RNPS No. 2142
ISSN 2072-6287
Versión Digital

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada

REPORTE TÉCNICO
Reconocimiento de Patrones

SERIE AZUL

Regresión multivariante y multi-vías

Gabriela Barcas e
Isneri Talavera Bustamante

RT_089 **febrero 2017**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada

REPORTE TÉCNICO
Reconocimiento de Patrones

Regresión multivariante y multi-vías

Gabriela Barcas e
Isneri Talavera Bustamante

RT_089 **febrero 2017**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

Tabla de contenido

Lista de figuras

Regresión multivariante y multi-vías

Gabriela Barcas e Isneri Talavera Bustamante

Equipo de Investigaciones en Reconocimiento de Patrones, CENATAV - DATYS, La Habana, Cuba
{gbarcas,italavera}@cenatav.co.cu

RT_089, Serie Azul, CENATAV - DATYS
Aceptado: 5 de febrero de 2017

Resumen. La regresión de datos es una tarea de vital importancia dentro de la Quimiometría debido a la importancia que tiene el poder predecir el valor de una variable de interés altamente costosa de adquirir a través de un conjunto de otras variables accesibles que tienen relación con esta. Uno de los aspectos importantes dentro de esta tarea y cualquier otra en el Reconocimiento de Patrones es la representación adecuada de los datos para poder aplicar análisis óptimos. Comúnmente los datos espectrales son representados en el espacio vectorial que tiene desventajas tales como la alta dimensionalidad. La Representación por Disimilitudes y el Análisis Funcional de Datos son enfoques que tienen grandes ventajas en el trabajo con datos espectrales y entre sus ventajas está el que eliminan la alta dimensionalidad al tener en cuenta información más discriminativa de los objetos. Por ello el siguiente trabajo abarcará el estudio de los algoritmos de regresión existentes para datos multivariados y multivías en el espacio vectorial, funcional y de las disimilitudes.
Palabras clave: regresión, análisis multivariado, multivías, representación por disimilitud, análisis funcional de datos.

Abstract. Regression is a vitally important task within Chemometrics due to the importance of being able to predict the value of a highly expensive variable of interest through a set of other accessible variables that are related to it. One of the important aspects in this task and any other in Pattern Recognition is the adequate representation of the data in order to apply optimal analysis. Commonly, the spectral data are represented in a vector space having disadvantages such as a high dimensionality. The Dissimilarity Representation and Functional Data Analysis are approaches that have great advantages in the work with spectral data and among their advantages, an important one is that they eliminate the high dimensionality since they take into account more discriminating information of the objects. Therefore the following work will cover the study of existing regression algorithms for multivariate and multi-way data in the feature, functional and dissimilarities space.
Keywords: regression, multivariate analysis, multi-way, dissimilarity representation, funcional data analysis.

1. Introducción

En los últimos años el análisis de datos se ha ido extendiendo a esferas tan diversas como son: la industria alimenticia, la medicina, la química, la psicología, la biometría, etc. Tal diversidad provoca el aumento de la información contenida en los datos y por ende su complejidad. Técnicas utilizadas en la Quimiometría como la Espectroscopia UV, NIR o Cromatografía Gaseosa, entre otras, en algunos casos no son suficientes para analizar las muestras, por tanto la búsqueda de nuevas técnicas y combinaciones de estas para llevar a cabo un análisis óptimo se hace necesaria. En el procesamiento de datos y más específicamente en la esfera del Reconocimiento de Patrones, uno de los problemas a resolver es la posibilidad de poder predecir los valores de una (o varias) variable a partir de otro conjunto debido a que las variables de interés en

la mayoría de los casos son difíciles de obtener o altamente costosas. Esta tarea se conoce como regresión y se distingue del resto de los análisis estadísticos debido a que tiene como objetivo expresar un conjunto de variables respuestas en función de un conjunto de variables predictoras. En Quimiometría la regresión es de gran importancia debido a la necesidad de poder predecir valores como el alcohol en sangre, concentración de cocaína en una muestra, etc. Los datos espectrales normalmente son representados a través de vectores a pesar de ser ploteados como funciones. Con esta representación, la información funcional de los datos no es tenida en cuenta y algunas de las características de los datos que pueden resultar escenciales se ignoradan. La Representación por Disimilitud (DR) y el Análisis de Datos Funcionales (FDA) son enfoques que tienen en cuenta esta información y sobre las cuales se han obtenido muy buenos resultados. FDA, es una extensión del análisis multivariado tradicional para datos de naturaleza funcional y considera el espectro observado como una función continua de valores reales en lugar de un vector de observaciones individuales. Por otra parte DR, fue propuesta para trabajar en el espacio de las proximidades entre los objetos en lugar del espacio definido por sus caraterísticas[1]. El espectro es representado a partir de sus disimilitudes con otros espectros, con lo cual se tiene en cuenta mayor información discriminativa. Uno de los principales problemas en Quimiometría es que se cuenta con pocas cantidades de muestras con alta dimensionalidad. Este problema es resuelto en ambas representaciones debido a que incorpora información más discriminativa como lo es la estructura del objeto.

El uso de la Representación por Disimilitud ha sido pobremente abordada dentro de la regresión, sin embargo, en la clasificación se han reportado excelentes resultados tanto con datos multivariados como multi-vías [2] lo cual constituye un buen indicio para su aplicación en esta área. Por esto, el principal objetivo de este trabajo es realizar un estudio sobre los algoritmos de regresión existentes tanto para datos multivariados y multi-vías en la representación vectorial, funcional y por disimilitud.

2. Regresión

La regresión es un enfoque que relaciona dos o más conjuntos de variables entre sí. Modela un conjunto de variables Y, sobre la base de un conjunto bien seleccionado de variables X. Es importante destacar que a pesar de que lo más común es tener solo dos conjuntos (y sobre el cual está dirigida nuestra atención), muchos de los problemas son formulados usando múltiples conjuntos[3].

En el contexto de la regresión, la variable a ser modelada se conoce como variable dependiente o Y-variable y las variables predictoras como variables independientes o X-variables. Entre estos dos conjuntos de variables debe existir una relación cuantitativa, si las variables independientes cambian, el valor de la variable dependiente debe cambiar consecuentemente. Las variables dependientes generalmente son variables cuyas mediciones son costosas y difíciles de obtener, sin embargo las variables independientes son variables que se obtienen de forma más eficiente y fácil, que se usan para predecir los valores de las variables dependientes [3]. En el proceso de regresión, tienen lugar dos etapas importantes: la etapa de calibración y la de predicción.

Etapa de calibración:
La calibración es el uso de datos y conocimiento previo para determinar cómo predecir información desconocida Y a partir de medidas disponibles de X basándose en una función matemática [4]. De manera general, la observación Y puede tomar valores de distintos tipos de espacios muestrales. Específicamente, puede ser un dato univariado, un dato multivariado o un dato funcional.

Durante esta etapa se crea el modelo de regresión que describe la relación (X,Y) o más bien, se estiman los parámetros de dicho modelo usando valores conocidos de X e Y (ver Figura 1).

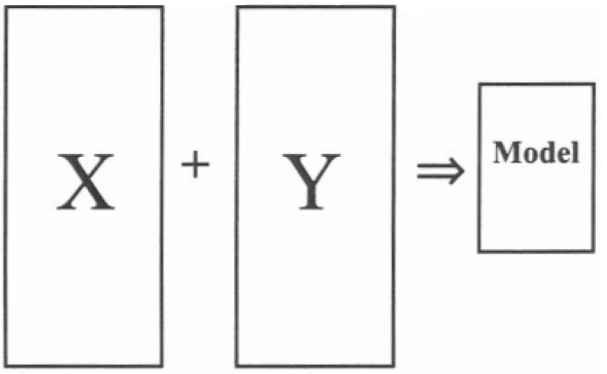

Fig. 1. Calibración: creación de un modelo a partir de valores conocidos de X e Y.

La calibración consiste raramente en solo en encontrar un modelo para describir las relaciones entre X e Y, sino que también se utiliza para predicciones futuras, como parte de la segunda etapa del proceso. La regresión busca una aproximación de las respuestas a partir de los predictores, la cual se ajusta de acuerdo a determinado criterio.

$$Y = f(X) + E = \hat{Y} + E, \tag{1}$$

donde f indica la relación matemática existente de manera general y E es la matriz que contiene los residuales, que no son más que la diferencia que existe entre el valor real de Y y el valor que se predice a partir del modelo[3].

La relación existente entre ambos conjuntos puede ser lineal o no, esta puede ser descrita a partir de una ecuación relativamente simple o puede implementar un algoritmo con una estructura menos evidente [5]. La relación se considera lineal cuando el comportamiento de ambas variables puede ser explicado a partir de una recta. Una recta no siempre es suficiente para modelar la relación entre las variables X e Y porque la relación exhibe algún grado de curvatura como se muestra en la Figura 2 , por lo que un modelo lineal resulta inadecuado y es necesario aplicar un modelo no lineal.

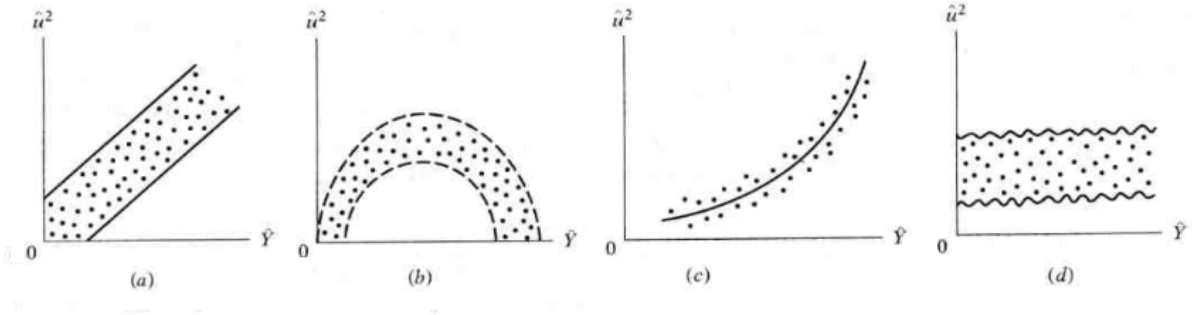

Fig. 2. Tipos de relaciones entre variables dependientes e independientes. a) Relación lineal, b),c) y d) Relación no lineal.

Los algoritmos de regresión lineal, funcionan razonablemente bien con grandes conjuntos de datos y son eficientes, pero son incapaces de detectar relaciones complejas dentro de estos. Por otro lado, los métodos no lineales presentan características opuestas, son buenos detectando relaciones complejas en los datos, pero no resultan eficientes. Una solución a este dilema, son las funciones kernel, que permiten conjugar la eficiencia de los algoritmos lineales con la flexibilidad de los no lineales.

Es relativamente simple crear el modelo que se ajuste a los datos de calibración, pero en muchos casos es inservible para predecir nuevas muestras. Este efecto tiene el nombre de *overfitting* o sobreajuste (Ver Anexos 7) y es un aspecto de peso en la creación del modelo [5]. La regresión puede realizarse directamente con los valores de las variables, pero los métodos más potentes usan un pequeño grupo de variables latentes (ver Anexos 7), el cual tiene las siguientes ventajas:

· Pueden ser usados los datos donde existe gran correlación con las variables independientes.

· La regresión puede ser aplicada sobre datos con más variables que muestras.

· La complejidad del modelo es controlado con el número de variables, así se puede evitar el sobreajuste y se puede alcanzar un mayor desempeño en la predicción.

Etapa de predicción:

Una vez obtenido el modelo de regresión, este es utilizado sobre un nuevo conjunto de mediciones de X para predecir nuevos valores de Y como se muestra en la Figura 3, en lugar de realizar nuevas mediciones de esta última.

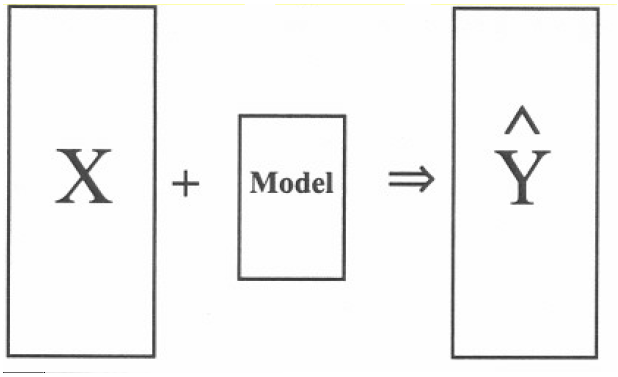

Fig. 3. Predicción: obtención nuevos valores de Y utilizando el modelo de regresión y nuevos valores de X.

