Revista de la Facultad de Ciencias Universidad Nacional de Colombia, Sede Medellín
V 5 N°1 Enero-Junio de 2016 • ISSN 0121-747X / ISSN-e 2357-5749• Artículo Investigación • Páginas 148 a 159
DOI: https://doi.org/10.15446/rev.fac.cienc.v5n1.54616

REGRESIÓN POR MÍNIMOS CUADRADOS PARCIALES *PLS* CON DATOS DE INTERVALO[a]

PARTIAL LEAST SQUARES REGRESSION *PLS* ON INTERVAL DATA

CARLOS GAVIRIA PEÑA[b*], RAÚL PÉREZ AGÁMEZ[c], MARÍA EUGENIA PUERTA YEPES[d]

Recibido 11-12-2015, aceptado 16-05-2016, versión final 30-06-2016.
Artículo Investigación

RESUMEN: La incertidumbre en los datos puede ser considerada mediante un intervalo numérico en el cual una variable puede asumir sus posibles valores, esto se conoce como datos de intervalo. En este artículo se extiende la metodología de regresión *PLS* al caso donde tanto las variables explicativas como la variable respuesta y los coeficientes de regresión son del tipo intervalo. Se propone una metodología de regresión que resuelve tres problemas que se presentan con los datos de tipo real: en primer lugar problemas de multicolinealidad tanto en las variables explicativas como en la variable respuesta, en segundo lugar problemas cuando los datos no pertenecen a un espacio Euclídeo y por último problemas cuando la incertidumbre en los datos se representa por medio de intervalos. Hoy en día existen tareas del común, tales como planificación y operación de sistemas eléctricos, planificación de producción, logística del transporte, inventarios, gestión de carteras de valores, entre otras, que involucran incertidumbre. Por lo anterior se requieren modelos que tengan en cuenta dicha incertidumbre y puedan dar la posibilidad de tomar decisiones para resultados óptimos desde una gama de posibilidades o escenarios posibles. Por otro lado, el análisis de datos reales a menudo se ve afectado por diferentes tipos de errores tales como: errores de medición, errores de cálculo e imprecisión relacionada con el método adoptado para la estimación de los datos. Este artículo es una propuesta metodológica de tipo teórico y está fundamentada en los desarrollos teóricos sobre optimización matemática sobre los conjuntos de multi-intervalos y multi-matrices.

PALABRAS CLAVE: Regresión por componentes principales, mínimos cuadrados parciales *PLS*, optimización intervalo-valuada, intervalo valores y vectores propios.

[a]Gaviria Peña, C.; Pérez Agámez R. & Puerta Yepes, M. E. (2016). Regresión por mínimos cuadrados parciales *PLS* con datos de intervalo *Revista de la Facultad de Ciencias*, 5 (**1**), 148–159. DOI: https://doi.org/10.15446/rev.fac.cienc.v5n1.54616

[b]M. Sc. en Matemáticas. Docente investigador. Facultad de Ingeniería. Universidad de San Buenaventura
[*]cagavirip@unal.edu.co
[c]Ph. D. en Estadística. Profesor Asociado. Escuela de Estadística. Facultad de Ciencias. Universidad Nacional de Colombia, sede Medellín
[d]Ph. D. en Ciencias Matemáticas. Profesora Asociada. Escuela de Matemáticas. Universidad EAFIT.

ABSTRACT: Uncertainty in the data can be considered as a numerical interval in which a variable can assume its possible values, this has been known as interval data. In this paper the PLS regression methodology is extended to the case where explanatory, response variables and coefficients regression are intervals. In this way a regression methodology solves three problems encountered with actual data type is proposed: first multicollinearity in explanatory and response variables, second real data does not belong to a Euclidean space and finally, problems when uncertainty in the data is represented by intervals. Today there are common tasks, such as planning and operation of electrical systems, production planning, transport logistics, inventory, management of securities portfolios; among others, involving uncertainty; this way models that take into account and the ability to make decisions for optimal results from a range of possibilities or scenarios are required. Furthermore, the analysis of real data is affected by different types of errors as measurement errors, miscalculations and imprecision related to the method adopted for estimating data. This paper is a methodological proposal of theoretical type and is based on development about mathematical optimization on multi-interval and multi-matrix spaces.

KEYWORDS: Principal components regression, partial least squares regression PLS, interval-valued optimization, interval eigen values and eigen vectors.

1. INTRODUCCIÓN

La regresión por mínimos cuadrados parciales (PLS) es una técnica de relación de variables introducida por Wold (1972,1975,1985) y extendida posteriormente al campo de la quimiometría por su hermano Wold *et al.* (1984), Wold (2001). Es conocido que la regresión lineal múltiple ordinaria generalmente manipula variables controlables o fácilmente medibles para predecir el comportamiento de otras variables y es usual cuando las variables explicativas son pocas, cuando no existen problemas de multicolinealidad y cuando existe una relación clara entre las variables. Si alguna de estas tres condiciones falla entonces éste tipo de regresión no es eficaz. Por otro lado, la regresión lineal múltiple puede utilizar varias variables explicativas, pero cuando el número de variables es demasiado grande se puede generar un modelo que ajuste muy bien los datos, pero que falla en la predicción de nuevos datos. En estos casos, donde existen muchas variables explicativas, puede que existan pocas variables latentes que recogen la mayor variabilidad de la variable respuesta. El objetivo general de la regresión PLS es extraer estas variables latentes, recogiendo la mayor variación de las variables explicativas de manera que sirvan para modelar la variable respuesta de la mejor manera posible.

