UNIVERSIDAD AUTÓNOMA DE QUERÉTARO

FACULTAD DE INGENIERÍA

EL ANÁLISIS DE PROYECCIÓN A ESTRUCTURAS LATENTES Y SU USO EN CLASIFICACIÓN

TESIS

QUE COMO PARTE DE LOS REQUISITOS PARA OBTENER EL TÍTULO DE

LICENCIADO EN MATEMÁTICAS APLICADAS

PRESENTA

OSCAR JORDAN PARRA SAN JUAN

DIRIGIDA POR

DR. EDUARDO CASTAÑO TOSTADO

SANTIAGO DE QUERÉTARO, QUERÉTARO, 2021

UNIVERSIDAD AUTÓNOMA DE QUERÉTARO

FACULTAD DE INGENIERÍA

LICENCIATURA EN MATEMÁTICAS APLICADAS

EL ANÁLISIS DE PROYECCIÓN A ESTRUCTURAS LATENTES Y SU USO EN CLASIFICACIÓN

TESIS

Que como parte de los requisitos para obtener el título de Licenciado en Matemáticas Aplicadas

Presenta

Oscar Jordan Parra San Juan

Dirigida por

Dr. Eduardo Castaño Tostado

Dr. Eduardo Castaño Tostado Presidente	Firma
Dr. Mario Santana Cibrian Secretario	Firma
M.I.M Wilfrido Jacobo Paredes García Vocal	Firma
M.I.M Juan Antonio Villeda Reséndiz Suplente	Firma

Centro Universitario
Querétaro, Qro.
Febrero 2021
México

Resumen

Desde mediados del siglo XX, a medida que se obtenían mejores instrumentos de medición en campos como la Química y la Biología, los conjuntos de datos han crecido en volumen, particularmente en el número de variables (p) medidas sobre sobre un número de muestras de interés (n). En muchos casos $n < p$, hecho que puso en conflicto los supuestos de las técnicas estadísticas más utilizadas para el análisis de regresión o para la clasificación.

Uno de los campos de estudio donde más ocurre $n < p$ es en el área de la metabolómica, el estudio de procesos químicos de los metabolitos de una muestra biológica. Uno de los métodos que se utilizan en el estudio de estos conjuntos de datos es la Proyección a Estructuras Latentes (PLS), también llamado Mínimos Cuadrados Parciales, con el objetivo de encontrar una estructura latente entre variables respuesta continuas versus variables predictoras, como en el caso de regresión multivariada, pero con $n < p$.

A pesar de que PLS no fue diseñado inicialmente como técnica de clasificación, su uso se ha extendido en el área de clasificación donde la variable respuesta es categórica en el contexto de clasificación. Una de las áreas de aplicación de PLS para propósitos de clasificación se ha dado en lo que se conoce como “metabólica” que comprende el estudio de causas de efectos en el metabolismo humano. PLS ha sido desarrollado fundamente desde una perspectiva quimiométrica como una herramienta con resultados relevantes en el análisis diferencial de perfiles metabolómicos ante tratamientos alternativos. Esto ha abierto el desarrollo computacional variado de PLS, hecho que en sí es útil, pero que hace difuso en México a los usuarios en estas áreas de aplicación, el cómo seleccionar el paquete en código libre adecuado a sus necesidades.

En el presente trabajo, en un inicio se estudian los detalles algebraicos de PLS. También se estudia a PLS y su relación con otras técnicas de clasificación. A continuación, se estudian técnicas de remuestreo, que se utilizan para evaluar el ajuste de un modelo. Entonces se exponen algunos de los paquetes que ofrecen el uso de PLS y sus características en el ambiente del software R. Finalmente, se realiza una aplicación en un estudio experimental que intenta dar, mediante PLS, una clasificación analizando un conjunto de datos de un estudio realizado en la UAQ en niños con un tratamiento basado en un subproducto de mango.

Se concluye revisando las capacidades y beneficios del uso de PLS en el contexto de clasificación, así como la evaluación de la paquetería que lo ofrece, que pueden servir de utilidad a investigadores en México que pretendan realizar un análisis de datos con las características antes mencionadas.

Summary

Since the middle of the 20th century, as better measuring instruments were obtained in fields such as Chemistry and Biology, data sets have grown in volume, particularly in the number of variables (p) measured over a number of samples of interest (n). In many cases $n < p$, a fact that put into conflict the assumptions of the statistical techniques most used for regression analysis or for classification.

One of the fields of study where $n < p$ occurs most is in the area of metabolomics, the study of chemical processes of the metabolites of a biological sample. One of the methods used in the study of these data sets is the Projection to Latent Structures (PLS), also called Partial Least Squares, with the aim of finding a latent structure between continuous response variables versus predictor variables, as in the case of multivariate regression but with $n < p$.

Although PLS was not initially designed as a classification technique, its use has been extended in the classification area where the response variable is categorical in the classification context. One of the areas of application of PLS for classification purposes has been in what is known as "metabolomics" which comprises the study of the causes of effects on human metabolism. PLS has been fundamentally developed from a chemometric perspective as a tool with relevant results in the differential analysis of metabolomic profiles in the face of alternative treatments. This has opened up the various computational development of PLS, a fact that in itself is useful, but inducing users in Mexico, in these areas of application, to doubt about how to select the appropriate open-source package for their needs.

In the present work, at the beginning, the algebraic details of PLS are studied. PLS and its relationship with other classification techniques are also studied. Next, resampling techniques are studied, which are used to evaluate the fit of a model. Then some of the packages that offer the use of PLS and its features in the R software environment are exposed. Finally, an application is made in an experimental study that tries to give, by means of PLS, a classification by analyzing a set of data from a study carried out at the UAQ in children with a treatment based on a mango by-product.

It concludes by reviewing the capabilities and benefits of using PLS in the classification context, as well as the evaluation of the packages that offer it, which may be useful to researchers in Mexico who intend to perform a data analysis with the aforementioned characteristics.

Dedicatorias

Para María Fernanda Flores Juárez,

éste y todos los futuros triunfos fruto de su continuo amor y apoyo.

"All Spartans are great warriors. We train from birth.
Our lives were discipline, duty, battle, and death.
Life was grim, and we greeted it grimly.
But Atreus of Sparta was unlike the rest of us.
He wore a smile even in the worst of times. He was… happy.
He inspired us to hope… that though we were machines of war,
yet there was humanity in us.
Goodness."

God of War (2018). SCE Santa Monica Studio

Índice

1 Introducción

En el área de la Estadística, uno de los objetivos principales es el estudio de fenómenos con el fin de encontrar estructuras de asociación, potencialmente causales, entre las variables observables; de hallarse tal tipo de estructuras, se puede tener una descripción de los fenómenos de manera más profunda. Las estructuras que se pueden hallar pueden variar respecto a los posibles usos de las mismas o a la complejidad de la información que se posee. En muchos estudios, la información se suele estructurar en dos conjuntos de datos: el conjunto de variables *respuesta* $\boldsymbol{Y}_{n\times q}$ y el conjunto de *variables* explicatorias $\boldsymbol{X}_{n\times p}$. Los subíndices n y p representan el número de *registros* u *observaciones*, y el número de *variables* contenidas en $\boldsymbol{X}$**,** respectivamente. El subíndice q representa el número de variables contenidas en $\boldsymbol{Y}$. Las p variables también se suelen denominar *variables independientes*, mientras que las q variables se pueden denominar *variables dependientes*. Es plausible tener como un objetivo principal el encontrar una estructura entre las p variables de $\boldsymbol{X}$ que expliquen de manera satisfactoria las variaciones de las q variables de $\boldsymbol{Y}$; cumplir este objetivo se conoce como *aprendizaje supervisado,* dado que lo que interesa es como recuperar la estructura en X que mejor explique a Y. Otro contexto de aprendizaje, donde no haya la diferenciación entre respuestas y explicatorias, se conoce como *aprendizaje no supervisado*.

En el estudio de las técnicas de la Estadística se suelen requerir supuestos que permitan construir modelos matemático - estadísticos. Por ejemplo, en el caso de la regresión lineal múltiple, se requiere que la matriz $\boldsymbol{X}_{n\times p}$ sea no singular. Es decir, que la matriz inversa $(X^TX)^{-1}$ exista. Para que esta condición se cumpla, los subíndices de esta matriz deben satisfacer $n > p$. En los últimos años, especialmente ya en el siglo 21, con el surgimiento de herramientas de cómputo, recolección y almacenamiento de datos más especializadas y con la reducción de costos en la obtención de los mismos, el número de variables por tabla o base de datos ha aumentado considerablemente, teniendo casos en los que $n < p$ e incluso $n \ll p$. En estas condiciones, se hace necesario el uso de técnicas o metodologías apropiadas para el estudio de estas nuevas bases de datos.

Entre las estrategias disponibles para trabajar con estas grandes cantidades de variables, están las técnicas clásicas de selección de variables. Por supuesto la selección de variables en buena parte debe apoyarse en conocimiento causal por parte del investigador, pero este conocimiento causal es escaso en muchas de las aplicaciones actuales que se embarcan, dada la democratización en el uso de paquetería estadística, en la solución de problemas sin mayor teoría causal que la sustente. Con esto, la selección de variables mediante sólo algoritmos corre el

riesgo de incluir variables que solo aportan ruido al modelo, resultando en un ajuste pobre.

El uso de técnicas como la regresión escalonada (Stepwise Regression), partiendo de un conjunto razonable de variables, tiene la ventaja de que es un proceso más eficiente que buscar entre todos los posibles modelos que se pueden formar con las variables disponibles, especialmente cuando el número de variables crece ($p > 10$). Sin embargo, cuando el número de variables es muy grande, el costo computacional crece también.

Otro tipo de técnicas a utilizar cuando el número de variables aumenta considerablemente, son las que permiten la descomposición de la matriz X en un número reducido de nuevas variables k, construidas como combinaciones lineales de las p variables originales. Estas nuevas variables se utilizan en lugar de las originales para disminuir la complejidad del modelo y simplificar la interpretación. La técnica más conocida que utiliza este procedimiento se conoce como Análisis de Componentes Principales PCA. Al utilizarse para análisis supervisado y regresión, esta técnica se conoce como Análisis de Regresión por Componentes Principales PCR.

Una tercera posibilidad en el contexto de regresión y clasificación que utiliza un enfoque de descomposición similar al de regresión por componentes principales, es la conocida como la Proyección a Estructuras Latentes PLS (mejor conocida como Mínimos Cuadrados Parciales). Este enfoque también intenta disminuir la dimensionalidad del problema, generando k nuevas variables de las p originales, de tal forma que capturen la mayor cantidad de información posible desde una perspectiva diferente al de regresión PCA. La diferencia entre la regresión por PCA y la regresión por PLS es el tipo de información a maximizar por cada nuevo componente. El uso de PCA permite construir los k componentes principales con la matriz X de tal manera que maximiza la extracción de varianza de la misma al obtener cada componente con lo que no garantiza una buena capacidad para explicar la variabilidad en Y como lo hacen en X. Por otro lado, PLS construye los k componentes principales, pero maximizando la *covarianza* entre X e Y, de tal manera que las nuevas k variables tengan una mayor capacidad explicativa de Y utilizando la información en X.

El modelo de regresión PLS se originó con el algoritmo NIPALS (Mínimos Cuadrados Parciales Iterativos No lineales) desarrollado por Herman Wold en 1973. En dicho trabajo se muestra un método para calcular componentes principales utilizando el método de mínimos cuadrados ordinario de manera iterativa, así como para computar correlaciones (llamadas canónicas) con una secuencia iterativa de

regresiones múltiples (Wold, 1973). De estos trabajos surgió en 1977 el algoritmo PLS, un método iterativo para encontrar *variables latentes* (Mateos-Aparicio, 2011).

Los objetivos del presente trabajo de tesis, fueron los siguientes:

1. Conocer en detalle de las bases matemático estadísticas de la técnica PLS (capítulo 2).
2. Revisión de técnicas de remuestreo para estudiar de evaluación de modelos estimados el contexto de PLS (capitulo 3).
3. Describir los paquetes disponibles a la fecha del PLS en el ambiente del software / proyecto R (capítulo 4).
4. Aplicar PLS en el análisis de un conjunto de datos reales provenientes de estudio experimental en metabolómica y consumo de un complemento alimentario (capítulo 5).