2.1. Regresión univariada y multivariada

La forma más simple de regresión es la que relaciona una propiedad y a una sola variable independiente (como se muestra en la Figura 4) a partir de un modelo y toma el nombre de regresión univariada. Este tipo de regresión es un caso particular de la regresión multivariada y asume que no existen errores en las X-variables [5].

Sea X el vector que contiene n valores (objetos) de variables independientes y el vector Y contiene los n valores (respuestas) correspondientes a las variables dependientes. El modelo lineal que relaciona X e Y está definido como:

$$y = b_0 + bx + e, \tag{2}$$

donde, b y b_0 son los parámetros o coeficientes de regresión, b_0 es el intercepto y b es la pendiente. Dado que los datos en general no tienen una relación lineal perfecta, el vector e contiene los residuales(errores)[5].

La siguiente tarea es obtener los valores de los coeficientes de regresión para obtener un modelo confiable que relación los valores de x e y, que significa minimizar la función de los errores y variará en la forma de hacerlo según el algoritmo.

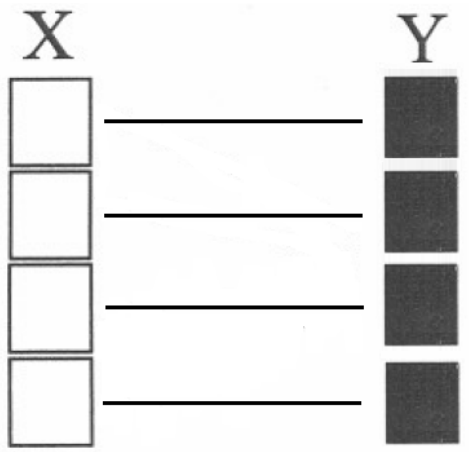

Fig. 4. Relación que existe entre las variables dependientes e independientes en la regresión univariada.

En regresión univariada solo existe una variable X para modelar Y, sin embargo, es posible medir varias variables X, para la misma propiedad (se conoce como regresión múltiple). Cuando se tiene más de una propiedad relevante, entonces se tienen conjuntos multivariados para X y Y. Si las propiedades están altamente correlacionadas, la combinación de todas las propiedades es admisible, de otro modo cada propiedad debe ser tratada de manera independiente.

Un dato multivariado consiste en una matriz X que con n filas y m columnas donde cada celda contiene un valor numérico. Cada fila corresponde a un objeto y cada columna una características particular del objeto (Figura 5)[5].

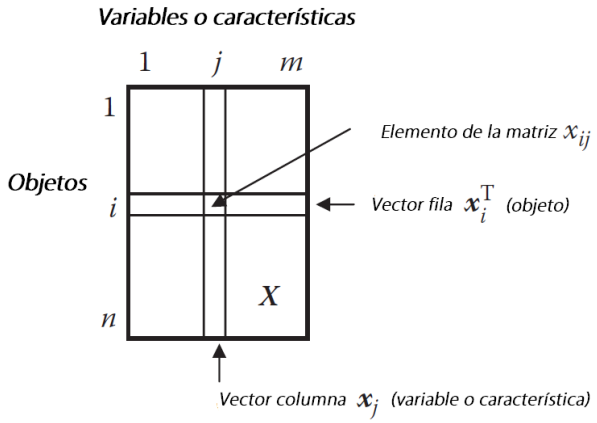

Fig. 5. Ejemplo de un datos multivariado, donde las filas representan los objetos o muestras y las columnas las variables o características medidas.

Si para la misma cantidad de objetos se tiene una medida de la propiedad (Figura 6) y la relación entre los conjuntos es lineal entonces se puede usar un modelo lineal que relacione todas las variables X con Y:

$$y_1 = b_0 + b_1 x_{11} + b_2 x_{12} + \ldots + b_m X_{1m} + e_1, \tag{3}$$

donde y está relacionado a la combinación lineal de las variables x más el término de error e. La diferencia con regresión univariada es que para cada variable x adicional se necesita un nuevo coeficiente de regresión.

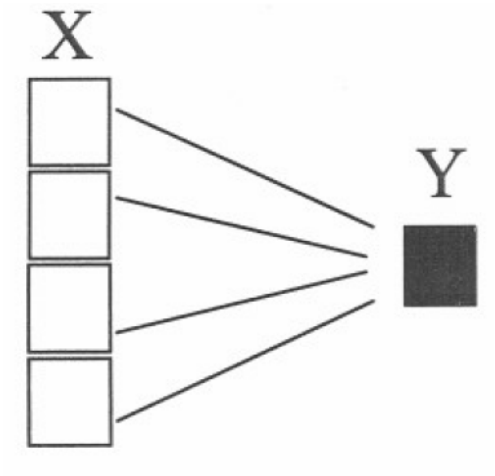

Fig. 6. Relación entre las variables dependientes e independientes en la regresión múltiple.

La regresión múltiple puede ser considerada como un caso especial de regresión multivariada ya que esta consiste en tener para varias mediciones de variables independientes, varias mediciones de variables dependientes, La cual resulta en un conjunto de ecuaicones de la siguiente forma:

$$\begin{aligned} y_1 &= b_0 + b_1 x_{11} + b_2 x_{12} + ... + b_m X_{1m} + e_1, \\ y_2 &= b_0 + b_1 x_{21} + b_2 x_{22} + ... + b_m X_{2m} + e_2, \\ &... \\ y_n &= b_0 + b_1 x_{n1} + b_2 x_{n2} + ... + b_m X_{nm} + e_n. \end{aligned} \tag{4}$$

El hecho de trabajar con varias variables en lugar de hacerlo con una como supone la regresión univariada constituyó un avance crucial debido a la cantidad de información que se dispone en el caso de regresión multivariada.

2.2. Regresión multi-vía

En los últimos años ha habido un incremento en la cantidad de problemas de regresión donde tanto los datos predictores como respuesta tienen estructuras multi-vía. Para hacer frente a esto se han desarrollado dos principales alternativas: una es redimensionar la estructura usando el desdoblado (Ver Anexo 7) y aplicar algoritmos tradicionales de análisis multivariado o generalizar estos algoritmos para los casos de estructuras multi-vías[6].

El análisis de datos multi-vías es una extensión del análisis multivariado para los datos con una estructura multi-vía. Un dato multivariado generalmente está dado por una matriz (estructura *two-way*) donde hay un número de objetos (filas) descritos por un conjunto de propiedades o características (columnas). Para una gran variedad de problemas la estructura de los datos puede ser más compleja, se pueden llevar a

tener por ejemplo, un grupo de propiedades para diferentes objetos medidas a diferentes instantes de tiempo como se puede ver en la Figura 7. Para este tipo de situaciones, es más apropiado una representación de mayor orden que los vectores o las matrices, los cuales son llamados arreglos multi-vías[2][7].

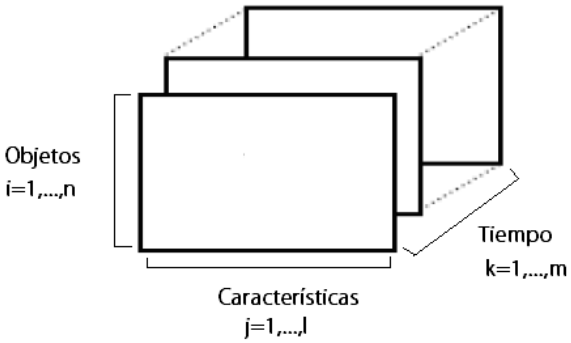

Fig. 7. Ejemplo de la estructura de un arreglo *three-way*.

La estructura multi-vías más común son los arreglos tridimensionales o *three-way* $X(I \times J \times K)$ que consisten en $objetos \times muestras \times condiciones$, pero es siempre posible generar datos de dimensiones aún mayores. Como se muestra en la Figura 8, el primer modo (a lo largo del eje horizontal) con índice i corresponde a los objetos, el segundo modo (a lo largo del eje vertical) con índice j es el de las variables y el tercer modo (a lo largo del eje de profundidad) con índice k es el de las condiciones [4].

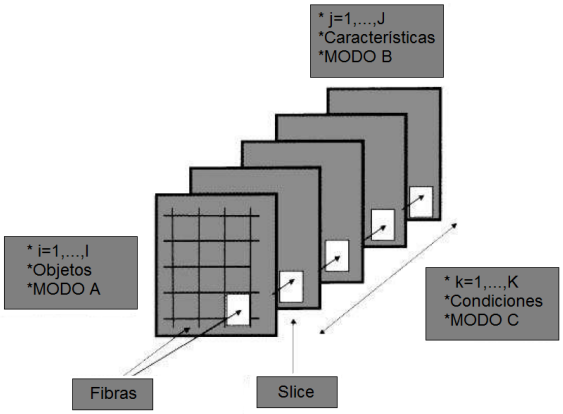

Fig. 8. Modos, *slices* y fibras que componen un arreglo *three-way*.

Los términos en este tipo de datos son un poco distintos a los datos *two-ways*, cada dimensión del arreglo multi-vías es llamado modo o *way*. Una estructura *three-way* también puede ser vista como una colección de matrices (Figura 9), conocidas como *slices* o *slab*. A su vez la estructura puede ser dividida en vectores que toman el nombre de *fibers* y dependiendo del tipo conforman filas, columnas y tubos como se puede ver en la Figura 10.

Dependiendo de en qué modo se encuentren las características de los objetos, existen diferentes diseños de este tipo de estructuras. El diseño más común es el definido por Kroonenberg [8], donde los objetos siempre se encuentran en el primer modo y las propiedades que los describen siempre se encuentran en los restantes modos.

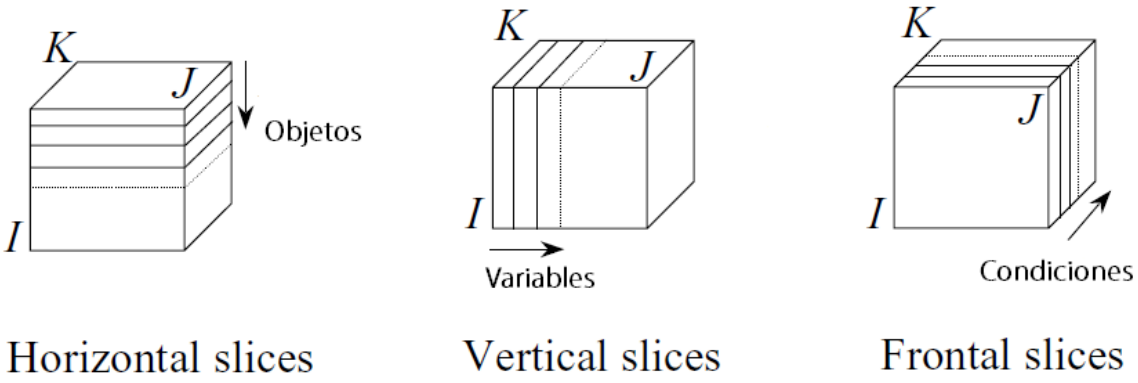

Fig. 9. Tipos de *slices* en un arreglo *three-way*.

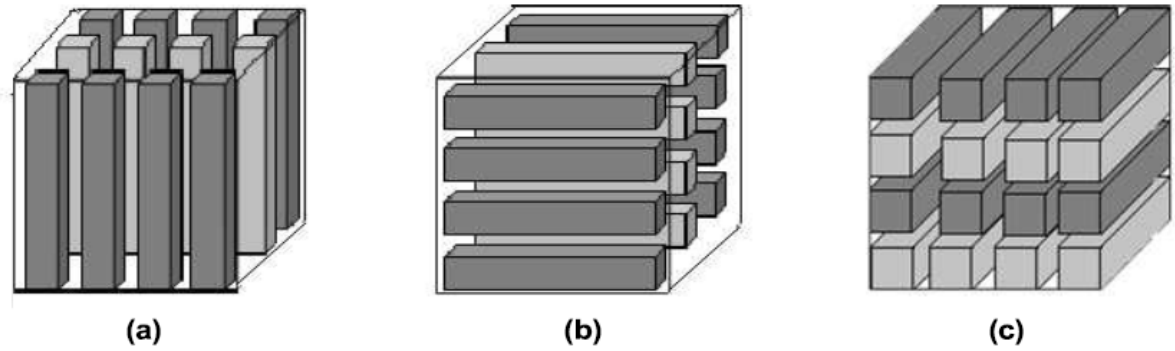

Fig. 10. Fibers en una estructura tridimensional:(a)columnas, (b)filas y (c) tubos.

Los datos multi-vías son recolectados debido a que todas las dimensiones son necesarias para dar respuesta a la interrogantes de la investigación[7]. La estructura multi-vías contiene información sobre la relación entre las propiedades, lo cual puede ser muy útil para una mejor comprensión de problema. La información obtenida de este tipo de estructuras es muy ventajosa para propósitos como regresión o clasificación, siempre que se utilicen las herramientas apropiadas para aprovechar sus beneficios.