Hoy en día la planificación y operación de sistemas eléctricos y de la producción, la logística del transporte, inventarios, la gestión de carteras de valores, entre otras, son tareas que involucran incertidumbre, ocasionada en algunos casos por carencia de datos fiables, por errores de medida ó parámetros que representan información sobre el futuro. En cualquiera de las tareas mencionadas se requieren modelos que acojan estas situaciones y que tengan en cuenta la incertidumbre y la posibilidad de tomar decisiones para resultados óptimos desde una gama de situaciones o escenarios posibles. El análisis de datos del mundo real a menudo se ve afectada por diferentes tipos de errores como: errores de medición, errores de cálculo, imprecisión relacionada con el método adoptado para la estimación de los datos. La incertidumbre en los datos, que está estrictamente relacionada con los errores anteriores, puede ser considerada, en lugar de un único valor para cada uno de los datos, como un intervalo de valores en los que la variable puede caer, esto se conoce como datos de intervalo. Las unidades estadísticas descritas por los datos de intervalo se puede asumir como

un caso especial de objetos simbólicos. En el análisis simbólico de datos, estos objetos se representan como cajas. Basados en la extensión del análisis de componentes principales a datos de intervalo, la intención del presente artículo es extender la metodología de regresión por mínimos cuadrados parciales PLS a datos de intervalo, que no constituyen un espacio Euclídeo. Para lograr dicho objetivo se utilizan, cuando es posible, los instrumentos de álgebra de intervalos para adaptar los modelos matemáticos, sobre la base de PLS clásico, para el caso en que se da una matriz de datos de intervalos.

El presente artículo está estructurado de la siguiente manera: primero se presenta la extensión del análisis de componentes principales con datos de intervalo, luego se presentan varios enfoques de regresión con datos de intervalo y por último se presenta la metodología de regresión PLS con datos de intervalo.

2. ELEMENTOS PRELIMINARES

Se han tratado diferentes enfoques para llevar a cabo el análisis de regresión lineal para los datos de intervalo, desde que se presentó el primer enfoque Billard & Diday (2000). Primero se hace una propuesta donde se ajusta un modelo de regresión lineal para el punto central de los intervalos, luego, se aplica el modelo ajustado a los límites inferior y superior de las variables independientes para hacer predicciones de los límites inferior y superior, respectivamente. Neto *et. al.* (2004) y De Carvalho *Et. al.* (2004) transforman las variables de intervalo originales en variables de punto central y rango y luego llevan a cabo un análisis de regresión clásica en cada una de las variables de punto central y variables de rango por separado. Posteriormente Billard & Diday (2007) extienden los conceptos de Lima Neto y de Carvalho considerando el punto central y el rango del intervalo de manera simultánea. Neto *et. al.* (2005,2010) mejoran su propuesta proponiendo un problema de programación lineal restringida. La principal desventaja de todos estos métodos es la pérdida de información al realizar las regresiones, dado que no trabajan con el intervalo sino que pasan de intervalos a puntos.

Por otro lado, teniendo en cuenta que los modelos de regresión lineal son la solución a un problema de optimización, se han hecho propuestas de modelos de regresión lineal donde el problema de optimización que conlleva al modelo de regresión lineal, es un problema de optimización intervalo-valuado Gioia & Lauro (2005).

2.1. Estimación de parámetros con optimización intervalo-valuada

A continuación se presentan algunos resultados obtenidos en Gallego-Posada & Puerta-Yepes (2015). En dichos resultados se muestra la estimación de parámetros intervalos usando una metodología análoga a mínimos cuadrados sobre el conjunto $\mathcal{I}$.

2.1.1. Ajuste polinomial generalizado

Considere $c_j = [c_j^L, c_j^U] \in \mathcal{I}$ para $j = 0, 1, 2, \ldots, n$. Se dice que $p(x)$ es un polinomio generalizado si puede ser expresado de la forma:

$$p(x) = \sum_{j}^{n} c_j x^j = \sum_{j=0}^{n} [c_j^L, c_j^U] x^j.$$

Considere un conjunto de observaciones $y_i = [y_i^L, y_i^U] \in \mathcal{I}$ para $i = 1, 2, \cdots, p$. Se puede modelar el fenómeno usando un polinomio de grado n en forma matricial como sigue:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & x_2^3 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_p & x_p^2 & x_p^3 & \cdots & x_p^n \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

o de manera compacta como $y = Vc + \varepsilon$, donde V es la matriz de Vandermonde y el vector de error $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p)^T$ se supone media cero, varianza constante, no correlación, independencia con las variables independientes en la matriz de Vandermonde.

En Gallego-Posada & Puerta-Yepes (2015) se considera un polinomio de grado 10 con coeficientes de intervalo. Se toma una muestra aleatoria de intervalos teóricos y con esta información se busca estimar los valores originales de los coeficientes que generan este comportamiento. En la figura 1 se muestra el polinomio intervalo-valuado, donde las bandas verticales representan la muestra aleatoria de intervalos.