2 El método de Proyección a Estructuras Latentes

2.1 Modelo

Denotemos con $\boldsymbol{X} \subset \mathbb{R}^N$ un espacio de variables N dimensional representando un primer bloque de variables, y $\boldsymbol{Y} \subset \mathbb{R}^M$ un espacio de variables M dimensional representando un segundo bloque de variables. PLS modela las relaciones entre los dos bloques $\boldsymbol{X}$ e $\boldsymbol{Y}$ por medio de vectores de *puntajes*. Al tener n observaciones de cada bloque de variables, PLS descompone las matrices $\boldsymbol{X}_{n\times N}$ y $\boldsymbol{Y}_{n\times M}$ de variables (con media cero) de la siguiente forma:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}' + \boldsymbol{E}$$

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}' + \boldsymbol{F}$$

donde $\boldsymbol{T}, \boldsymbol{U}$ son matrices $n \times p$ de vectores de puntajes (comúnmente llamados componentes); las matrices $\boldsymbol{P}, \boldsymbol{Q}$ son las matrices de *saturaciones* $N \times p$, $M \times p$ respectivamente; y las matrices $\boldsymbol{E}, \boldsymbol{F}$ son las matrices de residuales $n \times N$, $n \times M$ respectivamente. A partir de estas expresiones, se puede decir que PLS tiene como objetivo la aproximación simultánea de $\boldsymbol{X}$ e $\boldsymbol{Y}$ a partir de factorizaciones que maximizan la estructura de correlación entre $\boldsymbol{X}$ e $\boldsymbol{Y}$. El método PLS, en su forma clásica basada en el algoritmo NIPALS, encuentra vectores $\boldsymbol{w}, \boldsymbol{c}$ de *pesos* que cumplen:

$$\mathrm{cov}(\boldsymbol{t}, \boldsymbol{u})^2 = \big(\mathrm{cov}(\boldsymbol{X}\boldsymbol{w}, \boldsymbol{Y}\boldsymbol{c})\big)^2 \max_{|\boldsymbol{r}|=|\boldsymbol{s}|=1} \big(\mathrm{cov}(\boldsymbol{X}\boldsymbol{r}, \boldsymbol{Y}\boldsymbol{s})\big)^2$$

donde $\boldsymbol{t}, \boldsymbol{u}$ son vectores columna de las matrices $\boldsymbol{T}, \boldsymbol{U}$ respectivamente,

$$\boldsymbol{X}\boldsymbol{w} = \boldsymbol{t},$$

$$\boldsymbol{Yc} = \boldsymbol{u}\,,$$

y los vectores $\boldsymbol{w}$, $\boldsymbol{c}$ son los que satisfacen el criterio de maximización. Los vectores de *saturaciones* $\boldsymbol{p}$ y $\boldsymbol{q}$ son calculados como los coeficientes de la regresión de $\boldsymbol{X}$ en $\boldsymbol{t}$ y $\boldsymbol{Y}$ en $\boldsymbol{u}$ respectivamente,

$$\boldsymbol{p} = \frac{\boldsymbol{X}'\boldsymbol{t}}{\boldsymbol{t}'\boldsymbol{t}}\,,$$

$$\boldsymbol{q} = \frac{\boldsymbol{Y}'\boldsymbol{u}}{\boldsymbol{u}'\boldsymbol{u}}\,.$$

En otras palabras, PLS se enfoca en la maximización de la *covarianza* entre combinaciones lineales de las matrices $\boldsymbol{X}$ e $\boldsymbol{Y}$, donde

$$\mathrm{cov}(\boldsymbol{t},\boldsymbol{u}) = \frac{\boldsymbol{t}'\boldsymbol{u}}{n}$$

denota la covarianza muestral entre los vectores $\boldsymbol{t}$ y $\boldsymbol{u}$.

Se puede mostrar que el vector $\boldsymbol{w}$ corresponde al primer eigenvector del siguiente problema de eigenvalores (Rosipal & Kramer, 2006):

$$\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{Y}'\boldsymbol{X}\boldsymbol{w} = \lambda\boldsymbol{w}\,.$$

Si el primer par $(\lambda, \boldsymbol{w})$ no extrae una cantidad relevante de información respecto a la $\mathrm{cov}(\boldsymbol{X},\boldsymbol{Y})$, se procedería de manera secuencial a calcular nuevos pares $(\lambda, \boldsymbol{w})$ hasta recuperar tal covarianza en un porcentaje alto. Este proceso se describe en más detalle en la sección 2.3.

PLS, como método de proyección a variables latentes no es el único método disponible. Entre otros, los más usados que entran en esta categoría es el Análisis de Componentes Principales PCA y el Análisis de Correlaciones Canónicas CCA. Sus diferencias y similitudes se pueden identificar por medio de sus objetivos de optimización (y de las variables que consideran).

En PCA, el objetivo es la maximización de la varianza en una sola matriz de variables $\boldsymbol{X}$:

$$\max_{\boldsymbol{a}\neq\boldsymbol{0}} \frac{\boldsymbol{a}\boldsymbol{\Sigma}\boldsymbol{a}'}{\boldsymbol{a}\boldsymbol{a}'}\,,$$

donde $\boldsymbol{\Sigma}$ denota la matriz de covarianza de $\boldsymbol{X}$. Es importante notar que en PCA no se tiene en consideración la matriz $\boldsymbol{Y}$, razón por la que este método se suele clasificar como un modelo no supervisado.

Por otra parte, CCA tiene como objetivo encontrar la dirección de máxima correlación entre dos matrices $\boldsymbol{X}_{\boldsymbol{n}\times\boldsymbol{p}}$ y $\boldsymbol{Y}_{n\times q}$:

$$\mathrm{Corr}(\boldsymbol{X},\boldsymbol{Y}) \;=\; \max_{\boldsymbol{a},\boldsymbol{b}\neq\mathbf{0}} \frac{\boldsymbol{a}'\boldsymbol{\Sigma}_{\mathbf{12}}\boldsymbol{b}}{\sqrt{\mathbf{a}'\boldsymbol{\Sigma}_{\mathbf{11}}\boldsymbol{a}}\sqrt{\boldsymbol{b}'\boldsymbol{\Sigma}_{\mathbf{22}}\boldsymbol{b}}},$$

donde $\boldsymbol{\Sigma}_{\mathbf{11}}$ denota la matriz de covarianza de $\boldsymbol{X}$, $\boldsymbol{\Sigma}_{\mathbf{22}}$ denota la matriz de covarianza de $\boldsymbol{Y}$, $\boldsymbol{\Sigma}_{\mathbf{12}}$ denota la matriz de covarianza entre $\boldsymbol{X}$ e $\boldsymbol{Y}$. E problema de optimización de PLS, la maximización de la covarianza entre $\boldsymbol{X}$ e $\boldsymbol{Y}$, se puede reescribir como:

$$\mathrm{Cov}(\mathbf{X},\mathbf{Y}) = \sqrt{\mathrm{Var}(\boldsymbol{X})} * \mathrm{Corr}(\boldsymbol{X},\boldsymbol{Y}) * \sqrt{\mathrm{Var}(\boldsymbol{Y})}\,.$$

Lo anterior se puede interpretar como una variante de CCA, donde el objetivo de optimización de máxima correlación está restringido con la necesidad de explicar simultáneamente la varianza en $\boldsymbol{X}$ y en $\boldsymbol{Y}$ (Frank & Friedman, 1993). En el caso de que $\boldsymbol{Y}$ sea univariada, solo se considera la varianza en $\boldsymbol{X}$**.**

2.2 Uso de PLS en clasificación y relación con CCA, FDA

El análisis PLS no fue diseñado como una herramienta para clasificación, cuando las variables de respuesta en $\boldsymbol{Y}$ son categóricas (Liu & Rayens, 2007). A pesar de esto, y a partir de una codificación adecuada por medio de variables artificiales que representen a las variables respuesta categóricas de interés, PLS se ha utilizado de manera rutinaria para tales propósitos, y existe evidencia empírica de su buen desempeño en este rol (Barker & Rayens, 2003). Cuando se utiliza PLS para clasificación, recibe el nombre de PLS – DA (Partial Least Squares - Discriminant Analysis). Por otra parte, se ha mostrado que PLS tiene una conexión con el análisis de discriminante de Fisher (este análisis se denotará por FDA) y con el Análisis de Correlación Canónica CCA. Dicha conexión provee de fundamento teórico a PLS para ser utilizado con propósitos de clasificación en casos en los que FDA no puede ser aplicado, o cuando la reducción de dimensionalidad sea requerida (Barker & Rayens, 2003; Rosipal & Kramer, 2006).

Consideremos la matriz $\boldsymbol{X} \in \mathbb{R}^p$ que representa n observaciones de g clases (o grupos). Sea $\boldsymbol{Y}^*$ la matriz de pertenencia $n \times (g-1)$ como

$$\boldsymbol{Y}^* = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \mathbf{1}_{n_{g-1}} \\ \mathbf{0}_{n_g} & \mathbf{0}_{n_g} & \cdots & \mathbf{0}_{n_g} \end{pmatrix},$$

donde $\{n_i\}_{i=1}^{g}$ es el número de muestras en cada clase, $\sum_{i=1}^{g} n_i = n$ y donde $\mathbf{0}_{n_i}$, $\mathbf{1}_{n_i}$ son vectores $n_i \times 1$ todos cero o uno respectivamente. Sean

$$\overline{\boldsymbol{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \boldsymbol{x}_i^j \,,$$

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i} \boldsymbol{x}_i^j \,,$$

$$\boldsymbol{S_X} = \frac{1}{n-1}\boldsymbol{X'X} \,,$$

$$\boldsymbol{S_{Y^*}} = \frac{1}{n-1}\boldsymbol{Y^{*\prime}Y^*} \,,$$

$$\boldsymbol{S_{XY^*}} = \frac{1}{n-1}\boldsymbol{X'Y^*} \,,$$

las estimaciones muestrales de las matrices de covarianza $\boldsymbol{\Sigma_X}$, $\boldsymbol{\Sigma_{Y^*}}$ respectivamente, y la matriz de covarianza de productos cruzados $\boldsymbol{\Sigma_{XY^*}}$ ($\boldsymbol{X}$ e $\boldsymbol{Y^*}$ se asumen con media cero). Más aún, sean

$$\boldsymbol{H} = \sum_{i=1}^{g} n_i(\overline{\boldsymbol{x}}_i - \overline{\boldsymbol{x}})(\overline{\boldsymbol{x}}_i - \overline{\boldsymbol{x}})' \,,$$

$$\boldsymbol{E} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(\boldsymbol{x}_i^j - \overline{\boldsymbol{x}}_i\right)\left(\boldsymbol{x}_i^j - \overline{\boldsymbol{x}}_i\right)' \,,$$

las denominadas matrices de *suma de cuadrados entre grupos* y de *suma de cuadrados dentro de grupos*, donde $\boldsymbol{x}_i^j$ representa un vector $N-$ dimensional de la muestra j que pertenece a la clase i, $\{n_i\}_{i=1}^{g}$ denota el número de muestras en cada clase, $\sum_{i=1}^{g} n_i = n$. FDA fue desarrollado por Fisher en 1936, diseñado para maximizar la variabilidad entre grupos con respecto a una medida de variabilidad dentro de grupos. Las direcciones en las que se proyectan los datos están dadas por los eigenvectores del siguiente problema de eigenvalores (Barker & Rayens, 2003)

$$\boldsymbol{E^{-1}Ha} = \lambda\boldsymbol{a} \,.$$

Las direcciones resultantes en FDA son idénticas a las dadas por CCA utilizando una codificación de la matriz $\boldsymbol{Y}$ en $\boldsymbol{Y^*}$ representando pertenencia de grupo (Bartlett, 1938).

n el caso de que la discriminación sea el objetivo del análisis PLS (y utilizando $\boldsymbol{Y^*}$), el objetivo de minimización de PLS se puede modificar como sigue:

$$\frac{\mathrm{Cov}(\mathbf{X},\mathbf{Y})}{\sqrt{\mathrm{Var}(\boldsymbol{Y})}} = \sqrt{\mathrm{Var}(\boldsymbol{X})} * \mathrm{Corr}(\boldsymbol{X},\boldsymbol{Y})\,.$$

Es decir, eliminando el peso de la varianza en $\boldsymbol{Y}$ de la función objetivo (Barker & Rayens, 2003). De esta manera, el problema de eigenvalores se reduce al siguiente (Rosipal & Kramer, 2006):

$$\boldsymbol{X'Y(Y'Y)^{-1}Y'Xw = X'\widetilde{Y}\widetilde{Y}'Xw = \lambda w}\,,$$

donde

$$\boldsymbol{\widetilde{Y} = Y(Y'Y)^{-1/2}}$$

representa una matriz de variables de salida normalizadas y no correlacionadas. Al usar la siguiente relación

$$(n-1)\boldsymbol{S_{XY}S_Y^{-1}S'_{XY} = H}\,.$$

Los eigenvectores resultantes del problema anterior son equivalentes a los eigenvectores de

$$\boldsymbol{Hw = \lambda w}\,.$$

Así, este método PLS modificado con matriz $\boldsymbol{Y}$ codificada se basa en las soluciones del problema de eigenvalores de la matriz $\boldsymbol{H}$, lo que conecta la técnica con CCA y, de manera equivalente, a FDA (Bartlett, 1938).

