Ciencia en Desarrollo, Vol. 8 No. 2
ISSN 0121-7488 – Julio-Diciembre de 2017, pp. 9-24

Una metodología para el tratamiento de la multicolinealidad a través del escalamiento multidimensional

A methodology for treating multicolinearity via multidimensional scaling

Sara C. Guerrero [a*]
Oscar O. Melo [b]

Recepción: 30 de julio de 2016
Aceptación: 01 de mayo de 2017

Resumen

Se presenta el escalamiento multidimensional como estrategia alternativa para tratar el problema de multicolinealidad en el análisis de regresión múltiple, cuando las variables regresoras son cualitativas, cuantitativas o mixtas (cuantitativas y cualitativas) y la variable respuesta es continua. El propósito es obtener la matriz de coordenadas principales usando como métrica la distancia de Gower si las variables predictoras son mixtas o, en caso contrario, otra distancia de tipo Euclideana, y a partir de esta matriz estimar el modelo de regresión. Para observar las bondades del método propuesto, se realizan dos casos de simulación: el primero sin presencia de multicolinealidad y el segundo con presencia de multicolinealidad. Se muetran dos casos de aplicación analizados por [46] mediante regresión múltiple, en los casos simulados y en las aplicaciones se utilizó el paquete estadístico R. Los resultados de las simulaciones y aplicaciones se comparan con la regresión múltiple clásica y la basada en componentes principales. El análisis propuesto es una alternativa de modelamiento que corrige la colinealidad y permite trabajar con variables explicativas sin pérdida de información; además, esta técnica al transformar las variables originales en coordenadas, en su modelamiento logra ocultar el efecto de las variables observadas, de manera que no se manipulen los resultados.

***Palabras clave*:** Colinealidad, Coordenadas Principales, Distancia de Gower, Regresión Múltiple, Componentes Principales.

Abstract

We present the multidimensional scaling analysis as an alternative strategy to treat the multicollinearity problem in the multiple regression analysis, when the regressor variables are qualitative, quantitative or mixed (quantitative and qualitative) and the response variable is continuous. Our purpose is to obtain the matrix of the principal coordinates, using as a metric the Gower distanc when the predictives variables are mixed, or otherwise, the researcher must select an appropriate Euclidean distance and with this matrix to estimate the regression model. To observe the kindness of the proposed method, two cases of simulation are realized: the first one without presence of multicolinearity and the second one with presence of multicolinearity. Two application cases are illustrated, which were analyzed by [46] using multiple regressions. In both cases simulated and in the applications, the R package was used. The results of the simulations and applications are compared with the classical multiple regression and regression based on principal component . The analysis strategy proposal is an alternative modeling that corrects collinearity, and allows work with predicted variables without loss of information, Additionally, this technique when transforming the original variables into coordinates, in its modeling hides the effect of the observed variables, so that the results are not manipulated.

***Keywords*:** Collinearity, Principal coordinates, Gower Distance, Multiple Regression, Principal Components.

[a] Departamento de Matemáticas y Física, Facultad de Ciencias Básicas e Ingeniería, Universidad de los LLanos, Villavicencio, Colombia.

[*] Autor de correspondencia: sguerrero@unillanos.edu.co

[b] Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Sede Bogotá, Colombia. oomelom@unal.edu.co

1. Introducción

Un tema de creciente interés es el estudio y modelamiento de la asociación entre variables. En las diferentes áreas del conocimiento como biología, econometría, ecología, medicina, psicología, en general las ciencias humanas, entre otras, surgen situaciones donde el investigador está interesado en ajustar un modelo incluyendo cada una de las variables regresoras observadas. Con frecuencia, al modelar este tipo de información, el cumplimiento de las premisas necesarias para ajustar una regresión lineal múltiple no se satisfacen, en particular, el problema de colinealidad entre las variables regresoras.

En la literatura especializada existen métodos alternativos para tratar el modelamiento cuando hay presencia de multicolinealidad entre los predictores. La multicolinealidad implica la existencia de una dependencia lineal entre las variables regresoras (columnas de la matriz del modelo), trayendo consigo problemas de no estimación única de los parámetros y por lo tanto, una falsa relación entre las variables explicativas y la variable respuesta. En este sentido se tiene que la matriz diseño no es de rango completo y por lo tanto, no cualquier hipótesis que quiera plantearse es estimable y juzgable. Además, las varianzas de los estimadores son muy grandes, al efectuar contrastes individuales no se rechaza la hipótesis nula, mientras que al realizar contrastes conjuntos si, los coeficientes estimados serán muy sensibles ante pequeños cambios en los datos y finalmente, un coeficiente de determinación elevado.

Al respecto, [1] plantean la regresión Ridge para corregir los problemas de multicolinealidad, donde la estimación de los parámetros presentan menor sesgo que si la estimación se hiciere por mínimos cuadrados, esta metodología ha sido estudiada por [2, 3]. [4] estudió la bondad de la regresión ridge aplicada en datos de lluvias en el Rio Temporal en México.

El análisis de componentes principales (ACP) es otra alternativa para corregir la multicolinealidad. La metodología propuesta por [5] ha sido aplicada en diversos estudios [6, 7, 8, 9]. El ACP maximiza la correlación entre las variables originales, encontrando nuevas variables incorrelacionadas que eliminan las complicaciones generadas por la colinealidad, donde cada nueva variable se correlaciona con máximo un componente principal. La ventaja con respecto a otros métodos es que cada una de las variables objeto de estudio se involucran en el análisis. [10] analizaron la regresión ridge y la regresión sobre componentes principales, como técnicas efectivas para atenuar la colinealidad y para describir con exactitud y precisión los estimadores en el modelo de regresión lineal múltiple.

[11] comparan la eficiencia de la regresión basada en el ACP y la regresión desde mínimos cuadrados parciales como alternativas para solucionar los inconvenientes de colinealidad. En regresión logística, [12] a partir de un estudio sobre pruebas para detectar demencia, propusieron un estimador para tratar la colinealidad y la separación de los datos. Posteriormente, [13] mediante procesos de simulación analizaron cómo se afectan los estimadores que tratan la multicolinealidad mediante los procedimientos ridge iterativos y la separación de los datos.

Otros estudios tratan la multicolinealidad eliminando variables [14, 15, 16], o incluyendo información externa a los datos originales [6]. [17] recurren a la transformación de las variables, a la técnica stepwise y a la realización de todas las regresiones posibles, métodos que permiten hacer la selección de los predictores.

[18] plantean una metodología fundamentada en el ajuste de múltiples modelos de regresión con suficiente poder explicativo, donde se selecciona un único modelo, el que tenga un mejor nivel de predicción ante la presencia de un gran número de variables explicativas con una fuerte multicolinealidad. [19] estudiaron el efecto de la colinealidad en presencia de valores atípicos por medio del estimador de mínimos cuadrados ordinarios y evaluaron el comportamiento del error cuadrático medio como una alternativa para identificar los regresores colineales.

En estudios más recientes, [20] plantearon como método alternativo para corregir el problema de colinealidad el uso del análisis de conglomerados. [21] sugieren omitir la variable que es teóricamente menos importante para el investigador o la que presenta más valores faltantes o de alguna manera es menos satisfactoria para el análisis, o crear nuevas combinaciones de variables con diversas categorías o con escalas más elaboradas. [22] realizaron una investigación sobre la prevalencia y pronóstico en pacientes con infarto de miocardio en alto riesgo, trataron

el problema de la colinealidad excluyendo las variables que estaban más relacionadas. [23] para atenuar la multicolinealidad en un estudio de datos meteorológicos, crearon una nueva variable estableciendo el diferencial entre temperatura del aire y la temperatura del suelo.

Para tratar el problema de la colinealidad, se plantea una metodología que tiene en cuenta las técnicas de regresión múltiple trabajadas a través del análisis de escalamiento multidimensional (EM), cuando las variables explicativas son cuantitativas, cualitativas o mixtas y la variable dependiente es continua. La estrategia está fundamentada en la técnica de regresión basada en distancias (RBD) propuesta por [24] y en este artículo es estudiada como alternativa útil para corregir la colinealidad. Inicialmente se halla la matriz de distancias calculada a partir del coeficiente de similaridad de Gower [25] o la distancia Euclideana adecuada de acuerdo al tipo de variables intervinientes en el análisis. Posteriormente, se aplica la descomposición espectral a fin de obtener la matriz de coordenadas principales (obtenidas en el EM), y con ella, estimar la ecuación de regresión y los parámetros.

El EM es una técnica multivariada usada frecuentemente con fines descriptivos o de clasificación o en ocasiones es empleada para disminuir la dimensionalidad de un conjunto de datos. En este trabajo es utilizada como estrategia para corregir el problema de multicolinealidad entre los predictores en el análisis de regresión múltiple, donde el número de regresores (coordenadas principales) puede en ocasiones ser mayor al número de variables incluidas en modelamiento, pues la selección de éstos dependerá del porcentaje de variabilidad explicada por las coordenadas principales.

La metodología planteada es una alternativa que permite corregir la multicolinealidad en los predictores, garantizando la no pérdida de información al incluir todas las variables de interés para el investigador. Además, se propone como herramienta de análisis que oculta el efecto de las variables cuando se realizan predicciones, de manera que al aplicar el modelo, este no puede manipularse, puesto que no se conoce con certeza cuál de las variables regresoras influye más sobre la variable respuesta. Por ejemplo es aplicable en entidades bancarias o gubernamentales cuando hay que tomar decisiones en función de un sinnúmero de variables; el profesional encargado de realizar los análisis no podrá identificar cuáles son las variables que más influyen sobre la variable respuesta, dando un parte de garantía a las instituciones en la toma de la decisión adecuada.

