3 %, siendo LDA y MLR desempeños muy similares mientras que el desempeño de NDA solo estuvo mejor que MLR en tres ocasiones pero nunca mejor que LDA.

Capítulo 5

Conclusiones

Las técnicas de clasificación que mejor desempeño tuvieron en el estudio de simulación fueron LDA y MLR, su desempeño fue tan similar que en la mayoría de las figuras las líneas asociadas se superponen, mostrando esto que las diferencias son mínimas; la técnica NDA en la gran mayoría de casos obtuvo tasas de clasificación superiores a LDA y MLR. El único caso donde NDA no obtuvo las tasas de clasificación errónea más altas fue cuando se consideraron poblaciones distribuidas Lognormal y $Sinh^{-1}$-normal, en esta situación el peor desempeño lo presentó LDA.

En situaciones prácticas donde se presenté un problema de clasificación de nuevas observaciones a grupos ya definidos, teniendo en cuenta varias variables explicativas, se recomienda utilizar principalmente la técnica MLR seguida de la técnica LDA, siempre y cuando la distribución de probabilidad de los datos sea muy cercana a una situación de normalidad multivariada, el supuesto de homogenidad de varianzas puede violarse ligeramente y los resultados obtenidos con cualquiera de las técnicas serán similares, adicionalmente se sugiere utilizar este criterio siempre y cuando todas las variables explicativas sean de tipo cuantitativo.

Posibles trabajos futuros podrían estar encaminados a comparar el desempeño de las técni-

cas considerando otro tipo de escenarios en los cuales se pueden estudiar aspectos como: mayor número de grupos a clasificar, tamaños muestrales mayores, estructuras de matrices de varianzas y covarianzas diferentes, otros tipos de distribuciones para los grupos, medidas de desempeño diferentes a la tasa de clasificación errónea y algoritmos de búsqueda para determinar el vector de clasificación en la técnica NDA.

Apéndice A

Documentación de las funciones creadas en R

Funciones NDA y NDA.CLA en R para Análisis Discriminante no Métrico

1. Función NDA

   - Descripción: esta función permite encontrar el vector de clasificación para el método de Análisis Discriminante no Métrico usando el algoritmo sugerido por Choulakian y Almhama (2001).
   - Sintaxis de la función: NDA(datos , epsilon=10^{-5} , n.iteraciones=100)
   - Argumentos
     - datos: marco de datos (conjunto de entrenamiento) donde la primera columna debe ser un factor el cual contiene los nombres de los grupos a los cuales pertenece cada observación.

- epsilon: valor sugerido por Choulakian y Almhama (2001) para detener el proceso iterativo.
- n.iteraciones: número máximo de iteraciones se cree va a necesitar el algoritmo para la convergencia. Si el número total de iteraciones que tomó el algoritmo para converger es igual al que el usuario dió (n.iteraciones) se recomienda volver a correr la función y aumentar este parámetro; sin embargo, la ventaja de este algoritmo es que converge en no más de treinta iteraciones por lo general.

- Valores de salida
  - NETA: corresponde al vector que sirve como función de clasificación.
  - NETACERO: corresponde al vector de clasificación de Análisis Discriminante Canónico que sirve como semilla para el algoritmo.
  - DISCO: valor máximo del discriminante cuando converge el algoritmo.
  - Niteraciones: número de iteraciones que tomó el algoritmo para alcanzar la convergencia.
  - Adicionalmente se obtiene un gráfico donde se puede ver el comportamiento del discriminante, en el eje horizontal van las iteraciones y en el vertical el valor de DISCO.

- Ejemplo

  En este ejemplo se tomará la base de datos Iris, se obtendrá la función de clasificación con NDA.

```
# Tomando 20 observaciones de cada categorías
# para construir el conjunto de entrenamiento
k<-20
samp <- c(sample(1:50,k), sample(51:100,k), sample(101:150,k))
datos<-data.frame(grupo=factor(iris[samp,5]),iris[samp,-5])
# Aplicando NDA()
resultado<-NDA(datos=datos,epsilon=10^-5,n.iteraciones=200)
```

```
resultado
```

2. NDA.CLA

   - Descripción: esta función tiene como objetivo clasificar un conjunto de observaciones en uno de varios grupos utilizando el resultado obtenido por la función NDA.
   - Sintaxis de la función:

     ```
     NDA.CLA(datos= , datos.predic= , resultado= )
     ```

   - Argumentos
     - datos: marco de datos (o conjunto de entrenamiento) que se sirvieron para encontrar el vector de clasificación con la función NDA.
     - datos.predict: marco de datos que se quiere clasificar.
     - resultado: objeto que correponde a la salida de la función NDA.
   - Valores
     - TABLA: aparece una tabla donde se muestran las clasificaciones correctas y erradas.
     - N.clas.erradas: número de clasificaciones erradas.
     - tasa.clas.incorr: tasa de clasificaciones incorrectas.
   - Ejemplo

     En este ejemplo se tomará la base de datos Iris, se obtendrá la función de clasificación con NDA y luego se validará la clasificación con el mismo conjunto.