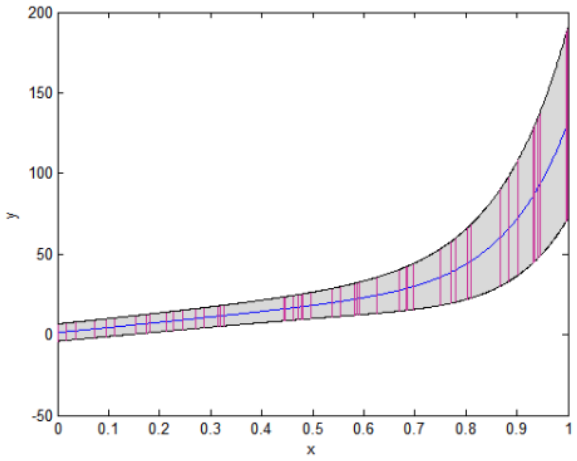

Figura 1: Gráfico polinomio intervalo-valuado.

Para estimar los coeficientes deseados se utilizaron varias técnicas de optimización. La primera técnica que se utiliza es la metodología de mínimos cuadrados ordinarios, donde la estimación de c está dada por $\widehat{c} = (V^T V)^{-1} V^T y$ siempre y cuando la inversa de la matriz $V^T V$ exista. Estos resultados pueden ser obtenidos minimizando la norma ℓ_2 de los residuales entre los puntos medios del modelo estimado y las medidas reales

$$\text{mín} \sum_{i=1}^{m} (m(y_i) - m(\widehat{y}_i))^2$$

donde $m(y_i)$ representa el punto medio de las medidas reales y $m(\widehat{y}_i)$ representa el punto medio del modelo estimado.

La segunda metodología que se implementa en Gallego-Posada & Puerta-Yepes (2015) para estimar los coeficientes del polinomio intervalo-valuado, es un algoritmo evolutivo llamado Differential Evolution (DE)

desarrollado originalmente por Price al tratar de resolver un problema de ajuste del polinomio de Chebychev propuesto por Storn. Una completa descripción de DE aparece en Storn & Price (1997). Las estimaciones obtenidas utilizando la metodología DE muestran que no hay una mejora sustancial en la calidad de la estimación en relación con los valores reales de los parámetros. La mayoría de las estimaciones caen casi en el punto medio de los intervalos reales. Sin embargo, algunos de los coeficientes se subestiman o sobrestiman en la longitud del intervalo. Además, dada la naturaleza de la heurística, la calidad de las estimaciones no es muy uniforme y, en algunos casos, la búsqueda no converge a valores adecuados de los parámetros.

Como otra alternativa, en Gallego-Posada & Puerta-Yepes (2015), se utiliza la implementación en el software CVX para optimización convexa desarrollado en I.CVX Research (2012) y Grant & Boyd (2008). Para evitar la sobre estimación de la longitud de los intervalos, la métrica inducida en $\mathcal{I}$ por la norma ℓ_2 se utiliza para medir los residuos, que se pueden expresar en términos de la distancia Hausdorff en $\mathcal{I}$. de esta manera, el problema de optimización puede ser expresado como:

$$\text{mín} \sum_{i=1}^{m} d_H(y_i, \widehat{y}_i)$$

donde d_H es la métrica de Hausdorff (1914) sobre el conjunto $\mathcal{I}$. Para $A = [a^L, a^U]$ y $B = [b^L, b^U]$, intervalos de $\mathcal{I}$, se tiene que:

$$d_H(X, Y) = \text{máx}\left\{|a^L - b^L|, |a^U - b^U|\right\}.$$

Los resultados de esta metodología muestran que las estimaciones coinciden con éxito con los valores reales de los parámetros, con errores de magnitud de 10^{-9} en relación a los puntos finales teóricos. De esta manera puede inferirse la potencia de la metodología y como esta captura la incertidumbre dada por las mediciones.

Con el fin de mostrar la potencia de la metodología, en Gallego-Posada & Puerta-Yepes (2015), se toma una función no tan suave como un polinomio.

2.1.2. Función de Weierstrass

Se considera la función de Weierstrass Hardy (1916) dada por:

$$f(x) = \sum_{n=0}^{\infty} a^n cos(b^n \pi x).$$

En este caso la intención es estimar el coeficiente a que es tomado como un intervalo, basados en un conjunto de medidas para $x \in [0, 1]$. Basados en la métrica ℓ_2 inducida en $\mathcal{I}$, se resuelve el problema de optimización mencionado previamente. En la Figura 2 se muestra la estimación de la función de Weierstrass junto con los intervalos estimados.

Como se puede ver en la figura, los coeficientes estimados para este modelo son capaces de manejar el comportamiento caótico y ruidoso de esta función, así como la extrema sensibilidad que existe en el parámetro.