Un procedimiento común usado sobre datos multi-vías es el desdoblado [8][7]. Este método consiste en llevar un arreglo multi-vías a una matriz, a partir de colocar todas las matrices de uno de los modos a continuación de la otra. De esta forma pueden ser aplicados métodos de análisis multivariado, sin embargo estos no son apropiados ya que no respetan el diseño multi-vía de los datos, se pierde la información de la estructura en sí y aumenta considerablemente la dimensionalidad del problema.

Un problema de regresión multi-vías consiste en encontrar una conexión entre una estructura multi-vías X de variables independientes y un vector, matriz o estructura multi-vías Y de variables dependientes [4]. Los algoritmos para este tipo de problemas se pueden dividir en dos grupos:

· Los métodos secuenciales calculan un componente a la vez.

· Los métodos simultáneos calculan todos los componentes simultáneamente minimizando cierto criterio, usualmente la función de suma del cuadrado de los residuales.

3. Algoritmos de regresión para datos vectoriales

3.1. Algoritmos de regresión lineal

3.1.1. OLS

Uno de los métodos más simples para la regresión lineal es el OLS (Ordinary Least-Squares Regression), este consiste en minimizar la suma del cuadrado de los residuales $\sum_{i=1}^{n} e_i^2$ para estimar los parámetros b y b_0. Una de las ventajas que tiene este modelo es el tener una ecuación explícita para estimar estos parámetros:

$$b = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \tag{5}$$

$$b_0 = \bar{Y} - b \cdot \bar{X}. \tag{6}$$

Note que la suma en el numerador de la Ecuación 5 es proporcional a la covarianza entre x e y, y el denominador es proporcional a la varianza de X. Este enfoque no es confiable si existen *outliers* en los datos, para ello es recomendable minimizar otra función de error que resulte en una estimaciones más robusta [5]. Para poder aplicar las ecuaciones 5 y 6 se asume que:

· Los errores existen en Y pero no en X.

· Los errores no están correlacionados y siguen una distribución normal con media cero y varianza σ^2.

3.1.2. Estimador clásico o máximo verosímil

Al aplicar el método de máxima verosimilitud para estimar los parámetros desconocidos θ y x_0 se obtiene lo que se conoce como estimador clásico $\widehat{x_C}$ de x_0, introducido por Eisenhart [9]:

$$\hat{x}_C = \frac{y_0 - \hat{\alpha}}{\hat{\beta}} = \bar{x} + \frac{1}{\hat{\beta}}(y_0 - \bar{y}), \tag{7}$$

donde $\hat{\alpha}$, $\hat{\beta}$ son las estimaciones máximo verosímiles de los coeficientes de regresión α, β a partir de las variables dependientes e independientes, $\bar{x}$ e $\bar{y}$ son los valores de las medias de x_i y y_i $(i = 1, ..., n)$, respectivamente. El estimador $\hat{x}_C$ puede obtenerse también invirtiendo la recta de regresión ajustada $y_0 = \hat{\alpha} + \hat{\beta}x_0$ con respecto a x_0. Si se tuvieran varias observaciones futuras $y_{01}, ..., y_{0m} (m > 1)$ de Y , entonces $\hat{x}_C$ se define reemplazando y_0 en la expresión 7 por el valor de la media $\overline{y_0}$ de $y_{01}, ..., y_{0m}$.

El estimador clásico es consistente y su error cuadrático medio asintótico es la función constante σ_2/β_2 en todo el rango de posibles valores de x_0. Sin embargo, este estimador ha sido muy criticado porque no tiene momentos finitos y por tanto su error cuadrático medio es infinito [10][11].

3.1.3. Estimadores iinverso

El estimador inverso $\hat{x}_I$ propuesto por Krutchkoff [12][13], es una alternativa al estimador clásico, se construye a considerando el modelo de regresión lineal de X como función de Y, es decir

$$X = \gamma + \delta Y + \eta, \tag{8}$$

$\hat{\gamma}$ y $\hat{\delta}$ son las estimaciones de los coeficientes γ y δ de este modelo respectivamente y se calculan por mínimos cuadrados, evaluando la recta de regresión en $Y = y_0$ y se obtiene

$$\hat{x}_I = \hat{\gamma} + \hat{\delta} y_0. \tag{9}$$

3.1.4. Estimadores de Halperin y Hagwood

Estos estimadores son muy similares al estimador inverso, cuando el número de *m* observaciones de Y que corresponden al valor desconocido $X = x_0$ es igual a 1.

El estimador de Halperin está dado por:

$$\hat{x}_H = (1 - R_M)\bar{x} + R_M \hat{x}_C, \tag{10}$$

donde

$$R_M = \frac{\hat{\beta}^2}{\frac{n-2\hat{\sigma}^2}{nM\sigma_x^2}}. \tag{11}$$

3.1.5. Estimadores basados en correcciones al estimador clásico

Ali y Singh [14] propusieron una suma ponderada del estimador clásico $\hat{x}_C$ y $\bar{x}$, cuya idea subyacente es mejorar el comportamiento del estimador clásico en la interpolación, contrayéndolo hacia $\bar{x}$ cuando x_0 está cercano a $\bar{x}$.

$$\hat{x}_{AS} = \lambda \hat{x}_C + (1 - \lambda)\bar{x}, \tag{12}$$

donde,

$$\hat{\lambda} = \frac{\hat{\beta}_2 \hat{\varsigma}_2}{\hat{\beta}_2 \hat{\varsigma}_2 + \hat{\sigma}_2}. \tag{13}$$

Esta contracción es menor cuando x_0 está alejada de $\bar{x}$, por tanto se espera que este estimador tenga el mismo comportamiento que el estimador clásico en la extrapolación.

Otro estimador de suma ponderada fue propuesto por Srivastava y Singh[15] sobre la base de la teoría asintótica con respecto a perturbaciones pequeñas.

$$\hat{x}_{SS} = (1 - \lambda)\hat{x}_C + \lambda \hat{x}_I, \tag{14}$$

donde,

$$\lambda = \frac{1}{n-2}. \tag{15}$$

Este estimador puede considerarse como una corrección leve al estimador inverso cuando el tamaño de muestra *n* es pequeño y una corrección leve al estimador clásico cuando *n* es grande.

3.1.6. Estimador predictivo no bayesiano

En [16] se propone una generalización de la densidad predictiva no Bayesiana propuesta por Harris[17] para modelos de regresión.

$$\bar{x}_P = \arg\max_{x \in R} \bar{L}_P(x), \tag{16}$$

donde,

$$\hat{L}_P = \hat{f}_P(\bar{y}_0; x_0, D) = f_{N(u_0\hat{\gamma}^T, u_0(U^TU)^{-1}u_0^T\hat{\sigma}^2 + v/m)}(\bar{y}_0) f_{\hat{\sigma}_2}(v; \hat{\sigma}_2) dv, \tag{17}$$

siendo *D* los datos de entrenamiento $(x_0; y_0), ..., (x_n; y_n)$.

Este estimador se comporta mejor para la extrapolación que el estimador inverso y el estimador clásico. En la interpolación su comportamiento conduce a mejoras respecto al estimador clásico aunque no mejora el estimador inverso. La superioridad de este estimador con respecto al estimador clásico se hace mayor

en situaciones en que σ^2 es grande y β^2 es pequeña. En [16] se muestra como este enfoque no solo provee estimaciones puntuales de x_0 sino que también verosimilitudes predictivas para todos los posibles valores de este parámetro.

3.1.7. MRL

MLR es el clásico algoritmo en regresión múltiple, es la extensión del algoritmo OLS para este tipo de regresión y su modelo puede ser expresado en notación matricial como se muestra a continuación:

$$y = Xb + e. \tag{18}$$

Al igual que en regresión univariada, los errores son calculados:

$$e = y - \bar{y}. \tag{19}$$

El cálculo de los coeficientes en regresión múltiple para OLS se puede expresar de la siguiente forma:

$$b = (X^T X)^{-1} X^T y. \tag{20}$$

Una diferencia importante con respecto a la regresión univariada es que se necesita la inversa de la matriz $X^T X$ para calcular los coeficientes de regresión. Dado que esta matriz relaciona la covarianza de las variables x pueden existir problemas con valores de x altamente correlacionados. En los casos en los que exista colinealidad entre las variables x, la inversa puede ser inestable e incluso imposible de calcular. En el caso en el que existan más variables regresoras que objetos, la inversa no podrá ser calculada y no se podrán hacer predicciones de y. Para ello se propone reducir el número de variables independientes o utilizar métodos como PCR (sección 3.1.8) y PLS (seccion 3.1.9). MLR puede fallar cuando existe colinealidad, ruido o errores en las variables X y cuando existen más variables que muestras [5].

OLS multivariado

La regresión múltiple consiste en relacionar un conjunto de X-variables con una única variable Y, mientras que la regresión multivariada relaciona un conjunto de X-variables con un conjunto de Y-variables. Teniendo n observaciones para una cantidad q de y-variables y m x-variables, resulta en una matriz Y de $n \times q$ y una matriz X de $n \times (m+1)$ (incluyendo el intercepto). El modelo de regresión puede ser descrito por:

$$Y = XB + E, \tag{21}$$

donde B es la matriz $(m+1) \times q$ de coeficientes de regresión y E es la matriz $n \times q$ que contiene los errores de regresión. Para estimar los coeficientes de regresión usando OLS es:

$$B = (X^T X)^{-1} X^T Y. \tag{22}$$

Note que los coeficientes de regresión para todas las y-variables pueden ser calculados simultáneamente usando la ecuación anterior, pero solo para los casos en los no exista colinealidad en las x-variables y la cantidad de muestras sea mayor que la cantidad de variables. Debido a que los métodos OLS tienen un pobre desempeño cuando existe colinealidad en los datos una gran variedad de algoritmos han sido desarrollados para resolver este problema. La diferencia que existirá entre uno y otro se centra en el criterio sobre el cual se basan para elegir los componentes en X para la predicción [5].

3.1.8. PCR

PCR (Principal Component Regression) resuelve el problema de la colinealidad de los datos y reduce el número de variables regresoras, pero estas ya no son las x-variables originales, sino una combinación de ellas como se muestra en la Figura 11. La combinación lineal que usa PCR son los componentes principales de las x-variables, por tanto se puede decir que es una combinación entre PCA y MRL [5].

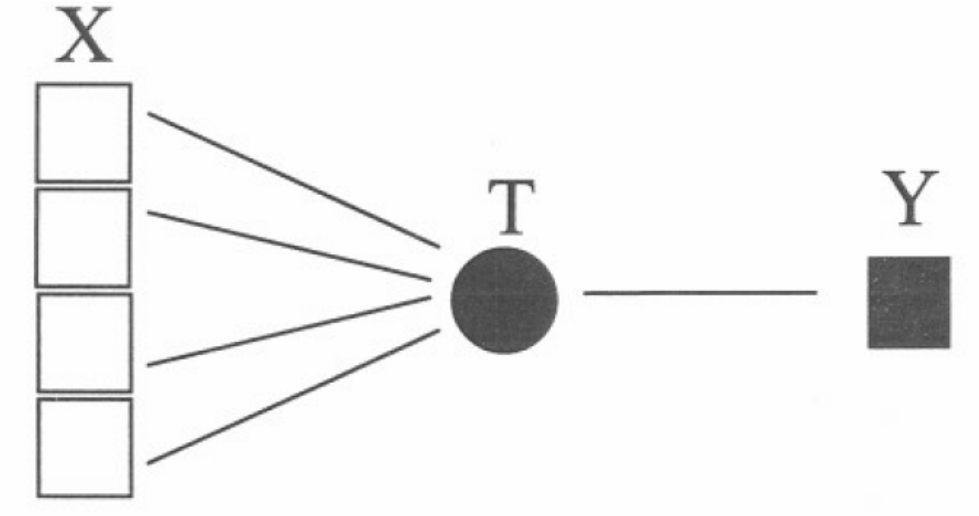

Fig. 11. Relación que establece PCR entre las variables dependientes e independientes.

PCA descompone la matriz X en los *scores* T y los *loadings* P. Para una cantidad a de componentes que usualmente es menos que la cantidad de variables de la matriz. Por tanto, el modelo de regresión descrito por PCR es:

$$X = TP^T + E, \tag{23}$$

donde E es la matriz de errores. Note que hasta este momento las y no son tenidas en cuenta. El modelo de regresión lineal múltiple para PCR es:

$$y = Tg + e_T, \tag{24}$$

con los nuevos coeficientes de regresión $g = P^T b$ y el término de error e_T. De esta forma se resuelve el problema de la colinealidad en los datos porque la información de la alta correlación en X se comprime en los vectores de *score* que no están correlacionados.