Por otro lado, aunque una de las soluciones al problema de multicolinealidad es eliminar una o más de las variables que ocasionan la colinealidad, en muchas investigaciones prácticas el investigador no desea eliminar alguna de las variables explicativas ya que estima conveniente dejarlas todas en su análisis, inclusive no es de su interés describir con precisión la relación que hay entre las variables explicativas y su variable respuesta; siendo para éste más importante obtener una buena predicción sin perder información. Por ello, un método de regresión como el basado en distancias o el EM, resulta ser una muy buena alternativa para solucionar esta clase de problemas, permitiéndole al investigador considerar en su análisis toda la información disponible e inclusive obteniendo mejores predicciones ya que puede incluir más coordenadas que las variables explicativas planteadas para la realización de un modelo de regresión lineal múltiple.

Por lo tanto, esta metodología ofrece mejores predicciones, ya que se tienen más coordenadas que variables explicativas, las cuales son multicolineales. Además, debido a que en las variables objeto de análisis hay variables cualitativas, continuas o discretas; otra ventaja de esta estrategia de análisis es que no es necesario recurrir al uso de variables indicadoras para ajustar la ecuación de regresión. No obstante, cuando en un estudio se toman muestras de tamaño muy grande se presenta una limitante o dificultad, pues al estimar las coordenadas principales puede tenerse tantas variables como individuos, esto generaría modelos superparametrizados causando inconvenientes en los procesos de estimación y en algunas ocasiones no es posible realizarlos.

2. Consideraciones generales

El EM busca representar las proximidades de los individuos en un espacio de dimensionalidad mínima. El EM se originó a partir de los trabajos en coordenadas principales por [26] y que posteriormente, [27] demostró que si se conocía la ordenación de las distancias entre los puntos (individuos), existe un espacio euclídeo que permite reproducir la ordenación

original de los elementos. “En 1966, Gower propuso el método de análisis de coordenadas principales, que puede considerarse un método métrico de escalamiento multidimensional, y que evita resolver los procesos iterativos de las técnicas no métricas” [28]. [29] aplicaron EM bidimensional, el coeficiente de similitud de Gower y el ACP en el ordenamiento de material genético, donde el EM permite adecuar no sólo la ordenación sino también la clasificación de las accesiones (variable de estudio). [30] aplicó EM a variables mixtas usando coeficiente de similitud de Gower [25].

El análisis de información basado en distancias es ampliamente utilizado en técnicas de ACP, análisis de conglomerados, análisis de coordenadas principales, análisis discriminante, estudios para analizar proximidades y en regresión. Al respecto, [31] en una investigación en hongos usó métodos multivariados basados en distancias para estudiar el agrupamiento de aislamientos del *Colletotrichum spp*. en función de características morfológicas y culturales haciendo uso del coeficiente de similitud de Gower.

La aplicación de distancias en regresión es un método planteado por [32]. [24] proponen el método de RBD utilizando métricas para trabajar variables regresoras cualitativas y continuas. Posteriormente, [33] la emplearon en el modelamiento basado en distancias cuando las variables explicativas son mixtas haciendo uso del coeficiente de similitud de Gower. [34] la aplicaron a regresión en el caso lineal y no lineal, donde la variable predictora es continua y las variables independientes son de tipo mixto o continuo.

[35] aplicó la regresión basada en distancias en regresión clásica, no lineal y en análisis discriminante. [36] propusieron una metodología para seleccionar los predictores en modelos de regresión. [37] la aplicaron en el modelamiento de variables aleatorias continuas y categóricas haciendo uso de los modelos lineales generalizados espaciales mixtos. [38] abordaron el modelamiento de datos longitudinales usando distancias. Igualmente, [39] propusieron una metodología en regresión beta basada en distancias donde las variables predictoras son mixtas y su aplicabilidad trabajada con información faltante, sin tener que recurrir a imputar datos faltantes.

La estrategia de análisis planteada, emplea la metodología seguida en EM, teniendo en cuenta la distancia de Gower aplicable a variables mixtas, pero en situaciones donde haya otro tipo de variable, la literatura especializada define las distancias respectivas [35, 40], en la sesión 3 se citan algunas. El propósito inicial es obtener la matriz de coordenadas principales (Z) trabajada por [41, 25] y en función de ella ajustar la ecuación de regresión. Se parte de los datos observados, luego se transforman teniendo en cuenta las distancia entre los individuos i e i' ($d(i,i')$), a fin de hallar la matriz de coordenadas principales. Posteriormente, a partir de ésta ajustar el modelo, corrigiéndose los problemas de colinealidad y además, involucrando cada una de las variables de interés para el investigador.

3. Metodología fundamentada en el escalamiento multidimensional

La matriz de datos $\boldsymbol{X}$ de orden $n \times p$ dada en (1), se conforma al observar p variables explicativas asociadas a n individuos; puede darse el caso en que las variables sean cuantitativas o cualitativas o mixtas, teniéndose p_1 variables continuas, p_2 variables dicotómicas y p_3 variables categóricas ($p_1+p_2+p_3=p$). A partir de la matriz de datos, se define la matriz de distancias $D_{n\times n}$, entonces el propósito es representar esta matriz mediante un conjunto de variables ortogonales llamadas coordenadas principales; de manera que las distancias sean lo más próximas posibles a las distancias o disimilaridades de la matriz original,

$$\boldsymbol{X} = \left(\boldsymbol{x}_1^t, \boldsymbol{x}_2^t, ..., \boldsymbol{x}_i^t, ..., \boldsymbol{x}_{i'}^t, ..., \boldsymbol{x}_n^t\right) \qquad (1)$$

donde cada $\boldsymbol{x}_i^t$ corresponde al i-ésimo vector fila de la matriz $\boldsymbol{X}$.

En la matriz $\boldsymbol{D} = (\delta_{ii'}) = (d(i,i'))$, cada $\delta_{ii'}$ corresponde a la distancia entre los individuos i y i', que satisface las siguientes propiedades: $d(i,i') \cong 0$ si $\boldsymbol{x}_i \cong \boldsymbol{x}_{i'}$, y si además, $d(i,i') \leq d(i',k) + d(k,i')$ se dice que la distancia es una métrica. Para el caso de estudio, si las variables son mixtas, las distancias se estimarán a través del coeficiente de similaridad propuesto por [25] y se define:

$$s_{ii'} = \frac{\sum_{k=1}^{p_1}\left(1 - \frac{|x_{ik} - x_{i'k}|}{r_k}\right) + c_{1ii'} + m_{ii'}}{p_1 + (p_2 - c_{0ii'}) + p_3} \qquad (2)$$

donde $c_{1ii'}$ número coincidencias de la forma (1,1) y $c_{0ii'}$ número de coincidencias (0,0), para las p_2 varia-

bles dicotómicas, $m_{ii'}$ es el número de coincidencias para las p_3 variables cualitativas y r_k es el rango (o distancia) para la k-ésima variable cuantitativa.

En el caso donde las p variables sean de tipo binario las similaridades entre dos individuos i e i' se definen a través de los índices:

$$s_{ii'} = \frac{c_{1ii'} + c_{0ii'}}{p} \quad \text{(Sokar-Michene)} \tag{3}$$

$$s_{ii'} = \frac{c_{1ii'}}{c_{1ii'} + c_{2ii'} + c_{3ii'}} \quad \text{(Jaccard)}$$

siendo $c_{0ii'}$, $c_{1ii'}$, $c_{2ii'}$ y $c_{3ii'}$ las frecuencias de (0,0), (1,1), (1,0), y (0,1), respectivamente, verificándose $p = c_{0ii'} + c_{1ii'} + c_{2ii'} + c_{3ii'}$.

Cuando las variables predictoras sean de tipo continuo, la distancia entre los individuos se halla a partir de la distancia Euclídea, Mahalanobis, Manhatan, o valor absoluto, entre otras, de acuerdo a las características de las variables (escalas de medición o correlación).

$$\delta_{ii'} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{i'k}\right)^2} \quad \text{(Euclídea)} \tag{4}$$

$$\delta_{ii'} = \sqrt{\left(\boldsymbol{x}_i - \boldsymbol{x}_{i'}\right)^t \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{x}_i - \boldsymbol{x}_{i'}\right)} \quad \text{(Mahalanobis)}$$

donde $\boldsymbol{x}_i$ y $\boldsymbol{x}_{i'}$ son vectores asociados al i-ésimo e i'-ésimo individuo, respectivamente y Σ la matriz de varianzas y covarianzas.