2.3 Algoritmo NIPALS

La aplicación de PLS – DA parte de la estructura del modelo PLS usando una matriz de pertenencia $\boldsymbol{Y}$ como la mencionada en la sección anterior, de manera que la estimación de los vectores y matrices se realiza de la misma manera que en PLS. Dada esta similitud, el algoritmo NIPALS es el algoritmo que se suele emplear para los cálculos mencionados (Chiang et al., 2000).

Denotemos por $\boldsymbol{X} \in \mathbb{R}^p$ un espacio $p-$dimensional que representa las variables predictoras, y sea $\boldsymbol{Y} \in \mathbb{R}^{g-1}$ el espacio $g-1$ dimensional que representa la pertenencia de la observación a uno de los g grupos. De manera similar a PCA, PLS no es invariante a la escala de los datos, por lo que se asume que las matrices $\boldsymbol{X}$ e $\boldsymbol{Y}$ están estandarizadas para tener que las variables incluidas ya propiamente en el algoritmo tengan media 0 y varianza 1.

El algoritmo NIPALS construye cada componente por cada iteración, de manera que, a excepción del primer componente, la información que explican no viene influido por los componentes calculados previamente. El algoritmo es como sigue:

Para $h = 1, \dots, k$:

1. Sea $\boldsymbol{u_h} = \boldsymbol{y_j}$, donde $\boldsymbol{y_j}$ es una variable en $\boldsymbol{Y}$ escogida de manera aleatoria.
2. $\boldsymbol{w'_h} = \boldsymbol{u'_h X / u'_h u_h}$
3. Se normaliza $\boldsymbol{w'_h}$ i.e. $\boldsymbol{w'_h} = \boldsymbol{w'_h} / \|\boldsymbol{w'_h}\|$
4. $\boldsymbol{t_h} = \boldsymbol{X w_h} \, / \, \boldsymbol{w'_h w_h}$
5. $\boldsymbol{q'_h} = \boldsymbol{t'_h Y} \, / \, \boldsymbol{t'_h t_h}$
6. Se normaliza $\boldsymbol{q'_h}$ i.e. $\boldsymbol{q'_h} = \boldsymbol{q'_h} / \|\boldsymbol{q'_h}\|$
7. $\boldsymbol{u_h} = \boldsymbol{Y q_h} \, / \, \boldsymbol{q'_h q_h}$
8. Se compara el valor $\boldsymbol{t}$ con el obtenido en la iteración anterior. Si $|\boldsymbol{t_h} - \boldsymbol{t_{h-1}}|$ es menor a una tolerancia especificada, se continúa al paso 9. En caso contrario, se regresa al paso 2. Cuando $\boldsymbol{Y}$ es univariada, los pasos 5 – 8 se pueden omitir asignando $\boldsymbol{q_h} = 1$ y no se requieren más iteraciones.
9. $\boldsymbol{p'_h} = \boldsymbol{t'_h X / t'_h t_h}$
10. $\boldsymbol{t_h} = \boldsymbol{t_h} \|\boldsymbol{p'_h}\|$
11. $\boldsymbol{w'_h} = \boldsymbol{w'_h} \|\boldsymbol{p'_h}\|$
12. Se normaliza $\boldsymbol{p'_h}$ i.e. $\boldsymbol{p'_h} = \boldsymbol{p'_h} / \|\boldsymbol{p'_h}\|$
13. $\boldsymbol{b_h} = \boldsymbol{u'_h t_h / t'_h t_h}$
14. $\boldsymbol{E_h} = \boldsymbol{E_{h-1}} - \boldsymbol{t_h p'_h}$; $\boldsymbol{E_0} = \boldsymbol{X}$
15. $\boldsymbol{F_h} = \boldsymbol{F_{h-1}} - \boldsymbol{b_h t_h q'_h}$; $\boldsymbol{F_0} = \boldsymbol{Y}$
16. $\boldsymbol{F^*_h} = \boldsymbol{F^*_{h-1}} - \boldsymbol{u_h q'_h}$; $\boldsymbol{F^*_0} = \boldsymbol{Y}$

El algoritmo se repite reemplazando las matrices $\boldsymbol{X}$ e $\boldsymbol{Y}$ en los pasos 2, 4, 5, 7, 9 por las matrices residuales $\boldsymbol{E_h}$ y $\boldsymbol{F_h}$ respectivamente. El algoritmo se puede iterar hasta que el número de componentes es igual al número de variables originales, es decir,

$$k \leq p \, .$$

El número de componentes a utilizar depende del estudio en cuestión, dado que a medida que se agregan componentes se aumenta la cantidad de información que el modelo explica, al mismo tiempo que aumenta la complejidad del mismo.

3 Métodos de remuestreo

Los métodos de remuestreo son una herramienta indispensable en la estadística moderna. Consisten en tomar muestras de manera repetida de un conjunto de datos

de entrenamiento y ajustar en cada muestreo un modelo de interés para obtener información adicional acerca del mismo. Estos métodos solían ser costosos en el ámbito computacional, ya que se requiere el ajuste repetido de un mismo modelo con subconjuntos distintos. Con los avances recientes en el poder de cómputo, los requerimientos en los métodos de remuestreo generalmente no son prohibitivos (James et al., 2013).

La utilidad de los métodos de remuestro está en la evaluación del desempeño de un modelo estimado en la predicción de datos no utilizados en la estimación del mismo. Al ajustar un modelo a un conjunto de datos de entrenamiento, se puede calcular el error que presentan dichos datos con respecto a los estimados por tal modelo. Este tipo de error se denomina *error de entrenamiento*, dado que se calcula utilizando los valores y las estimaciones de los datos que se utilizaron para la construcción del modelo. Sin embargo, este cómputo no representa la bondad de ajuste de dicho modelo si se utilizan datos que no hayan sido utilizados en la estimación del modelo, y generalmente se llega a conclusiones erróneas si sólo se toma en cuenta este error. Otro tipo de error, el *error de prueba*, se obtiene con el cómputo de valores que no hayan formado parte de la estimación inicial del modelo, y es una mejor representación de la capacidad de generalización del modelo ante datos nuevos. El objetivo principal de los métodos de remuestro es la estimación del error de prueba utilizando subconjuntos de los datos de entrenamiento con el fin de elegir el modelo que logre una mejor predicción de los datos futuros.

3.1 Validación cruzada (CV)

El procedimiento de validación cruzada es quizá el método de remuestreo más conocido. Denotemos al conjunto completo de datos como

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\},$$

donde x_i es la observación i y y_i es el valor de respuesta de dicha observación. En términos generales, consiste en dividir los datos en dos subconjuntos: uno de entrenamiento y otro de validación. Los datos de entrenamiento se utilizan para ajustar el modelo deseado, y éste se utiliza para predecir los valores de los datos de validación. El *error de validación* (*test error*) se calcula para esta separación de los datos, y se prosigue a realizar un nuevo muestreo. Se pueden identificar dos enfoques distintos.

3.1.1 Leave One Out Cross Validation

El enfoque de LOOCV involucra dividir el conjunto completo de datos de manera que el conjunto de validación tiene solo *una* observación (también conocido como *jackknife*), y las demás $n-1$ observaciones conforman el conjunto de entrenamiento. El modelo entonces se ajusta en los $n-1$ datos, y se estima el valor

de la observación que se “dejó fuera”. El proceso inicia con el conjunto de validación $\{(x_1, y_1)\}$ y el conjunto de entrenamiento $\{(x_2, y_2), \dots, (x_n, y_n)\}$ con el que se ajusta el modelo. Posteriormente se estima $\hat{y}_1$ por medio de x_1 y se calcula el error cuadrado $MSE_1 = (y_1 - \hat{y}_1)^2$. El proceso continúa reemplazando el conjunto de validación por $\{(x_2, y_2)\}$ y el conjunto de entrenamiento por $\{(x_1, y_1), (x_3, y_3), \dots, (x_n, y_n)\}$ obteniendo $MSE_2 = (y_2 - \hat{y}_2)^2$ y así sucesivamente hasta obtener n errores cuadrados $MSE_1, MSE_2, \dots, MSE_n$. El valor estimado por LOOCV del error de entrenamiento MSE es

$$\frac{1}{n}\sum_{i=1}^{n} MSE_i \, .$$

Una de las ventajas de este procedimiento es que los resultados no se ven afectados por un factor aleatorio, debido a la manera específica en la que se construyen los conjuntos de validación y de prueba. Una desventaja importante, sin embargo, es que la variabilidad de cada MSE_i es alta debido a que se basa en una sola observación. Otra posible desventaja es que el costo computacional aumenta a medida que el número de observaciones n en el conjunto de datos inicial crece.

3.1.2 k – fold Cross Validation

El procedimiento de k – fold Cross Validation es una alternativa a LOOCV descrito anteriormente. Consiste en dividir el conjunto de datos en k subconjuntos (llamados k - *folds*) de un tamaño aproximadamente igual entre sí. El primero de estos grupos se considera como el conjunto de validación, y el resto de los grupos conforman el conjunto de entrenamiento. Se estima entonces el MSE_1 con los datos de validación y se repite el procedimiento k iteraciones, reemplazando en cada iteración el conjunto de validación por otro de los subconjuntos formados. El estimado del MSE se calcula como

$$\frac{1}{k}\sum_{i=1}^{k} MSE_i \, .$$

En el caso en el que $k = n$, se tiene que $k -$ fold CV es equivalente a LOOCV. Comúnmente se utiliza este método con $k = 5$ o $k = 10$. La ventaja de usar estos valores para k con respecto a LOOCV es el tiempo de cómputo requerido para estimar MSE.

Al realizar CV, la meta es determinar el desempeño esperado de un modelo estadístico en datos independientes (en los ejemplos anteriores, la estimación de MSE es el dato de interés), pero también es utilizado para encontrar el *punto mínimo en la curva MSE estimada*. Esto es útil cuando se requiere escoger uno de varios

modelos posibles, o un modelo con diferentes grados de flexibilidad) que tenga el menor error de entrenamiento. En este caso es el punto donde se alcanza el mínimo valor del MSE, en lugar del valor específico del mismo.

El error MSE utilizado en los procedimientos anteriores es utilizado en el contexto de regresión: cuando la variable de respuesta $\boldsymbol{Y}$ es cuantitativa. En los casos de clasificación los métodos de validación cruzada también son utilizados, con la diferencia en el uso del número de observaciones mal clasificadas en lugar del error cuadrático medio, denominado *tasa de error:*

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{Err}_i \,,$$

donde $\mathrm{Err}_i = I(y_i \neq \hat{y}_i)$.

4 Paquetería disponible en R para realizar PLS-DA

4.1 mixOmics

El paquete mixOmics, disponible para el software estadístico R, está dedicado al análisis multivariado de conjuntos de datos biológicos, con un enfoque específico a la exploración de datos, reducción de la dimensionalidad y la visualización (Rohart et al., 2017). Este paquete incluye una implementación del modelo PLS – DA, y tiene además la capacidad de integrar datos de múltiples fuentes o a través de estudios independientes. El contexto en el que trabaja mixOmics es en el análisis supervisado, buscando clasificar grupos, identificando las características que predicen la pertenencia a dichos grupos de mejor manera.

