Acta Biol. Venez., Vol. 37(2):171-186 **ARTÍCULO**
Julio-Diciembre, 2017

DESEMPEÑO DE CATORCE ALGORITMOS DISCRIMINANTES: UN CASO DE ESTUDIO CON DATOS MORFOMÉTRICOS MULTIVARIADOS

Enrique Quintero-Torres[*1], *Beatriz López-Sánchez*[1] *y Carlos Méndez*[2]

1. Laboratorio de Ecología Acuática, Centro de Ecología, Instituto Venezolano de Investigaciones Científicas (IVIC), Apdo. 20632 Caracas 1020-A, Venezuela.
2. Laboratorio de Ecosistemas y Cambio Global, Centro de Ecología, Instituto Venezolano de Investigaciones Científicas (IVIC), Apdo. 20632 Caracas 1020-A, Venezuela. *equinter@ivic.gob.ve; equintero1981@gmail.com.

RESUMEN

El análisis lineal discriminante (LDA) es una de las herramientas multivariadas más usadas en estudios morfométricos, sin embargo requiere el cumplimiento de supuestos que rara vez son comprobados, produciendo sesgo y malas interpretaciones de los resultados. Los más discutidos en la literatura son la normalidad multivariada, la homocedasticidad de las matrices de covarianza, ausencia de colinealidad entre variables predictoras, baja dimensionalidad y la sensibilidad ante datos atípicos multivariados. El objetivo principal de este trabajo fue comparar el desempeño de catorce algoritmos discriminantes en un caso de estudio particular de datos morfométricos reales. La evaluación de los supuestos reveló importantes desviaciones de la normalidad multivariada y homocedasticidad en las matrices de covarianza, también se encontraron potenciales problemas de colinealidad entre variables predictoras y gráficamente se evidenció la presencia de datos atípicos multivariados. El análisis discriminante localizado (LocLDA) mostró el menor porcentaje de error de clasificación (0,85 %), seguido por los algoritmos discriminante regularizado (RDA), discriminante diagonal (DDA) ambos con 1,96 % y el discriminante cuadrático (2,35 %). Estos resultados permiten resaltar la factibilidad de utilizar los algoritmos LocLDA, RDA y DDA ante violaciones de los supuestos del LDA, en datos morfométricos similares a los presentados en este estudio.

Palabras clave: Análisis lineal discriminante, Análisis discriminante localizado, Colinealidad, Homocedasticidad de matrices de covarianza, Normalidad multivariada.

Performance of fourteen discriminant algorithms: a case study with multivariate morphometric data

Abstract

Linear discriminant analysis (LDA) is one of the most widely used multivariate tools in morphometric studies, however its underlying assumptions are usually not verified, producing bias or misinterpretation of the results. The most discussed assumptions in the literature include multivariate normality, homoscedasticity of covariance matrices, no collinearity among predictor variables, low dimensionality and sensitivity to multivariate outliers. The main objective of this work was to compare the performance of fourteen discriminant algorithms in a particular case study of real morphometric data. The evaluation of the assumptions revealed important deviations of the multivariate normality and homoscedasticity in the covariance matrices. Potential collinearity problems were also found between predictor variables and the presence of multivariate outliers was graphically evidenced. The localized discriminant analysis

Número Especial "*Primer Encuentro Venezolano de Métodos de Cuantificación Morfológica*"
Recibido: enero 2017
Aceptado: diciembre 2017

(LocLDA) showed the lowest percentage of classification error (0,85 %), followed by the algorithms regularized discriminant (RDA), diagonal discriminant (DDA) both with 1,96 % and the quadratic discriminant (2,35 %). These results highlight the feasibility of using LocLDA, RDA and DDA algorithms in case of violations of the LDA assumptions, in morphometric datasets similar to those presented in this study.

Keywords: Lineal discriminant analysis, Localized discriminant analysis, Collinearity, Homocedasticity of covariance matrix, Multivariate normality.

INTRODUCCIÓN

En los estudios de morfometría, los análisis estadísticos frecuentemente empleados son el análisis de componentes principales (PCA), el análisis factorial y el análisis lineal discriminante (LDA) (Marcus, 1990; Adams *y col.*, 2004). Recientemente, en búsqueda de mayor robustez de las clasificaciones, varios algoritmos discriminantes han sido desarrollados y ampliamente usados en el campo del aprendizaje automático supervisado (Kumar y Andreou, 1998; Lim *y col.*, 2000; Tarca *y col.*, 2007; Bandos *y col.*, 2009; Rousseeuw y Hubert, 2011; Rosa *y col.*, 2016). No obstante, el LDA es una de las herramientas multivariadas más usadas, la cual consiste en un método canónico que permite hallar la combinación lineal de variables independientes que maximizan las diferencias entre clases establecidas *a priori*, usando distancias Mahalanobis (Fisher, 1936).

La utilidad del LDA en los estudios de morfometría radica en que se pueden determinar las características morfológicas responsables de las diferencias entre dos o más especies o poblaciones de una especie, obtener un valor de la probabilidad de clasificar correctamente las clases predeterminadas y clasificar nuevos individuos a partir de muestras desconocidas (Legendre y Legendre, 1998; Quinn y Keough, 2002). Sin embargo, el LDA requiere el cumplimiento de supuestos que rara vez son comprobados en estudios morfométricos, produciendo sesgo y malas interpretaciones en los resultados. Los más discutidos en la literatura son la normalidad multivariada, la homocedasticidad de las matrices de covarianza, ausencia de colinealidad entre variables predictoras, baja dimensionalidad y la sensibilidad ante datos atípicos multivariados (Williams, 1983; Hastie *y col.*, 1994; Næs y Bjørn-Helge, 2001; Quinn y Keough, 2002).

Una alternativa al LDA es el análisis discriminante cuadrático (QDA), el cual puede ser empleado cuando las matrices de covarianza son heterocedásticas. Pero, al igual que el LDA, postula que las observaciones provienen de una distribución normal multivariada (Lachenbruch y Goldstein, 1979; Nakanishi y Sato, 1985; Quinn y Keough, 2002). Por otra parte, los métodos robustos para el análisis discriminante lineal (rLDA) y

cuadrático (rQDA) utilizan un estimador de gran robustez ante la presencia de valores atípicos y dispersión multivariada, denominado Matriz de Covarianza de Determinante Mínimo, en el cual se utilizan distancias Mahalanobis robustas (Rousseeuw, 1984; Rousseeuw y Van Driessen 1999; Hubert y Debruyne, 2010).