Usando la matriz T de *score* en lugar de las X originales, se puede aplicar OLS para estimar los coeficientes de la siguiente forma:

$$g = (T^T T)^{-1} T^T y. \tag{25}$$

Debido a la no correlación entre los *scores*, $T^T T$ es una matriz diagonal, por lo cual su inversa es fácil de calcular y es numéricamente estable. El modelo PCR a menudo tiene un desempeño similar al de PLS(ver sección 3.1.9). Normalmente este necesita más componentes que PLS porque no usa información de y para obtener los *scores* de PCA, esta característica no es necesariamente una desventaja debido a que considera mayor varianza de X con lo cual el modelo gana en estabilidad [5].

Un punto crucial en la construcción del modelo de predicción con PCR es determinar la cantidad de componentes a utilizar. En principio se pudiera utilizar la selección de variables(ver Anexos 7), pero para ganar en simplicidad, los componentes principales se ordenarán de manera decreciente de acuerdo con la varianza y calculando el error del modelo de regresión para los primeros componentes nos dirá la cantidad óptima de componentes.

La obtención de la óptima cantidad de componentes es realmente importante. Si se toman muy pocos no se ajusta X y no se predice bien Y, pero si se toman muchos componentes, se sobreajusta X e Y por tanto se predicen nuevos valores de Y inestables. Una estrategia para seleccionar un buen conjunto de *scores* de PCA para el caso de PCR es: seleccionar los primeros *scores* de PCA de manera que expliquen un por ciento determinado de la varianza total de X y luego de estos seleccionar los *scores* que tengan máxima correlación con y. Otra opción recomendable es usar métodos de validación para estimar la cantidad óptimo de componentes [5] [4].

3.1.9. PLS

PLS(Partial Least-Squares Regression) significa Mínimos Cuadrados Parciales y/o Proyección de Estructuras Latentes por Mínimos Cuadrados. Es un método ampliamente utilizado en Quimiometría para calibración multivariada. Este método utiliza la información de Y explícitamente para definir el espacio de las variables latentes. Busca componentes que ofrezcan un compromiso entre la varianza explicada de X y la predicción de las respuestas en Y. La propiedad más importante del algoritmo PLS es que la descomposición se lleva a cabo de forma tal que los *scores* tienen la máxima covarianza con las variables dependientes. Esta es la característica principal por lo cual difiere con PCR [18].

PLS es un método que relaciona la matriz X ya sea con un vector o una matriz Y. El concepto matemático de este método es menos estricto que OLS o PCR. La estructura es la misma para PLS que para PCR: la matriz X es transformada en un conjunto intermedio de variables latentes y estas nuevas variables son las que se usan para la regresión con las variables dependientes. PCR usa los *scores* de los componentes principales como factores mientras que PLS usa componentes que estén relacionados con y. El criterio mayormente usado por PLS para las variables latentes es el de máxima covarianza entre *scores* e y. La covarianza combina las grandes varianzas de X y la alta correlación con y, por lo que PLS puede ser considerado como un compromiso entre PCR y OLS [5]. PLS y PCR son modelos lineales donde las variables latentes finales predicen la propiedad y son combinaciones lineales de las variables originales. PLS admite la colinealidad entre las variables así como grandes cantidades de estas. PCR puede ofrecer errores de predicción tan bajos como los de PLS, pero casi siempre haciendo uso de más componentes.

Para el caso del modelado de una sola variable respuesta Y, el algoritmo se conoce bajo el nombre de PLS-1 y el cálculo del primer vector PLS latente para identificar la dirección del espacio multivariado definido por el vector de pesos w_1, de forma tal que los *scores* de esa dirección t_1 tengan la máxima covarianza con y, se puede expresar de la siguiente manera:

$$\begin{gathered} \max_{w_1}[cov(t_1, y)] = \max_{w_1}(t_1^T y), \\ t_1 = X w_1, \\ \|w_1\|^2 = 1. \end{gathered} \tag{26}$$

PLS asume que la regresión entre los bloques dependientes e independientes ocurre a nivel de *scores*, por lo que el siguiente paso del algoritmo es encontrar el coeficiente c_1 que relacione t_1 con y:

$$c_1 = \frac{y^T t_1}{t_1^T t_1}. \tag{27}$$

Una vez que la primera componente es calculada, se aplica el paso conocido como *deflaction* que consiste en eliminar tanto de X como de y la parte de la varianza ya modelada. Dado que los pesos w_1 describen la covarianza entre los bloques X e y, el paso de *deflaction* para la matriz X se basa en un segundo conjunto de coeficientes p, el cuál se asemeja a los *loadings* de PCA y se calculan como:

$$p_1 = \frac{X^T t_1}{t_1^T t_1}. \tag{28}$$

Los *loadings* p_1, en conjunto con los *scores* correspondientes, son utilizados para hacer *deflaction* de la matriz independiente:

$$E_{x,1} = X - t_1 p_1^T, \tag{29}$$

donde la matriz E_1 contiene la variación residual después de la substracción de la contribución de la primera componente PLS. De manera análoga, también se le hace *deflate* a la variable dependiente:

$$e_{Y,1} = y - c_1 t_1. \tag{30}$$

Aunque en el caso de regresión univariada no es obligatorio hacer *deflate* a las variables dependientes, ya que lo realizado en la Ecuación 30 es suficiente para hacer al bloque independiente X ortogonal a la variación de Y ya explicada. Después del paso de *deflaction* es posible calcular la segunda componente de PLS de la misma forma que la primera, con la única excepción de que X e y son sustituidas por $E_{X;1}$ y $e_{Y;1}$ respectivamente.

A continuación se calcula un segundo conjunto de *scores* t_2 y pesos w_2 y un segundo coeficiente de regresión c_2. Entonces se calculan unos nuevos *loadings* p_2, para realizar un nuevo paso de *deflaction* y el proceso continúa hasta que la cantidad de componentes deseadas son extraídas [5].

$$E_{X,2} = E_{X,1} - t_2 p_2^T, \tag{31}$$

$$e_{Y,2} = e_{Y,1} - c_2 t_2. \tag{32}$$

PLS2

El algoritmo PLS para regresión multivariada es denominado PLS2. El propósito sigue siendo el mismo, crear un modelo de calibración para las variables respuestas a partir de las variables predictoras X que maximice la covarianza entre los *scores* x e y.

PLS usualmente es introducido como un algoritmo matemático que maximiza una función objetivo bajo ciertas restricciones. La función objetivo es la covarianza entre los *scores* de x e y y la restricción es la ortogonalidad de los *scores* [5] [4].

En la regresión usando PLS2, se asumen que los datos de X e Y son multivariados, con dimensiones de $n \times m$ y $n \times q$ respectivamente. Por lo que el modelo PLS2 busca encontrar la relación lineal entre las variables X e Y usando la matriz B de $m \times q$ de coeficientes de regresión y una matriz de error como se describe en la siguiente ecuación

$$Y = XB + E. \tag{33}$$

Lejos de encontrar la relación directamente, X e Y son modelas como variables latentes.

$$X = TP^T + E_x, \tag{34}$$

$$Y = UQ^T + E_y, \tag{35}$$

donde, E_x y E_y son las matrices de errores, T y U son las matrices de *scores* y la matrices P y Q son las matrices de *loadings* con a columnas, donde $a \leq min(m,q,n)$ es la cantidad de componentes. Los *scores* en T y U son las combinaciones lineales de las variables X e Y respectivamente. La relación lineal entre los *scores* X e Y, se describe

$$u_j = d_j t_j + h_j, \tag{36}$$

donde, h_j son los residuales y d_j los parámetros de regresión. Si la relación entre u_1 y t_1 es fuerte (el valor de h_1 es pequeño) entonces el *score-x* del primer componente PLS es bueno para predecir los s*cores-y* y por tanto para predecir los datos y [5].

Para la construcción del modelo PLS la estructura de correlación del bloque dependiente es tenida en cuenta y utilizada de manera explícita, mientras que en MLR (OLS múltiple) y como consecuencia PCR, asumen que las respuestas son independientes. PLS es un método adecuado para analizar datos donde hay varias Y y en algunos casos ofrecer mejores resultados si Y es colinear debido a que hace uso de toda la información disponible en Y. Existen algunos casos donde PLS2 no logra modelar correctamente algunas variables Y, en casos como estos, la solución es aplicar PLS-1 separados, lo que provoca que se deba interpretar cada modelo construido de forma separada.

3.1.10. PCovR

El método Principal Covariates Regression fue propuesto en [19] y es una combinación entre PCA (Análisis de Componentes Principales=) de X y regresión de Y minimizando una función de pérdida de mínimos cuadrados bien definida.

Este método es nombrado PCovR porque: principal como PCA para enfatizar que los componentes deben tener la mayor varianza, covariables, para puntualizar que estos componentes también se relacionan con la varianza de Y y regresión, porque se tiene los conjuntos de variables dependientes e independientes. Sea X una matriz $n \times p$ de variables predictoras y Y una matriz de $n \times m$ de variables respuesta. No existen restricciones con respecto al tamaño de X y Y pero se tiene que $m < p$. Se tiene un sub-espacio de X conformado por a componentes $t_i (i = 1, a)$ que tenga en cuenta la máxima variación en X y Y, es decir, los componentes t_i deben ser capaces de contruir X y Y de la mejor forma posible.

$$T = XW, \tag{37}$$

$$X = TP_x + E_x, \tag{38}$$

$$Y = TP_y + E_y, \tag{39}$$

donde, T es la matriz de *scores* para los a componentes, W es la matriz de $p \times a$ de los pesos de los componentes, P_x y P_y contiene los parámetros de regresión relacionados con las variables X e Y respectivamente. E_x y E_y contienen los valores únicos de X e Y que no están correlacionados con los *scores* de T. Como criterio a maximizar en PCovR se propone:

$$\sigma \cdot R^2_{XT} + (1 - \sigma) \cdot R^2_{YT}, \tag{40}$$

donde, R^2_{XT} es el por ciento de varianza de X que se incluye en T y R^2_{YT} es el por ciento de varianza de Y que es explicada por T. Existen dos casos especiales: $\sigma = 0$ o $\sigma = 1$. Si $\sigma = 0$, el énfasis recae en el

ajuste de Y o si $\sigma = 1$, el énfasis recae completamente en reconstruir X. Claramente $\sigma = 0$ o $\sigma = 1$ nunca serán los valores óptimos, aunque es difícil la elección de un valor correcto para σ, es también de vital importancia para que ambos conjuntos tengan igual de importancia. Utilizar validación cruzada puede ser una solución razonable para obtener un óptimo valor de σ.

3.1.11. Unfold-PLS y Undold-PCR

La forma más fácil de convertir un problema de regresión multi-vías en un problema de regresión multivariado tradicional es aplicando desdoblado(Ver Anexo 7) en la estructura multi-vías y aplicando cualquiera de las técnicas de regresión existentes. Puesto que la matriz obtenida a partir del desdoblado tiene un alto grado de colinealidad, surgieron alternativas como Unfold-PLS (U-PLS) y Unfold-PCR (U-PCR)[20].

Como se menciona al principio, estos algoritmos consisten en aplicar desdoblado en los conjuntos (independiente o dependiente) que tengan estructura multi-vías para obtener matrices y aplicar sobre este resultado PLS o PCR. Estas variantes describen una cantidad igual o mayor de covarianza usando la misma cantidad de parámetros o más en el modelo. Sin embargo la interpretación de los resultados puede ser un poco más difícil debido a que las variables de los modos no son modeladas por separado sino mezcladas durante el desdoblado [6].

Para problemas de calibración donde las variables independientes tienen una estructura tridimensional y las variables dependientes son un vector, estos algoritmos plantean aplicar desdoblado en las variables independientes en el primer modo (en el de los objetos) y construir a partir de esta matriz el modelo de regresión usando PLS2 o PCR [18].

La desventaja que tiene el usar estos método es el desdoblado en sí. Los algoritmos que trabajan con la estructura multi-vías son más fáciles de interpretar y menos potencialmente propenso a ruido, debido a que la información de todos los modos es usada para la descomposición.

3.1.12. N-PLS

N-PLS o Multilinear PLS, propuesto por Bro en [18], es una generalización del algoritmo PLS para datos multi-vías [6]. Los métodos PLS consisten básicamente en dos pasos: primero se descomponen los datos de calibración y segundo se establece la relación entre los datos descompuestos para las variables dependientes e independientes [6] [21].