Además del modelo PLS – DA, este paquete incluye otras técnicas como el análisis de componentes principales PCA, de componentes independientes ICA, PLS y el análisis de correlaciones canónicas regularizado rCCA.

En el momento en que fue escrito este documento, el paquete no se encuentra disponible en el repositorio CRAN. La descarga se puede realizar a través de la página de internet www.bioconductor.org

4.2 ropls

El paquete ropls implementa los métodos PCA y PLS(DA) con las versiones originales del algoritmo basado en NIPALS. Incluye el cálculo de los métricos R^2 y Q^2, diagnósticos de permutación, el cálculo de los valores VIP, así como varios gráficos (puntajes, saturaciones, predicciones, diagnósticos, atípicos, etc.).

4.3 MetaboAnalystR

MetaboAnalystR es un paquete que contiene todas las funciones y bibliotecas utilizadas en el servidor web MetaboAnalyst (Xia et al., 2009). MetaboAnalyst es una herramienta de procesamiento de datos metabolómicos, normalización de datos, análisis estadístico multivariado, generación de gráficas, entre otros. De los métodos que soporta, se incluye PCA, PLS-DA, pruebas t, agrupamiento jerárquico (hierarchical clustering). El paquete MetaboAnalystR permite utilizar las funcionalidades del servicio web de manera local.

Las funciones que incorpora el paquete detalladas en la documentación disponible indican que los métodos de PLS-DA usados son adaptaciones de funciones en los paquetes mixOmics y ropls, por lo que el ejemplo a realizar no tomará en cuenta el uso de este paquete.

5 Ejemplo: Tratamiento con subproducto de mango

5.1 Descripción de los datos

El conjunto de datos a utilizar como ejemplo de la utilización del modelo PLS – DA es el conjunto de datos analizados en un estudio realizado en la Universidad Autónoma de Querétaro cuyo propósito fue evaluar el efecto de un subproducto de jugo de mango en síntomas de infecciones del tracto gastrointestinal y respiratorio superior en niños de 6 a 8 años. El estudio fue realizado por medio de una intervención aleatorizada, doble ciego, paralela y controlada. 91 niños fueron reclutados de una escuela primaria en Querétaro, México. De los 91 niños, 11 no se presentaron al estudio y 15 fueron descartados dada la presencia de parásitos patógenos. El grupo de tratamiento ($n = 33$) recibió diariamente el subproducto de mango disuelto en 50 mililitros de agua durante 2 meses, mientras que el grupo de control ($n = 32$) recibió agua con saborizante (Anaya-Loyola et al.,2020). Del estudio anterior se registraron un total de 160 variables, que se componen de las siguientes categorías:

- Variable de grupo: Identifica si el menor forma parte del grupo de control o tratamiento. Se considera como variable categórica binaria.
- Variables antropométricas: Para evaluar el crecimiento y nutrición general de los niños durante el estudio. Comprenden variables nominales (edad) y variables continuas.

- Variables hematológicas. Información relacionada con células sanguíneas. Las variables son todas continuas.
- Variables proteínicas. Contienen información respecto a proteínas sanguíneas. Todas las variables son continuas.

De manera general, se exponen las variables a utilizar y la categoría en la que se encuentran en la Tabla 5.1.1

Variables			Clasificación
Grupo	IgE	IL-6	Variable de grupo
Edad	IgG	IL-8	Antropométrica
Peso	IgM	IL-12p70	Hematológica
Talla	CD40	Lipocalin-2	Proteínica
Cintura	CRP	MCP-1	
Cintura/Estatura	E-Selectin	MCP-2	
Cadera	IL-1a	MIF	
ZP/E	IL-1b	MIP-1a	
ZT/E	IL-2Ra	MIP-1b	
ZIMC/E	IL-10	OPN	
WBCx	IL-13	PAI-I	
RBCx	IL-18	PF4	
HGBx	ST2	Procalcitonin	
HCTx	TNFa	RAGE	
MCVx	CD14	Resistin	
MCHx	CD163	Thrombomodulin	
PLTx	FAS	TREM-1	
LYMx	FASL	uPAR	
MXDx	G-CSF	VCAM-1	
NEUTx	ICAM-1	VEGF	
MPV	IL-2		
IgA	IL-4		

Tabla 5.1.1 Número de componentes óptimo por distancia

Las variables fueron medidas en cada niño antes y después de aplicar el tratamiento. Para realizar el análisis PLS – DA, la matriz $\boldsymbol{X}$ se conformará por las variables hematológicas y las variables proteínicas medidas luego de la ingesta del subproducto de mango, mientras que el vector $\boldsymbol{Y}$ tendrá la información de pertenencia al grupo de control o tratamiento de cada niño. El objetivo en este

ejemplo será analizar si se puede discriminar a los grupos de control y tratamiento con los resultados de las variables consideradas, y en caso afirmativo, cuáles de las variables son las que más cambian al aplicar el tratamiento. En este contexto, el modelo sirve como un análisis exploratorio para identificar posibles cambios en las variables expuestas en la Tabla 5.1.1 al aplicar el tratamiento, así como la obtención de un modelo de clasificación cuya precisión puede sugerir el grado de separación que se puede obtener.

5.2 mixOmics: Modelo PLS – DA

El ajuste del modelo se realiza con la función **plsda**, especificando las matrices de variables $\boldsymbol{X}$ e $\boldsymbol{Y}$. El número de componentes que se utilizará en el modelo a usar se obtendrá utilizando el método de validación cruzada; para iniciar, se ajustará un modelo con 10 componentes. La función estandariza los datos por defecto (media 0 y varianza 1). La estandarización se realiza en ambas matrices $\boldsymbol{X}$ e $\boldsymbol{Y}$, lo que es posible dado que la matriz $\boldsymbol{X}$ se conforma de variables continuas, y la matriz $\boldsymbol{Y}$ es la matriz de pertenencia. Esta transformación se puede omitir con el argumento `scale = FALSE`.

El resultado del modelo PLS – DA ajustado contiene información acerca de la varianza recuperada por cada componente para $\boldsymbol{X}$ e $\boldsymbol{Y}$. Es importante aclarar que, a diferencia de PCA, los componentes no se construyen con el objetivo de recuperar la mayor varianza.

Componente	Varianza Recuperada (X)
1	0.09068
2	0.07564
3	0.04812
4	0.03618
5	0.0402
6	0.02951
7	0.03944
8	0.04062
9	0.02071
10	0.03322

Tabla 5.2.2 Varianza de X explicada por cada componente

En la Tabla 5.2.2 se muestra que la varianza recuperada de X por los componentes no es muy grande; el primer componente es el que más proporción recupera: 8.9%.

La varianza de Y recuperada por los componentes como reporta mixOmics se muestra en la Tabla 5.2.3.

Componente	Varianza Recuperada (Y)
1	1
2	0.47364
3	0.31791
4	0.24975
5	0.21861
6	0.19939
7	0.18196
8	0.17069
9	0.16146
10	0.14762

Tabla 5.2.3 Varianza de Y explicada por cada componente

Para calcular el número de componentes óptimo en PLS-DA, se utilizará validación cruzada de 10 iteraciones (10 – fold cross validation), repitiendo dicho proceso 50 veces, con el fin de estimar la tasa de error general de clasificación y la tasa de error balanceado (BER), y verificar qué número de componentes minimizan dichos errores. El error general de clasificación está definido como:

$$\frac{FP + FN}{FP + FN + TP + TN},$$

donde FP y FN son predicciones incorrectas de positivos y negativos respectivamente, y TP, TN son predicciones correctas de positivos y negativos respectivamente. En otras palabras, el error general de clasificación es la proporción de predicciones erróneas respecto al total de predicciones.

La tasa de error balancead BER está definido como:

$$\frac{FPR + FNR}{2},$$

donde

$$FPR = \frac{FP}{FP + TN},$$

$$FNR = \frac{FN}{FN + TP},$$

son las tasas de falsos positivos y falsos negativos (False Positive Rate, False Negative Rate) respectivamente. La tasa BER es apropiada cuando el número de muestras por clase o grupo no es balanceado (Rohart et al., 2017). En nuestro ejemplo, el número de muestras en el grupo de control es de 32, y para el grupo de tratamiento son 33, por lo que el uso de la tasa BER no ofrece un beneficio importante respecto a la tasa de error general de clasificación. Para realizar el análisis, se utiliza la función **perf**. Los resultados de la validación cruzada se muestran en la Figura 5.2.1.

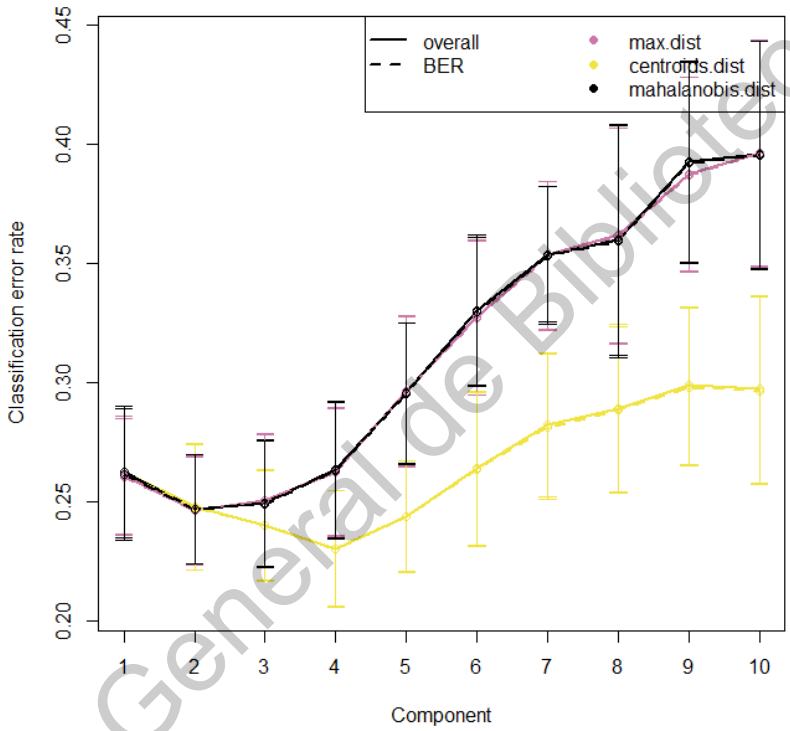

Figura 5.2.1 Tasa de error de clasificación por componente, calculado usando validación cruzada

Las tasas de error (overall y BER), disminuyen a medida que el número de componentes aumenta. Considerando los errores con la distancia máxima y la distancia Mahalanobis, los errores empiezan a aumentar después del tercer componente. En el caso de la distancia a centroides, los errores aumentan después del cuarto componente. En la figura 5.2.1 se muestra también la desviación estándar para cada estimación. Una revisión visual sugiere que el número óptimo de componentes a utilizar son tres.

El paquete también incluye una función para indicar el número óptimo de componentes a utilizar. Según el paquete mixOmics, el número óptimo de componentes a utilizar se muestra en la Tabla 5.2.4.

	Max dist	Centroids dist	Mahalanobis dist
Overall	2	3	2
BER	2	3	2

Tabla 5.2.4 Número de componentes óptimo por distancia

Los resultados que muestra el paquete sugieren utilizar dos componentes (tres si se utiliza la distancia a centroides), lo que es consistente con lo observado con la gráfica de las estimaciones de error de clasificación. Además de la información anterior, el proceso de validación cruzada también reporta los valores AUC para cada número de componentes.

Componente	Media AUC	SD AUC
1	0.82648	0.0128
2	0.83491	0.0182
3	0.82355	0.0153
4	0.80616	0.0201
5	0.7813	0.0219
6	0.74172	0.0263
7	0.71443	0.0297
8	0.70517	0.0351
9	0.67949	0.0346
10	0.67263	0.0456

Tabla 5.2.5 Área bajo la curva (AUC) para cada componente, por media y desviación estándar

Como se muestra en la Tabla 5.2.5, se puede ver que el número de componentes que logra un valor AUC más alto es con dos componentes, lo que contrasta con los resultados óptimos indicados por el paquete. La desviación estándar es también menor, por lo que se ajustará el modelo final con los primeros dos componentes.