Los algoritmos de regularización como el análisis discriminante regularizado (RDA), han sido satisfactorios cuando el número de parámetros iguala o excede el número de muestras (en estos casos las estimaciones de los parámetros pueden ser inestables, dando lugar a una alta varianza), ya que intentan mejorar las estimaciones de los parámetros al distorsionar sus valores basados en las muestras, haciéndolos más plausibles (Friedman, 1989). La versión robusta del análisis discriminante regularizado (rRLDA) calcula una matriz de covarianza inversa de baja densidad a partir de las observaciones dadas, mediante la maximización de una función de verosimilitud ponderada, en el cual la dispersión se controla mediante un parámetro de penalización y los valores atípicos se tratan con un parámetro de robustez que especifica la cantidad de observaciones para las cuales se maximiza la función de verosimilitud (Gschwandtner *y col.*, 2012). Mientras que el análisis discriminante heterocedástico (HDA), que incluye la regularización en su algoritmo, calcula una matriz a partir de la transformación lineal de las cargas (*loadings*), para la discriminación de clases con matrices de covarianza desiguales (Burget, 2004; Szepannek, 2016).

Otros algoritmos discriminantes incluyen modificaciones del LDA, con aplicaciones para datos con alta dimensionalidad (donde el número de variables es muy grande y el número de observaciones es limitado). Este es el caso del análisis discriminante de alta dimensionalidad (HDDA), el cual supone que los datos de alta dimensionalidad se ubican en diferentes subespacios con baja dimensionalidad (Bouveyron *y col.*, 2007; Berge *y col.*, 2012); mientras que en el análisis discriminante diagonal (DDA), los elementos fuera de la diagonal de la matriz de covarianza de las muestras agrupadas se establecen en cero (Dudoit *y col.*, 2002; Ramey, 2016).

La alta correlación entre variables predictoras (colinealidad), tal como generalmente sucede con las variables morfométricas, representa un problema para el LDA, ya que pueden causar sobreajuste del modelo y afectar el desempeño de clasificación de nuevas observaciones (Næs y Bjørn-Helge, 2001). Para solucionar este problema se han propuesto algoritmos como el análisis discriminante penalizado (PDA), donde la clasificación se modela en el marco de la regresión penalizada usando la aproximación de puntuaciones óptimas, equivalente a una versión asimétrica del análisis de correlación canónica (Hastie *y col.*, 1995). Así mismo, los análisis discriminantes flexible y de mezclas (FDA y MDA, respectivamente) utilizando puntuaciones óptimas, reemplazan la regresión lineal por un

método de regresión no paramétrico, y en el MDA además se ajustan las funciones de densidad Gaussiana mixta a cada clase para facilitar una clasificación efectiva en escenarios de ausencia de normalidad multivariada (Hastie *y col.*, 1994; Hastie y Tibshirani, 1996). Por otro lado, el análisis discriminante de contracción o encogimiento (shrinkage-DA), es otro algoritmo para matrices de alta dimensionalidad que a su vez presentan alta correlación entre las variables predictoras, en el cual se utilizan puntuaciones t ajustadas por correlación (Ahdesmaki *y col.*, 2015).

En los clasificadores globales (LDA, QDA, análisis discriminante logístico, entre otros) un conjunto de parámetros se estima a partir de la muestra total de observaciones, donde la clasificación de las observaciones individuales se obtiene mediante transformaciones de los valores predictores en los que se basan estos parámetros estimados, a diferencia de la aproximación de observaciones específicas donde los algoritmos se adaptan a cada observación al calcular una nueva regla de clasificación para cada una de las mismas (Tutz y Binder, 2005). Este principio es empleado por el análisis discriminante localizado (LocLDA), donde la localización hace necesario construir una regla de decisión individual para cada observación (Roever *y col.*, 2014).

El objetivo principal de este trabajo es comparar el desempeño de estos catorce algoritmos en un caso de estudio particular de datos morfométricos reales de cinco poblaciones de cangrejos (*Aratus pisonii*, Crustacea: Sesarmidae) cuyas diferencias han sido establecidas *a priori*, los cuales no cumplen con los supuestos del análisis lineal discriminante. Este tipo de estudios empíricos son de gran importancia tanto para los investigadores en el área de aprendizaje automático, como para aquellos que utilizan algoritmos de clasificación en sus problemas de estudio, ya que les permite concentrarse en algoritmos más útiles (Wainer, 2016).

MATERIALES Y MÉTODOS

Base de datos morfométricos: la base de datos utilizada en este trabajo proviene de un estudio previo (López-Sánchez *y col.*, 2016), el cual consideró la variabilidad morfológica del cangrejo de mangle *Aratus pisonii* a partir de nueve variables morfométricas estandarizadas por el ancho del caparazón (las siglas utilizadas en López-Sánchez *y col.*, 2016 para estas variables se mantienen en el presente trabajo), recolectados en poblaciones de cinco sectores del estado Falcón, Venezuela. Estos sectores se caracterizaron por presentar diferencias ambientales y estructurales contrastantes: Boca de Ricoa (BR), Tiraya Marina (TIM), Laguna de Tiraya (TIA y TIB) y Tacuato (TAC). La base de datos completa está disponible gratuitamente en el repositorio de datos Mendeley, a través del siguiente enlace: http://dx.doi.org/10.17632/967pm9yfdh.2; para objeto de este estudio se

utilizaron sólo datos de los machos (nTotal= 257, BR= 51, TIM= 39, TIA= 59, TIB= 54, TAC= 54), los cuales a diferencia de las hembras presentaron claras diferencias multivariadas entre las poblaciones evaluadas (López-Sánchez *y col.*, 2016). Información adicional sobre el área de estudio y el gradiente ambiental evaluado puede ser consultada en López *y col.* (2011) y López-Sánchez y Quintero-Torres (2015).