El punto de partida de la estrategia de análisis para remediar el problema de colinealidad es estimar la matriz de distancias euclidiana $\boldsymbol{D}$ [25, 32, 42], donde cada $\delta_{ii'}$ puede ser transformado a partir del coeficiente de similaridad:

$$\delta_{ii'} = \sqrt{1 - s_{ii'}} \tag{5}$$

En general al conformar la matriz de similaridades, los elementos de su diagonal pueden ser $s_{ii'} \neq 1$. La transformación que permite pasar de similaridad a distancia es:

$$\delta_{ii'} = \sqrt{s_{ii} + s_{i'i'} + 2s_{ii'}} \tag{6}$$

Una vez estimado $\boldsymbol{D}$, se define la matriz $\boldsymbol{A} = -\frac{1}{2}\boldsymbol{D}^{(2)}$, donde cada $a_{ii'} = -\frac{\delta_{ii'}^2}{2}$. Luego se procede a aplicar doble centrado sobre $\boldsymbol{A}$, se conforma la matriz $\boldsymbol{B}_{n\times n}$ simétrica y semidefinida positiva [43, 44, 35, 45] puesto que ha sido construida sobre una distancia Euclídea.

$$\boldsymbol{B} = \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{A}\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right) \tag{7}$$

donde $\boldsymbol{J}$ es la matriz de unos, $\boldsymbol{I}$ es la matriz identidad y $\boldsymbol{B}$ es de rango m, $(m \leq n-1)$. Como $\boldsymbol{B}$ ha sido construida sobre una métrica Euclídea es posible obtener su descomposición espectral:

$$\boldsymbol{B} = \boldsymbol{L}\Lambda\boldsymbol{L}^t \tag{8}$$

donde $\boldsymbol{L}$ es la matriz de vectores propios de $\boldsymbol{B}$, $\Lambda_{n\times npena}$ es la matriz diagonal de los valores propios de $\boldsymbol{B}$. Además, se verifica que $\boldsymbol{B} = \boldsymbol{Z}\boldsymbol{Z}^t$ y $\boldsymbol{Z}^t\boldsymbol{Z} = \Lambda$. Por conveniencia la matriz de valores propios se ordenan en forma descendente $\lambda_1 \geq \lambda_2 \geq, ..., \geq \lambda_m \geq 0$, y finalmente, se obtiene la matriz de coordenadas principales:

$$\boldsymbol{Z} = \boldsymbol{L}\Lambda^{\frac{1}{2}} \tag{9}$$

Obteniendo la matriz de coordenadas principales, se procede a plantear el modelo de regresión.

4. Planteamiento del modelo

Suponga que se tiene en general, un conjunto de datos con p variables predictoras mixtas, (p_1 continuas, p_2 dicotómicas y p_3 variables cualitativas con más de dos estados) y la variable respuesta observada de tipo continuo. El modelo de regresión clásico expresado en términos de las p variables regresoras corresponde:

$$y_i = \theta_0 + \sum_{j=1}^{p} \theta_j x_{ij} + \varepsilon_i, \qquad i = 1, 2, ..., n \tag{10}$$

donde θ_0 es el intercepto, $\theta_1, \theta_2, ..., \theta_p$ son los parámetros desconocidos asociados a las variables de los datos originales y ε_i es el término del error $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. Altenativamente, expresándolo matricialmente, se tiene:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\theta} + \varepsilon \tag{11}$$

donde $\boldsymbol{Y}_{n\times 1} = (y_1, y_2, ..., y_n)^t$, $\boldsymbol{X} = (\boldsymbol{1}, \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_p)$ con $\boldsymbol{1}$ un vector de unos, $\boldsymbol{\theta} = (\theta_0, \theta_1, ..., \theta_p)$ y $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$. Bajo condiciones de rango completo de la matriz $\boldsymbol{X}$, el vector de parámetros estimados esta dado por $\hat{\boldsymbol{\theta}} = \left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{Y}$.

Al plantear la matriz de coordenadas principales definida en (9), las columnas de la matriz $\boldsymbol{Z}$,

$\boldsymbol{Z}_1, \boldsymbol{Z}_2, ..., \boldsymbol{Z}_m$ son las nuevas variables predictoras interpretadas en términos de las coordenadas principales ($m =$ rango de $\boldsymbol{B}$). El modelo de RBD se define como:

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j z_{ij} + \xi_i \qquad (12)$$

donde β_0 es el intercepto, β_j es el coeficiente de regresión para la i-ésima coordenada principal, $z_{ii'}$ denota el i'-ésimo valor de la coordenada principal para el i-ésimo individuo y ξ_i es el termino del error, $\xi_i \sim N(0, \sigma^2_\xi)$.

El modelo (12) matricialmente se expresa como:

$$\boldsymbol{Y} = \beta_0 \mathbf{1} + \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\xi} \qquad (13)$$

donde $Y_{n\times 1}$ es el vector de observaciones de la variable dependiente, $\boldsymbol{\beta}_{m\times 1}$ es el vector de parámetros desconocidos, $\boldsymbol{Z}$ de orden $n \times m$ en sus columnas se encuentran las coordenadas principales, y $\boldsymbol{\xi}$ el vector de errores aleatorios que se distribuye $N(\mathbf{0}, \sigma^2_\xi \boldsymbol{I})$.

Teniendo en cuenta que el número de coordenadas puede ser muy grande, supóngase el tamaño adecuado es tal que $\boldsymbol{Z} = \left(\boldsymbol{Z}_{(q)}, \boldsymbol{Z}_{(m-q)}\right)$, siendo q la dimensión sugerida para el modelo, las $m-q$ columnas restantes de $\boldsymbol{Z}$ se eliminan ya que sus valores propios tienden a cero. Por lo tanto, el modelo de RBD asociado es de la forma:

$$\boldsymbol{Y} = \beta_0 \mathbf{1} + \boldsymbol{Z}_{(q)}\boldsymbol{\beta}_{(q)} + \boldsymbol{\xi}_{(q)} \qquad (14)$$

Debido a que $\boldsymbol{Z}_{(q)} = (\mathbf{1}, \boldsymbol{Z}_1, \boldsymbol{Z}_2, ..., \boldsymbol{Z}_q)$ son los vectores propios de $\boldsymbol{B}$ con valores propios asociados $0, \lambda_1, \lambda_2, ..., \lambda_q$, respectivamente, el criterio para seleccionar las coordenadas se presenta en la sección 4.1.

La estimación de los parámetros por mínimos cuadrados para el modelo (14) de acuerdo a lo planteado por [24] es:

$$\hat{\beta}_0 = \overline{Y}, \qquad \hat{\boldsymbol{\beta}}_{(q)} = \boldsymbol{\Lambda}^{-1}_{(q)} \boldsymbol{Y} \qquad (15)$$

con $\boldsymbol{\Lambda}_{(q)} = diag(\lambda_1, \lambda_2, ..., \lambda_q)$.

El coeficiente de determinación, para el modelo (14) se obtiene mediante la siguiente expresión de acuerdo a lo planteado por [24], [25] y [34]:

$$R^2_q = \sum_{j=1}^{q} r^2\left(\boldsymbol{Y}, \boldsymbol{Z}_j\right) \qquad (16)$$

donde $r^2\left(\boldsymbol{Y}, \boldsymbol{Z}_j\right)$ es el coeficiente de correlación simple entre $\boldsymbol{Y}$ y las variables predictoras en términos de las coordenadas principales.

El modelo de regresión planteado en (12) y (13) es el modelo completo, en éste se involucran cada una de las columnas de $\boldsymbol{Z}$ teniendo en cuenta el rango de $\boldsymbol{B}$. Dado que la matriz de coordenadas principales es de orden $n \times m$ ($m \leq n-1$). Cuando n es muy grande, el proceso de estimación de los parámetros se dificulta o en ocasiones no se puede realizar. Además, el coeficiente de determinación tiende a uno en presencia de muchos parámetros, siendo no necesariamente todos significativos, generando inconvenientes ya que el número de coordenadas principales (nuevas variables regresoras) puede ser tan grande como $n-1$ obteniéndose un modelo superparametrizado. Para evitar este problema es necesario definir el número de coordenadas adecuado para el análisis a fin de estimar (15).

4.1. Determinación de la dimensionalidad de las coordenadas principales

Para seleccionar el número de coordenadas adecuadas al modelo definido en la ecuación (13), [42] cita que a partir de los q valores propios asociados a las coordenadas escogidas, se debe medir el grado de ajuste de las m_q coordenadas, en términos del porcentaje de la variabilidad explicada utilizando, por ejemplo,

$$m_q = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{m} |\lambda_i|} \times 100\,\%, \qquad q = 1, 2, ..., m \quad (17)$$

Una segunda alternativa para seleccionar las columnas de la matriz de coordenadas principales es: tomar las variables que estén más correlacionadas con la variable respuesta [33], es decir $r^2(\boldsymbol{Y}, \boldsymbol{Z}_1) > r^2(\boldsymbol{Y}, \boldsymbol{Z}_2) > ... > r^2(\boldsymbol{Y}, \boldsymbol{Z}_m)$, donde el coeficiente de correlación entre $\boldsymbol{Y}$ y $\boldsymbol{Z}_j$ se define como:

$$r^2(\boldsymbol{Y}, \boldsymbol{Z}_j) = \frac{(\boldsymbol{Y}^t \boldsymbol{Z}_j)^2}{n\lambda_j S^2_Y}, \qquad j = 1, 2, ..., m \qquad (18)$$

donde S^2_Y es la varianza de la variable respuesta $\boldsymbol{Y}$.

Una tercera opción para elegir las q primeras coordenadas requiere realizar una partición de la matriz $\boldsymbol{Z}$:

$$\boldsymbol{Z} = \left(\boldsymbol{Z}_{(q)}, \boldsymbol{Z}_{(m-q)}\right) \qquad (19)$$

donde $\boldsymbol{Z}_{(q)}$ contiene las primeras columnas de $\boldsymbol{Z}$ asociados a los q primeros vectores propios de $\boldsymbol{B}$ ordenados respecto a sus valores propios. El objetivo es construir la representación gráfica de la secuencia de puntos de la forma $(q, 1-C(q))$, $C(q)$ se define en la ecuación (20) y mide la predictibilidad de las q dimensiones ponderadas por los correspondientes valores propios.

$$C(0) = 0, \qquad C(q) = \frac{\sum_{j=1}^{q} r^2(\boldsymbol{Y}, \boldsymbol{Z}_j)\lambda_j}{\sum_{j=1}^{m} r^2(\boldsymbol{Y}, \boldsymbol{Z}_j)\lambda_j} \quad (20)$$

donde $m = \text{rango}(\boldsymbol{B})$ y λ_j es el c-ésimo valor propio asociado a $\boldsymbol{Z}_j$ con $j = 1, ..., m$. Por lo tanto, las coordenadas $\boldsymbol{Z}_{q+1}, ..., \boldsymbol{Z}_m$ deben ser eliminadas. Este procedimiento se considera adecuado cuando n es lo suficientemente grande (ver [33]).