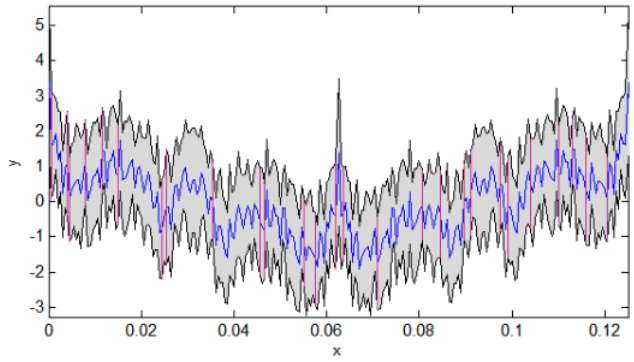

Figura 2: Estimación de la función de Weierstrass.

3. REGRESIÓN POLINOMIAL Y LINEAL MÚLTIPLE CON DATOS DE INTERVALO

3.1. Regresión lineal simple con datos de intervalo

En Gioia & Lauro (2005) se hace una extensión de la metodología de regresión lineal simple al caso intervalo valuado. A continuación se muestran de manera general dichos resultados.

Considere un conjunto de p pares $\left(X_1{}^{\mathcal{I}}, Y_1{}^{\mathcal{I}}\right)$, $\left(X_2{}^{\mathcal{I}}, Y_2{}^{\mathcal{I}}\right)$, $\cdots$, $\left(X_p{}^{\mathcal{I}}, Y_p{}^{\mathcal{I}}\right)$, donde:

$$X_j{}^{\mathcal{I}} = [\underline{x}_j, \overline{x}_j], \quad Y_j{}^{\mathcal{I}} = [\underline{y}_j, \overline{y}_j], \quad j = 1, 2, \ldots, p.$$

El propósito de la propuesta es considerar todas las posibles combinaciones de pares ordenados (x_i, y_i) con $x_i \in X_i^{\mathcal{I}}$ y $y_i \in Y_i^{\mathcal{I}}$ y determinar los estimadores $\widehat{\beta}_0^{\mathcal{I}}$, $\widehat{\beta}_1^{\mathcal{I}}$ de los parámetros $\beta_0^{\mathcal{I}}$, $\beta_1^{\mathcal{I}}$, tales que:

$$Y^{\mathcal{I}} = \beta_0^{\mathcal{I}} + \beta_1^{\mathcal{I}} X^{\mathcal{I}} + \varepsilon.$$

Para dicho propósito se consideran los conjuntos:

$$\widehat{\beta}_1 = \left\{ \widehat{\beta}_1(x_1, \cdots, x_p, y_1, \cdots, y_p) = \frac{\sum_{i=1}^{p} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{p} (x_i - \overline{x})^2} : x_i \in X_i^{\mathcal{I}}, y_i \in Y_i^{\mathcal{I}} \right\} \tag{1}$$

$$\widehat{\beta}_0 = \left\{ \widehat{\beta}_0(x_1, \cdots, x_p, y_1, \cdots, y_p) = \overline{y} - \widehat{\beta}_1 \overline{x} : x_i \in X_i^{\mathcal{I}}, y_i \in Y_i^{\mathcal{I}} \right\}. \tag{2}$$

Maximizando y minimizando las funciones 1 y 2, se obtienen los siguientes intervalos:

$$\widehat{\beta}_1^{\mathcal{I}} = \left[\min_{x_i \in X_i^{\mathcal{I}}, y_i \in Y_i^{\mathcal{I}}} \widehat{\beta}_1 \ , \quad \max_{x_i \in X_i^{\mathcal{I}}, y_i \in Y_i^{\mathcal{I}}} \widehat{\beta}_1 \right] \tag{3}$$

$$\widehat{\beta}_0^{\mathcal{I}} = \left[\min_{x_i \in X_i^{\mathcal{I}}, y_i \in Y_i^{\mathcal{I}}} \widehat{\beta}_0 \ , \quad \max_{x_i \in X_i^{\mathcal{I}}, y_i \in Y_i^{\mathcal{I}}} \widehat{\beta}_0 \right]. \tag{4}$$

En Gioia & Lauro (2005) se muestran varios ejemplos utilizando esta metodología.

3.2. Regresión de polinomios con datos de intervalo

En la subsección 2.1 se hace una primera extensión de la metodología de regresión por mínimos cuadrados ordinarios en el caso de variables explicativas en el conjunto de los números reales $\mathbb{R}$ y variable respuesta intervalo. Dicha extensión se hizo utilizando la métrica inducida en $\mathcal{I}$ por la norma ℓ_2; esto es, por medio de la distancia Hausdorff en $\mathcal{I}$. La intención ahora es extender dichos conceptos considerando tanto las variables explicativas como la respuesta de tipo intervalo.

El propósito es determinar los estimadores $\widehat{\beta}_0^{\mathcal{I}}$, $\widehat{\beta}_1^{\mathcal{I}}$, $\cdots$, $\widehat{\beta}_q^{\mathcal{I}}$ de los parámetros $\beta_0^{\mathcal{I}}$, $\beta_1^{\mathcal{I}}$, $\cdots$, $\beta_q^{\mathcal{I}}$ tales que:

$$Y^{\mathcal{I}} = \beta_0^{\mathcal{I}} + \beta_1^{\mathcal{I}} X^{\mathcal{I}} + \beta_2^{\mathcal{I}} (X^{\mathcal{I}})^2 + \cdots + \beta_q^{\mathcal{I}} (X^{\mathcal{I}})^q + \varepsilon,$$

con $(X^{\mathcal{I}})^k = X^{\mathcal{I}}(X^{\mathcal{I}})^{k-1}$ donde $k = 2, \ldots, q$.