Comprobación de supuestos: la normalidad multivariada dentro de las clases (sectores) se evaluó utilizando el índice de Mardia (Mardia, 1970), del cual se obtiene la significancia estadística de las medidas de asimetría y curtosis. La homocedasticidad de las matrices de covarianza fue comprobada a través de la prueba de Box's M, la cual se basa en una aproximación chi-cuadrado (x^2) (Box, 1949). Para evaluar la colinealidad entre cada par de variables, se realizó un análisis de correlación múltiple de Pearson, acompañado de gráficos de dispersión bivariada.

LDA y desempeño de los algoritmos discriminantes: a partir de la base de datos se obtuvo un LDA y el gráfico de ordenación multivariada, en el cual se pueden detectar la presencia de datos atípicos multivariados. Para evaluar el desempeño de los algoritmos discriminantes (LDA, QDA, rLDA, rQDA, RDA, rRLDA, HDA, HDDA, DDA, PDA, FDA, MDA, *shrinkage*-DA y LocLDA), se obtuvieron las matrices de confusión, utilizando el método de validación cruzada, a partir de las cuales se calculó el error de clasificación, la exactitud con sus intervalos de confianza (± IC 95 %) y la prueba de significancia (P). Esta última se basó en las probabilidades *a posteriori* de cada muestra, asignadas a cada una de las clases evaluadas, al realizar el ajuste a cada uno de los modelos discriminantes. Además se determinó para cada uno de los algoritmos el índice de Kappa, que evalúa la concordancia de clasificación en comparación con lo que se puede obtener simplemente por azar (Cohen, 1960); y para cada clase, la sensibilidad (tasa de verdaderos positivos) y especificidad (tasa de verdaderos negativos). Por haber más de dos clases, los resultados de éstos estadísticos son calculados utilizando una aproximación de "uno versus el resto", donde la primera clase es comparada con la segunda y luego con la tercera y así sucesivamente (Kuhn, 2008).

Programa y librerías: todos los análisis fueron realizados utilizando el lenguaje de programación R (R Core Team, 2016). Las librerías de R empleadas para la comprobación de supuestos fueron: "dprep" para el índice de Mardia (Acuna, 2015), "biotools" para ejecutar el Box's M (Rodrigo da Silva, 2016) y "Hmisc" para la correlación múltiple (Harrel, 2006). El gráfico del LDA se obtuvo con la librería "ggplot2" (Wickham, 2009). Las librerías de los algoritmos discriminantes fueron: "MASS" para LDA y QDA (Venables y Ripley, 2002), "rrcov" para las versiones robustas rLDA y rQDA (Todorov y Filzmoser, 2009), "klaR" para el RDA y LocLDA (Roever *y col.*, 2014), "rrlda" para ejecutar el rRLDA (Gschwandtner *y col.*, 2012), "hda"

para el HDA (Szepannek, 2016), "sparsediscrim" para el DDA (Ramey, 2016), "HDclassif" para el HDDA (Berge *y col.*, 2012), "mda" para MDA, FDA y PDA (Hastie *y col.*, 2016), y "sda" para el *shrinkage*-DA (Ahdesmaki *y col.*, 2015). Las matrices de confusión y estadísticos para evaluar el desempeño de cada algoritmo se obtuvieron utilizando la librería "caret" (Kuhn 2008; Kuhn *y col.*, 2016).

RESULTADOS

Comprobación de supuestos. Se encontró que la base de datos evaluada en este trabajo no cumple con los supuestos del LDA. El índice de Mardia dentro de las clases solo mostró evidencia de normalidad multivariada para el sector TIM, mientras que el resto de los sectores mostraron desviación significativa de asimetría y curtosis (Tabla 1). La prueba Box's M mostró el incumplimiento del supuesto de homocedasticidad de las matrices de covarianza ($X^2_{(approx.)}$= 1035; g.l.= 180; P= $2{,}2e^{-16}$). El análisis de correlación múltiple mostró altos valores de correlación entre varios pares de variables morfométricas, destacando aquellas con valores de r superiores a 0,90 como puede observarse en la relación entre la longitud del caparazón (CL) con la altura del cuerpo (HB) y la relación entre las longitudes y alturas de ambas quelas (RLCh, LLCh, RHCh, LHCh), todos los valores de correlación fueron significativos (P<0,01) (Figura 1).

Tabla 1. Asimetría, curtosis y valor de P del índice de Mardia, calculados a partir de la matriz de datos morfométricos de *Aratus pisonii* en cinco sectores del estado Falcón.

Sector	Asimetría	*P*	Curtosis	*P*
BR	496,70	0,000	9,31	0,000
TIM	175,66	0,27	-0,17	0,86
TIA	612,86	0,000	12,32	0,000
TIB	679,50	0,000	15,29	0,000
TAC	895,86	0,000	27,51	0,000

BR: Boca de Ricoa, TIM: Tiraya Marina, TIA: Tiraya A, TIB: Tiraya B TAC: Tacuato.

LDA. Los resultados del análisis lineal discriminante muestran, en primer lugar, las probabilidades *a priori* para las observaciones, las cuales reflejan la proporción de observaciones para cada clase (BR= 0,20; TIM= 0,15; TIA= 0,23; TIB= 0,21; TAC= 0,21). La mayor cantidad de variación entre las clases es explicada por la primera función discriminante (LD1= 0,93; LD2= 0,06). Los coeficientes estandarizados de las dos primeras funciones discriminantes (Tabla 2), muestran que las variables que poseen mayor influencia en la discriminación entre los sectores fueron la altura del cuerpo (HB), longitud del caparazón (CL) y las alturas de las quelas izquierda y derecha (LHCh y RHCh respectivamente). Por su parte la

ordenación multivariada de las dos primeras funciones discriminantes muestra una buena separación entre los sectores a lo largo del eje x (LD1), con un ligero solapamiento entre el sector TIB con los sectores BR y TIM; aunque las diferencias en la dispersión multivariada entre clases es evidente, así como también la presencia de datos atípicos multivariados para las clases TAC, TIA y TIB (Figura 2).