$C(q)$ mide la predictibilidad de las q primeras coordenadas, la selección de las coordenadas se hace representando en una gráfica la secuencia de puntos $(q, 1-C(q))$ con $q = 0, 1, ..., m^* < m$ donde m^* es tal que $1-C(q)$ esté muy próxima 0, entonces se descartaran los puntos que estén muy cercanos al eje horizontal. Teniendo en cuenta el valor inicial de $C(0) = 0$, interpretado como la falta de predictibilidad, la dimensión de q es aceptada o rechazada según si r_q^2 o λ_q sean grandes o pequeños.

4.2. Verificación multicolinealidad

Los valores propios de la matriz $\boldsymbol{X}^t\boldsymbol{X}$ permiten medir el grado de colinealidad de los datos. Si hay una o más dependencias casi lineales, una o más raíces características serán pequeñas o cercanas a cero, entonces, uno o más vectores propios pequeños implicarán dependencia entre las columnas de la matriz (1). Al respecto puede inspeccionarse el número de condición (IC):

$$IC = \frac{\lambda_{max}}{\lambda_{min}} \quad (21)$$

De acuerdo a [16, 46], sí $IC < 100$ no hay problemas de multicolinealidad, valores de $100 < IC < 1000$ implican multicolinealidad moderada a fuerte, y en caso que $IC > 1000$ es indicio de una fuerte multicolinealidad. Los índices de condición (IC_j) de la matriz $\boldsymbol{X}^t\boldsymbol{X}$, se hallan a partir de la expresión (22). Por cada índice de condición grande, cuando $IC_j \geq 1000$, indica la cantidad de dependencias casi lineales en $\boldsymbol{X}^t\boldsymbol{X}$.

$$IC_j = \frac{\lambda_{max}}{\lambda_j}, \qquad j = 1, 2, ..., p \quad (22)$$

Una segunda alternativa para detectar multicolinealidad es utilizar el factor de inflación de la varianza (FIV), el cual cuantifica si para cada término del modelo es confiable su estimación. De acuerdo a [16] cuando FIV asume valores superiores a 5 o 10 indica que los coeficientes asociados a la ecuación de regresión no están estimados de manera correcta, debido a la presencia de multicolinealidad.

Una tercera alternativa para detectar la multicolinealidad es inspeccionar la matriz de correlaciones $\boldsymbol{R} = (\boldsymbol{X}^*)^t\boldsymbol{X}^*$, donde $\boldsymbol{X}^*$ es la matriz de datos estandarizada [42, 40], las componentes fuera de la diagonal corresponden a los coeficientes de correlación simple entre dos variables k y l. Si se observan valores altos, hay dependencia lineal entre ellas, la inspección de estos valores no permite cuantificar la multicolinealidad.

4.3. Relación del modelo de RBD con el modelo de regresión clásico

En el modelo de RBD (13) cuando las variables regresoras son continuas es equivalente a la de regresión clásica (11), bajo el modelo ortogonal centrado; en este caso en la RBD la distancia Euclidiana es la métrica usual para estimar la distancia entre los individuos i e i', donde a su vez esta representa una transformación lineal que convierte las variables en coordenadas. De igual manera esta equivalencia se mantiene para el caso en el que las variables predictoras sean cualitativas o mixtas si se selecciona la medida de disimilaridad adecuada [24], [32], [33] [42].

Si todas las variables explicativas dadas en (1), la distancia Euclídea definida en (4) se puede expresar como:

$$\begin{aligned}\delta_{ii'}^2 &= (\boldsymbol{x}_i - \boldsymbol{x}_{i'})^t (\boldsymbol{x}_i - \boldsymbol{x}_{i'}) \\ &= \boldsymbol{x}_i^t\boldsymbol{x}_i + \boldsymbol{x}_{i'}^t\boldsymbol{x}_{i'} - 2\boldsymbol{x}_i^t\boldsymbol{x}_{i'}\end{aligned}$$

La matriz $\boldsymbol{D}^{(2)} = (\delta_{ii'}^2)$, entonces

$$\begin{aligned}\boldsymbol{A} = (a_{ii'}) &= -\frac{1}{2}(\delta_{ii'}^2) \\ &= -\frac{1}{2}\left[diag(\boldsymbol{X}\boldsymbol{X}^t)\boldsymbol{1}^t + \boldsymbol{1}(diag(\boldsymbol{X}\boldsymbol{X}^t))^t - 2\boldsymbol{X}\boldsymbol{X}^t\right]\end{aligned}$$

donde $diag(\boldsymbol{X}\boldsymbol{X}^t)$ son vectores que contienen los términos de la diagonal de $\boldsymbol{X}\boldsymbol{X}^t$. Por lo tanto haciendo $\boldsymbol{H} = \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)$, la matriz $\boldsymbol{B}$ definida en (8) se puede expresar como:

$$\boldsymbol{B} = \boldsymbol{H}\boldsymbol{A}\boldsymbol{H} = \boldsymbol{H}\boldsymbol{X}\boldsymbol{X}^t\boldsymbol{H} = \boldsymbol{Z}\boldsymbol{Z}^t$$

porque $diag(\boldsymbol{X}\boldsymbol{X}^t)\mathbf{1}^t\boldsymbol{H} = \boldsymbol{H}\mathbf{1}(diag(\boldsymbol{X}\boldsymbol{X}^t))^t = \mathbf{0}$ ya que las variables se han centrado en la media. Además, se ha realizado una transformación lineal que origina unas nuevas variables ortogonales (coordenadas principales), obteniéndose la equivalencia entre la regresión clásica con el modelo ortogonal centrado y la RBD.

Sin embargo, no es necesario considerar una distancia Euclidiana p-dimensional. Sea E el espacio generado por las columnas de $\boldsymbol{Z}$, donde $\boldsymbol{Z}$ son soluciones del escalamiento multidimensional obtenidas a partir de una distancia aplicada a los mismos datos. Entonces tomando $q > p$, es decir las columnas de $\boldsymbol{Z}$ más importantes, el modelo de RBD supera al modelo de regresión clásica cuando $(\boldsymbol{Y} - \hat{\beta}_0\mathbf{1}) \in E$. Observe que esto siempre sucede cuando $q = n-1$ con $q > p$.

4.4. El efecto de la multicolinealidad sobre la estimación de $\boldsymbol{B}$ y $\boldsymbol{Z}$

Trabajar la regresión múltiple basada en EM genera nuevas variables llamadas coordenadas principales, que por su construcción garantizan la no multicolinealidad entre los predictores al originarse unas nuevas variables que son independientes e incorrelacionadas.

El aplicar la metodología de EM origina que las columnas (variables) de la matriz de coordenadas principales $\boldsymbol{Z}_1, \boldsymbol{Z}_2, ..., \boldsymbol{Z}_q$ sean ortogonales [42, 40], entonces la multicolinealidad y la autocorrelación se descartan [33].

Teniendo en cuenta la estructura de las coordenadas principales presentada en la ecuación (9), el conjunto de vectores $\left\{\boldsymbol{Z}_1, \boldsymbol{Z}_2, ..., \boldsymbol{Z}_q\right\}$ son ortogonales, por tal razón estos son linealmente independientes. Además, se verifica que el producto punto entre dos vectores cualesquiera es cero, es decir: $\boldsymbol{Z}_j^t\boldsymbol{Z}_{j'} = 0$.

Como se mencionó inicialmente el ACP es una estrategia de análisis empleada para tratar el problema de multicolinealidad, el análisis de coordenadas principales está muy relacionado con el ACP; el objetivo de estas técnicas generalmente es reducir la dimensionalidad de los datos. El ACP parte de la matriz $\boldsymbol{X}^t\boldsymbol{X}$, se hallan sus valores propios y luego se proyectan para obtener los valores de las componentes, se definen las matrices de covarianzas $\boldsymbol{S} = \boldsymbol{X}^t\boldsymbol{X}/n$ y $\boldsymbol{X}^t\boldsymbol{X}$ que puede interpretarse como una matriz de similitud (covarianzas) entre los n elementos, estas matrices son cuadradas semidefinidas positivas. Entonces las componentes principales son idénticas a las coordenadas principales que se obtienen directamente de la matriz $\boldsymbol{X}\boldsymbol{X}^t$; además, si se garantiza que la matriz de distancias es Euclídea, los dos métodos conducirán al mismo resultado [42].

Se demuestra que los valores propios de la matriz $\boldsymbol{X}^t\boldsymbol{X}$ y la matriz $\boldsymbol{X}\boldsymbol{X}^t$ trabajada en coordenadas principales, son los mismos. Se parte de la matriz de datos originales expresión (1), donde las variables se centran tanto en fila como en columna; de manera que su media sea cero, las componentes principales son los valores propios de la matriz $\boldsymbol{X}^t\boldsymbol{X}/n$. Si $\boldsymbol{v}_j$ es un vector propio de $\boldsymbol{X}^t\boldsymbol{X}$ y λ_j el valor propio asociado a $\boldsymbol{v}_j$, entonces

$$\begin{aligned}\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{v}_j &= \lambda_j v_j \\ \boldsymbol{X}\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{v}_j &= \lambda_j \boldsymbol{X}\boldsymbol{v}_j\end{aligned}$$

donde $\boldsymbol{X}\boldsymbol{v}_j$ es un vector propio de $\boldsymbol{X}^t\boldsymbol{X}$ con el mismo valor propio λ_j. Si $n > p$ y la matriz $\boldsymbol{X}^t\boldsymbol{X}$ es de rango completo, tendrá p valores no nulos que son también los valores no nulos de $\boldsymbol{X}\boldsymbol{X}^t$. Los vectores propios de $\boldsymbol{X}\boldsymbol{X}^t$ son las proyecciones de la matriz $\boldsymbol{X}$ sobre la dirección de los vectores propios de $\boldsymbol{X}^t\boldsymbol{X}$. Como uno de los objetivos de EM es lograr describir lo más cercanamente posible la matriz de datos definida en (1) y además $\boldsymbol{B}$ se obtiene a partir de $\boldsymbol{B} = \boldsymbol{Z}\boldsymbol{Z}^t \cong \boldsymbol{X}\boldsymbol{X}^t$, al tenerse la equivalencia, las dos técnicas logran describir lo más cercanamente posible la matriz de datos observados.