N-PLS es un algoritmo secuencial que calcula un componente cada vez. Descompone los datos independientes y dependientes para que cada par de *scores* correspondientes a X y Y tengan la máxima covarianza. El modelo N-PLS se calcula ajustando al mismo tiempo el modelo multilineal para cada uno de los bloques y el modelo de regresión relacionado con los *scores*. En términos matemáticos, teniendo $X(I \times J \times K)$ y $Y(I \times L \times M)$ el modelo N-PLS está compuesto por las siguientes relaciones:

$$X^{(I \times JK)} = T(W^K \odot W^J)^T + E_x, \tag{41}$$

$$Y^{(I \times LM)} = T(Q^M \odot Q^L)^T + E_y, \tag{42}$$

$$U = TC, \tag{43}$$

donde, los T y U son los *scores* y W y Q se utilizan para indicar los pesos (*weights*) del modelo PLS para X e Y respectivamente, C es la matriz de los coeficientes de regresión.

3.1.13. MCovR

Con el objetivo de construir un modelo cuyos componentes tengan la capacidad de explicar tanto la varianza de los datos dependientes e independientes se propone MCovR (Multi-way Covariates Regression), por Smilde y Kiers en [22], que es la extensión del algoritmo PCovR [19].

Multi-way Covariates Regression tiene la ventaja de ser un modelo simultáneo, por lo cual todos los componentes se extraen al mismo tiempo. Este método no establece limitantes sobre el tipo de estructura que puede ser usada para descomponer los bloques dependientes e independientes, por lo que para calcular los conjuntos de *loadings* P_x y P_y, puede usarse tanto el modelo TUCKER (Ver Anexo 7) como PARAFAC (Ver Anexo 7) [23]:

La descomposición del bloque X mediante PARAFAC:

$$P_X^T = (C \odot B)^T. \tag{44}$$

La descomposición del bloque X mediante TUCKER:

$$P_X^T = G(C \otimes B)^T. \tag{45}$$

En este algoritmo los componentes que se extraen de forma que maximicen la suma ponderada de la varianza explicada por los dos conjuntos (variables dependientes e independientes) y la importancia de ambas es regulada a partir de la constante σ como se establece en PCovR(Ver Sección 3.1.10). El enfoque de Multi-way Covariates Regression para el caso en que $X(I \times J \times K)$ y $Y(I \times L)$ es de la siguiente forma:

$$T = X^{I \times JK} W, \tag{46}$$

$$X^{(I \times JK)} = T P_X^{(F \times JK)} + E_X, \tag{47}$$

$$Y = T P_Y + E_Y, \tag{48}$$

donde, F es la cantidad de componentes del modelo y la matriz de pesos W satisface el criterio de mínimos cuadrados [23].

3.1.14. SCREAM

Desde el punto de vista teórico, los métodos multi-vías asumen que cada muestra es descrita con la misma cantidad de *loadings*. Sin embargo, estos métodos se vuelven menos adecuados cuando la cantidad de *loadings* cambian de forma entre una muestra y otra (como se muestra en la Figura 12). En estos casos es aún posible obtener un buen modelo de descomposición utilizando una de las modificaciones del algoritmo PARAFAC, que es PARAFAC2 [24]. En los problemas de calibración no existen alternativas para esto. A partir de esta limitante Bro et al. en [25] proponen el método SCREAM (Shifted Covariates Regression Analysis for Multi-way data), que está basado en una combinación de PARAFAC2 y el algoritmo de regresión PCovR [19].

SCREAM tiene como objetivo desarrollar un modelo de regresión a partir de variables latentes que sean útiles para la predicción. Para ello, propone utilizar PARAFAC2 para descomponer el conjunto de datos X y utilizar PCovR para calcular los coeficientes de regresión.

Definen una función de pérdida que tiene en cuenta el ajuste del conjunto X y la predicción del conjunto Y:

$$\sigma \|X - CP^T\|_F^2 + (1-\sigma)\|y - Cr\|_F^2, \tag{49}$$

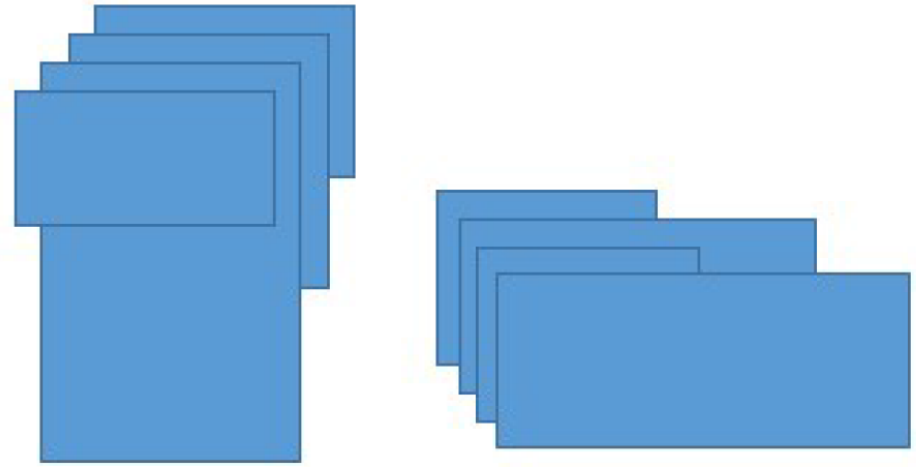

Fig. 12. Diferencias que pueden existir entre los *slices* de las filas y de las columnas en un arreglo three-way.

donde, C corresponde a los *loadings* del modo que contiene las muestras y P es una matriz que contiene los *loadings* de A y B_k:

$$\sum_{k=1}^{K}(X_k - AD_kB_k^T)^2 = \|X - CP^T\|_F^2. \tag{50}$$

Al mismo tiempo que se usa C para ajustar X al modelo PARAFAC2, se buscan los componentes en C que son relevantes para predecir las variables dependiente y el vector de regresión es explícitamente definido por C como un problema de MLR.

3.2. Algoritmos de regresión no lineal

Las Redes neuronales son un método popular de estimación de la regresión con alta dimensión y no linealidad; sin embargo, presentan algunas desventajas. En particular, su arquitectura tiene que ser determinada a *priori*, y además requieren estimar un número grande de parámetros no lineales, cuya determinación conduce a problemas de optimización con múltiples mínimos locales en los que resulta difícil hallar el mínimo global[26].

Los métodos de regresión basados en vectores de soportes como las Máquinas de Vectores Soportes (SVR) y las Máquinas de Vectores de Soporte de Mínimos Cuadrados (LSSVM) constituyen alternativas ventajosas, pues la cantidad de parámetros a estimar no aumenta con la dimensión de Y, presenta gran flexibilidad para modelar relaciones no lineales, tienen alta capacidad de generalización con pocos datos en el conjunto de calibración. A diferencia de las redes neuronales, la solución es única y global y su solución es rala. A pesar de estas ventajas, es necesario ajustar varios hiperparámetros, lo cual resulta costoso computacionalmente.

La Máquina de Vectores Relevantes [27][28] es un enfoque que elimina varias de las desventajas enteriormente mencionadas. Se basa en un modelo probabilístico ralo que tiene forma funcional idéntica a SVR y LSSVM, pero adopta un enfoque de aprendizaje Bayesiano para obtener las estimaciones. A pesar de que su aplicación en problemas de regresión es poco explorada, en [16] se evalúa su comportamiento.

Las Máquinas de Vectores Relevantes fueron propuestas para retomar las ideas principales de SVR en un contexto Bayesiano [27]. Dado un conjunto de datos de entrenamiento $D = x_i, y_{i=1}^n$, el modelo de regresión lineal generalizado que se usa para describir la relación de mapeo entre el vector de patrones de entrada y, y el escalar x es:

$$x_i = g(y_i; w) + e_i, \tag{51}$$

donde,

$$g(y, w) = \sum_{j=1}^{m} w_j \phi_j(y) + w_0, x = \Phi w, \tag{52}$$

$\Phi = [\phi_1, ..., \phi_m]$ es la matriz de diseño $n \times m$ cuyas columnas contienen el conjunto completo de los m vectores bases.

La Regresión Inversa Partida (SIR) es una metodología semi-paramétrica para la calibración multivariada. Fue introducida por Li[29] y es un método de reducción de dimensión que supone que toda la información en Y acerca de X puede obtenerse a través de su proyección en un subespacio de menor dimensión. La estimación del modelo se lleva a cabo mediante la estimación de la función de regresión inversa $E(Y/X)$, la cual subyace en dicho subespacio bajo ciertos supuestos.

Algoritmos de kernelización:

La kernelización se puede ver como la reformulación de un algoritmo de manera que la determinación de una pauta o regularidad lineal en los datos puedan llevarse a cabo exclusivamente a partir de la información recogida en los productos escalares calculados para todas las parejas de elementos del espacio. El conjunto de dichos productos escalares recoge, en esencia, la información existente en el conjunto de datos relativa a las normas de los elementos del espacio, así como a los ángulos que existen entre ellos. Aunque resulta evidente que prescindir de las coordenadas reales de los elementos en el espacio y limitarse a la información recogida en el conjunto de productos escalares supone una pérdida de información (por ejemplo, se pierde la información relativa a la orientación del conjunto de datos en el espacio o la información relativa a la alineación de los elementos con las variables originales), en muchas ocasiones esta pérdida no es relevante para alcanzar el objetivo de detección de patrones lineales [30].

Este enfoque se compone de [30]:

· Un conjunto de entrada X que no necesita tener una estructura algebraica particular. La clave de este enfoque consiste en que sea posible definir una función que a cada pareja de elementos de un espacio de entrada X, le haga corresponder un valor real y , que este definida sobre el producto cartesiano $X \times X$. No es necesario que X sea un espacio vectorial, la gran variedad de espacios de entrada sobre los que se puede aplicar esta metodología es una de sus propiedades más interesantes.

· Un conjunto de patrones F al que llamaremos Espacio de Características (Feature Space), que debe tener una estructura algebraica de Espacio de Hilbert. En particular, el espacio F debe estar dotado de un producto escalar:

$$<,>: F \times F \rightarrow \Re. \tag{53}$$

· Una función $\phi : X \rightarrow F$ que incrusta (*embeds*) cada elemento del espacio de entrada X en el espacio de características F . Esta función ϕ recibe el nombre de *embedding*.

· Un algoritmo de detección de patrones lineales en el *feature space* F que ha sido kernelizado, es decir, ha sido rediseñado de manera tal que para alcanzar su objetivo en F el algoritmo no necesita disponer de los valores concretos de $\phi(x) \forall x \in X$, sino tan solo de los productos escalares entre las imágenes de los elementos de X , es decir, $< \phi(x), \phi(x) > \forall x, z \in X$,siendo $<,> F \times F \rightarrow \Re$ el producto escalar definido en F.

· Una función real definida sobre el producto cartesiano del espacio de entrada con él mismo $k : X \times X \to \Re$, llamada función *kernel*, que a cada pareja de elementos del espacio de entrada X le hace corresponder el producto escalar en F de sus respectivas imágenes por la función ϕ, es decir, tal que:

$$k(x,z) =< \phi(x), \phi(z) > \forall x,z \in X. \tag{54}$$

Podemos entender la función kernel como una pseudocomposición del embedding ϕ y del producto escalar $<,>: F \times F \to \Re$ que para cada pareja (x,z) de elementos del conjunto X proporciona directamente $< \phi(x), \phi(z) >$ sin necesidad de transitar por el embedding *phi* ni por el producto escalar $<,>$. Este atajo de $X \times X \to \Re$ junto con la kernelización aludida en el punto anterior conforman el llamado *truco kernel* que es la base del método.

En efecto los algoritmos basados en funciones kernel son en esencia métodos lineales en el espacio F que tienen todas sus ventajas tales como: sencillez, estabilidad en la solución, eficiencia computacional, etc. Al mismo tiempo, gozan de la flexibilidad de los algoritmos no lineales gracias al cambio en la representación de los datos. Como resulta evidente, la elección de la función de embedding (función ϕ) hace que los patrones lineales detectados en el espacio de características correspondan a patrones potencialmente no lineales en el espacio de entrada. Así, mediante la detección de patrones lineales (mediante sencillos procesos de optimización convexa con garantías de localización de óptimos globales en tiempos razonables) en el espacio F, un algoritmo basado en funciones kernel está, en realidad, detectando patrones no lineales en el espacio original X.

Algunos de los algoritmos que han sido kernelizados son: PLS [31], PCR [32], PCA [33][34][33] y Ridge Regresion [35][36].