     ```
     # Tomando las 30 observaciones restantes
     # de cada categoría para realizar la validación
     datos.validacion<-data.frame(grupo=factor(iris[-samp,5]),iris[-samp,-5])
     # Aplicando NDA.CLA( )
     NDA.CLA(datos=datos,datos.predic=datos.validacion,resultado=resultado)
     ```

3. Código de las funciones

```
################################################################################
########### Librerias necesarias para la aplicación de las tecnicas ############
################################################################################
require(MASS)
require(nnet)
require(klaR)
################################################################################
########### Esta función sirve para la construcción de las matrices ############
############################### Bgh y Vgh  #####################################
restas<-function(x){
n<-nrow(x)
resultado<-matrix(,ncol=ncol(x))
  for (i in 1:n-1){
  z<-matrix(x[i,],nrow=n-i,byrow=T,ncol=ncol(x))
  temp<-x[(i+1):n,]-z
  resultado<-rbind(resultado,temp)
  }
resultado[-1:-(n+1),]
}
################################################################################
######### Esta función sirve para la construcción de las matrices ##############
############################   Bgh y Vgh  ######################################
################################################################################
producto<-function(x) x%*%t(x)
################################################################################
######### Esta función sirve para la construcción de las matrices ##############
#####################   B(neta) y V(neta)  #####################################
################################################################################
Vcoc<-function(x,neta,p) x/sqrt(t(neta)%*%matrix(x,ncol=p)%*%neta)
Vneta<-function(x,neta,p)  matrix(rowSums( apply(x,2,Vcoc,neta=neta,p=p) ) ,ncol=p)

Bcoc<-function(x,neta,p)  x / sqrt(t(neta)%*%matrix(x,ncol=p)%*%neta)
Bneta<-function(x,neta,p,Nij) {
matrix(rowSums( (apply(x,2,Bcoc,neta=neta,p=p))%*%diag(Nij) ),ncol=p)  }
################################################################################
#########################    Función DISCO   ###################################
```

```
##############################################################################
Disco<-function(B.neta,V.neta,neta) (t(neta)%*%B.neta%*%neta)/(t(neta)%*%V.neta%*%neta)
##############################################################################
#####################   Funcion NETA k   #####################################
##############################################################################
neta.sig<-function(neta.ant,disco,Vneta.ant,Bneta.ant){
t((1-2*disco)*neta.ant+2*solve(Vneta.ant,tol=1e-25)%*%Bneta.ant%*%neta.ant) }
##############################################################################
############################  Función Principal ##############################
##############################################################################
NDA<-function(datos,epsilon=1e-5,n.iteraciones=100)  {

g<-nlevels(datos[,1])
p<-ncol(datos)-1
N<-table(datos[,1])                        # N° de elementos por grupo
NN<-outer(N,N) ; Ngh<-NN[upper.tri(NN)]    # Multipli de los n° element