De acuerdo a la equivalencia entre las dos técnicas, en el caso de aplicación se comparará la metodología planteada con el ACP, a fin de contrastar los resultados obtenidos.

5. Regresión a partir de las componentes principales

Considérese la matriz (1), conformada por p variables mixtas observadas en n individuos, inicialmente se estandariza la matriz, para el caso de las

variables dicotómicas se realiza teniendo en cuenta la distribución binomial, donde la media de acuerdo a lo planteado por [42] para una variable binaria es la proporción de unos, $\mathscr{P}$, y para las variables categóricas multiestado se recurre a la distribución de una variable multinomial, usando variables indicadoras, en la cual para el i-ésimo nivel de la variable $\boldsymbol{X}_i$, la media asociada es $\mathscr{P}_{\mathrm{i}}$.

Posteriormente, se procede a generar las componentes principales. Con la matriz de datos estandarizados, $\boldsymbol{X}^*$, se estima la matriz de correlación $\boldsymbol{R}$, que corresponde a la matriz de correlación de los datos originales. $\boldsymbol{R}$ se descompone $\boldsymbol{R} = (\boldsymbol{L}^*)^t \boldsymbol{\Lambda}^* \boldsymbol{L}^*$, sean $\lambda_1^*, \lambda_2^*, , ..., \lambda_p^*$ los valores propios asociados a la matriz de correlación y Λ^* la matriz diagonal de valores propios y $\boldsymbol{L}^*$ la matriz ortogonal de vectores propios, donde se verifica $\boldsymbol{L}^*(\boldsymbol{L}^*)^t = \boldsymbol{I}$.

Se genera la matriz $\boldsymbol{Z}^*_{n\times p} = \boldsymbol{X}^*\boldsymbol{L}^*$ que contiene las componentes principales y cada una de las columnas de $\boldsymbol{Z}^*$ define el nuevo conjunto de variables predictoras ortogonales, donde $(\boldsymbol{Z}^*)^t\boldsymbol{Z}^* = (\boldsymbol{L}^*)^t(\boldsymbol{X}^*)^t\boldsymbol{X}^*\boldsymbol{L}^*$.

El objetivo de este tipo de modelamiento es encontrar unas nuevas variables regresoras, componentes principales, a partir de la matriz de observaciones donde los datos iniciales se transforman en otras variables independientes e incorrelacionadas; con ellas se ajusta la ecuación de regresión y se reescribe el modelo de regresión clásico (11) en términos de las componentes (la forma canónica del modelo):

$$\boldsymbol{Y} = \boldsymbol{Z}^*\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{23}$$

donde $\boldsymbol{\alpha} = \boldsymbol{L}^*\boldsymbol{\theta}$ es el vector de parámetros asociado a las p componentes principales, $(\boldsymbol{X}^*\boldsymbol{L}^*)^t\boldsymbol{X}^*\boldsymbol{L}^* = (\boldsymbol{Z}^*)^t\boldsymbol{Z}^* = \boldsymbol{\Lambda}^*$ y $\boldsymbol{\varepsilon}$ es el error.

De manera similar que en el EM, se seleccionan las q^* primeras componentes, como estas fueron generadas a partir de la matriz de correlaciones, se seleccionan las que presenten un valor propio mayor a 1, o decidir en función del porcentaje de variabilidad explicada satisfactorio al analista o en su defecto emplear la expresión (17), la cual permite seleccionar las q^* primeras componentes principales. El vector de parámetros asociado a las componentes principales retenidas, teniendo en cuenta el modelo definido en la expresión (23) se estima como:

$$\hat{\boldsymbol{\alpha}}_{q^*} = \left((\boldsymbol{Z}^*_{q^*})^t\boldsymbol{Z}^*_{q^*}\right)^{-1}(\boldsymbol{Z}^*_{q^*})^t\boldsymbol{Y} = (\boldsymbol{\Lambda}^*_{q^*})^{-1}(\boldsymbol{Z}^*_{q^*})^t\boldsymbol{Y}$$

6. Simulación

La propuesta de simulación se plantea en dos escenarios, inicialmente un conjunto de datos donde se garantiza el cumplimiento de cada uno de los supuestos necesarios para realizar una regresión clásica ($S1$) y el segundo donde las variables regresoras presentan multicolinealidad ($S2$) los demás requerimientos para ajustar una regresión se garantizan. En el escenario $S1$ los datos se generan a partir de la ecuación de regresión clásica con tres variables cuantitativas $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ y $\boldsymbol{X}_3$, una variable cualitativa $\boldsymbol{X}_4$ con dos categorías $\boldsymbol{D}_1$ y $\boldsymbol{D}_2$. A fin de garantizar el cumplimiento de los premisas, los errores fueron creados bajo una distribución normal con media cero y desviación estándar equivalente a 2.

Las variables $\boldsymbol{X}_1$ y $\boldsymbol{X}_2$ se generaron bajo una distribución uniforme (U) donde $\boldsymbol{X}_1 \sim U_n[0,10]$ y $\boldsymbol{X}_2 \sim U_n[10,15]$, respectivamente, $\boldsymbol{X}_3$ se generó bajo una distribución binomial con parámetros n y la probabilidad $\mathscr{P} = 0.4$. Para la variable categórica $\boldsymbol{X}_4$ con distribución multinomial (MN), donde $\boldsymbol{X}_4 \sim MN\,(n, \mathscr{P} = 0.4, 0.35, 0.25)$. El modelo sobre el cual se generan los datos corresponde: $y_i = 3 - 1.2x_{i1} + 1.5x_{i2} + 2.3x_{i3} + 2.7D_{i1} + 4D_{i2}$, $i = 1, ..., n$, a partir de esta ecuación se crean los valores de la variable respuesta para los tamaños de muestra $n = 50, 100, 500, 1000$, y a su vez para cada tamaño de muestra, se realizaron $m = 50, 100, 150, 200$ simulaciones con el propósito de verificar las estimaciones obtenidas en el proceso de simulación, y además se espera que al aumentar el tamaño de muestra las estimaciones tiendan a tomar el valor de los parámetros definidos inicialmente. Se garantizó el cumplimiento de los supuestos (normalidad y homocedásticidad) necesarios para realizar una análisis de regresión.

Posteriormente, se crea el escenario $S2$, donde se introduce la colinealidad a partir del modelo $x_1 = -2 + 1.5x_2$ agregando un ruido con distribución normal con media 0 y varianza 0.1, en este escenario también se valido el cumplimiento de los supuestos necesarios para ajustar la regresión.

Una vez realizadas las simulaciones, se evaluó la multicolinealidad mediante los valores propios de la matriz $(\boldsymbol{X}^*)^t\boldsymbol{X}^*$ en los dos escenarios propuestos. En $S2$ los correspondientes valores asociados fueron: $\lambda_1 = 3893.63$, $\lambda_2 = 2364.42$, $\lambda_3 = 1243.93$, $\lambda_4 = 95.97$ y $\lambda_5 = 0.038$, evidenciándose la presen-

Tabla 1. FIV en los escenarios $S1$ y $S2$ para los tamaños de muestra $n = 50, 100, 500, 1000$.

	Escenario $S1$				Escenario $S2$			
	Tamaño de muestra				Tamaño de muestra			
Variables	50	100	500	1000	50	100	500	1000
x_1	1.00	1.00	1.00	1.00	635.78	611.01	446.85	502.26
x_2	1.03	1.00	1.00	1.01	635.78	610.56	446.82	502.31
x_3	1.01	1.01	1.01	1.01	1.23	1.02	1.01	1.01
D_1	1.48	1.70	1.70	1.52	1.34	1.49	1.71	1.52
D_2	1.46	1.71	1.71	1.52	1.36	1.46	1.71	1.52

cia de un valor propio tendiente a cero, indicando la presencia de multicolinealidad. Así mismo, se estimaron los FIV presentados en la Tabla 1. En el caso $S1$, no hubo colinealidad en los predictores, en tanto que para $S2$ los FIV asumen valores muy grandes confirmando con ello la dependencia entre las variables regresoras.

El proceso de simulación realizado se corrió para cada uno de los tamaños de muestra estipulados anteriormente, pero a su vez en cada uno de ellos se realizó el proceso de simulación m veces. A fin de comparar los ajustes de cada uno de los modelos obtenidos se estiman los valores de: R^2_{ajus}, BIC, AIC y $logLik$, y se promediaron dichos resultados en las m simulaciones, sobre estos resultados se fundamenta el análisis en cada uno de los casos presentados.