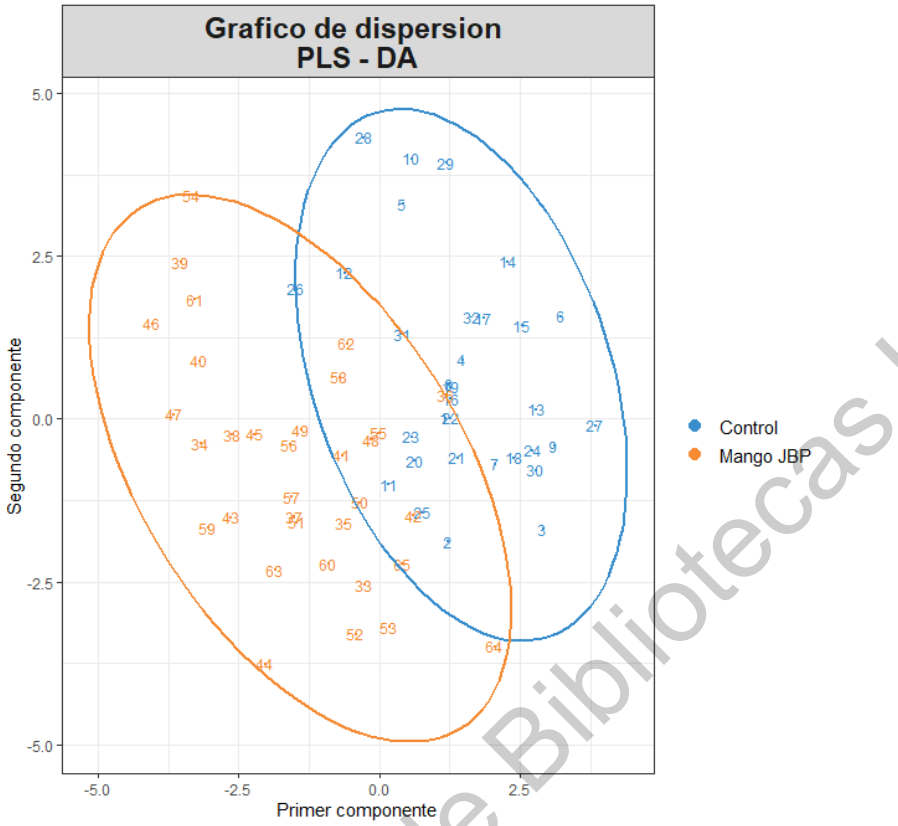

Figura 5.2.2 Gráfico de dispersión con respecto a los dos primeros componentes

En la Figura 5.2.2 se puede apreciar la separación lograda con los primeros dos componentes. El algoritmo PLS-DA parece hacer un buen trabajo en identificar los componentes que pueden discriminar los dos grupos. Para identificar las variables que más influyen en estos componentes se revisarán los coeficientes vip, cuyos valores reporta el paquete con la función **vip**.

Variable	Coeficiente vip	Clasificación
MIF	2.64668	Proteínica
NEUTx	1.87545	Hematológica
HGBx	1.8502	Hematológica
PAI.I	1.69611	Proteínica
IgM	1.60893	Proteínica
WBCx	1.47785	Hematológica
IL.2	1.40841	Proteínica
PLTx	1.39794	Hematológica
PF4	1.28965	Proteínica

IL.18	1.27046	Proteínica

Tabla 5.2.6 Coeficientes VIP del primer componente

En la tabla 5.2.6, se muestran las variables con mayor peso en el primer componente, de las cuales las principales cuatro son las variables MIF, NEUTx, HGBx y PAI.I. La variable IgM puede ser considerada también, pero su coeficiente vip no es tan elevado como las antes mencionadas. Las demás variables tienen un coeficiente vip cercano a 1, por lo que su impacto en el componente no es tan elevado. Podemos notar que las variables más importantes tienen loadings más alejados de 0. Es importante mencionar también que la variable MIF tiene un coeficiente vip bastante mayor a la variable que le sigue. Esto sugiere que la variable MIF tiene bastante peso por sí sola en este primer componente, y por tanto en general, para la separación de los grupos. Entre las variables mencionadas, vemos que no hay una presencia dominante en la clasificación de las mismas, ya que la proporción de variables proteínicas con respecto a las hematológicas es muy similar.

Variable	Coeficiente vip	Clasificación
MIF	2.54147	Proteínica
NEUTx	1.65137	Hematológica
HGBx	1.63515	Hematológica
PAI.I	1.48998	Proteínica
IgM	1.41356	Proteínica
PF4	1.36994	Proteínica
WBCx	1.13109	Hematológica
IgA	1.29181	Proteínica
IL.2	1.24409	Proteínica
Il.6	1.22955	Proteínica

Tabla 5.2.7 Coeficientes VIP del segundo componente

De los resultados del segundo componente mostrados en la Tabnla 5.2.7 notamos que las 5 variables más importantes, con base en el coeficiente vip, son exactamente las mismas que en el componente 1. En este componente las variables IL.18 y PLTx dejan de estar entre las 10 más importantes, siendo reemplazadas por

IgA e IL.6. Por último, los coeficientes vip de las 5 variables más importantes que se repiten en los dos componentes son distintos, siendo menores en el segundo componente. En este caso podemos apreciar que las variables más importantes son en su mayoría proteínicas.

Dado que las 5 variables más importantes en ambos componentes son las mismas, estudiaremos la distribución de las variables por medio de histogramas, con el fin de comparar la separación que logran dichas variables en los grupos de control y tratamiento.

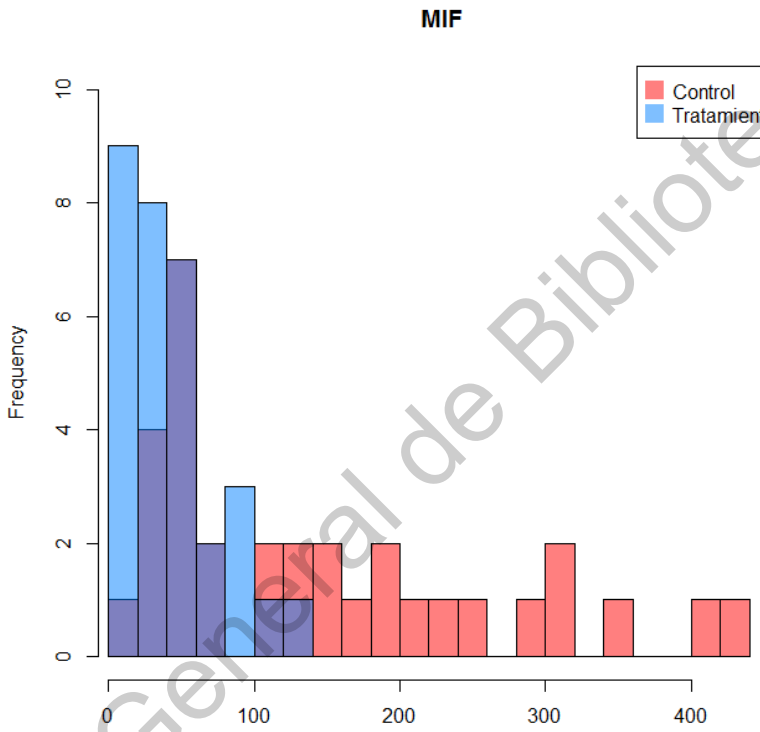

Figura 5.2.3 Histograma de la variable MIF

El histograma de la Figura 5.2.3 muestra una separación de los grupos de control y tratamiento. En el grupo de control, los registros tienen una distribución relativamente uniforme, con un ligero sesgo a la derecha. En el caso del grupo de tratamiento, los registros se concentran más a la izquierda. El diagrama de caja y brazos en la Figura 5.2.4 muestra otra comparación entre los dos grupos.

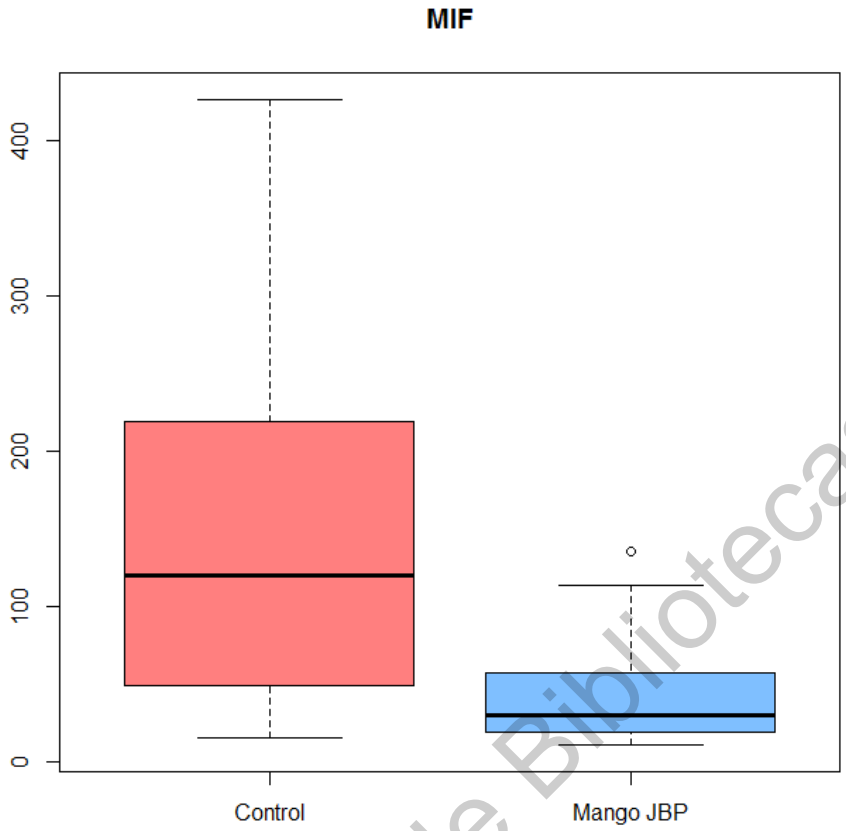

Figura 5.2.4 Diagrama de caja y brazos de la variable MIF, por grupo

Min	1er cuartil	Mediana	Media	3er cuartil	Max
15	49.92	120.15	146.79	213.45	426.91

Tabla 5.2.8 Estadísticos de la variable MIF para el grupo de control

Min	1er cuartil	Mediana	Media	3er cuartil	Max
10.54	19.09	30.2	42.57	57.16	134.97

Tabla 5.2.9 Estadísticos de la variable MIF para el grupo de control

En las Tablas 5.2.8 y 5.2.9 se muestran los estadísticos generales de la variable MIF. Los datos indican que el grupo de tratamiento tiene una media bastante menor a la media del grupo de control. De hecho, todos los registros del grupo de tratamiento son menores a la media del grupo de control.

Las distribuciones de los grupos no parecen seguir una distribución normal, por lo que no es recomendable utilizar pruebas de diferencia de medias. Para efectos de este ejemplo, se realizarán las pruebas de normalidad de Shapiro para determinar si los grupos siguen una distribución normal:

Grupo	Estadístico W	p-valor
Control	0.87833	0.001818
Tratamiento	0.85862	0.00077

Tabla 5.2.10 Resultados de la prueba de Shapiro para la variable MIF

La prueba de Shapiro test tiene como hipótesis nula que los datos considerados siguen una distribución normal. El histograma anterior sugería que los grupos no seguían una distribución normal, y la Tabla 5.2.10 muestra los resultados de la prueba de Shapiro, que sugiere coincidir con dicha sospecha. Considerando el p-valor, se asume que los datos no siguen una distribución normal.