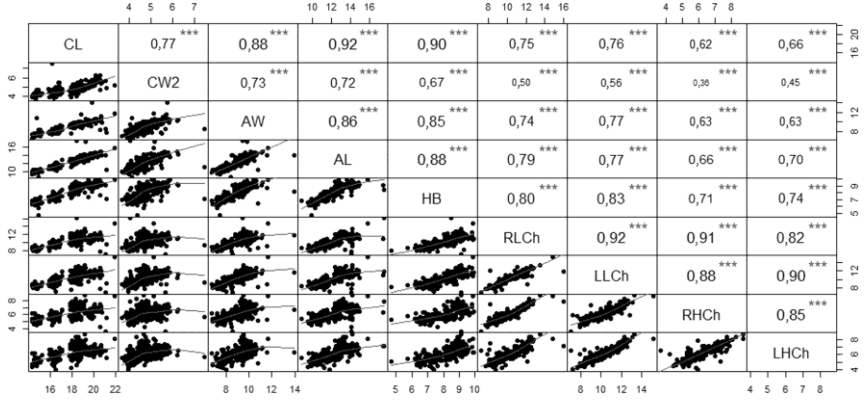

Figura 1. Correlación múltiple y gráficos de dispersión entre variables morfométricas de *Aratus pisonii*. CL: longitud del caparazón; CW2: ancho máximo del caparazón; AW: ancho del abdomen; AL: longitud del abdomen; HB: altura del cuerpo; RLCh: longitud de la quela derecha; LLCh: longitud de la quela izquierda; RHCh: altura de la quela derecha; LHCh: altura de la quela izquierda. *** P < 0,001.

Tabla 2. Coeficientes para cada variable morfométrica de la primera y segunda función discriminante (LD1 y LD2) del modelo LDA.

Variables[a]	LD1	LD2
CL	-1,007	-0,764
CW2	-0,270	-0,907
AW	-0,423	0,027
AL	-0,519	0,056
HB	-1,306	0,799
RLCh	-0,126	0,353
LLCh	0,113	-0,903
RHCh	0,237	1,070
LHCh	-0,309	1,479

[a] Las siglas de las variables fueron definidas en la Figura 1.

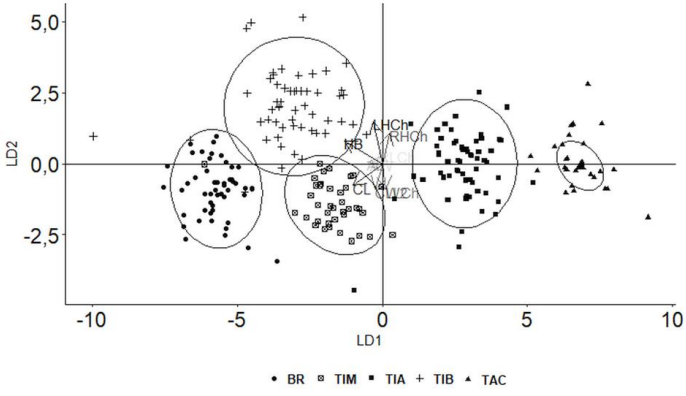

Figura 2. Proyección de las observaciones en la primera y segunda función discriminante (LD1 y LD2), a partir de una base de datos morfométricos de *Aratus pisonii* provenientes de cinco sectores (BR: Boca de Ricoa, TIM: Tiraya Marina, TIA: Tiraya A, TIB: Tiraya B TAC: Tacuato). Se muestran las elipses de confianza al 95%.

Desempeño. En la Tabla 3 se muestra el desempeño de cada algoritmo basado en el error de clasificación a partir de las matrices de confusión, así como el promedio de sensibilidad y especificidad entre las clases. El mejor desempeño se obtuvo con el algoritmo LocLDA, el cual presentó un menor error de clasificación (0,85%) y altos valores de sensibilidad y especificidad promedio (mayores a 0,99), seguido por los algoritmos RDA y DDA ambos con un porcentaje de error de clasificación de 1,96 %. En tercer lugar se pueden observar algoritmos con errores de clasificación mayores a 2 % y menores a 3 % (QDA, MDA y FDA), mientras que el desempeño más bajo se obtuvo con los algoritmos robustos (rLDA, rRLDA y rQDA), de éstos el rQDA destaca por su bajo desempeño con un error de clasificación del 19,60 %. En cuanto a la predicción de las clases, TAC fue la mejor clasificada con 0 % de error en 11 de los algoritmos evaluados (Tabla 3). La exactitud y el índice de Kappa fueron altos (superiores a 0,94 y 0,93 respectivamente) para todos los algoritmos evaluados, con excepción de rQDA (Tabla 4).

DISCUSIÓN

En el presente estudio se han comparado distintos algoritmos discriminantes, aplicados a una base de datos morfométricos multivariados. La base de datos evaluada se caracterizó por no cumplir los supuestos del LDA, ya que presentó ausencia de normalidad multivariada, heterocedasticidad de las matrices de covarianza, presencia de datos atípicos, y altos valores de correlación bivariada entre variables predictoras.

Tabla 3. Matrices de confusión (observados vs pronosticados), error de clasificación (%), sensibilidad y especificidad para cada clase. Los siguientes algoritmos discriminantes se han ordenado de mayor a menor desempeño: localizado (LocLDA), regularizado (RDA), diagonal (DDA), cuadrático (QDA), flexible (FDA), de mezclas (MDA), de alta dimensionalidad (HDDA), lineal discriminante (LDA), penalizado (PDA), discriminante de contracción (shrinkage-DA), heterocedástico (HDA) y las versiones robustas del lineal, regularizado y cuadrático (RLDA, RRLDA y RQDA respectivamente).