Al ajustar la regresión sobre las coordenadas principales en $S1$ y $S2$ se consideran las siguientes situaciones: tomar menos coordenadas principales que variables, ($q = 3$ y $q = 4$), la segunda igual número de coordenadas principales que variables ($q = 5$) y la tercera situación, tomar más coordenadas principales que variables ($q = 6$ y $q = 10$). Los resultados obtenidos se comparan con la regresión basada en componentes principales y con la regresión clásica, donde esta última es el referente para comparar los ajustes.

Para el análisis de los resultados se tendrá presente las siguientes consideraciones: al ajustar la regresión sobre las coordenadas principales, se evidenció en la mayoría de los casos simulados que cuatro de las coordenadas involucradas en el análisis son significativas, en otros casos 3 o 5, pero que no necesariamente recaía la significancia sobre las primeras coordenadas seleccionadas, razón por la cual al haber elegido pocas coordenadas, en particular en el caso $q = 3$ se obtuvo algunos ajustes muy bajos. Al comparar las corridas obtenidas en las componentes y las coordenadas, el número de coordenadas estimadas corresponde al número de observaciones incluidas, mientras que en las componentes el número es equivalente al número de variables contenidas. Además en los dos escenarios de simulación se verificó el cumplimento del supuesto de normalidad a través de la prueba *Shapiro Wilk* y en cada uno de los casos, los *valores p* fueron cercanos a uno. Adicionalmente, dentro de la premisas tenidas en cuenta para el escenario de simulación se garantizó que los errores se distribuyesen normal con un valor fijo de la media y varianza constante como se mencionó anteriormente.

En la Tabla 2 se muestran los resultados obtenidos al correr la regresión múltiple sobre las coordenadas principales, para los diferentes tamaños de muestra (n), al seleccionar las tres primeras coordenadas ($q = 3$) se evidencia que el porcentaje de variabilidad que explica la regresión es muy bajo, pero al aumentar q son parecidos a los obtenidos en la regresión clásica (ver Tabla 3). En los diferentes tamaños de muestra se evidencia que los mejores ajustes se tienen cuando $q = 5, 6$ y 10; los valores del R^2_{ajus} en $S1$ están cercanos al 87 %, mientras que en $S2$ son próximos al 70 %. En general, el porcentaje de variabilidad explicado por el R^2_{ajus} es mejor a medida que se involucran más coordenadas, siendo mejores cuando se seleccionan 5 y 6 coordenadas, debido a que en la mayoría de los casos en cada una estas regresiones simuladas están contenidas las coordenadas donde la mayoría de ellas fueron significativas. En $q = 10$ no ocurre lo mismo pues se han agregado variables donde no todas necesariamente son influyentes sobre la variable respuesta. En cuanto a los valores más bajos en el AIC y el BIC, generan también buenos ajustes, para $q = 4, 5$ y 6 coordenadas elegidas, además, los diferentes estadísticos para evaluar el ajuste al compararlos con los valores en la regresión clásica son similares (ver Tabla 3).

Tabla 2. Promedio de las *m* simulaciones de R^2_{ajus}, *AIC*, *BIC* y *logLik*, utilizando regresión por coordenadas principales en los escenarios *S*1 y *S*2 con $n = 50, 100, 500, 1000$.

		*S*1				*S*2			
n	*q*	R^2_{ajus}	*AIC*	*BIC*	*logLik*	R^2_{ajus}	*AIC*	*BIC*	*logLik*
50	3	42.53%	291.97	301.53	-140.99	18.77%	263.02	272.58	-126.51
	4	66.30%	265.88	277.35	-126.94	67.63%	217.15	228.62	-102.57
	5	86.96%	218.61	231.99	-102.30	67.78%	217.79	231.17	-101.89
	6	87.00%	219.28	234.58	-101.64	67.75%	218.65	233.94	-101.32
	10	87.37%	220.85	243.79	-98.42	67.56%	221.93	244.87	-98.96
100	3	29.11%	598.85	611.88	-294.43	49.38%	479.28	492.30	-234.64
	4	77.31%	483.59	499.22	-235.80	69.96%	427.81	444.10	-207.66
	5	86.02%	437.39	455.63	-211.70	69.98%	428.43	446.66	-207.21
	6	86.16%	437.28	458.12	-210.64	69.97%	429.40	450.24	-206.70
	10	86.98%	434.73	466.00	-205.37	70.00%	432.89	464.16	-204.45
500	3	37.52%	2903.04	2924.12	-1446.52	65.41%	2199.32	2220.39	-1094.66
	4	73.48%	2475.10	2500.38	-1231.55	70.47%	2121.04	2146.33	-1054.52
	5	86.11%	2152.46	2181.96	-1069.23	70.66%	2118.93	2148.43	-1052.47
	6	86.13%	2152.74	2186.45	-1068.37	70.68%	2119.58	2153.30	-1051.79
	10	86.86%	2129.60	2180.18	-1052.80	70.69%	2123.31	2173.88	-1049.65
1000	3	35.75%	5823.90	5848.44	-2906.95	62.27%	4448.20	4472.74	-2219.10
	4	80.17%	4648.47	4677.92	-2318.24	69.52%	4235.58	4265.03	-2111.79
	5	86.15%	4290.46	4324.81	-2138.23	69.56%	4235.56	4269.91	-2110.78
	6	86.15%	4291.50	4330.77	-2137.75	69.59%	4235.44	4274.70	-2109.72
	10	86.79%	4248.34	4307.23	-2112.17	69.59%	4239.45	4298.34	-2107.72

Tabla 3. Promedio de las *m* simulaciones de R^2_{ajus}, *AIC*, *BIC* y *logLik*, utilizando regresión clásica en los escenarios *S*1 y *S*2 con $n = 50, 100, 500, 1000$.

	*S*1				*S*2			
n	R^2_{ajus}	*AIC*	*BIC*	*logLik*	R^2_{ajus}	*AIC*	*BIC*	*logLik*
50	87.44%	216.70	230.08	-101.35	67.85%	217.69	231.07	-101.84
100	87.06%	429.63	447.87	-207.82	70.10%	428.05	446.28	-207.02
500	86.96%	2120.64	2150.15	-1053.32	70.76%	2117.07	2146.57	-1051.54
1000	86.92%	4233.11	4267.47	-2109.56	69.68%	4231.29	4265.65	-2108.65

Cuando $n = 100$ en el escenario *S*2, comparando los diferentes valores obtenidos en la regresión sobre las coordenadas (Tabla 2) al seleccionar $q = 4, 5$ y 6 son similares a las estimaciones mostradas en la regresión clásica, ver Tabla 3. De igual manera, ocurre para los tamaños de muestra 500 y 1000. Se evaluó el FIV y se observó que la regresión basada en EM subsana el problema de multicolinealidad, además, en el análisis se incluyeron cada una de las variables planteadas. Asimismo, se evidenció la imposibilidad de identificar cuáles son las variables que más influyen sobre la respuesta. Esto último puede ser muy bueno ya que no pierde representatividad en cuanto al ajuste del modelo y ayuda a que éste no se pueda manipular por personas externas.

En la Tabla 4 se muestra la regresión sobre las componentes principales, en los escenarios *S*1 y *S*2. De acuerdo a los valores de los R^2_{ajus}, *BIC*, *AIC* y *logLik*, los ajustes son mejores a medida que aumenta el número de componentes. Teniendo en cuenta que el número máximo de componentes estimadas en el ACP depende del número de variables incluidas en el análisis, cuando $q^* = 5$ (número total de componentes) los ajustes son iguales a la regresión clásica. En todo el proceso de simulación realizado se observó que por lo general tres componentes eran significativas; sin embargo, se incluyeron las tres primeras, las cuales no necesariamente son todas significativas, y al igual que las coordenadas, pueden quedar algunas variables no tan importantes en su relación con la variable respuesta.

Al comparar la regresión en componentes con coordenadas principales, teniendo en cuenta el caso donde $q = q* = 3$ en las Tablas 2 y 4 para el escenario *S*2, los valores *AIC* y *BIC* en las componentes principales son más bajos que los valores obtenidos en las coordenadas principales para cualquier tamaño de muestra simulados. Sin embargo, en el esce-

Tabla 4. Promedio de las *m* simulaciones de R^2_{ajus}, *AIC*, *BIC* y *logLik*, utilizando regresión en componentes principales en los escenarios *S*1 y *S*2 con $n = 50, 100, 500, 1000$.

		*S*1				*S*2			
n	q^*	R^2_{ajus}	*AIC*	*BIC*	*logLik*	R^2_{ajus}	*AIC*	*BIC*	*logLik*
50	3	35.78%	297.58	307.14	-143.79	65.86%	218.94	228.50	-104.47
	5	87.44%	216.70	230.08	-101.35	67.85%	217.69	231.07	-101.84
100	3	29.58%	598.18	611.20	-294.09	68.65%	430.92	443.94	-210.46
	5	87.06%	429.63	447.87	-207.82	70.10%	428.05	446.28	-207.02
500	3	33.40%	2934.97	2956.04	-1462.48	69.38%	2138.25	2159.32	-1064.12
	5	86.96%	2120.64	2150.15	-1053.32	70.76%	2117.07	2146.57	-1051.54
1000	3	31.91%	5882.01	5906.55	-2936.00	68.18%	4277.74	4302.28	-2133.87
	5	86.92%	4233.11	4267.47	-2109.56	69.68%	4231.29	4265.65	-2108.65

nario *S*1 para $q = q* = 3$, sucede lo contrario, es decir que las coordenadas principales funcionan de manera mas apropiada que las componentes principales. En general, los valores del *BIC* y el *AIC* son mas bajos en las coordenadas que en las componentes para el escenario *S*1, mientras que, en el escenario *S*2 las componentes tienen valores de *AIC* y *BIC* mas bajos que en coordenadas principales. Las dos metodologías son similares y a su vez, los diferentes valores de ajuste son parecidos a los obtenidos en la regresión clásica.