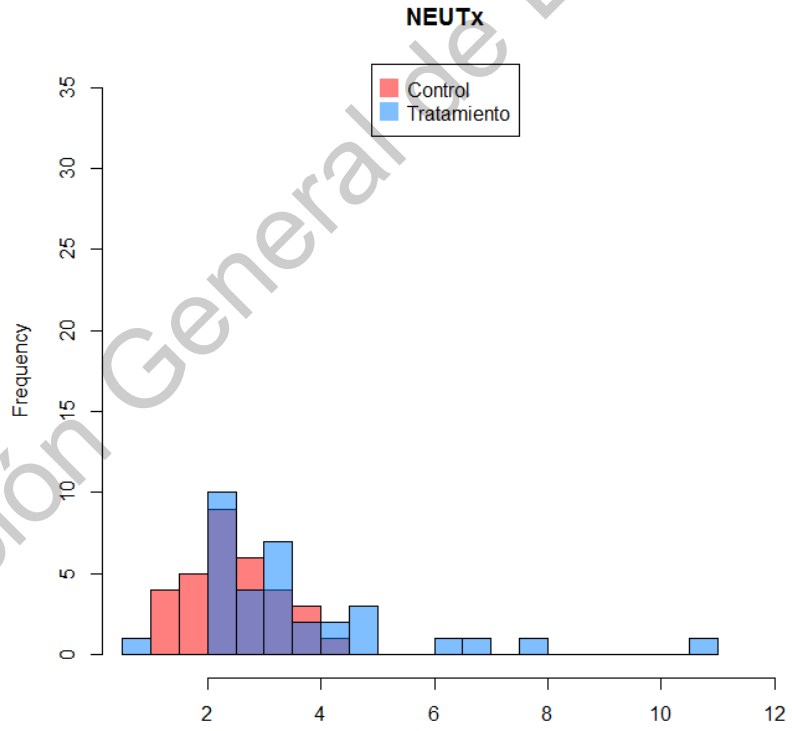

Figura 5.2.5 Histograma de la variable NEUTx

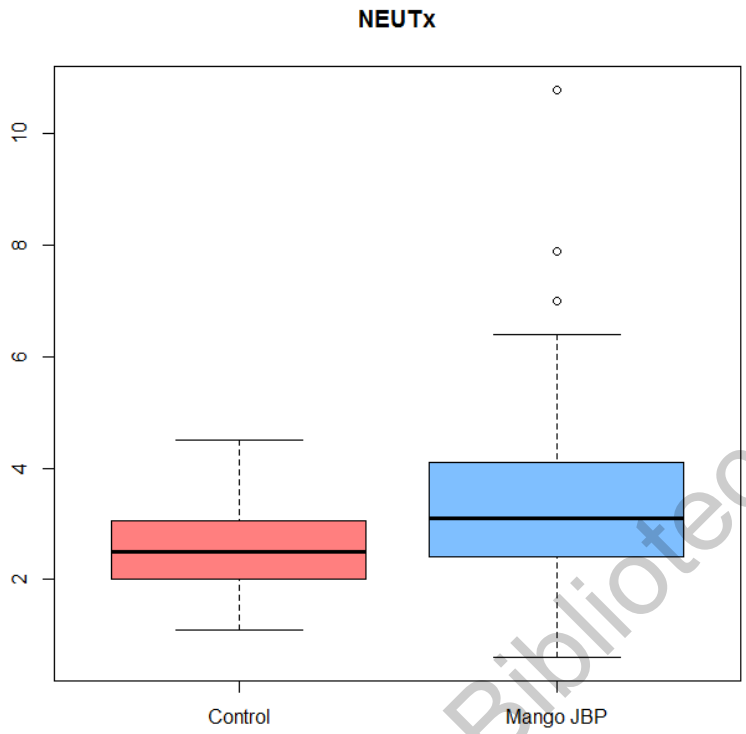

Figura 5.2.6 Boxplot de la variable NEUTx, por grupo

En la variable NEUTx, el grupo de control tiene los registros concentrados alrededor de 2, con un ligero sesgo a la derecha, como se muestra en las Figuras 5.2.5 y 5.2.6. En el caso del grupo de tratamiento, se presentan registros con valores muy elevados, aunque casi todos se concentran alrededor de 3. Esta variable no muestra una separación muy clara entre los grupos, algo que se sugería con los coeficientes vip. Pareciera que el tratamiento eleva un poco los valores de esta variable.

Min	1er cuartil	Mediana	Media	3er cuartil	Max
1.1	2	2.5	2.541	3.025	4.5

Tabla 5.2.11 Estadísticos de la variable NEUTx para el grupo de control

Min	1er cuartil	Mediana	Media	3er cuartil	Max
0.6	2.4	3.1	3.679	4.1	10.8

Tabla 5.2.12 Estadísticos de la variable NEUTx para el grupo de tratamiento

En las Tablas 5.2.11 y 5.2.12 vemos que, en el caso del grupo de control, el rango entre el mínimo y el máximo es de sólo 3.4, mientras que el grupo de tratamiento tiene un rango de 10.2, por lo que el tratamiento parece alterar la agrupación de los datos observados en el grupo de control, al mismo tiempo que los valores máximos son más altos. Vemos que la media y la mediana en el grupo de control son muy cercanos, lo que se ve reflejado en la simetría observada en el histograma, mientras que en el grupo de tratamiento pasa lo contrario: valores de mediana y media alejados entre si refleja una asimetría.

Grupo	Estadístico W	p-valor
Control	0.98263	0.8714
Tratamiento	0.8113	0.00005

Tabla 5.2.13 Resultados de la prueba de Shapiro para la variable NEUTx

El test de Shapiro muestra que el grupo de control sigue una distribución normal, mientras que el grupo de tratamiento no, de acuerdo con los resultados de la Tabla 5.2.13. Esto sugiere que el tratamiento cambia la distribución de la variable.

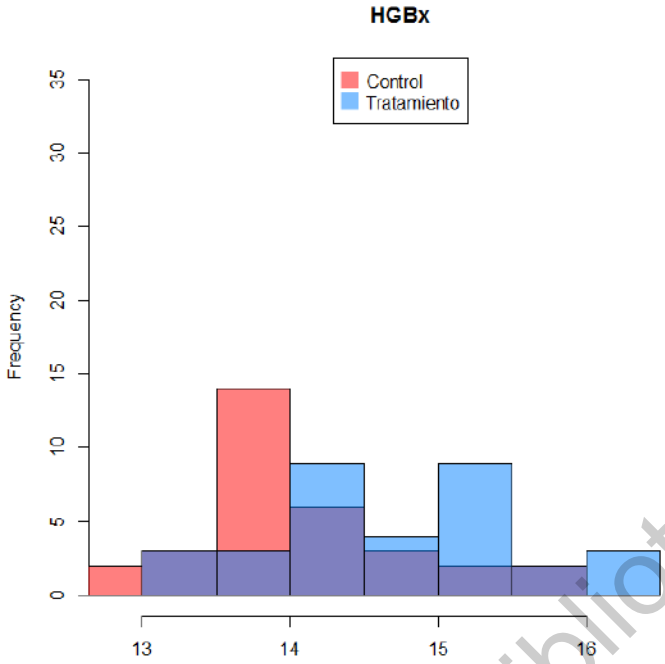

Figura 5.2.7 Histograma de la variable HGBx

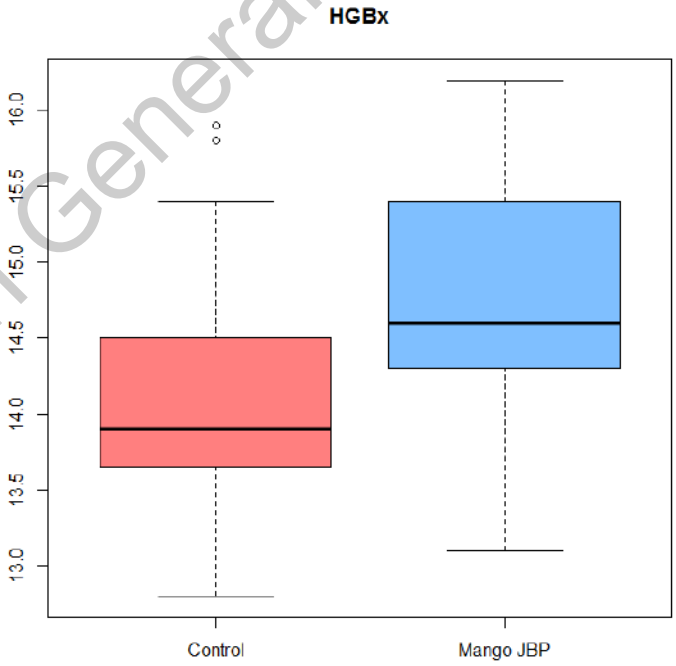

Figura 5.2.8 Boxplot de la variable NEUTx, por grupo

En el caso de HGBx, vemos que las distribuciones entre los grupos son similares, y el grupo de tratamiento tiene valores más altos que el grupo de control.

Min	1er cuartil	Mediana	Media	3er cuartil	Max
12.8	13.68	13.90	14.13	14.5	15.9

Tabla 5.2.14 Estadísticos de la variable HGBx para el grupo de control

Min	1er cuartil	Mediana	Media	3er cuartil	Max
13.1	14.3	14.6	14.74	15.4	16.2

Tabla 5.2.15 Estadísticos de la variable HGBx para el grupo de tratamiento

En las Tablas 5.2.14 y 5.2.15 se muestra una diferencia ligera entre los dos grupos en cada cuartil, siendo más elevado en el grupo de tratamiento. También notamos que la mediana y media en dicho grupo tienen valores más cercanos comparados con el grupo de control.

Grupo	**Estadístico W**	**p-valor**
Control	0.94586	0.1099
Tratamiento	0.96771	0.4197

Tabla 5.2.16 Resultados de la prueba de Shapiro para la variable HGBx

En las pruebas Shapiro de la Tabla 5.2.16 se muestra que ambos grupos siguen una distribución normal, aunque en el caso del grupo de control la decisión de conservar la hipótesis nula es menos clara que en el grupo de tratamiento. Para tener mayor información respecto a la diferencia en la distribución de los dos grupos, se realizará una prueba *t* de diferencia de medias, cuyos resultados se muestran en la Tabla 5.2.17.

Media grupo control	Media grupo tratamiento	Estadístico *t*	P - valor
14.13438	14.73939	-2.9844	0.004052

Tabla 5.2.17 Resultados de la prueba *t* para la variable HGBx

La prueba *t* de diferencia de medias muestran que las medias entre los grupos de control y tratamiento tienen una diferencia distinta de 0, con p valores bastante pequeños.