Algoritmo	Sector	BR	TIM	TIA	TIB	TAC	Error (%)	Sensibilidad	Especificidad
LocLDA	**BR**	51	0	0	0	0	0,00	0,981	1,000
	TIM	1	38	0	0	0	2,56	1,000	0,995
	TIA	0	0	58	0	1	1,69	1,000	0,995
	TIB	0	0	0	54	0	0,00	1,000	1,000
	TAC	0	0	0	0	54	0,00	0,982	1,000
	Promedio						**0,85**	**0,993**	**0,998**
RDA	**BR**	51	0	0	0	0	0,00	0,944	1,000
	TIM	1	38	0	0	0	2,56	0,974	0,995
	TIA	0	0	58	0	1	1,69	1,000	0,995
	TIB	2	1	0	51	0	5,56	1,000	0,985
	TAC	0	0	0	0	54	0,00	0,982	1,000
	Promedio						**1,96**	**0,980**	**0,995**
DDA	**BR**	51	0	0	0	0	0,00	0,962	1,000
	TIM	1	38	0	0	0	2,56	0,974	0,995
	TIA	0	0	58	0	1	1,69	0,983	0,995
	TIB	1	1	0	52	0	3,70	1,000	0,990
	TAC	0	0	1	0	53	1,85	0,982	0,995
	Promedio						**1,96**	**0,980**	**0,995**
QDA	**BR**	49	2	0	0	0	3,92	0,980	0,990
	TIM	1	38	0	0	0	2,56	0,905	0,995
	TIA	0	1	57	0	1	3,39	1,000	0,990
	TIB	0	1	0	53	0	1,85	1,000	0,995
	TAC	0	0	0	0	54	0,00	0,982	1,000
	Promedio						**2,35**	**0,973**	**0,994**
MDA	**BR**	51	0	0	0	0	0,00	0,944	1,000
	TIM	1	38	0	0	0	2,56	0,950	0,995
	TIA	0	1	56	0	2	5,08	1,000	0,985
	TIB	2	1	0	51	0	5,56	1,000	0,985
	TAC	0	0	0	0	54	0,00	0,964	1,000
	Promedio						**2,64**	**0,972**	**0,993**
FDA	**BR**	50	0	0	1	0	1,96	0,980	0,995
	TIM	1	38	0	0	0	2,56	0,905	0,995
	TIA	0	1	57	0	1	3,39	1,000	0,990
	TIB	0	3	0	51	0	5,56	0,981	0,985
	TAC	0	0	0	0	54	0,00	0,982	1,000
	Promedio						**2,69**	**0,970**	**0,993**
HDDA	**BR**	50	0	0	1	0	1,96	0,962	0,995
	TIM	1	37	0	1	0	5,13	0,949	0,991
	TIA	0	0	57	1	1	3,39	1,000	0,990
	TIB	1	2	0	51	0	5,56	0,944	0,985
	TAC	0	0	0	0	54	0,00	0,982	1,000
	Promedio						**3,21**	**0,967**	**0,992**
LDA PDA	**BR**	51	0	0	0	0	0,00	0,927	1,000
	TIM	1	38	0	0	0	2,56	0,905	0,995
	TIA	0	2	56	0	1	5,08	1,000	0,985
	TIB	3	2	0	49	0	9,26	1,000	0,976
	TAC	0	0	0	0	54	0,00	0,982	1,000
	Promedio						**3,38**	**0,963**	**0,991**

HDA	**BR**	51	0	0	0	0	0,00	0,911	1,000
	TIM	1	37	0	1	0	5,13	0,974	0,991
	TIA	1	0	57	0	1	3,39	0,983	0,990
	TIB	3	1	0	50	0	7,41	0,980	0,981
	TAC	0	0	1	0	53	1,85	0,982	0,995
	Promedio						**3,56**	**0,966**	**0,991**
***shrinkage*-DA**	**BR**	50	1	0	0	0	1,96	0,926	0,995
	TIM	1	38	0	0	0	2,56	0,9048	0,9953
	TIA	0	1	56	0	2	5,08	1,000	0,985
	TIB	3	2	0	49	0	9,26	1,000	0,976
	TAC	0	0	0	0	54	0,00	0,964	1,000
	Promedio						**3,77**	**0,959**	**0,990**
rLDA	**BR**	51	0	0	0	0	0,00	0,927	1,000
	TIM	1	38	0	0	0	2,56	0,864	0,995
	TIA	0	2	54	0	3	8,47	1,000	0,975
	TIB	3	4	0	47	0	12,96	1,000	0,967
	TAC	0	0	0	0	54	0,00	0,947	1,000
	Promedio						**4,80**	**0,948**	**0,987**
rRLDA	**BR**	49	2	0	0	0	3,92	0,925	0,990
	TIM	1	37	0	1	0	5,13	0,881	0,991
	TIA	0	1	54	0	4	8,47	1,000	0,975
	TIB	3	2	0	49	0	9,26	0,980	0,976
	TAC	0	0	0	0	54	0,00	0,931	1,000
	Promedio						**5,36**	**0,943**	**0,986**
rQDA	**BR**	45	0	0	6	0	11,76	0,957	0,971
	TIM	1	26	0	12	0	33,33	0,963	0,944
	TIA	0	0	54	5	0	8,47	0,720	0,973
	TIB	1	1	0	52	0	3,70	0,684	0,989
	TAC	0	0	21	1	32	40,74	1,000	0,902
	Promedio						**19,60**	**0,865**	**0,956**

Tabla 4. Exactitud, intervalos de confianza (IC) al 95%, valor de P e índice de Kappa obtenidos a partir de las matrices de confusión en cada algoritmo.