Al aumentar el número de coordenadas en el modelo, $q = 4, 5, 6, 10$ las estimaciones obtenidas tienden a ser parecidas a las componentes. Cuando *n* es más grande los ajustes mejoran y tienden a estabilizarse (ver Tablas 2, 3 y 4). Al aplicar la regresión basada en coordenadas y componentes principales las dos metodologías corrigen el incumplimiento del supuesto de multicolinealidad. Es de resaltar aquí que si se toman todas las coordenadas y componentes significativas, se obtiene un ajuste igual o superior (caso de coordenadas principales) que el modelo de regresión lineal clásico, haciendo de la regresión en coordenadas principales una muy buena alternativa para el manejo de la colinealidad e inclusive en el caso de no colinealidad entre variables predictoras.

7. Ejemplos de aplicación

En esta sección se ilustra la aplicación de la metodología basada en EM, a dos conjuntos de datos planteados en [46], los cuales fueron analizados aplicando regresión múltiple. Los resultados obtenidos son comparados con la regresión clásica y la regresión con componentes principales.

7.1. Ejemplo 1. Variables regresoras mixtas

Los pesos de 13 pavos fueron medidos en libras (y), el objetivo es relacionar estos con la edad de los pavos medida en semanas y una variable que indica el lugar donde fueron criados (Georgia, Virginia y Wisconsin), la cual en el análisis de regresión clásica se dicotomiza con las categorías D_1 y D_2 [46]. Se generó en los datos una variable colineal (x_c) a partir de la edad de los pavos mediante la expresión $\sqrt{3x/2}$, se estimó la matriz de distancias por medio del coeficiente de similitud de Gower, haciendo uso de la función *daisy* implementada en la *librería cluster* del paquete estadístico *R* [48]; a fin de obtener la matriz $\boldsymbol{B}$ y sus valores propios. Se selecciona $q = 7$ coordenadas principales teniendo en cuenta que el porcentaje de ajuste (m_q) definido en (17) fue del 97.4%, y para la regresión basada en componentes principales, se elige $q^* = 3$ componentes, $m^*_q = 99.99\%$.

Se evaluaron las premisas necesarias para ajustar la regresión clásica, a los datos originales planteados por el autor, a través de la prueba de *Shapiro-Wilk* ($p\ valor$=0.862) se constató que los datos satisfacen este requisito, así mismo se analizaron los residuales evidenciándose que son homocedásticos e independientes ($Durbin - Watson$ de 1.9), garantizándose el cumplimiento de los supuestos. En la Tabla 5, se muestran los resultados obtenidos al aplicar la regresión clásica a los datos originales, luego se involucró la variable colineal x_c, se halló el *FIV* haciendo uso de la *librería car* del paquete estadístico *R* [48], verificándose la presencia de multicolinealidad entre las variables regresoras. Así mismo se presentan los modelos ajustados en cada una de la regresiones, comparando la regresión sobre las componentes y coordenadas principales. En la regresión

Tabla 5. Ejemplo 1. Estadísticos de bondad de ajuste *BIC*, *AIC*, *logLik* y R^2_{ajus}, y los respectivos modelos basados en regresión clásica, coordenadas y componentes principales.

	Regresión				
	Clásica sin colinealidad		Coordenadas		Componentes
BIC	16.16		17.08		17.22
AIC	12.77		13.69		14.96
logLik	-0.38		-0.84		-3.48
R^2_{ajus}	96.93%		96.70%		96.04%
Modelo	$y=\theta_0+\theta_1x_1+\theta_3D_1+\theta_4D_2+\varepsilon$		$y=\beta_0+\beta_1z_1+\beta_2z_2+\beta_3z_3+\beta_5z_5+\varepsilon$		$y=\alpha_0+\alpha_1z_1^*+\alpha_3z_3^*+e$
Variable	x_1	D_1	D_2	x_c	
FIV	195.27	1.48	1.37	196.60	

basada en la metodología propuesta utilizando EM, se observa un buen ajuste de acuerdo a los valores del *BIC*, *AIC*, *logLik* y R^2_{ajus}. Este ajuste es muy cercano a los obtenidos en la regresión clásica cuando las variables no fueron contaminadas con el efecto de la multicolinealidad, reflejando confiabilidad en las estimaciones realizadas en la metodología basada en el EM. También se resalta que las estimaciones fueron muy parecidas a las obtenidas en la regresión basada sobre las componentes principales.

7.2. Ejemplo 2. Variables regresoras continuas

En una planta de agua un ingeniero de control es responsable de reducir los costos de una producción [46]. Debido a que uno de los factores más costoso es la cantidad de agua usada en las instalaciones cada mes, decide estudiar su consumo y las variables, x_1 temperatura mensual promedio, x_2 la producción promedio medida en libras, x_3 el número de días de funcionamiento de la planta en el mes, x_4 el número de personas en la nómina, x_5 números aleatorios, esta variable la incluyó el experimentador debido a que era escéptico con los resultados obtenidos en el análisis de regresión (x_5 variable reportada en el ejemplo [46]) y la variable dependiente consumo de agua mensual (galones).

En los datos se observó multicolinealidad moderada entre las variables regresoras. Se verificaron los supuestos necesarios para el análisis, la prueba de $Shapiro-Wilk$ ($p\ Valor = 0.537$) confirma el cumplimiento del supuesto de normalidad, a través del análisis de la gráfica de los residuales se evidenció que son independientes ($Durbin-Watson$=2.5) e incumplió el supuesto de homocedásticidad; razón por la cual se transformó la variable respuesta por el método $Box-Cox$ (y^λ) [16],[47] implementada en el paquete estadístico *R* [48] en la *librería MASS* y con un valor de $\lambda=-2$ se ajustó la regresión clásica, sobre coordenadas y componentes principales.

La matriz $\boldsymbol{D}$ se estima a partir de la distancia de Mahalanobis puesto que las variables regresoras son cuantitativas, con escalas de medición diferentes y además están correlacionadas. Posteriormente, se hallaron las coordenadas principales, se seleccionó $q=5$ teniendo en cuenta el valor de m_q, el porcentaje de ajuste es del 80%.Por lo tanto, se decide tomar las cinco primeras coordenadas.

Inicialmente se corre la regresión clásica sin transformar la variable respuesta, se evidenció multicolinealidad moderada entre las variables, los valores *FIV* se presentan en la Tabla 6. Posteriormente se transformó la variable respuesta y se realizó el ajuste mediante la regresión clásica. Al realizar la regresión sobre las coordenadas y las componentes principales se logra corregir la multicolinelalidad moderada entre las variables. Teniendo en cuenta que los valores del $AIC=-565.29$ y $BIC=-561.12$ son los más pequeños, $logLik=287.64$ corresponde al valor más grande y el $R^2_{ajus}=65.80\%$, el modelo ajustado con la metodología propuesta usando EM es más apropiado, logrando superar el ajuste con respecto a la regresión clásica y la basada en Componentes.

8. Conclusiones

La aplicación de la regresión basada en la metodología seguida en el escalamiento multidimensional es una estrategia alternativa de análisis que corrige el problema de multicolinealidad entre las variables regresoras.

Teniendo en cuenta los escenarios de simulación generados, la regresión trabajada a partir del escalamiento multidimensional presenta predicciones

Tabla 6. Ejemplo 2. Estadísticos de bondad de ajuste *BIC*, *AIC*, *logLik* y R^2_{ajus} y los respectivos modelos basados en regresión clásica, coordenadas y componentes principales.

	Regresión				
	Clásica transformando y		Coordenadas		Componentes
BIC	-557.4971		-561.12		-560.94
AIC	-563.3296		-565.29		-565.10
logLik	288.6648		287.64		287.55
R^2_{ajus}	64.15%		65.80%		65.42%
Modelo	$y^{-2} = \theta_0 + \sum_{j=1}^{5} \theta_j x_j + \varepsilon$		$y^{-2} = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_5 z_5 + \varepsilon$		$y^{-2} = \alpha_0 + \alpha_1 z_1^* + \alpha_4 z_4^* + \alpha_5 z_5^* + e$
Variable	x_1	x_2	x_3	x_4	x_5
FIV	1.26	6.70	1.27	6.74	1.05

muy similares a las obtenidas en la regresión basada en componentes principales, además, se evidenció que a medida que se incrementaban el número de coordenadas y componentes en el análisis, los ajustes en los modelos corridos fueron muy cercanas a la regresión clásica donde las variables no habían sido contaminadas con la colinealidad, esto garantiza que la metodología describe de manera adecuada los datos observados aún en presencia de multicolinealidad.

En los casos de aplicación se constató que la regresión sobre las coordenadas principales presentó mejor ajuste con respecto a la regresión clásica y la regresión basada en las componentes principales.

La regresión basada en el escalamiento multidimensional, al transformar las variables en coordenadas principales, es una estrategia que oculta el efecto de las variables, puesto que el analista no podrá identificar cuáles de las variables observadas son las que realmente influyen sobre la variable respuesta. En este sentido, no es evidente para el investigador influenciar la respuesta puesto que no logra identificar cuál de las variables presenta mayor peso en el modelamiento.

9. Agradecimientos

Agradecemos a los evaluadores por sus valiosas y oportunas observaciones que permitieron mejorar el artículo.

Referencias

[1] A. E. Hoerl and W. R. Kennard, "Ridge regression: applications to nonorthogonal problems", *Technometrics*, vol. 12, no. 1, pp. 69-82, 1970.

[2] S. Velilla, "Obtención simultánea de multicolinealidad y observaciones influyentes", *Estadística Española*, vol. 30, no. 17, pp. 83-98, 1988.