4. Algoritmos de regresión para datos funcionales

Como consecuencia del desarrollo vertiginoso de los instrumentos de medición, los cuales proveen una gran cantidad de datos como funciones digitalizadas de alta resolución, el Análisis de Datos Funcionales (FDA, del inglés Functional Data Analysis) se ha convertido en un campo de creciente investigación para la solución de problemas de calibración. FDA fue propuesto como forma de recuperar las características intrínsecas de la función subyacente de los datos funcionales discretos. En este enfoque las observaciones son vistas como una sola entidad continua. Sin embargo, los algoritmos que trabajan en espacios funcionales, tienen dimensiones infinitas que pueden conducir a dificultades teóricas y prácticas. Para contrarrestar este problema, se construyó un enfoque de filtrado para alcanzar una representación de dimensionalidad finita. En este enfoque, tenemos que seleccionar una familia adecuada de funciones base que coincida con la función o funciones subyacentes a estimar. En el caso de los datos espectrales, los b-splines son una opción apropiada, en este caso la función es explicada a partir de sus coeficientes y los métodos tomarán estos como la nueva representación de los datos en lugar de los puntos originales. Un aspecto muy importante a tener en cuenta en este tipo de datos es la alta dimensionalidad, que en el contexto de FDA es un problema resuelto, ya que la dimensión del espectro se reduce de una cantidad de mediciones a una cantidad de parámetros funcionales.

Los métodos empleados para el problema de calibración con datos funcionales han estado orientados a aproximar la función de regresión $\gamma(y) = E(X/Y = y)$. El modelo de regresión funcional está dado por

$$X = \gamma(Y) + e. \tag{55}$$

Los primeros trabajos estuvieron enfocados en modelos de regresión lineal donde la función de regresión tiene la forma:

$$\gamma(y) = c + < \beta, y >, \tag{56}$$

donde, $c \in R$ y $\beta \in Y$ son parámetros desconocidos.

Para solucionar el hecho de que los métodos de regresión lineal no pueden tratar con dependencias no lineales entre las variables predictoras y respuesta, se han propuesto varios enfoques no paramátricos.

Ferraty y Vieu [37], propusieron un enfoque basado en el uso de estimadores de regresión por núcleos funcionales,

$$\hat{\gamma}(y) = \frac{\sum_{i=1}^{n} K(d(y_i, y)/h) y_i}{\sum_{i=1}^{n} K(d(y_i, y)/h)}, \tag{57}$$

donde, $h > 0$ es el ancho de banda del núcleo, d es la semi-métrica en Y y $K : R_+ \to R_+$ es una función núcleo apropiada. Este tipo de estimadores permite gran flexibilidad para ajustar modelos no lineales, sin embargo, la selección del parámetro ancho del núcleo h sobre la base de los datos de calibración es un problema difícil, especialmente para los casos donde exista gran dimensionalidad.

Otro de los estimadores son las redes neuronales funcionales propuestas en [38], donde el perceptrón de una sola capa se define por:

$$\hat{\gamma}(y) = \sum_{j=1}^{q} \hat{a}_j T(\hat{u}_j + \hat{l}_j(y)), \tag{58}$$

donde, $T : R \to R$ es una funcion de activación dada, las $(l_j)_j$ son funcionales lineales a ser estimadas y $(a_j)_j, (u_j)_j$ son parámetros desconocidos que tambien deben estimarse. El perceptrón funcional tiene la propiedad de aproximación universal que hace posible representar una gran variedad de funcionales no lineales. Pero nótese que dependen de un número de parámetros bastante grande $(w_{j)}j, (a_j)_j, (u_j)_j$, los cuales aumentan con el número de neuronas q, y su estimación por mínimos cuadrados conduce a problemas de mínimos locales. Además, tiene que seleccionarse el número de neuronas, lo cual es una tarea computacionalmente difícil.

La aproximación de funciones en espacios de Hilbert con núcleos reproductores (RKHS) se ha usado también para introducir estimadores de regresión funcionales [39], los cuales tienen la forma general:

$$\hat{\gamma}(y) = \sum_{(}^{n} i = 1) \hat{a}_i K(y_i, y), \tag{59}$$

donde, $K : Y \times Y \to R$ es un núcleo reproductor en $Y, y(a_i)_i \in R$ son parámetros desconocidos.

Una ventaja importante de este enfoque es que el estimador resultante es lineal con respecto a los parámetros desconocidos $(a_i)_i$, por tanto su estimación por mínimos cuadrados se reduce a resolver un problema lineal algebraico. Las aplicaciones clásicas de los RKHS para métodos de regresión tratan situaciones en las que $Y \subset R^d$ y, por tanto, $H \subset R^Y$ está constituida por funciones multivariadas $F : Y \subset R^d \to R$. Este es el planteamiento para métodos de regresión multivariados, en los cuales el funcional de regresión Ψ a ser estimado es una función multivariada. Por el contrario, los modelos de regresión no paramétricos funcionales tratan casos en que $Y \subset R^T$ es un conjunto de funciones $y : T \to R$, donde T es un conjunto de dimensión infinita. Por tanto, en modelos de regresión con datos funcionales el funcional desconocido Ψ está definido en un espacio normado Y de funciones con valores en los reales.

De igual forma algunas versiones funcionales de aproximación de funciones de bases radiales han sido propuestas[39], que tienen la forma:

$$\hat{\gamma}(y) = \sum_{i=1}^{m} \hat{a}_i \phi(d(y, c_i)), \tag{60}$$

donde, $\gamma : R_+ \to R$ es la función de base radial adoptada, $c_1, ..., c_m \in Y$ son centros dados, d es la distancia definida en Y, y $(a_i)_i$ son parámetros desconocidos.

Otro de los enfoques es la generalización de k-vecinos más cercanos para datos funcionales, el cual conduce a la siguiente función de regresión:

$$\hat{\gamma}(y) = \frac{1}{k} \sum_{i=1}^{k} x_{(i,y)}, \tag{61}$$

donde, $x_{(i,y)}$ es el valor de X para el *i-ésimo* vecino más cercano de y dentro de la muestra $(y_i)_{i=1,...,n}$. este enfoque es igual que al estimador de núcleos, pero tiene la limitación que brinda soluciones menos suaves, especialmente cuando la cantidad de muestras de calibración es pequeña.

En [16] se propone un enfoque no paramétrico que conduce a la reducción de la dimensión, este consiste en una extensión del método de Máquinas de Vectores Soportes para datos funcionales denominado Máquina de Vectores Soportes Funcionales(FSVR). Este método es simple y requiere de pocos hiperparámetros en comparación con otros métodos basados en vectores de soporte, por lo que los modelos pueden ser optimizados de manera más exacta.

La Regresión Inversa Funcional es un enfoque que puede verse como un compromiso entre los métodos paramétricos muy restrictivos y los no paramétricos [40][41][42]. Este enfoque supone que se cumple el siguiente modelo:

$$X = g(\langle \beta_1, Y \rangle, ..., \langle \beta_d, Y \rangle) + e, \tag{62}$$

donde, d es la llamada dimensión efectiva y $g : Rd \to R$ es la función desconocida. Bajo algunos supuestos adicionales (que se garantizan si Y tiene distribución elíptica), las direcciones $(\beta_j)_j$ puede ser estimada de la descomposición espectral del operador de covarianza $V(Y)$ y $V(E(Y/X))$. Esto último involucra ajustar la media del modelo inverso

$$Y = \mu(X) + e, \tag{63}$$

donde, e es el modelo aleatorio en Y con media cero, no correlacionado con X.

5. Algoritmos de regresión para representación por disimilitud

En el Reconocimiento de Patrones un aspecto de vital importancia es encontrar una óptima representación de los datos de forma tal que la información estructural pueda ser incluida en el proceso de aprendizaje. La representación por disimilitud fue propuesta como una representación más flexible que la representación basada en características, con el propósito de tener mayor información sobre la estructura de los objetos. Esta se basa en el importante rol que juega el concepto de proximidad en cualquier problema de reconocimiento de patrones [1].

La (di) similitud puedes ser vista como una función que asigna (pequeños) grandes valores a objetos parecidos y (grandes) valores a objetos con distintas características. Por tanto, una gran similitud y una pequeña disimilitud significan lo mismo con respecto a la comparación de objetos.

Esta representación consiste básicamente en representar los objetos a partir de sus disimilitudes con respecto a otros objetos, por lo que en lugar de tener una matriz $X(m \times n)$, donde m representa en la cantidad de objetos y n las variables medidas para cada uno, el conjunto estará representado por una matriz $D(m \times q)$. Esta matriz contiene los valores de disimilitud entre cada objeto $x \in X$ y los objetos del conjunto representativo $R(p_1, p_2, , p_q)$ basados en una medida de disimilitud. Los elementos de R son llamados prototipos y preferentemente deben ser seleccionados con algún método de selección de atributos [43]. Estos prototipos son usualmente los objetos más representativos de cada clase $R \subseteq X$, puede utilizarse todo el conjunto X, una parte de este o un conjunto completamente distonto de X. En el caso de utilizarse el conjunto X se obteniene una matriz de disimilitud cuadrada $D(m \times m)$.

La matriz de disimilitud registra en cada elemento X_{ij}, la medida de disimilitud entre la muestra i y la j como se muestra en la Figura 13, donde los valores de la diagonal serán igual a 0 debido a que no hay diferencias entre un objeto y el mismo[44].

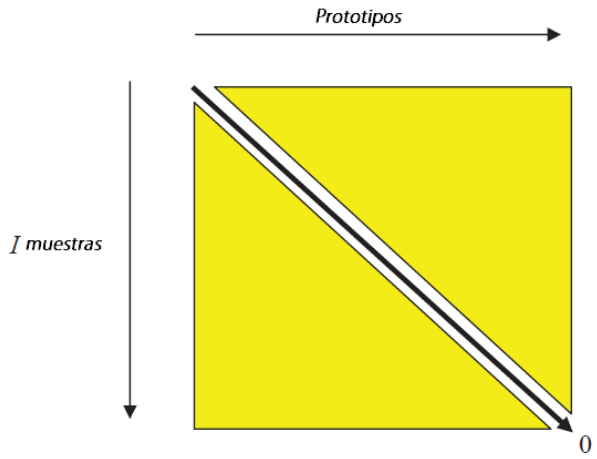

Fig. 13. Ejemplo de una matrix de disimilitud.

El espacio de disimilitud es un espacio vectorial en el que las dimensiones se definen por los vectores de disimilitud obtenidos a partir de las disimilitudes entre la muestra y cada objeto del conjunto representativo [45]. La medida de disimilitud entre cada par de muestras del conjunto indica que mientras más pequeño es este valor, más similares son ellos y mientras mas grande sea, menos similares serán [44].

Una medida de disimilitud lo suficientemente general para cualquier tipo de dato, no existe, Para cada problema debe ser seleccionada la medida que más se adapte al tipo de dato en cuestión.

Para ser una métrica, la medida de disimilitud tiene que obedecer las siguientes reglas:

· La distancia siempre tiene que ser mayor o igual que cero.

· $d_{ik} = d_{ki}$, la distancia entre la muestra i y k debe ser la misma sin importar desde donde son medidas.

· $d_{ii} = 0$, la distancia de una muestra a ella misma siempre debe ser cero.

· $d_{kl} \leq d_{ik} + d_{il}$, dada tres muestras, la distancia entre dos de ellas no puede ser mayor que la suma entre los dos restantes pares.

Una medida de disimilitud que no obedezca todas las reglas no es estrictamente una distancia, pero puede ser empleada como una semi-métrica [44]. Cuantas más propiedades de las mencionadas una medida cumpla, mejor descripción de ella y su comportamiento se tiene. La medida de disimilitud a utilizar debe ser elegida o diseñada de tal forma que la información de los datos sea incluida. De hecho, en muchas ocasiones resulta difícil encontrar la medida que logre esto. Por tanto, en la práctica, cumplir con cada una de estas propiedades y otras, no es posible. Una ventaja de la representación por disimilitud, es que esta

puede ser generada a partir de cualquier representación de objetos, mientras se tenga la medida adecuada. La reducción de dimensionalidad puede lograrse tanto como se desee, todo dependerá de la cantidad de prototipos escogidos para establecer las comparaciones entre espectros. Los problemas no linealmente separables en el espacio vectorial son convertidos en problemas lineales en el espacio de las disimilitudes.

5.1. D-PLS

D-PLS(Dissimilarity Partial Least Squares), propuesto en [46], está orientado a resolver problemas de no linealidad, basandose en el enfoque clásico de PLS, donde las X-variables es una matriz de disimilitud que representa las disimilitudes entre los datos. Es rápido y puede ser aplicado sobre conjuntos de datos que contienen gran cantidad de objetos y variables.