Algoritmo	Exactitud	95% IC	P	Kappa
LocLDA	0,9922	0,9722 - 0,9991	$2{,}20e^{-16}$	0,990
RDA **DDA**	0,9805	0,9552 - 0,9937	$2{,}20e^{-16}$	0,976
QDA	0,9767	0,9499 - 0,9914	$2{,}20e^{-16}$	0,971
MDA **FDA**	0,9728	0,9447 - 0,989	$2{,}20e^{-16}$	0,966
HDDA	0,9689	0,9396 - 0,9865	$2{,}20e^{-16}$	0,961
LDA **PDA** **HDA**	0,9650	0,9346 - 0,9839	$2{,}20e^{-16}$	0,956
***shrinkage*-DA**	0,9611	0,9296 - 0,9812	$2{,}20e^{-16}$	0,951
RLDA	0,9494	0,9151 - 0,9728	$2{,}20e^{-16}$	0,937
RRLDA	0,9455	0,9103 - 0,9699	$2{,}20e^{-16}$	0,932
RQDA	0,8132	0,7601 - 0,859	$2{,}20e^{-16}$	0,764

Los resultados muestran una desviación muy alta de la asimetría y considerable para la curtosis en cuatro de las cinco clases evaluadas. Reyment (1971), analizó diferentes matrices multivariadas usando datos morfométricos con diferentes tamaños muestrales, así por ejemplo con datos

de tres variables morfométricas de brachiópodos fósiles del Devónico y n= 248, encontró desviación significativa de la asimetría con un valor de 2,71 y para de la curtosis con 20,4; otro ejemplo con datos morfométricos multivariados es el de la salamandra común europea en el cual se analizaron cuatro variables con 299 observaciones, se encontró una desviación significativa de la asimetría multivariada con un valor de 65,45. Sin embargo, en los resultados del presente trabajo la clase TIM, mostró un valor de asimetría superior al de los ejemplos y aún así no fue significativa, lo que claramente indica que los valores de estos parámetros estimados por el índice de Mardia se ven afectados por el tamaño muestral. Adicionalmente al bajo tamaño muestral de TIM, este también presentó menos valores atípicos en comparación con otras clases, característica que tiene una fuerte influencia en los resultados de normalidad multivariada (Oppong y Agbedra, 2016).

Con respecto al supuesto de homocedasticidad de matrices de covarianza, el valor obtenido en la prueba Box's M es extremadamente alto, por ejemplo Joachimsthaler y Stam (1988) reportan diferencias significativas en la dispersión de las matrices de covarianza con valores de Box's M= 22,9 y 71,4 (n = 50); sin embargo es conocida la sensibilidad de esta prueba ante desviaciones de la normalidad multivariada y presencia de valores atípicos (Quinn y Keough, 2002; Huberty y Olejnik, 2006). Asimismo, aunque el análisis de correlación múltiple no mostró colinealidad exacta ($r \geq 0,99$) en la relación entre variables morfométricas, sin embargo los altos valores de correlación ($r > 0,90$) pudieran estar indicando potenciales problemas de multicolinealidad (Quinn y Keough, 2002), como puede observarse en los resultados CL y HB están altamente correlacionadas al igual que LHCh y RHCh por lo que el peso de estos pares de variables sobre las dos primeras funciones discriminantes es redundante.

Aunque la violación de los supuestos es crítica para el LDA, los resultados revelaron un alto porcentaje de varianza explicada y una clara separación entre grupos en la ordenación multivariada. Estos resultados fueron similares a los obtenidos por López-Sánchez *y col.* (2016), en donde se utilizaron los mismos datos, pero se empleó un análisis canónico de coordenadas principales con el propósito de discriminar las clases evaluadas, encontrándose un porcentaje de varianza explicada (96 %) y un patrón observado en la ordenación multivariada similar al obtenido por el LDA en este trabajo. Los resultados indican un buen desempeño del LDA ante la violación de los supuestos evaluados en este caso en particular. Diferentes autores coinciden en que el LDA es robusto ante la ausencia de normalidad multivariada (Joachimsthaler y Stam, 1988; Huberty y Olejnik, 2006; Quinn y Keough, 2002). Sin embargo, el supuesto más importante y que limita la interpretación del LDA es la homocedasticidad de covarianza dentro de los grupos, en especial para la clasificación y obtención del valor de P, ya que es muy sensible a varianzas heterogéneas (Quinn y Keough, 2002; Huberty y Olejnik, 2006).

La evaluación del desempeño de los algoritmos discriminantes aplicados a este caso particular de datos morfométricos, reflejó que el algoritmo LocLDA fue el mejor clasificador, seguido por los métodos RDA y DDA en segundo lugar y QDA en tercer lugar. Algunos estudios han demostrado que los métodos de localización permiten obtener un sesgo de clasificación menor en comparación a los métodos globales. Por ejemplo en un estudio de simulación para un problema de dos clases con dos subclases cada una, se obtuvo que la tasa de error de clasificación del LocLDA y MDA fueron menores al del LDA, en donde el MDA fue elegido como un fuerte competidor del LocLDA (Czogiel *y col.*, 2006); resultados similares fueron obtenidos por Schiffner *y col.* (2012) basados en 26 conjuntos de datos reales y artificiales, donde estos algoritmos presentaron un sesgo menor en comparación con el LDA, aunque con un ligero incremento en las varianzas. Tutz y Binder (2005) señalan que el método de localización para clasificadores globales, en cierto modo es similar a la técnica de aprendizaje automático conocida como Boosting, en la cual un procedimiento global es usado repetidamente con diferentes pesos en las observaciones, y de esta manera la localización en combinación con la reducción apropiada de dimensionalidad es capaz (frecuentemente) de mejorar los algoritmos globales.

Se ha mencionado que el RDA puede considerarse como método intermedio entre LDA y QDA cuando no se cumple con el supuesto de similaridad en las matrices de covarianza (Wu *y col.*, 1996; Guo *y col.*, 2005), sin embargo en algunos casos su desempeño es mejor, tal como se encontró en el presente estudio. Friedman (1989), al proponer este método, evaluó su desempeño mediante simulaciones basado en diferentes escenarios de estructuras de las matrices de covarianza, clases de distribución poblacional y relaciones entre los tamaños muestrales con el número de variables, y encontró que el método de regularización tiene el potencial de incrementar el poder del análisis discriminante, en situaciones donde el tamaño de la muestra es bajo y existe un gran número de variables predictoras. Por su parte el DDA originalmente fue probado en tres bases de datos de expresión génica para la clasificación de tumores (Dudoit *y col.*, 2002), es importante señalar que este tipo de datos generalmente presentan alta dimensionalidad y correlación entre las variables predictoras, en ese estudio los autores encontraron que el DDA y el algoritmo de vecinos más cercanos (Nearest-Neighbor Classifiers) se desempeñaron notablemente bien en comparación con otros algoritmos más complejos. Sin embargo, los autores de este trabajo, aunque destacan la simplicidad del DDA muestran su reserva ante el hecho de que este algoritmo ignora la correlación entre variables predictoras las cuales desde el punto de vista biológico pueden ser relevantes, e indican que ignorarlas no es deseable ya que estas interacciones entre variables (en ese caso genes), pueden contribuir a la distinción entre clases (Dudoit *y col.*, 2002). Con respecto al QDA, se encontró un buen desempeño de este algoritmo, aunque se ha mencionado que el éxito de clasificación es bajo cuando las densidades de probabilidad

de las clases son muy diferentes de la distribución normal (Bose *y col.*, 2015). En el presente trabajo tal como se esperaría para este caso particular de datos morfométricos, se confirma la ventaja del QDA sobre el LDA cuando las matrices de covarianza son muy diferentes.