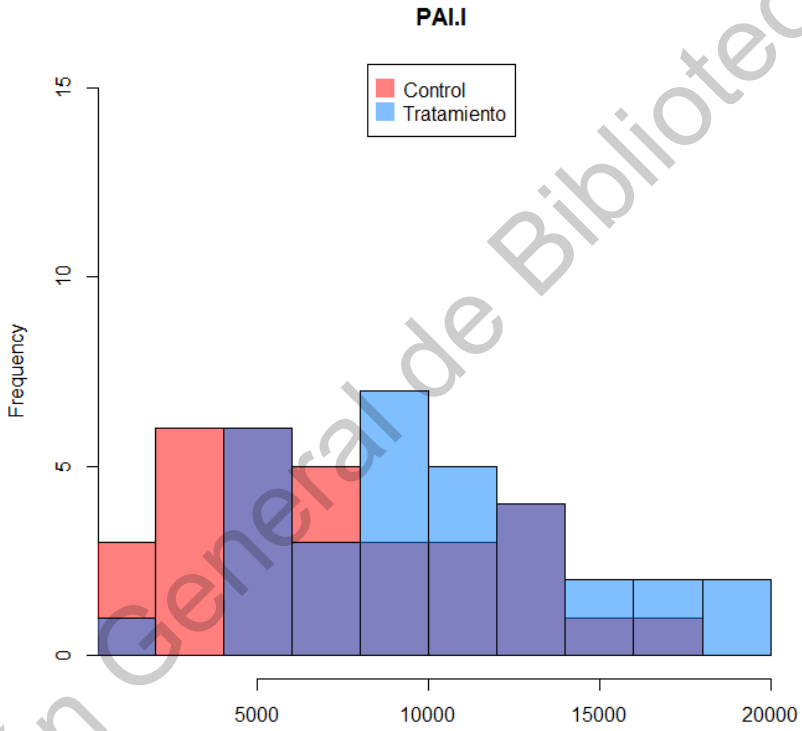

Figura 5.2.9 Histograma de la variable PAI.I

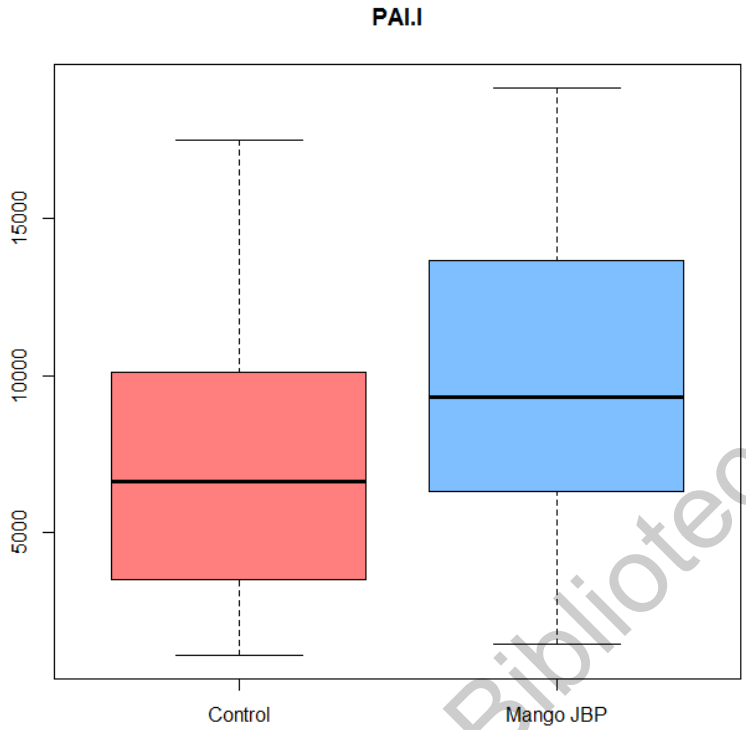

Figura 5.2.10 Boxplot de la variable PAI.I, por grupo

En el caso de la variable PAI.I, vemos que sucede algo similar al caso anterior: ambos grupos tienen una distribución similar, y el grupo de tratamiento tiene valores mayores al grupo de control.

Min	1er cuartil	Mediana	Media	3er cuartil	Max
1096	3664	6622	7093	10069	17484

Tabla 5.2.18 Estadísticos de la variable PAI.I para el grupo de control

Min	1er cuartil	Mediana	Media	3er cuartil	Max
1476	6325	9329	10094	13507	19174

Tabla 5.2.19 Estadísticos de la variable PAI.I para el grupo de tratamiento

Las Tablas 5.2.18 y 5.2.19 muestran también un comportamiento similar al caso de NEUTx. Es importante señalar que el valor mínimo del grupo de tratamiento está alejado de los demás valores, algo que se muestra comparando el primer cuartil de los dos grupos.

Grupo	Estadístico W	p-valor
Control	0.94799	0.1264
Tratamiento	0.96877	0.4661

Tabla 5.2.20 Resultados de la prueba de Shapiro para la variable PAI.I

Realizando la prueba de Shapiro en ambos grupos, vemos que siguen una distribución normal, aunque no se tiene mucha evidencia para el grupo de control de manera similar al caso anterior, conforme a los resultados de la tabla 5.2.20.

Media grupo control	Media grupo tratamiento	Estadístico *t*	P - valor
14.13438	14.73939	-2.9844	0.004052

Tabla 5.2.21 Resultados de la prueba *t* para la variable PAI.I

Los resultados de las pruebas *t* en la Tabla 5.2.21 muestran que la diferencia entre las medias de los grupos de tratamiento y control son distintas a 0. La similitud en las distribuciones observada en el histograma se ve reflejada en la similitud de los resultados entre las pruebas.

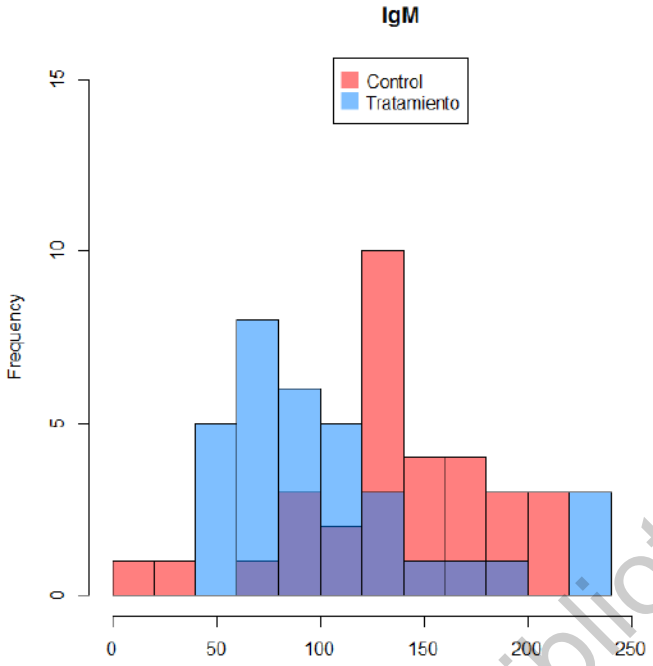

Figura 5.2.11 Histograma de la variable IgM

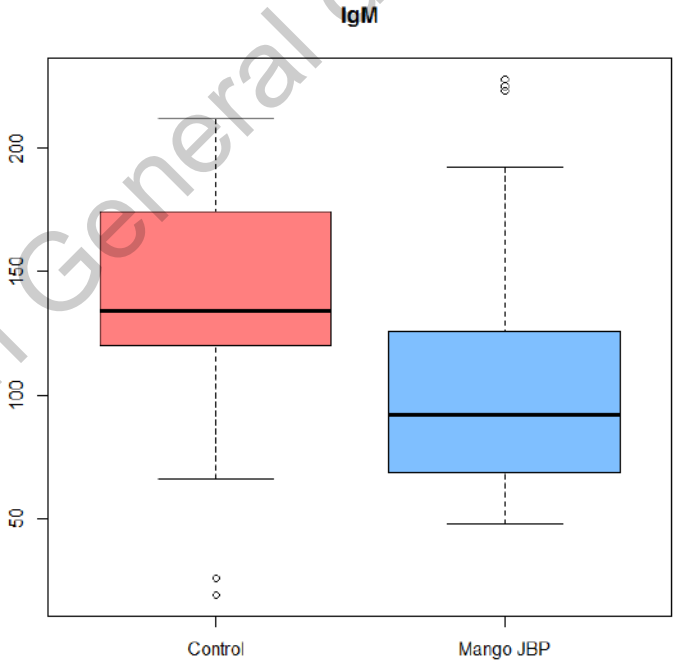

Figura 5.2.12 Boxplot de la variable IgM, por grupo

En el caso de la variable IgM, observamos que el grupo de tratamiento se concentra en valores ligeramente menores a los valores en los que se concentra el grupo de control, como se muestra en las Figuras 5.2.11 y 5.2.12. No parece que los grupos tengan una distribución simétrica, pero se aprecia fácilmente la separación de los grupos. El tratamiento parece disminuir los valores de la variable IgM. Se presentan valores alejados en ambos grupos.

Min	1er cuartil	Mediana	Media	3er cuartil	Max
19	121.5	134	136.6	173.8	212

Tabla 5.2.22 Estadísticos de la variable IgM para el grupo de control

Min	1er cuartil	Mediana	Media	3er cuartil	Max
48	69	92	105.6	126	228

Tabla 5.2.23 Estadísticos de la variable IgM para el grupo de tratamiento

El comportamiento de los valores atípicos se muestra en la diferencia entre el mínimo y el primer cuartil para el caso del grupo de control, mientras que en el grupo de tratamiento se nota entre el 3er cuartil y el máximo. La prueba de Shapiro – Wilk arroja los resultados siguientes:

Grupo	**Estadístico W**	**p-valor**
Control	0.94961	0.1405
Tratamiento	0.85772	0.0005

Figura 5.2.24 Resultados de la prueba de Shapiro para la variable IgM

El grupo de control parece seguir una distribución normal, pero el grupo de tratamiento claramente no tiene dicho comportamiento.

El efecto de las variables antes mencionadas en la separación de los registros se puede visualizar también por medio de un *Mapa de calor*, mostrado en la Figura 5.2.13

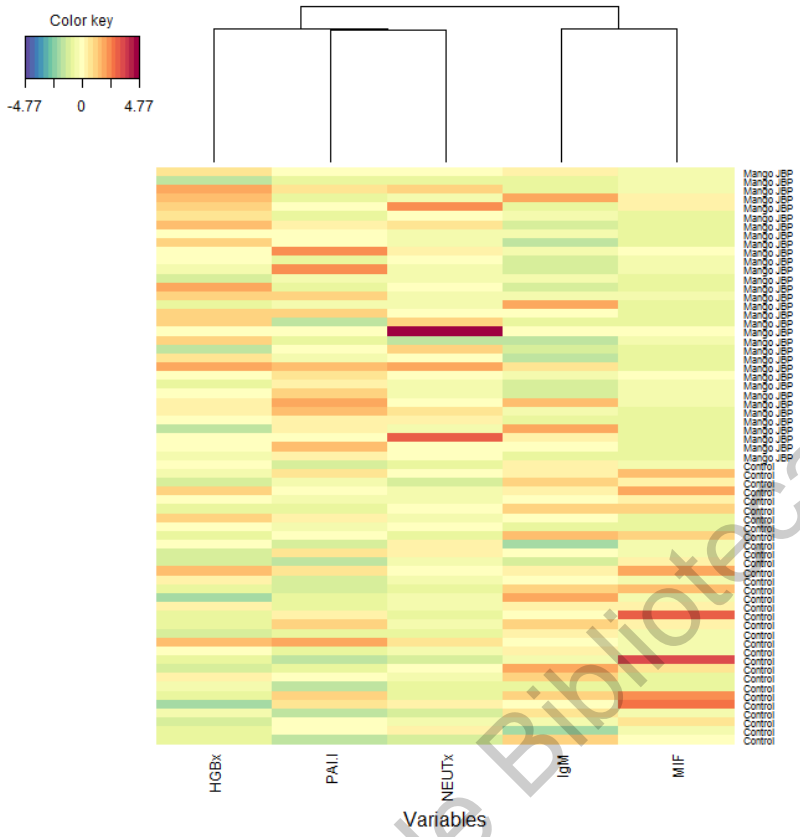

Figura 5.2.13 Mapa de calor con las 5 variables más importantes

En la Figura 5.2.20 se puede apreciar que las variables MIF y NEUTx, que son las dos más importantes para los dos componentes, tienen una separación más clara en cuanto a los registros para cada grupo. Las otras tres variables: HGBx, PAI.I e IgM no muestran el mismo comportamiento con la misma claridad, ya que los registros se distribuyen de manera más uniforme.

Con el análisis realizado con mixOmics, se puede notar que el modelo PLS-DA logra cierta separabilidad con el uso de combinaciones lineales de las variables originales, misma que contrasta con el análisis más general realizado con las variables con mayor coeficiente vip. Estos resultados son útiles en contextos donde se requiere reducir el número de variables necesarias para futuros análisis.

5.2.1 Modelo de regresión logística

Teniendo en cuenta la separación que logran las cinco variables obtenidas con el método PLS – DA, se realizará una modelación con regresión logística, con el fin de determinar cuáles de las variables más importantes pueden ser de utilidad para separar los grupos de control y tratamiento, además de comparar la capacidad de

clasificación de ambos modelos. De los posibles modelos que se pueden formar con las cinco variables, se seleccionará el mejor modelo por medio del método de regresión escalonada (stepwise regression), con el objetivo de obtener el modelo con el menor AIC.