Dicho propósito se logra resolviendo el problema de optimización:

$$\text{mín} \sum_{i=1}^{n} d_H(Y_i^{\mathcal{I}}, \widehat{Y}_i^{\mathcal{I}}).$$

3.3. Regresión lineal múltiple con datos de intervalo

El objetivo de la regresión lineal múltiple con datos de intervalos, es construir un modelo que relacione una variable dependiente de tipo intervalo $Y^{\mathcal{I}}$ con un conjunto de variables explicativas de tipo intervalo $X_1^{\mathcal{I}}$, $X_2^{\mathcal{I}}$, $\cdots$, $X_q^{\mathcal{I}}$. La relación de tipo lineal mencionada está dada por:

$$Y^{\mathcal{I}} = \beta_0^{\mathcal{I}} + \beta_1^{\mathcal{I}} X_1^{\mathcal{I}} + \beta_2^{\mathcal{I}} X_2^{\mathcal{I}} + \cdots + \beta_q^{\mathcal{I}} X_q^{\mathcal{I}} + \varepsilon \tag{5}$$

De manera análoga a la regresión de polinomios con datos de intervalo, la estimación de los parámetros $\widehat{\beta}_0^{\mathcal{I}}$, $\widehat{\beta}_1^{\mathcal{I}}$, $\cdots$, $\widehat{\beta}_q^{\mathcal{I}}$, se logra resolviendo el problema de optimización:

$$\text{mín} \sum_{i=1}^{n} d_H(Y_i^{\mathcal{I}}, \widehat{Y}_i^{\mathcal{I}}).$$

3.4. Metodología de regresión por componentes principales con datos de intervalo

En Federica & Carlo (2006) se hace una extensión del álgebra lineal en relación al cálculo de valores y vectores propios para el caso de matrices con entradas de intervalos. Por otro lado, se muestra como se utiliza el análisis de componentes principales cuando se tiene este tipo de datos. Si bien el cálculo de valores y vectores propios está propuesto de manera teórica, existen actualmente algoritmos numéricos que permiten llevar a cabo los cálculos de una manera más sencilla (Hladik *et al.*, 2008; Hladik *et al.*, 2009; Hladik, *et al.*, 2011; Stoyanov, 2014; Rhon, 1993).

Utilizando los elementos desarrollados en Federica & Carlo (2006) y la metodología de regresión lineal múltiple con datos de intervalo propuesta en la expresiob dada en (3.3), se propone a continuación la metodología de regresión por componentes principales con datos de intervalo. Considere una variable

dependiente de tipo intervalo $Y^{\mathcal{I}}$ y un conjunto de variables explicativas de tipo intervalo $X_1^{\mathcal{I}}$, $X_2^{\mathcal{I}}$, $\cdots$, $X_q^{\mathcal{I}}$, que presentan multicolinealidad.

1. Se utiliza el análisis de componentes principales desarrollada en Federica & Carlo (2006) y se calculan las componentes principales $C_1^{\mathcal{I}}$, $C_2^{\mathcal{I}}$, $\cdots$, $C_k^{\mathcal{I}}$, donde $k < p$, que recogen la mayor variabilidad.
2. Con las componentes principales calculadas en el punto 1, se utiliza la metodología de regresión lineal múltiple con datos de intervalo propuesta en la ecuación (3.3); esto es, se determinan estimadores $\widehat{\gamma}_0^{\mathcal{I}}$, $\widehat{\gamma}_1^{\mathcal{I}}$, $\cdots$, $\widehat{\gamma}_k^{\mathcal{I}}$ de los parámetros $\gamma_0^{\mathcal{I}}$, $\gamma_1^{\mathcal{I}}$, $\cdots$, $\gamma_k^{\mathcal{I}}$ tales que:

$$Y^{\mathcal{I}} = \widehat{\gamma}_0^{\mathcal{I}} + \widehat{\gamma}_1^{\mathcal{I}} C_1^{\mathcal{I}} + \widehat{\gamma}_2^{\mathcal{I}} C_2^{\mathcal{I}} + \cdots + \widehat{\gamma}_k^{\mathcal{I}} C_k^{\mathcal{I}} + \varepsilon. \tag{6}$$

Como sucede en el caso clásico, la metodología de regresión por componentes principales con datos de intervalo, sólo tiene en cuenta las variables explicativas para resolver el problema de multicolinealidad y en ningún paso del algoritmo tiene en cuenta la variable respuesta. De esta manera entonces, tiene sentido pensar en una metodología de regresión que resuelva el problema de la multicolinealidad pero teniendo en cuenta la variable respuesta. Dicha metodología se lleva a cabo en la siguiente sección.