if(g==2) Ngh=as.matrix(Ngh)

list.datos<-split(datos[,-1],datos[,1])
medias<-t(sapply(list.datos,mean))
medias
Restas.medias<-restas(medias)
if(g==2) Restas.medias=t(as.matrix(Restas.medias))
B<-apply(Restas.medias,1,producto)
V<-apply(restas(as.matrix(datos[,-1])),1,producto)

neta.cero<-coefficients(lda(grupo ~ . ,data=datos))[,1]

neta.cero<-neta.cero/max(abs(neta.cero))

V.cero<-Vn<-Vneta(x=V,neta=neta.cero,p=p)
B.cero<-Bn<-Bneta(x=B,neta=neta.cero,p=p,Nij=Ngh)
Disco.cero<-Disco(B.neta=B.cero,V.neta=V.cero,neta=neta.cero)

### Aquí inician las iteraciones

iter<-matrix(0,ncol=(p+1),nrow=n.iteraciones)
iter[1,1]<-Disco.cero ; iter[1,2:(p+1)]<-t(neta.cero)
```

```
convergencia<-matrix(0,ncol=1,nrow=n.iteraciones)
convergencia[1,1]<-1

i<-2
while(convergencia[i-1,1]>0)  {
iter[i,2:(p+1)]<-neta.sig(neta.ant=iter[i-1,-1],disco=iter[i-1,1],Vneta.ant=Vn,Bneta.ant=Bn)
Bn<-Bneta(x=B,neta=iter[i,2:(p+1)],p=p,Nij=Ngh)
Vn<-Vneta(x=V,neta=iter[i,2:(p+1)],p=p)
iter[i,1]<-Disco(B.neta=Bn,V.neta=Vn,neta=iter[i,2:(p+1)])
convergencia[i,1]<-ifelse(abs(iter[i,1]-iter[(i-1),1])>epsilon,1,0)
i<-i+1
}

nconver<-sum(convergencia,na.rm =T)
par(bg='lightyellow',cex.axis=0.7)
plot(x=iter[1:nconver,1],ylab='Disco',xlab='Iteración',
main='Comportamiento de Disco \n  NDA',pch=20,ylim=c(0,1.01),col='blue')
abline(a=1,b=0,col='red')

iter
list(NETACERO=t(neta.cero),NETA=iter[nconver,-1],
DISCOCERO=iter[1,1],DISCO=iter[nconver,1],Niteraciones=nconver)
}



############################################################
# El conjunto de datos A CLASIFICAR se denota por "datos.predic"
# El conjunto de datos de entreno de denota por "datos"

NDA.CLA<-function(datos,datos.predic,resultado){
vector<-resultado$NETA   # Vector o Función Discriminante
scores.clas<-as.matrix(datos[,-1])%*%as.matrix(vector)
# Scores para clasificar
n.grupos<-nlevels(datos[,1])
G<-datos[,1]
S<-scores.clas
x<-data.frame(G=factor(datos[,1]),S=scores.clas)
y.clas<-x[order(S,G),]
```

```
# Y es la matriz con los grupos y los scores de clasificac

plot(x,ylab='Scores',main='Distribución de los Scores por grupo',xlab='Grupos')

# Distribución de los scores por grupo
prediccion<-datos.predic[,1]
scores.predic<-as.matrix(datos.predic[,-1])%*%as.matrix(vector)
# Scores para las nuevas observaciones

cortes<-lapply(split(y.clas[,2],y.clas[,1]),mean,3)

distancias<-matrix(0,ncol=n.grupos,nrow=nrow(datos.predic))
for (i in 1:nrow(datos.predic)){ distancias[i,]= abs (unlist(t(as.matrix(cortes))) - scores.predic[i,1]) }
distancias
predicciones<-names(cortes[max.col(-distancias)])

tabla<-errormatrix(true=datos.predic[,1], predicted=predicciones, relative = FALSE)
# Tabla de clasificacion
n.clas.erradas<-errormatrix(true=datos.predic[,1], predicted=predicciones,
relative = FALSE)[n.grupos+1,n.grupos+1]  # N° de clasificaciones que fueron clasificadas incorrectamente
tasa.clas.incorr<-n.clas.erradas/nrow(datos.predic)

list(TABLA=tabla,N.clas.erradas=n.clas.erradas,tasa.clas.incorr=tasa.clas.incorr)

}
################################################################################
################################################################################
################################################################################
```