Por otro lado, el menor éxito de clasificación de los análisis discriminantes robustos (rLDA, rRLDA y rQDA) muestran que, aunque son resistentes a la presencia datos atípicos, su desempeño parece verse afectado por la ausencia de normalidad y la heterocedasticidad de covarianzas. Al respecto se ha mencionado que los modelos robustos no son apropiados cuando la normalidad multivariada no se cumple (Álvarez y Avendaño, 2015).

Si bien en el presente trabajo casi todos los algoritmos discriminantes obtuvieron un buen desempeño (con excepción del rQDA) y aunque no es adecuado dar recomendaciones generales a partir de casos particulares, los resultados permiten destacar el potencial de los algoritmos LocLDA, RDA y DDA ante escenarios similares de desviación de los supuestos de los datos morfométricos evaluados. Es importante reiterar que la evaluación rigurosa de los supuestos es esencial antes de realizar e interpretar un LDA en estudios morfométricos, y a partir de los resultados de esta evaluación elegir un algoritmo discriminante que se ajuste mejor a las características de los datos multivariados.

AGRADECIMIENTOS

Esta investigación fue financiada por el Instituto Venezolano de Investigaciones Científicas (IVIC) a través del proyecto No. 467. Agradecemos a Marcos Manzanares y Elizabeth Gordon, organizadores del Primer Encuentro Venezolano: Métodos de Cuantificación Morfológica, de donde surge la idea de desarrollar la presente investigación. También queremos agradecer a Hendrik Sulbaran Pineda y a dos revisores anónimos por sus sugerencias y observaciones que contribuyeron a mejorar la versión final del manuscrito.

LITERATURA CITADA

Acuna, E. 2015. dprep: Data Pre-Processing and Visualization Functions for Classification. R package. <https://cran.r-project.org.

Adams, D. C., F. J. Rohlf y D. E. Slice. 2004. Geometric morphometrics: Ten years of progress following the "revolution." *Ital. J. Zool.* 71(1): 5-16.

Ahdesmaki, M., V. Zuber, S. Gibb y K. Strimmer. 2015. sda: Shrinkage Discriminant Analysis and CAT Score Variable Selection. R package. <https://CRAN.R-project.org.

Álvarez, H.R. y G. Avendaño. 2015. Comparación de las metodologías de análisis

discriminante robusto y redes neuronales. *Rev. Ontare* 2(2): 35-64.
Bandos, T.V., L. Bruzzone y G. Camps-Valls. 2009. Classification of Hyperspectral Images With Regularized Linear Discriminant Analysis. IEEE Trans. *Geosci. Remote Sens.* 47(3): 862-873.
Berge, L., C. Bouveyron y S. Girard. 2012. HDclassif: An R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data. *J. Stat. Softw.* 46: 1-29.
Bose, S., A. Pal, R. Saharay y J. Nayak. 2015. Generalized quadratic discriminant analysis. *Pattern Recognit.* 48(8): 2676-2684.
Bouveyron, C., S. Girard y C. Schmid. 2007. High-Dimensional Discriminant Analysis. *Commun. Stat. - Theory Methods* 36(14): 2607–2623.
Box, G.E. 1949. A General Distribution Theory for a Class of Likelihood Criteria. *Biometrika* 36(3/4):317–346.
Burget, L. 2004. Combination of speech features using smoothed heteroscedastic linear discriminant analysis. *Proceedings of Interspeech*. pp. 2549–2552. <http://www.fit.vutbr.cz.
Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20 (1):37-46.
Czogiel, I., K. Luebke, M. Zentgraf, y C. Weihs. 2006. Localized linear discriminant analysis. En: *Proceedings of the 30th Annual Conference of the Gesellschaft Für Klassifikation. Advances in Data Analysis.* (Decker, R. y H.J. Lenz, Eds.), Freie Universität Berlin. Springer, Berlin. Pp. 133-140.
Dudoit, S.; J. Fridlyans y T. P. Speed. 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, 97 (457): 77-87.
Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 7: 179-188.
Friedman, J. H. 1989. Regularized Discriminant Analysis. *J. Am. Stat. Assoc.*, 84 (405): 165–175.
Gschwandtner, M., P. Filzmoser, C. Croux y G. Haesbroeck. 2012. Robust Regularized Linear Discriminant Analysis. R package. https://cran.r-project.org.
Guo, Y., T. Hastie y R. Tibshirani. 2007. Regularized discriminant analysis and its application in microarrays. *Biostatistics* 8(1): 86-100.
Harrell, F. 2006. Hmisc package. R package. <http://biostat.mc.vanderbilt.edu.
Hastie, T., A. Buja y R. Tibshirani. 1995. Penalized Discriminant Analysis. *Ann. Stat.* 23(1): 73-102.
Hastie, T. y R. Tibshirani. 1996. Discriminant analysis by Gaussian mixtures. *J. R. Stat. Soc.* 58(1): 155-176.
Hastie, T., R. Tibshirani, y A. Buja. 1994. Flexible Discriminant Analysis by Optimal Scoring. *J. Am. Stat. Assoc.* 89(428): 1255-1270.
Hastie, T., R. Tibshirani, F. Leisch, K. Hornik y B. D. Ripley. 2016. mda: Mixture and Flexible Discriminant Analysis. R package. <https://cran.r-project.org.
Hubert, M. y M. Debruyne. 2010. Minimum covariance determinant. Wiley Interdiscip. *Rev. Comput. Stat.* 2(1): 36-43.
Huberty, C. J. y S. Olejnik. 2006. Applied MANOVA and Discriminant Analysis. Segunda Edición. John Wiley & Sons, Inc., Hoboken, New Jersey.
Joachimsthaler, E. A. y A. Stam. 1988. Four Approaches to the Classification Problem in Discriminant Analysis: an Experimental Study. *Decis. Sci.* 19(2): 322–333.
Kuhn, M. 2008. Building Predictive Models in R Using the caret Package. J. *Stat. Softw.* 28(5): 1-26.
Kuhn, M., J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang y C. Candan.