4. REGRESIÓN PLS CON DATOS DE INTERVALO

El propósito de la presente sección es extender la metodología de regresión por mínimos cuadrados parciales PLS sobre espacios euclídeos, al caso de metodología de regresión por mínimos cuadrados parciales PLS con datos de intervalo. Como se menciona en Moore *et al.* (2009), el conjunto de intervalos $\mathcal{I}$ no es un espacio vectorial; sin embargo se puede inmersar (Fernandez, 2008) en un espacio vectorial usando el teorema de Rådström (1953). De esta manera se extiende la metodología de regresión PLS sobre un espacio no Euclídeo. La extensión de la metodología PLS al caso intervalo-valuado se presenta de manera teórica; resultados numéricos con datos reales y simulados se presentarán en trabajos futuros.

Considere dos multi matrices $X^{\mathcal{I}}$ y $Y^{\mathcal{I}}$, cuyos datos en las filas provienen de n individuos u objetos, donde $X^{\mathcal{I}}$ contiene la información de p características y $Y^{\mathcal{I}}$ describe q propiedades. El objetivo es determinar una relación lineal:

$$Y^{\mathcal{I}} \approx X^{\mathcal{I}} B^{\mathcal{I}}$$

En lugar de determinar esta relación directamente, se tiene que tanto $X^{\mathcal{I}}$ como $Y^{\mathcal{I}}$ son modelados mediante variables latentes con base en los modelos de regresión:

$$X^{\mathcal{I}} = T^{\mathcal{I}} (P^{\mathcal{I}})^T + E \quad \text{y} \quad Y^{\mathcal{I}} = U^{\mathcal{I}} (Q^{\mathcal{I}})^T + F,$$

donde las matrices E y F son las matrices de error y la relación entre los scores está dada por:

$$U^{\mathcal{I}} = T^{\mathcal{I}} D^{\mathcal{I}}.$$

A partir de los métodos del centro, del centro y el rango, bivariante y restringido, se extiende de manera natural la metodología de regresión por mínimos cuadrados parciales PLS al caso intervalo-valuado. De esta manera se da lugar a los siguientes métodos cuando hay presencia de multicolinealidad en el caso intervalo-valuado:

1. Método del centro para la metodología PLS con datos de intervalo.

2. Método del centro y el rango para la metodología PLS con datos de intervalo.
3. Método bivariante del centro y el rango para la metodología PLS con datos de intervalo.
4. Método restringido para la metodología PLS con datos de intervalo.

Estas metodologías son una propuesta inicial para resolver el problema intervalo-valuado; sin embargo su principal inconveniente es que botan la información de los intervalos desde el principio.

A continuación se propone la metodología de regresión por mínimos cuadrados parciales con datos de intervalos; donde la información de los intervalos no se bota en principio.

4.1. Algoritmo Kernel para PLS con datos de intervalo

Por razones técnicas, se utilizan otros vectores de cargas, $w^{\mathcal{I}}$ para los $x^{\mathcal{I}}$-variables y $c^{\mathcal{I}}$ para las $y^{\mathcal{I}}$-variables; esto es $t^{\mathcal{I}} = X^{\mathcal{I}}w^{\mathcal{I}}$ y $c^{\mathcal{I}} = Y^{\mathcal{I}}c^{\mathcal{I}}$. De manera análoga a la metodología clásica, se tiene el siguiente algoritmo:

1. Se hallan $w_1^{\mathcal{I}}$ es el eigenvector asociado al eigenvalor más grande de $(X^{\mathcal{I}})^T Y^{\mathcal{I}}(Y^{\mathcal{I}})^T X^{\mathcal{I}}$ y el eigenvector, $c_1^{\mathcal{I}}$, , al eigenvalor más grande de $(Y^{\mathcal{I}})^T X^{\mathcal{I}}(X^{\mathcal{I}})^T Y^{\mathcal{I}}$.
2. Se calculan los scores de las direcciones encontradas, dadas por las proyecciones:
$$t_1^{\mathcal{I}} = X^{\mathcal{I}}w_1^{\mathcal{I}}, \qquad u_1^{\mathcal{I}} = Y^{\mathcal{I}}c_1^{\mathcal{I}}$$
3. Se calculan las siguientes variables latentes:
   3.1 $p_1^{\mathcal{I}}$ es calculada en relación al modelo $X^{\mathcal{I}} = T^{\mathcal{I}}(P^{\mathcal{I}})^T$, utilizando la metodología de regresión lineal múltiple con datos de intervalo propuesta en 3.3.
   3.2 $q_1^{\mathcal{I}}$ es calculada en relación al modelo $Y^{\mathcal{I}} = U^{\mathcal{I}}(Q^{\mathcal{I}})^T$, utilizando la metodología de regresión lineal múltiple con datos de intervalo propuesta en 3.3.
4. A partir de las variables latentes $t_1^{\mathcal{I}}$, $p_1^{\mathcal{I}}$, $u_1^{\mathcal{I}}$ y $q_1^{\mathcal{I}}$; se construyen las matrices desinfladas $X_1^{\mathcal{I}}$ y $Y_1^{\mathcal{I}}$, dadas por:
$$X_1^{\mathcal{I}} = X^{\mathcal{I}} \ominus t_1^{\mathcal{I}}(p_1^{\mathcal{I}})^T, \quad Y_1^{\mathcal{I}} = Y^{\mathcal{I}} \ominus u_1^{\mathcal{I}}(q_1^{\mathcal{I}})^T,$$
   donde la diferencia Hakuhara $\ominus$ es un tipo especial de sustracción entre intervalos de $\mathcal{I}$. Si $A = [a^L, a^U]$, $B = [b^L, b^U] \in \mathcal{I}$, entonces:
$$A \ominus B = [a^L - b^L, a^U - b^U], \text{ sii } a^L - b^L \leq a^U - b^U.$$
   Utilizando las matrices $X_1^{\mathcal{I}}$ y $Y_1^{\mathcal{I}}$, se tiene que $w_2^{\mathcal{I}}$ es el eigenvector asociado al eigenvalor más grande de $(X_1^{\mathcal{I}})^T Y_1^{\mathcal{I}}(Y_1^{\mathcal{I}})^T X_1^{\mathcal{I}}$. De manera análoga, $c_2^{\mathcal{I}}$ es el eigenvector asociado al eigenvalor más grande de $(Y_1^{\mathcal{I}})^T X_1^{\mathcal{I}}(X_1^{\mathcal{I}})^T Y_1^{\mathcal{I}}$,
5. El proceso continúa de manera análoga y se calculan $w_1^{\mathcal{I}}, w_2^{\mathcal{I}}, \cdots, w_a^{\mathcal{I}}$ y $c_1^{\mathcal{I}}, c_2^{\mathcal{I}}, \cdots, c_a^{\mathcal{I}}$ o de manera compacta, las matrices $W^{\mathcal{I}}$ y $W^{\mathcal{I}}$. De esta manera, se propone:
$$B^{\mathcal{I}} = W^{\mathcal{I}}\left((P^{\mathcal{I}})^T W^{\mathcal{I}}\right)^{-1}(C^{\mathcal{I}})^T,$$
   donde la inversa de una multimatriz se calcula según Rhon (2011).