[3] M. R. Piña, M. A. Rodríguez, y J. Aguirre, "Regresión Ridge y la distribución central t", *CIENCIA ergo-sum*, vol. 14, no. 2, pp. 191-196, 2007.

[4] D. F. Campos, "Transferencia regional de información hidrológica mediante regresión lineal múltiple de tipo ridge", *Agrociencia, México*, vol. 47, no. 5, pp. 411-427, 2013.

[5] E. R. Mansfield, J. T. Webster and R. F. Gunst, "An analytic variable selection technique for principal component regression", *Applied statistics*, vol. 26, no. 1, pp. 34-40, 1977.

[6] E. López, "Tratamiento de la colinealidad en regresión múltiple", *Psicothema*, vol. 10, no. 2, pp. 491-507, 1998.

[7] S. A. Abdul, C. S. Bakheit and S. M. Al-Alawi, "Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations", *Environmental Modelling & Software*, vol. 20, no. 10, pp. 1263-1271, 2005.

[8] O. Navarro, "Selección de variables en regresión componentes principales", *Seventh Latin American and Caribbean Conference for Engineering and Technology*, San Cristobal, 2009.

[9] J. M. Rajab, M. Z. MatJafri and H. S. Lim, "Combining multiple regression and principal component analysis for accurate predictions for column ozone in Peninsular Malaysia", *Atmospheric Environment*, vol. 71, pp. 36-43, 2013.

[10] J. M. Del Valle, and W. B. Guerra, "La Multicolinealidad en modelos de Regresión Lineal Múltiple", *Revista Ciencias*

Técnicas Agropecuarias,Universidad Agraria de La Habana, vol. 21, no. 4, pp. 80-83, 2012.

[11] J. C. Vega-Vilca y J. Guzmán,"Regresión PLS y PCA como solución al problema de multicolinealidad en regresión múltiple", *Revista de Matemática Teoría y Aplicaciones*, vol. 18, no. 1, pp. 09-20, 2011.

[12] J. Shen, and S. Gao,"A solution to separation and multicollinearity in multiple logistic regression", *Journal of data science: JDS*, vol. 6, no. 4, pp. 515-531, 2008.

[13] F. Godínez-Jaimes, R. Reyes-Carreto, F. J. Ariza-Hernandez y E. Barrera-Rodriguez, "La colinealidad y la separación en los datos en el modelo de regresión logística", *Agrociencia, México*, vol. 46, no. 4, pp. 411-425, 2012.

[14] M. Sáez y M. A. Barceló, "Un criterio para omitir variables superfluas en modelos de regresión", *Gaceta Sanitaria*, vol. 12, no. 6, pp. 281-283, 1998.

[15] J. Llorca, "Omisión de variables en modelos de regresión con alta multicolinealidad", *Gaceta Sanitaria*, vol. 13, no. 3, pp. 243-244, 1999.

[16] D. Montgomery, E. Peck and G. Vining, "Introduction to linear regression analysis", *John Wiley & Sons*, 2015.

[17] M. Rosas, F. Chacín, J. García, M. Ascanio y M. Cobo, "Modelos de regresión lineal múltiple en presencia de variables cuantitativas y cualitativas para predecir el rendimiento estudiantil", *Revista de la Facultad de Agronomía de La Universidad del Zulia, Venezuela*, vol. 23, no. 2, pp. 197-214, 2006.

[18] M. Ueki and Y. Kawasaki, "Multiple choice from competing regression models under multicollinearity based on standardized update", *Computational Statistics & Data Analysis*, vol. 63, pp. 31-41, 2013.

[19] D. Villegas, W. Ascanio y M. Cobo, "Evaluación de la multicolinealidad en modelos de regresión lineal múltiple con presencia de valores atípicos", *Revista de la Facultad de Agronomía UCV*, vol. 39, no. 3, pp. 134-143, 2013.

[20] I. Méndez-Ramíreza, H. Moreno-Macías, I. Méndez Gómez-Humarán y Ch. Muratad, "Conglomerados como solución alternativa al problema de la multicolinealidad en modelos lineales", *Revista de Ciencias Clínicas*, vol. 15, no. 2, pp. 39-46, 2014.

[21] D. Garcés y F. Jaimes, "Ronda clínica y epidemiológica. Introducción al análisis multivariable (parte II)", *Iatreia*, vol. 28, no. 1, pp. 87-96, 2015.

[22] J. Marrugat, R. Elosua, M. Grau, S. Sayols-Baixeras y I. R. Dégano,"Prevalencia y pronóstico de los pacientes con infarto de miocardio de alto riesgo candidatos a doble tratamiento antiagregante prolongado", *Revista Española de Cardiología*, vol. 69, no. 5, pp. 480-487, 2016.

[23] G. Garcíat, M. Brogioni, V. Venturini, L. Rodríguez, G. Fontanelli E. Walker, S. Graciani, y G. Macelloni, "Determinación de la humedad de suelo mediante regresión lineal múltiple con datos TerraSAR-X", *Revista de la Asociación Española de Teledetección*, vol. 46, pp.73-81, 2016.

[24] C. Cuadras and C. Arenas, "A distance based regression model for prediction with mixed data", *Communications in Statistics A. Theory and Methods*, vol. 19, pp. 2261-2279, 1990.

[25] J. C. Gower, "Some distance properties of latent root and vector methods used in multivariate analysis", *Biometrika*, vol. 53, pp. 325-338, 1966.

[26] W. S. Torgerson, "Multidimensional scaling: I. Theory and method", *Psychometrika*, vol. 17, no. 4, pp. 401-419, 1952.

[27] R. N. Shepard, "The analysis of proximities: Multidimensional scaling with an unknown distance function. I.", *Psychometrika*, vol. 27, no. 2, pp. 125-140, 1962.

[28] G. Linares, "Escalamiento multidimensional: conceptos y enfoques", *Revista Investigación Operacional*, Editorial Universitaria , vol. 22, no. 2, pp. 173-183, 2009.

[29] A. Arroyo, C. Bruno, J. Di Rienzo y M. Balzarini,"Árboles de expansión mínimos: ayudas para una mejor interpretación de ordenaciones en bancos de germoplasma", *Interciencia*, vol. 30, no. 9, pp. 550-554, 2005.

[30] P. Parés, "Estudio de razas de palomas españolas a partir del análisis de caracteres morfoló-

gicos cualitativos", *Revista MVZ Córdoba*, vol. 15, no. 3, pp. 2158-2164, 2010.

[31] G. Correa, L. Lavalett, M. Galindo, y L. Afanador, "Uso de métodos multivariantes para la agrupación de aislamientos de Colletotrichum spp. con base en características morfológicas y culturales", *Revista Facultad Nacional de Agronomia Medellin*, vol. 60, no. 1 pp. 3671-3690, 2007.

[32] C. Cuadras, "Distancias estadísticas", *Estadística Españolah*, vol. 30, pp. 295-378, 1998.

[33] C. Cuadras, C. Arenas and J. Fortiana "Some computational aspects of a distance-based model for prediction", *Communications in Statistics-Simulation and Computation*, vol. 25, no. 3, pp. 593-609, 1996.

[34] C. Arenas and C. Cuadras, "Recent statistical methods based on distances", *Contributions to Science*, vol. 2, no. 2, pp. 183-191, 2002.

[35] J. Fortiana, "Enfoque basado en Distancias de algunos Métodos Estadísticos Multivariates", *Tesis doctoral, Universität de Barcelona, España*, 2001.

[36] E. Boj, J. M. Claramunt, A. Esteve y J. Fortiana "Criterios de selección de modelo en el credit scoring: aplicación del análisis discriminante basado en distancias", *Anales del Instituto de Actuarios Españoles*, vol. 15, pp. 209-230, 2009.

[37] O. O. Melo, J. Mateu and C. E. Melo "Spatial generalised linear mixed models based on distances", *Statistical Methods in Medical Research*, vol. 45, no. 10, pp. 2010-2030, 2013.

[38] S. Melo, and O. O. Melo "Distance-based approach in univariate longitudinal data analysis", *Journal of Applied Statistics*, vol. 40, no. 3, pp. 674-692, 2013.

[39] O. O. Melo, C. E. Melo and J. Mateu, "Distance-based beta regression for prediction of mutual funds", *AStA Advances in Statistical Analysis*, vol. 99, no. 1, pp. 83-106, 2015.

[40] L. G. Díaz y M. Morales, "Análisis estadístico de datos multivariados", *Universidad Nacional de Colombia*, Bogotá, 2012.

[41] W. S. Torgerson, "Theory and methods of scaling", *New York: John Wiley and Sons*, 1958.

[42] D. Peña, "Análisis de datos multivariantes", *McGraw-Hill Madrid*, vol. 24, 2002.

[43] K. Mardia, J. Kent and J. M. Bibby, "Análisis de datos multivariantes", *Academic Press London*, 2002.

[44] C. Cuadras, "Nuevos métodos de análisis multivariante", *CMC Editions*, 2007.

[45] C. Cuadras and J. Fortiana, "Aplicación de las distancias en estadística", Institut d'Estadística de Catalunya, vol. 17, pp. 39-74, 1993.

[46] N. Draper and H. Smith, "Applied regression analysis", *John Wiley & Sons*, 2014.

[47] J. W. Osborne, "Improving your data transformations: Applying the Box-Cox transformation", *Practical Assessment, Research & Evaluation, Citeseer*, vol. 15, no. 12, pp. 1-9, 2010.

[48] R Development Core Team, "R: A Language and Environment for Statistical Computing", *Vienna, Austria, http://www.R-project.org/*, 2016.