2016. caret: Classification and Regression Training. R package. <https://cran.r-project.org.
Kumar, N. y A. G. Andreou. 1998. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Commun* 26(4): 283–297.
Lachenbruch, P. A. y M. Goldstein. 1979. Discriminant Analysis. *Biometrics* 35(1): 69-85.
Legendre, P. y L. Legendre. 1998. *Numerical Ecology*. Segunda Edición. Elsevier Science, Amsterdam.
Lim, T., W. Loh y Y. Shih. 2000. A comparison of prediction accuracy, complexity, and training time of thirty three old and new classification algorithms. *Mach. Learn.* 40(3): 203–229.
López-Sánchez, B. y E. Quintero-Torres. 2015. Inversión reproductiva de *Aratus pisonii* (Decapoda: Sesarmidae): diferencias entre hábitats y análisis de rutas. *Rev. Biol. Trop.* 63(2): 385-399.
López-Sánchez, B., E. Quintero-Torres y A. Oliveiras-Durand. 2016. Can contrasting environmental conditions of mangroves induce morphological variability in *Aratus pisonii* (Crustacea: Brachyura: Sesarmidae)? *Sci. Mar.* 80(3): 349-358.
López, B., M. B. Barreto y J. E. Conde. 2011. Caracterización de los manglares de zonas semiáridas en el noroccidente de Venezuela. *Interciencia* 36(12): 888-893.
Marcus, L. F. 1990. Traditional morphometrics. En: *Proceedings of the Michigan morphometrics workshop.* (F. J. Rohlf y F. L. Bookstein, Eds.). The University of Michigan Museum of Zoology, Michigan. Pp. 77–122.
Mardia, K. V. 1970. Measures of Multivariate Skewness and Kurtosis with Applications. *Biometrika* 57(3): 519-530.
Næs, T. y M. Bjørn-Helge. 2001. Understanding the collinearity problem in regression and discriminant analysis. *J. Chemom.* 15(4): 413-426.
Nakanishi, H. y Y. Sato. 1985. The performance of the linear and quadratic discriminant functions for three types of non-normal distribution. *Commun. Stat. - Theory Methods* 14(15): 1181–1200.
Oppong, F. y Agbedra, S. 2016. Assessing Univariate and Multivariate Normality, A Guide For Non-Statisticians. *Math. Theory Model* 6(2): 26–33.
Quinn, G. P. y M. J. Keough. 2002. *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge.
R Core Team. 2016. R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna. Austria <https://www.R-project.org.
Ramey, J. A. 2016. sparsediscrim: Sparse and Regularized Discriminant Analysis. R package. <https://cran.r-project.org.
Reyment, R. A. 1971. Multivariate normality in morphometric analysis. *J. Int. Assoc. Math. Geol.* 3(4): 357–368.
Rodrigo da Silva, A. 2016. biotools: Tools for Biometry and Applied Statistics in Agricultural Science. R package. <https://cran.r-project.org.
Roever, C., N. Raabe, K. Luebke, U. Ligges, G. Szepannek y M. Zentgraf. 2014. Package "klaR": Classification and visualization. R package. <http://www.statistik.tu-dortmund.de.
Rosa, I. M., A. T. Marques, G. Palminha, H. Costa, M. Mascarenhas, C. Fonseca y J. Bernardino. 2016. Classification success of six machine learning algorithms in radar ornithology. *Ibis* 158(1): 28-42.
Rousseeuw, P. J. 1984. Least Median of Squares Regression. *J. Am. Stat. Assoc.* 79(388): 871-880.
Rousseeuw, P. J. y K. Van Driessen. 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics* 41(3): 212–223.

Rousseeuw, P. J. y M. Hubert. 2011. Robust statistics for outlier detection. *WIREs Data Mining Knowl Discov.* 1(1): 73–79.

Schiffner, J., B. Bischl y C. Weihs. 2012. Bias-Variance Analysis of Local Classification Methods. En: *Challenges at the Interface of Data Analysis, Computer Science, and Optimization. Studies in Classification, Data Analysis, and Knowledge Organization.* (Gaul, W.; A. Geyer-Schulz, L. Schmidt-Thieme y J. Kunze, Eds.). Springer, Berlin.

Szepannek, G. 2016. hda: Heteroscedastic Discriminant Analysis. R package. <https://cran.r-project.org.

Tarca, A. L., V. J. Carey, X. Chen, R. Romero y S. Drăghici. 2007. Machine learning and its applications to biology. *PLoS Comput. Biol.* 3(6): e116.

Todorov, V. y P. Filzmoser. 2009. An object oriented framework for robust multivariate analysis. *J. Stat. Softw.* 32(3): 1-47.

Tutz, G. y H. Binder. 2005. Localized classification. *Stat. Comput.* 15(3): 155-166.

Venables, W. N. y B. D. Ripley. 2002. *Modern Applied Statistics with S.* Cuarta Edición. Springer, New York.

Wainer, J. 2016. Comparison of 14 different families of classification algorithms on 115 binary datasets. ArXiv e-prints 1606.00930. arXiv:1606.00930.

Wickham, H. 2009. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag, New York.

Williams, B. K. 1983. Some observations of the use of discriminant analysis in ecology. *Ecology* 64(5): 1283–1291.

Wu, W., Y. Mallet, B. Walczak, W. Penninckx, D. Massart, S. Heuerding y F. Erni. 1996. Comparison of regularized discriminant analysis linear discriminant analysis and quadratic discriminant analysis applied to NIR data. *Anal. Chim. Acta* 329 (3): 257-265.