4.2. Algoritmo NIPALS para PLS con datos intervalo

A continuación se muestra una versión del algoritmo NIPALS para PLS con datos intervalo, con los principales pasos. Si se quiere calcular la primera componente PLS se procede así:

1. Inicialice $u_1^{\mathcal{I}}$, por ejemplo, con la primera fila de la matriz $Y^{\mathcal{I}}$.

2. Calcule $w_1^{\mathcal{I}}$ resolviendo $X^{\mathcal{I}} = u_1^{\mathcal{I}}(w_1^{\mathcal{I}})^T$.

3. $t_1^{\mathcal{I}} = X^{\mathcal{I}} w_1^{\mathcal{I}}$.

4. Calcule $c_1^{\mathcal{I}}$ resolviendo $Y^{\mathcal{I}} = t_1^{\mathcal{I}}(c_1^{\mathcal{I}})^T$.

5. $(u_1^*)^{\mathcal{I}} = Y^{\mathcal{I}} c_1^I$.

6. $\Delta u^{\mathcal{I}} = d_H((u_1^*)^{\mathcal{I}}, u_1^{\mathcal{I}})$.

7. Si $\Delta u^{\mathcal{I}} < \varepsilon$, entonces pare; sino $u_1^{\mathcal{I}} = (u_1^*)^{\mathcal{I}}$ y vuelva al paso 2.

8. Calcule $p_1^{\mathcal{I}}$ resolviendo $X^{\mathcal{I}} = t_1^{\mathcal{I}}(p_1^{\mathcal{I}})^T$.

9. Calcule $q_1^{\mathcal{I}}$ resolviendo $Y^{\mathcal{I}} = u_1^{\mathcal{I}}(q_1^{\mathcal{I}})^T$.

10. Calcule $d_1^{\mathcal{I}}$ resolviendo $u_1^{\mathcal{I}} = t_1^{\mathcal{I}}(d_1^{\mathcal{I}})^T$.

11. $X_1^{\mathcal{I}} = X^{\mathcal{I}} \ominus t_1^{\mathcal{I}}(p_1^{\mathcal{I}})^T$ y $Y_1^{\mathcal{I}} = Y^{\mathcal{I}} \ominus d_1^{\mathcal{I}} t_1 (c_1^{\mathcal{I}})^T$

Finalmente se tiene que:

$$B^{\mathcal{I}} = W^{\mathcal{I}} \left((P^{\mathcal{I}})^T W^{\mathcal{I}}\right)^{-1} (C^{\mathcal{I}})^T.$$

5. CONCLUSIONES

En la presente propuesta se presentan, de manera general, los elementos básicos que se han desarrollado hasta el momento para abordar el problema de regresión lineal múltiple con datos de intervalo. Como se mencionó antes, la gran mayoría de propuestas metodológicas no consideran toda la información aportada por los datos sino que desechan la información pasando de intervalos a puntos. Basados en estas construcciones, hemos propuesto una extensión de la metodología de regresión por mínimos cuadrados parciales PLS al caso intervalo-valuado donde se resuelve, al menos de manera teórica, el problema de la multicolinealidad para este tipo de datos. A modo de trabajo futuro se harán simulaciones que permiten evaluar la pertinencia de la metodología propuesta en el presente artículo y la comparación con otras metodologías existentes. Por ejemplo, la regresión PLS sobre Espacios Euclídeos se puede comparar con metodologías que resuelven el problema de la multicolinealidad, tales como: regresión por componentes principales, regresión de Ridge, regresión de Lasso, análisis y correlación canónica, entre otros. De esta manera, además de hacer simulaciones para evaluar el desempeño de la metodología propuesta en este artículo, también se puede pensar en proponer las metodologías mencionadas para el caso intervalo-valuado.