Según [47] [46] los problemas de regresión no lineal pueden ser resueltos cuando la representación vectorial de los datos es reemplazada por la representación por disimilitud. Manteniendo la nomenclatura utilizada en el algoritmo PLS descrito anteriormente y reemplazando las variables independientes originales por la matrix de disimilitud, el modelo según D-PLS se puede expresar:

$$D = TP^T + E, \tag{64}$$

$$Y = UC^T + E, \tag{65}$$

siendo T y U las matrices de *scores*, P y C las matrices de *loadings* y E la matriz que contiene los residuales.

D-PLS puede lidiar eficientemente con datos no homogéneos. La elección de una medida de disimilitud tendrá un impacto en la información contenida en la matriz. La distancia euclidiana es seleccionada debido a que representa bien los datos.

6. Conclusiones

En el presente trabajo se abarcó la teoría fundamental sobre la regresión de datos, las características de los datos que relaciona este enfoque y los tipos de regresión que existen. Posteriormente se explicaron los aspectos más importantes de los métodos existentes en cada uno de los tipos de regresión. En algunos casos se establecieron comparaciones entre ellos y se derivaron análisis donde los más importantes son:

· Utilizando para el análisis variables latentes en lugar de las variables originales representa una ventaja importante al reducir dimensionalidad, que es característica clave en los conjuntos de datos dentro de la Quimiometría.

· Usando los algoritmos PLS y PCR se pueden obtener soluciones bastante parecidas, solo que con el primero esta solución se obtiene con una menor cantidad de componentes.

· En datos multi-vías, se obtienen mejores resultados al aplicar algoritmos adaptados a esta estructura como N-PLS o SCREAM que aplicar desdoblado sobre los datos y luego algún algoritmo multivariante como U-PLS. Esto se debe a que en la segunda opción no se aprovecha toda la información contenida en la estructura multi-vía.

· El cambio de representación de los datos, específicamente la transformación al espacio de las disimilitudes, es un enfoque prometedor por características como son la reducción de dimensionalidad y la incorporación de información valiosa en el análisis.

A pesar de que la regresión no es de los campos menos estudiados, hay líneas de investigación que no han sido muy investigadas y merecen la pena explorar, como son:

· A pesar de los resultados tan alentadores que existen usando Representación por Disimilitudes en la clasificación supervisada se hace necesario probar su comportamiento en la regresión de datos.

· Teniendo en cuenta que el acierto de la Representación por Disimilitudes depende de la medida que se aplique, se hace necesaria la selección, optimización y validación de estas medidas en el contexto de la regresión de acuerdo a cada una de las técnicas analíticas.

· Los problemas de detección de relaciones no lineales entre los conjuntos en la regresión sigue siendo un problema abierto.

· A pesar de que en [46] [47] se plantea que aparentemente con el cambio a la Representación por Disimilitud se eliminan los problemas de no linealidad, esto es algo que no está probado y estudiado. Podría resultar interesante explorar el comportamiento de la disimilitud y su combinación con soluciones existentes como la kernelización.

· El análisis multi-vía a pesar de haber surgido ante la necesidad de procesar la información que brinda el incorporar nuevas dimensiones al estudio, no ha sido un área altamente explotada y sobre la cual queda mucho por hacer tanto en la Representación vectorial como en la Representación por Disimilitudes donde no hay nada hecho hasta el momento en regresión.

Referencias bibliográficas

1. Duin, R.P., Pekalska, E.: The dissimilarity representation for pattern recognition: A tutorial. Technical report, Technical Report (2009)
2. Munoz, D.P.: Classification of Continuous Multi-way Data Via Dissimilarity Representation. PhD thesis (2013)
3. Marini, F.: Chemometrics in food chemistry. Volume 28. Newnes (2013)
4. Smilde, A., Bro, R., Geladi, P.: Multi-way analysis with applications in the chemical sciences. 2004
5. Esbensen, K.H., Guyot, D., Westad, F., Houmoller, L.P.: Multivariate data analysis-in practice: an introduction to multivariate data analysis and experimental design. Multivariate Data Analysis (2002)
6. Gurden, S.P., Westerhuis, J.A., Bro, R., Smilde, A.K.: A comparison of multiway regression and scaling methods. Chemometrics and Intelligent Laboratory Systems **59**(1) (2001) 121–136
7. Kroonenberg, P.M.: Applied multiway data analysis. Volume 702. John Wiley & Sons (2008)
8. Kiers, H.A.: Towards a standardized notation and terminology in multiway analysis. Journal of chemometrics **14**(3) (2000) 105–122
9. Eisenhart, C.: The interpretation of certain regression methods and their use in biological and industrial research. The Annals of Mathematical Statistics **10**(2) (1939) 162–186
10. Berkson, J.: Estimation of a linear function for a calibration line; consideration of a recent proposal. Technometrics **11**(4) (1969) 649–660
11. Williams, E.J.: Regression methods in calibration problems. Bull. Int. Statist. Inst **43**(1) (1969) 17–28
12. Krutchkoff, R.G.: Classical and inverse regression methods of calibration in extrapolation. Technometrics **11**(3) (1969) 605–608
13. Krutchkoff, R.: Classical and inverse regression methods of calibration. Technometrics **9**(3) (1967) 425–439
14. Ali, M., Singh, N.: An alternative estimator in inverse linear regression. Journal of Statistical Computation and Simulation **14**(1) (1981) 1–15
15. Srivastava, V., Singh, N.: Small-disturbance asymptotic theory for linear-calibration estimators. Technometrics **31**(3) (1989) 373–378
16. Hernández González, N.: Nuevos métodos para la calibración estadística basada en datos univariados, multivariados y funcionales. PhD thesis (2010)
17. Harris, I.R.: Predictive fit for natural exponential families. Biometrika **76**(4) (1989) 675–684
18. Bro, R.: Multiway calibration. multilinear pls. Journal of chemometrics **10** (1996) 47–61
19. De Jong, S., Kiers, H.A.: Principal covariates regression: part i. theory. Chemometrics and Intelligent Laboratory Systems **14**(1-3) (1992) 155–164
20. Wold, S., Geladi, P., Esbensen, K., Öhman, J.: Multi-way principal components-and pls-analysis. Journal of chemometrics **1**(1) (1987) 41–56
21. St, L., et al.: Aspects of the analysis of three-way data. Chemometrics and Intelligent Laboratory Systems **7**(1-2) (1989) 95–100
22. Smilde, A.K., Kiers, H.A., et al.: Multiway covariates regression models. Journal of Chemometrics **13**(1) (1999) 31–48
23. Amigo, Jose, M., Marini, F.: Multiway methods. In: Data handling in science and technology chemometrics in food chemistry. Volume 28. Elsevier (2013) 265–313
24. Harshman, R.A.: Parafac2: Mathematical and technical notes. UCLA working papers in phonetics **22**(3044) (1972) 122215
25. Marini, F., Bro, R.: Scream: A novel method for multi-way regression problems with shifts and shape changes in one mode. Chemometrics and Intelligent Laboratory Systems **129** (2013) 64–75
26. Bishop, C.M.: Neural networks for pattern recognition. Oxford university press (1995)
27. Tipping, M.E.: The relevance vector machine. in advances in neural information processing systems. (2000)
28. Tipping, M.E.: Sparse bayesian learning and the relevance vector machine. Journal of machine learning research **1**(Jun) (2001) 211–244
29. Li, K.C.: Sliced inverse regression for dimension reduction. Journal of the American Statistical Association **86**(414) (1991) 316–327
30. Gibaja Martíns, J.J.: Aprendizaje estadístico con funciones kernel. (2010)
31. Rosipal, R., Trejo, L.J.: Kernel partial least squares regression in reproducing kernel hilbert space. Journal of machine learning research **2**(Dec) (2001) 97–123
32. Rosipal, R., Trejo, L.J., Cichocki, A.: Kernel principal component regression with em approach to nonlinear principal components extraction. University of Paisley (2000)
33. Scholkopf, B., Smola, A., Muller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural computation **10**(5) (1998) 1299–1319
34. Rosipal, R., Girolami, M.: An expectation-maximization approach to nonlinear component analysis. Neural Computation **13**(3) (2001) 505–510

35. Saunders, C., Gammerman, A., Vovk, V.: Ridge regression learning algorithm in dual variables. In: (ICML-1998) Proceedings of the 15th International Conference on Machine Learning, Morgan Kaufmann (1998) 515–521
36. Cristianini, N., Shawe-Taylor, J.: An introduction to support vector machines and other kernel-based learning methods. Cambridge university press (2000)
37. Ferraty, F., Vieu, P.: Nonparametric functional data analysis: theory and practice. Springer Science & Business Media (2006)
38. Rossi, F., Conan-Guez, B.: Functional multi-layer perceptron: a non-linear tool for functional data analysis. Neural networks **18**(1) (2005) 45–60
39. Preda, C.: Regression models for functional data by reproducing kernel hilbert spaces methods. Journal of Statistical Planning and Inference **137**(3) (2007) 829–840
40. Dauxois, J., Ferré, L., Yao, A.F.: Un modele semi-paramétrique pour variables aléatoires hilbertiennes. Comptes Rendus de l'Académie des Sciences-Series I-Mathematics **333**(10) (2001) 947–952
41. Ferré, L., Villa, N.: Multilayer perceptron with functional inputs: an inverse regression approach. Scandinavian Journal of Statistics **33**(4) (2006) 807–823
42. Ferré, L., Yao, A.F.: Functional sliced inverse regression analysis. Statistics **37**(6) (2003) 475–488
43. Pekalska, E., Duin, R.P.: Prototype selection for finding efficient representations of dissimilarity data. In: Pattern Recognition, 2002. Proceedings. 16th International Conference on. Volume 3., IEEE (2002) 37–40
44. Brereton, R.: Chemometrics for pattern recognition. John Wiley & Sons (2009)
45. Duin, R.P., Pekalska, E.: The dissimilarity space: Bridging structural and statistical pattern recognition. Pattern Recognition Letters **33**(7) (2012) 826–832
46. Zerzucha, P., Daszykowski, M., Walczak, B.: Dissimilarity partial least squares applied to non-linear modeling problems. Chemometrics and Intelligent Laboratory Systems **110**(1) (2012) 156–162
47. Martin, Y.C., Lin, C.T., Hetti, C., DeLazzer, J.: Pls analysis of distance matrixes to detect nonlinear relationships between biological potency and molecular properties. Journal of medicinal chemistry **38**(16) (1995) 3009–3015
48. Bro, R.: Multi-way analysis in the food industry: models, algorithms, and applications. PhD thesis, Københavns UniversitetKøbenhavns Universitet, LUKKET: 2012 Det Biovidenskabelige Fakultet for Fødevarer, Veterinærmedicin og NaturressourcerFaculty of Life Sciences, LUKKET: 2012 Institut for FødevarevidenskabDepartment of Food Science, LUKKET: 2012 Kvalitet og TeknologiQuality & Technology (1998)
49. Bro, R.: Parafac. tutorial and applications. Chemometrics and intelligent laboratory systems **38**(2) (1997) 149–171
50. Harshman, R.A., Lundy, M.E.: The parafac model for three-way factor analysis and multidimensional scaling. Research methods for multimode data analysis **46** (1984) 122–215
51. Harshman, R.A., Lundy, M.E.: Parafac: Parallel factor analysis. Computational Statistics & Data Analysis **18**(1) (1994) 39–72
52. Porro-Munoz, D., Talavera, I., Duin, R.: Multi-way data analysis. Technical report, Technical report, CENATAV (2009)
53. Kroonenberg, P.M.: Three-mode principal component analysis: Theory and applications. Volume 2. DSWO press (1983)
54. Kroonenberg, P.M., De Leeuw, J.: Principal component analysis of three-mode data by means of alternating least squares algorithms. Psychometrika **45**(1) (1980) 69–97

7. Anexo

Overfitting

Mientras más complejo es el modelo, es mejor su capacidad de ajuste a los datos dados. El error de predicción para el conjunto de calibración en general disminuye a medida que aumenta la complejidad del modelo. Por tanto un modelo altamente complicado puede ajustar casi cualquier dato con casi cero desviaciones entre la *y* experimental y la *y* modelada. Evidentemente este tipo de modelos no serán necesariamente útiles en los casos de nuevas objetos, porque probablemente están sobre ajustados, esto quiere decir que está muy bien ajustado al conjunto de calibración y por tanto no posee suficiente generalización [5]. Los errores de predicción para nuevas muestras son grandes para modelos pequeños o sobre ajustados. Determinar la complejidad óptima del modelo es muy importante pero no siempre es una tarea fácil. En Quimiometría, la complejidad es usualmente controlada por el número de componentes y la complejidad óptima se estima utilizando validación cruzada.