Bibliografía

[1] Anderson T. W. and Bahadur R. R. (1962). "Classification into two Multivariate Normal Distributions with Different Covariance Matrices". *The Annals of Mathematical Statistics*, Vol. 33, No. 2. pp. 420-431.

[2] Anderson, J.A. (1972). "Separate sample logistic discrimination". *Biometrica*. 59, 19-35.

[3] Choulakian, V. and Almhana, J (2001). "An Algorithm for Nonmetric Discriminant Analysis". *Computational Statistics & Data Analysis*, 35, 253-264.

[4] Carroll, R. and Pederson, S. (1993). "On Robustness in the Logistic Regression Model". *Journal of the Royal Statistical Society*. Vol. 55, pp. 693-706.

[5] Castrillon, F. (1998). "Comparación de la discriminación normal lineal y Cuadrática con la regresión logística para clasificar vectores en dos poblaciones". Medellín. Tesis Magíster en Estadística. Facultad de Ciencias. Universidad Nacional de Colombia, Sede Medellín.

[6] Cheng, T., Pia. M. and Feser, V. (2002). "High-breakdown estimation of multivariate mean and covariance with missing observations". *British Journal of Mathematical and Statistical Psychology*. Vol. 55, 317-335.

[7] Cherkaoui, O. and Cleroux, R. (1991). "Comparative study of six classification methods for mixtures of variables" . *Computing Science and Statisctics*. 23, 233-236.

[8] Chernoff, H. (1972). "The selection of effective attributes for deciding between hypotheses using linear discriminant functions". *Frontiers of pattern recognition*, New York: Academic Press, pp. 55-60.

[9] Clunies, C. W. and Riffenburgh, R. H. (1960). "Geometry and linear discrimination". *Biometrics* Vol. 47, No. 1/2. pp. 185-189.

[10] Cornfiel, J. (1962). "Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis.". *Proceedings of the Federal American Society of Experimental Biology*. Vol 21, 58-61.

[11] Cox, D.R. (1996). "Some procedures associated with the logistic qualitative response curve". *Research papers in statistics: Festschrift for J. Newman.*, New York Wiley.

[12] Crawley, D. R. (1979). "Logistic discrimination as an alternative to Fisher's linear function". *New Zealand Statistician*. 14, 21-25.

[13] Croux, C. and Dehon, C. (2001). "Robust Linear Discriminant Analysis Using S-Estimators". *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*. Vol. 29, No. 3. pp. 473-493.

[14] Díaz,L. (2002). "Estadística multivariada: inferencia y métodos". Universidad Nacional de Colombia.

[15] Day, N.E. and Kerridge, D.F. (1967). "A general maximum likelihood discriminant". *Biometrics*. 23, 313-323.

[16] Efron, B. (1975). "The Efficiency of Logistic Regresion compared to Normal Discriminant Analysis". *Journalof the American Statistical Association*, 70, 892-898.

[17] Estrada, J., Camacho, J., Restrepo, M. y Parra, C. (1998). "Parámetros antropométricos de la población laboral colombiana 1995 (acopla95)". *Revista de la Facultad Nacional de Salud Pública*, 15(2): 112-139.

[18] Fisher R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems". *Annual Eugenics* 7: 179-188.

[19] Guttman, L. (1998). "Eta, disco, odisco and F.". *Psychometrika*. 53, 393-405.

[20] Harrell, F.E. and Lee, K.L. (1985). "A comparison of discriminant analysis and logistic regression under multivariate normality" . *Biostatistic: Statistics in Biomedical*. Amsterdam: Elsevier Science Publishers.

[21] Hawkins, D.M. and McLachan J. (1997). "High-Breakdown linear discriminant analysis". *Journal of American Statistical Asociation*. 92, 136-146.

[22] Johnson, M.(1987). "Multivariate statistical simulation". New York: John Wiley and Sons.

[23] Johnson, R. y Wichern, D. (1998). "Applied Multivariate Statistical Analysis". London: Prentice-Hall.

[24] Little, R. J. and Smith, P. J. (1987). "Editing and imputing for quantitative survey data". *Journal of the American Statistical Association*. 82, 58-68.

[25] Pregibon, D. (1981). "Logistic Regression Diagnostics". *The Annals of Statistics*. Vol. 9, pp. 705-724.

[26] Rao, C.R. (1948). "The utilization of multiple meausurements in problems of biological classification". *J.Royal Statistic.*, Vol. 10, 159-193.

[27] Raveh, A. (1983). "Preference structure analysis: A nonmetric approach" .*Patter Recognition* 16, 253-259.

[28] Raveh, A. (1989). "A Nonmetric Approach to Linear Discriminant Analysis" . *Journal of the American Statistical Association*. 84, 176-183.

[29] Shelley, B. and Donner, A. (1987). "The Efficiency of Multinomial Logistic Regression compared with Multiple Group Discriminant Analysis". *Journal of American Statistical Association*. 82, 1118-1122.

[30] Trevor, F. And Ferry, G. (1991). "Robust Logistic Discrimination". *Biometrika*,Vol. 78, No 4. pp. 841-849.

[31] Usuga, O. (2006). "Comparación entre Análisis de Discriminante no métrico y Regresión Logística". *Proceedings of the Federal American Society of Experimental Biology*, 21, 58-61.

[32] Welch, B.L. (1939). "Note on Discriminant Functions". *Biometrika*, Vol. 31, No. 1/2, 218-220