Referencias

Billard, L. & Diday, E. (2000). Regression Analysis for Interval-Valued Data. Data analysis, Classification, and Related Methods. *eds. H.A.L. Kiers, J.-P. Rassoon, P.J.F.Groenen, and M. Schader*, Springer-Verlag, Berlin. 369–374.

Billard, L. & Diday, E. (2007). Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley, Chichester. 295–306.

De Carvalho, F, Neto, E & Tenorio, C. (2004). A New Method to Fit a Linear Regression Model for Interval-valued Data, Springer-Verlag, Berlin. 295–306.

Federica, G. & Carlo, N. (2006). Principal components analysis on interval data . *Computational Statistic*. 21. 343–363

Fernandez, J.P. (2008). Optimización multi-objetivo intervalo valuada. Thesis of Master. Universidad EAFIT.

Gallego-Posada, J.D & Puerta-Yepes, M.E. (2015). Interval Analysis and Optimization Applied to Parameter Estimation under Uncertainty. *Preprint*.

Gioia, F. & Lauro, C. (2005). Basic Statistical Methods for Interval Data. *Statistica Applicata*, 17, In press.

Grant, M. & Boyd, S. (2000). Graph implementations for nonsmooth convex programs Recent Advances in Learning and Control, Springer-Verlag, Limited.

Hardy, G. (1916). Weierstrass non-differentiable function. *Transactions of the American Mathematical Society*, 17, 301.

Hausdorff. (1914). Grundzuege der mengenlehre. *Leipzig: Veit and Company.*

Hladik, M., Daney, D. & Tsigaridas, E. (2008). An Algorithm for the Real Interval Eigenvalue Problem. *Institut National of de Recherche en Informatique et en Automatique*, 6680, 1–28.

Hladik, M., Daney, D. & Tsigaridas, E. (2009). Bounds on eigenvalues and singular values of interval matrices. *Institut National of de Recherche en Informatique et en Automatique*, 1234, 1–18.

Hladik, M, Daney, D. & Tsigaridas, E. (2011). Characterizing and approximating eigenvalue sets of symmetric interval matrices. *Computers and Mathematics with Applications*, 62, 3152–3163.

I.CVX Research (2012). CVX: Matlab software for disciplined convex programming, version 2.0. *http://cvxr.com/cvx*.

Moore, R., Baker, R. & Claud, M. (2009). Introduction to Interval Analysis. Society for Industrial and Applied Mathematics, Philadelphia.

Neto, E.A., De Carvalho & Tenorio, C. (2004). Univariate and Multi-variate Linear Regression Methods to Predict Interval-valued Features, Springer-Verlag, Berlin. 526–537.

Neto, E, De Carvalho & Tenorio, C. (2010). Constrained Linear Regression Models for Symbolic Interval-valued Variables, *Computational Statistics and Data Analysis*. 54**(2)**. 333–347.

Neto, E, De Carvalho & Tenorio, C.(2005). Applying Constrained Linear Aggression Models to Predict Interval-Valued Data, Springer-Verlag,Berlin. 92–106.

Rådström, H. (1953). An embedding theorem for spaces of convex sets. *American Mathematical Society*, 3, 165-169.

Rhon, J. (1993). Interval Matrices: Singularity and Real Eigenvalues. *Society for Industrial and Applied Mathematics*, 14, 82–91.

Rhon, J. (2011). Inverse Interval Matrix: A Survey. *Electronic Journal in Linear Algebra*, 22, 704–719.

Storn, R. & Price, K. (1997). Differential evolution-a simple and effcient heuristic for global optimization over continuous spaces. *Jornal of Global Optimization*, 11, 341–359.

Stoyanov. (2014). Eigenvalues of Symmetric Interval Matrices. Thesis of Master. Charles University in Praga.

Wold, H. (1975). Soft Modeling by Latent Variables; The Non-linear Iterative Partial Least Squares Approach. *Perspectives in Probability and Statistics*, 1–2.

Wold, H. (1985). Partial Least Squares. *Encyclopedia of Statistical Sciences*, 6, 581–591.

Wold, H. (2001). Personal Memories of the early PLS Development. *Chemometrics and Intelligent Laboratory Sistems*, 58, 109–130.

Wold, H. (1982). *Estimation of Principal Components and Related Models by Iterative Least Squares*. In Krishnaiah, P(ed.), Multivariate Analysis, Academic Press. New York. 391–420.

Wold, S, Albano, C, DunnIII,J, Edlund, U, Esbensen, K, Geladi, P, Hellberg, S, Johansson, E & Lindberg, W. (1984). *Multivariate Data Analysis in Chemestry, in Chemometrics, Mathematics and Statistics in Chemestry*. Reidel Publishing Company. Dordrecht. 17–18.