Variables latentes

Un concepto muy importante en el análisis multivariado de datos es el de variables latentes que consiste en combinar matemáticamente varias variables para formar una nueva que posee cierta propiedad. Esta nueva variable es nombrada como variable latente, componente o factor y su valor se conoce como *score* [5]. Dependiendo del objetivo que se persiga en el análisis de los datos, existen diferentes criterios matemáticos para definir una variable latente:

· En el Análisis de Componentes Principales, se usa el criterio de máxima varianza de los *scores*, que proporciona una óptima representación de las distancias euclidianas entre los objetos.

· En clasificación multivariada, las variables latentes son variables discriminativas que poseen la capacidad de separar dos clases.

· En calibración multivariada, las variables latentes tienen los máximos coeficientes de correlación o covarianza con la propiedad, por lo que pueden ser usados para predecir esta propiedad.

Selección de variables

En regresión múltiple son usadas todas las variables predictoras disponibles para construir un modelo lineal para predecir las variables respuestas. Este enfoque es muy útil mientras que la cantidad de variables regresoras sea pequeño (no más de 10). Sin embargo en muchos problemas, especialmente en Quimiometría, se debe lidiar con cientos de variables regresoras. Esto puede resultar un problema debido a que la regresión OLS no se puede llevar a cabo si las variables regresoras están altamente correlacionadas o la cantidad de objetos es menor que la cantidad de variables regresoras.

El uso de todas las variables puede llevar a un mejor ajuste del modelo para los datos de entrenamiento, pero usualmente el interés recae en aumentar el desempeño para los datos de validación. Los modelos con gran cantidad de variables son imposible de interpretar, por esto la reducción de las variables regresoras puede evitar el sobre ajuste, resultar en un mejor desempeño en la predicción y reducir el tiempo computacional considerablemente.

El desempeño de los métodos de selección de variables depende de los datos y del interés que se tenga en la estructura interna de los datos. Esto hace que no exista una guía para elegir el método que se ajuste mejor a los datos. En regresión, la selección de variables es una forma de reducir el número de variables

regresoras y eliminar la multicolinealidad. Gracias a esto se puede conseguir un modelo de regresión con buena interpretabilidad, pero sacrificando el costo computacional ya que en los datos en los que se tiene una gran cantidad de variables, puede ser bastante costoso.

Una estrategia simple para la selección de variables se basa en la información de otros métodos multivariantes como PCA o regresión por PLS. Estos métodos forman nuevas variables latentes a partir de la combinación lineal de las variables regresoras $b_1x_1 + b_2x_2 + ... + b_mx_m$. Los coeficientes o *loadings*, $b_1, b_2, ..., b_m$ reflejan la importancia de las variables en las nuevas variables latentes. Los coeficientes cercanos a cero indican menos importancia de la variable. Por lo que el valor absoluto de los coeficientes puede ser utilizado como un criterio de selección de variables. En el caso de PCA, los coeficientes solo utilizan la información de las variables x, mientras que en PLS también se tiene en cuenta su relación con las variables respuesta y. En [5] se exponen otros criterios de selección de variables, así como algunas estrategias a seguir que no son de nuestro interés.

Desdoblado o unfold

El término *unfolding* es usualmente conocido como desdoblado y es un concepto importante en el análisis multi-vías. Consiste en convertir un arreglo multi-vías en una matriz o arreglo *two-way* concatenando cada *slice* de uno de los modos al lado del otro, como se puede ser en la Figura 14 [48].

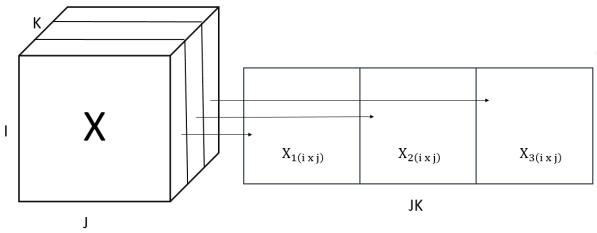

Fig. 14. Proceso de desdoblado en una estructura tridimensional.

Note que la dimensión de la columna generada es mucho más grande en el modo que está formado por dos de los anteriores (modo JK). Esto ocurre porque ya sean las características o las condiciones de los modos originales son combinados en un único modo, no hay una nueva variable que se refiera a una variable original sino a un conjunto de ellas[48]. Este procedimiento puede efectuarse en cada uno de los modos de la estructura multi-vías, en el caso de los arreglos *three-way* se realiza como se muestra en la Figura 15.

Una vez que la estructura multi-vías es redimensionada a un arreglo *two-way* los métodos de análisis multivariado pueden emplearse para entender los datos. Esta transformación, ignorando la estructura multi-vías al aplicar los métodos tradicionales puede causar pérdida de información y mala interpretación de los datos, especialmente si los datos contienen ruido el modelo puede ser: menos robusto, menos interpretable y menos predictivo. Por tanto, entre las desventajas principales de este principio de desdoblado está la posible optención de modelos poco robustos, menos interpretables y poco predictivos.

PARAFAC

PARAFAC es un método de descomposición para datos multi-vías, el cual puede ser comparado con PCA. Para datos *three-way*, la descomposición se realiza a partir de componentes trilineales. En lugar de obtener un vector de *scores* y uno de *loadings* como en PCA, cada componente consiste en un vector de

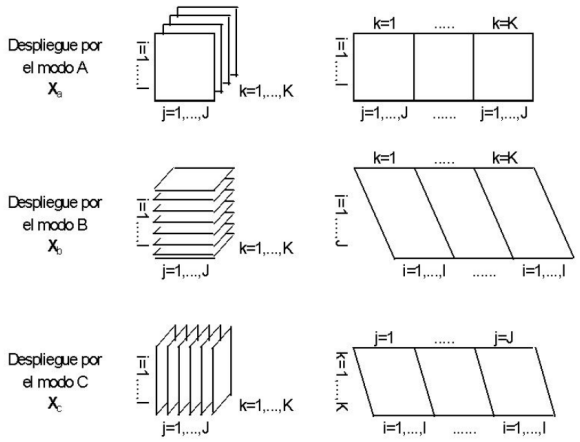

Fig. 15. Modos de aplicar desdoblado en una estructura three-way.

scores y dos de *loadings*. Una práctica común en este tipo de datos es no distinguir entre *scores* y *loadings* ya que estos son tratados por igual numéricamente. La estructura del modelo PCA es:

$$\hat{X_{ij}} = \sum_{f=1}^{F} a_{if} b_{jf}. \tag{66}$$

Del mismo modo el modelo PARAFAC para arreglos *three-way* está dado por tres matrices de *loadings* A, B y C.

$$\hat{X_{ijk}} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf}. \tag{67}$$

En la Figura 16 se muestra un modelo PARAFAC de dos componentes para un arreglo *three-way* X.

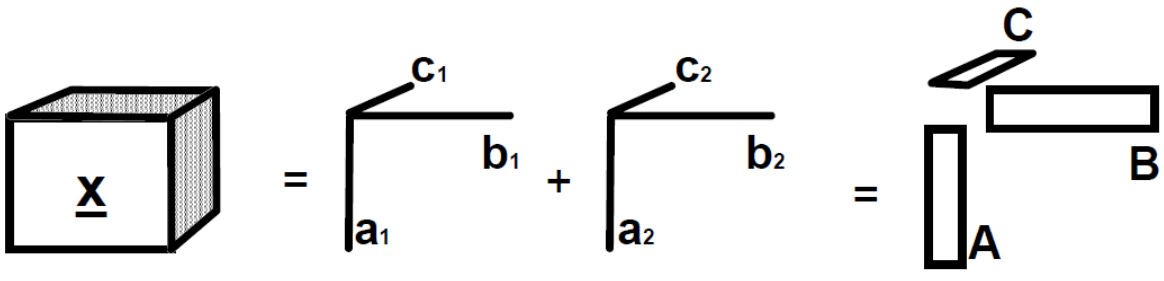

Fig. 16. Representación de la descomposición de los datos aplicando PARAFAC.

Al igual que en el algoritmo PCA tradicional los vectores de *loadings* de PARAFAC solo son combinados con sus correspondientes factores en otros modos, de manera que a_1 interactúa solo con b_1 y c_1. Más información y especificidades del algoritmo pueden encontrarse en [49][50][51].

PARAFAC2

En el modelo PARAFAC se asume que todos los *slices* en la estructura multi-vías tienen las mismas dimensiones, lo cual no es posible en todos los casos, debido a que las muestras o las características no pueden ser medidas en todas las condiciones o instantes de tiempo. Dada esta restricción en el modelo PARAFAC surge PARAFAC2 para los casos en los que esta restricción impide la aplicación del modelo [52]. Por lo que el modelo se puede expresar como:

$$X_k = A_k D_k B' + E_k, \tag{68}$$

donde, X_k es la matriz de $N_k \times J$ correspondiente al k-esimo *slice*, A_k es el factor de *scores* del mismo *slice*, el resto de los elementos de la ecuación tienen la misma interpretación que el modelo PARAFAC.

TUCKER

Los modelos Tucker son otro grupo de modelos para descomponer arreglos multi-vías [53][54]. El más importante de estos modelos es Tucker3 ya que el modelo Tucker2 es considerado como un caso especial de este y Tucker1 consiste en aplicar PCA sobre la estructura desdoblada.

A diferencia de PARAFAC y PARAFAC2 este algoritmo permite la interacción entre los factores de todos los modos. Y como consecuencia de esto, el número de factores en los diferentes modos no tiene por qué ser el mismo y el algoritmo incorpora un arreglo G (conocido como *core array*) cuyas dimensiones corresponden a la cantidad de factores de cada uno de los modos, que contiene las magnitudes de la interacción entre los factores. El modelo Tucker para un arreglo *three-way* puede ser expresado matemáticamente por:

$$X^{(I \times JK)} = A G^{D \times EF} (C \otimes B)^T + E^{I \times JK}, \tag{69}$$

donde, $G^{D \times EF}$ corresponde al *core array* desdoblado en una matriz de dimensiones $D \times EF$, D, E y F son la cantidad de factores para cada uno de los modos y $A(I \times D)$, $B(J \times E)$ y $C(K \times F)$ corresponden a las matrices de *loadings*.

El modelo PARAFAC puede ser considerado como un modelo Tucker $(F \times F \times F)$ donde todos los elementos de la superdiagonal del *core array* son cero.

ALS

ALS(Alternating Least Squares), fue introducido en 1933 por Yates y su idea fundamental es la de resolver grandes problemas de optimización con pequeños sub-problemas de forma iterativa.

En el algoritmo los parámetros a estimar están separados en diferentes conjuntos (la menor cantidad posible), esta separación hace posible el uso de algoritmos simples para estimar los parámetros. En cada iteración son ajustados los conjuntos de parámetros excepto uno de los conjuntos, con el cual se va a minimizar una nueva función de pérdida. El algoritmo va a iterar alternando entre un conjunto y otro hasta que no se observen cambios en la función de pérdida o en los parámetros, o si la variación de estos es menor que el criterio de convergencia definido. Mientras menor sea la cantidad de conjuntos de parámetros, disminuye la posibilidad de encontrar un mínimo local o de converger lentamente. Si el algoritmo converge a un mínimo local, el modelo de mínimos cuadrados es encontrado.

Dado un arreglo X y un modelo general $X = f(A,B,C,) + E$, los pasos para el algoritmo ALS para estimar los parámetros A, B, C, etc. son:

1. Inicializar los parámetros.
2. A es el resultado de $\min_A \|X - f(A,B,...,C)\|_F^2$.

3. B es el resultado de mín$_B \|X - f(A,B,...,C)\|_F^2$.
4. C es el resultado de mín$_C \|X - f(A,B,...,C)\|_F^2$.
5. Se estiman todos los parámetros de igual forma
6. Realizar los pasos del 2-5 hasta que converja

donde f es el modelo de X y es la función de los parámetros A, B, C, etc. $\|.\|f$, se refiere a la norma Frobenius. Este algoritmo tiene la ventaja de ser fácil de implementar y es simple comparado con los algoritmos que trabajan simultáneamente, también puede manejar datos perdidos, puede ser extendido a datos multi-vías y garantiza la convergencia. No obstante tiene problemas de convergencia lenta en los casos en los que exista colinealidad y tampoco soporta la presencia de *outliers*, el cual es muy común en muchas líneas de investigación [52][48].

RT_089, febrero 2017
Aprobado por el Consejo Científico CENATAV

Editor: Lic. Lucía González Bayona
Diseño de Portada: Di. Alejandro Pérez Abraham
RNPS No. 2142
ISSN 2072-6287
Indicaciones para los Autores:
Seguir la plantilla que aparece en www.cenatav.co.cu
C E N A T A V
7ma. A No. 21406 e/214 y 216, Rpto. Siboney, Playa;
La Habana. Cuba. C.P. 12200
Impreso en Cuba