Una de las hipótesis que deben cumplirse para realizar la regresión, es la ausencia de multicolinealidad entre las variables independientes, es decir, dichas variables no están correlacionadas linealmente entre sí. El gráfico de dispersiones mostrado en la Figura 5.2.1.1 para las cinco variables más importantes permitirá identificar posibles correlaciones entre las variables:

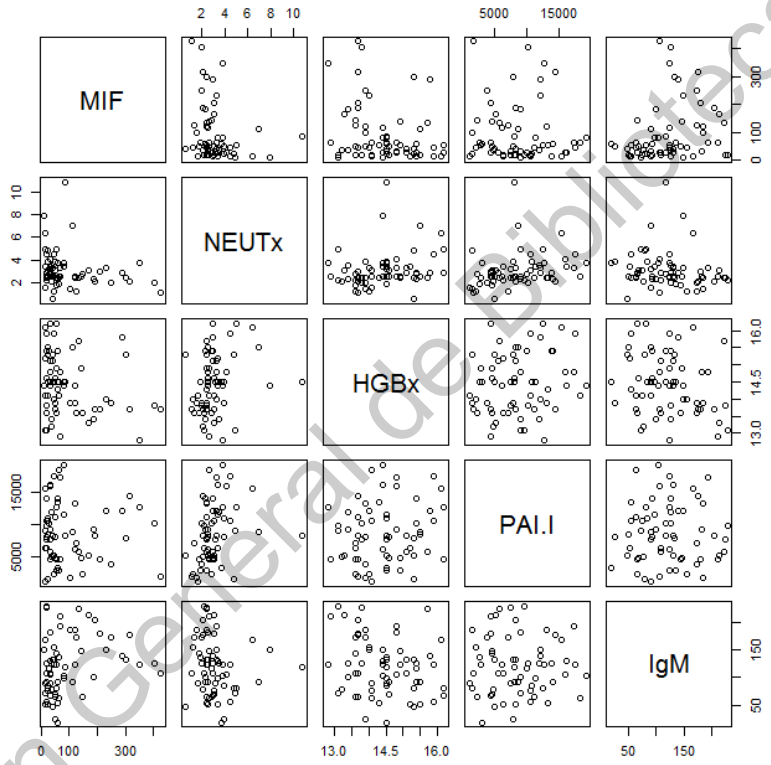

Figura 5.2.1.1 Gráfico de dispersión para las 5 variables más importantes

En los gráficos de dispersiones de la Figura 5.2.1.1 no se alcanza a apreciar una correlación lineal evidente entre las variables consideradas. La siguiente matriz de correlaciones muestra el nivel de correlación lineal de dichas variables:

	MIF	NEUTx	HGBx	PAI.I	IgM
MIF	1	- 0.208	- 0.2658	- 0.0384	0.2584
NEUTx	- 0.208	1	0.2654	0.2744	- 0.1389

HGBx	- 0.2658	0.2654	1	0.1756	- 0.251
PAI.I	- 0.0384	0.2744	0.1756	1	0.0007
IgM	0.2584	- 0.1389	- 0.251	0.0007	1

Tabla 5.2.1.1 Matriz de correlaciones de MIF, MEUTx, HGBx, PAI.I, IgM

Con la matriz de correlaciones de la Tabla 5.2.1.1 vemos que no hay multicolinealidad clara entre las variables. Con esto, se pueden utilizar estas variables en un modelo de regresión logística. La codificación de la variable de respuesta es 1 para el grupo de tratamiento y 0 para el grupo de control.

Con el método de regresión escalonada se obtendrán distintos modelos posibles. Para seleccionar el mejor modelo, usaremos como parámetro a comparar el Criterio de Información de Akaike (AIC),

Respecto a los datos, para poder comparar el desempeño entre los modelos propuestos, se separará la base en datos de entrenamiento y datos de validación. Dicha separación se hará en una proporción de tres a uno para los datos de entrenamiento, es decir, el 75% de los datos conformarán la base de datos con la que se realizarán las regresiones escalonadas necesarias.

Utilizando la regresión escalonada hacia adelante obtenemos los resultados mostrados en la Tabla 5.2.1.2.

Coeficientes	Estimaciones	AIC
Intercepto	-13.99	45.37
MIF	-0.0243	
HGBx	1.012	
PAI.I	0.0001	

Tabla 5.2.1.2 Resultado de la regresión escalonada hacia adelante

Que corresponde al modelo:

$$Grupo = \beta_0 + \beta_1 MIF + \beta_2 HGBx + \beta_3 PAI.I$$

Al utilizar la regresión escalonada hacia atrás, se obtienen los siguientes resultados (Tabla 5.2.1.3):

Coeficientes	Estimaciones	AIC
Intercepto	-13.99	45.37
MIF	-0.0243	
HGBx	1.012	
PAI.I	0.0001	

Tabla 5.2.1.3 Resultado de la regresión escalonada hacia atrás

El cual es el mismo modelo obtenido con la regresión escalonada hacia adelante, con los mismos coeficientes para cada variable. Al utilizar la misma técnica con ambas direcciones (Tabla 5.2.1.4):

Coeficientes	Estimaciones	AIC
Intercepto	-13.99	45.37
MIF	-0.0243	
HGBx	1.012	
PAI.I	0.0001	

Tabla 5.2.1.4 Resultado de la regresión escalonada con ambas direcciones

Por tanto, la regresión escalonada, en cualquiera de sus variantes, indica que el mejor modelo (utilizando el Criterio de Información de Akaike) es:

$$Grupo = -13.99 - 0.0243 * MIF + 1.012 * HGBx + 0.0001 * PAI.I$$

Para la evaluación del modelo, se utilizan los 17 registros de prueba, cuyos resultados de clasificación se muestran en las Tablas 5.2.1.5 y 5.2.1.6

Resultados del modelo	
Precisión	0.8235
Intervalo Confianza 95%	(0.5657, 0.962)
P – valor	0.0121
Kappa	0.6383
Sensibilidad	1
Especificidad	0.625
Valor predictivo positivo	0.75

Valor predictivo negativo	1
Prevalencia	0.5294
Tasa de detección	0.5294
Prevalencia de detección	0.7059
Precisión balanceada	0.8125

Tabla 5.2.1.5 Resultados del modelo de regresión logística

Matriz de confusión del modelo			
$n = 17$		Real	
		Control	Tratamiento
Predicción	Control	5	0
	Tratamiento	3	9

Tabla 5.2.1.6 Matriz de confusión de los datos de validación

La clase positiva que reporta R es la clase de tratamiento. De los 8 registros en el grupo de control, tres fueron mal clasificados como pertenecientes al grupo de tratamiento. Por otro lado, todos los nueve registros del grupo de tratamiento fueron clasificados correctamente. El modelo obtenido muestra una buena capacidad de identificar a niños que reciben el tratamiento con las tres variables especificadas, aunque se muestra con dificultades al identificar correctamente a niños que no han consumido el tratamiento.

La Tabla 5.2.1.7 contiene los estadísticos que se emplean para comprobar el desempeño del modelo de regresión logística, mismos que se emplean para la definición de los valores reportados en la Tabla 5.2.1.5

Parámetros de resultados en clasificación	
Número de casos positivos	$P = 9$
Número de casos negativos	$N = 8$
Predicciones positivas correctas	$TP = 9$
Predicciones negativas correctas	$TN = 5$
Predicciones positivas incorrectas	$FP = 3$
Predicciones negativas incorrectas	$FN = 0$

Tabla 5.2.1.7 Parámetros de resultados de un modelo de clasificación y su valor en la regresión logística ajustada

De los resultados del modelo en la Tabla 5.2.1.5, el primer valor reportado, la precisión, es la proporción de predicciones correctas respecto al número total de registros:

$$Precisión = \frac{TP + TN}{P + N}.$$

El intervalo *95% CI* es el intervalo de confianza al 95% para la precisión, calculado usando el método por defecto del intervalo de confianza binomial (Kuhn, 2008).

El valor de *Kappa* mide el nivel de concordancia entre las clasificaciones y los valores reales. Un valor de Kappa de 1 representa concordancia perfecta, mientras que un valor de 0 representa no concordancia. Se calcula como sigue:

$$Kappa = \frac{Precisión - Precisión\ aleatoria}{1 - Precisión\ aleatoria},$$

donde

$$Precisión\ aleatoria = \frac{(TN + FP) * (TN + FN) + (FN + TP) * (FP + TP)}{(P + N)^2}.$$

La sensibilidad es la proporción de predicciones positivas con respecto al número total de positivos. Se calcula como

$$Sensibilidad = \frac{TP}{P}.$$

Se puede interpretar también como la probabilidad de obtener una predicción positiva de una observación que realmente es positiva. En el contexto de los datos que se tienen, representa la capacidad del modelo de detectar correctamente a los niños que reciben el tratamiento

La especificidad es la proporción de predicciones negativas con respecto al total de casos negativos:

$$Especificidad = \frac{TN}{N}.$$

Se refiere a la capacidad del modelo de detectar correctamente casos negativos; en este caso, la capacidad de detectar niños que no recibieron el tratamiento. Con los resultados mostrados en la Figura 5.2.1.5, se aprecia que el modelo no es muy

efectivo en detectar a niños que obtienen el tratamiento, mientras que es más efectivo en detectar a niños que no lo siguen.

El valor predictivo positivo es la proporción de predicciones positivas correctas con respecto al total de predicciones positivas, y toma valores entre 0 y 1. En otras palabras, la proporción de predicciones correctas para la clase positiva (con tratamiento). Se calcula como

$$Valor\ predictivo\ positivo = \frac{TP}{TP + FP}.$$

El valor predictivo negativo es la proporción de predicciones negativas correctas con respecto al número total de predicciones negativas, cuyo valor está entre 0 y 1. Se calcula como

$$Valor\ predictivo\ negativo = \frac{TN}{TN + FN}.$$

La prevalencia es la proporción de registros positivos del total de registros:

$$Prevalencia = \frac{P}{P + N}.$$

La tasa de detección es la proporción de predicciones positivas correctas del número total de registros:

$$Tasa\ de\ detección = \frac{TP}{P + N}.$$

La prevalencia de detección es la proporción de predicciones positivas con respecto al total de registros:

$$Prevalencia\ de\ detección = \frac{TP + FP}{P + N}.$$

Por último, la precisión balanceada se calcula como el promedio entre la sensibilidad y especificidad:

$$Precisión\ balanceada = \frac{Sensitivity + Specificity}{2}.$$

La distribución de los residuales, considerando los datos de entrenamiento, para el modelo se muestra en la Figura 5.2.1.2.

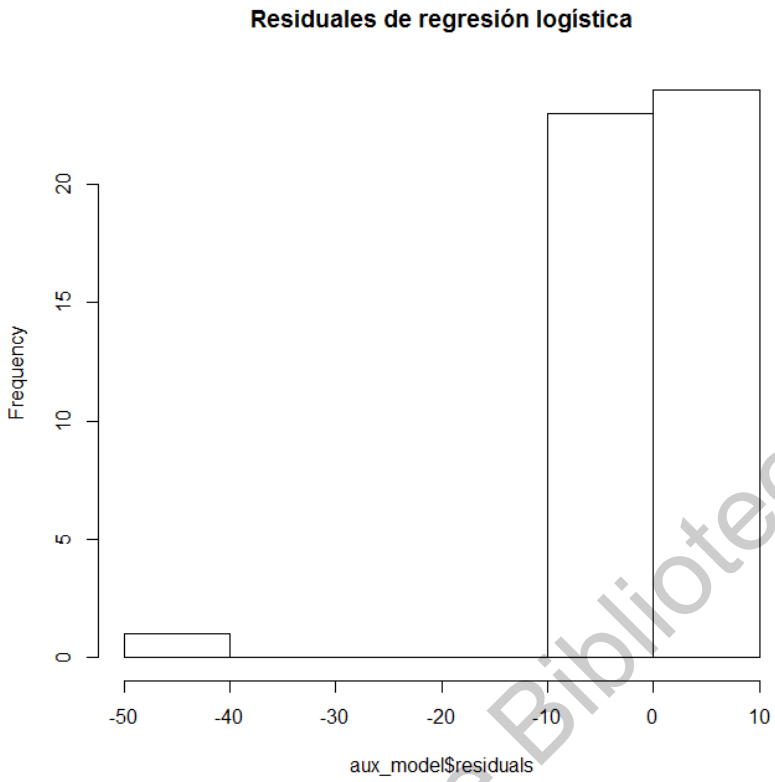

Figura 5.2.1.2 Residuales del modelo logístico

La mayoría de los residuales se encuentran alrededor de cero, pero se puede identificar un valor atípico muy por debajo de los demás. La gráfica de residuales sugiere que se puede ajustar un mejor modelo. Para el modelo PLS con mixOmics, la misma observación se muestra en la Figura 5.2.1.3.

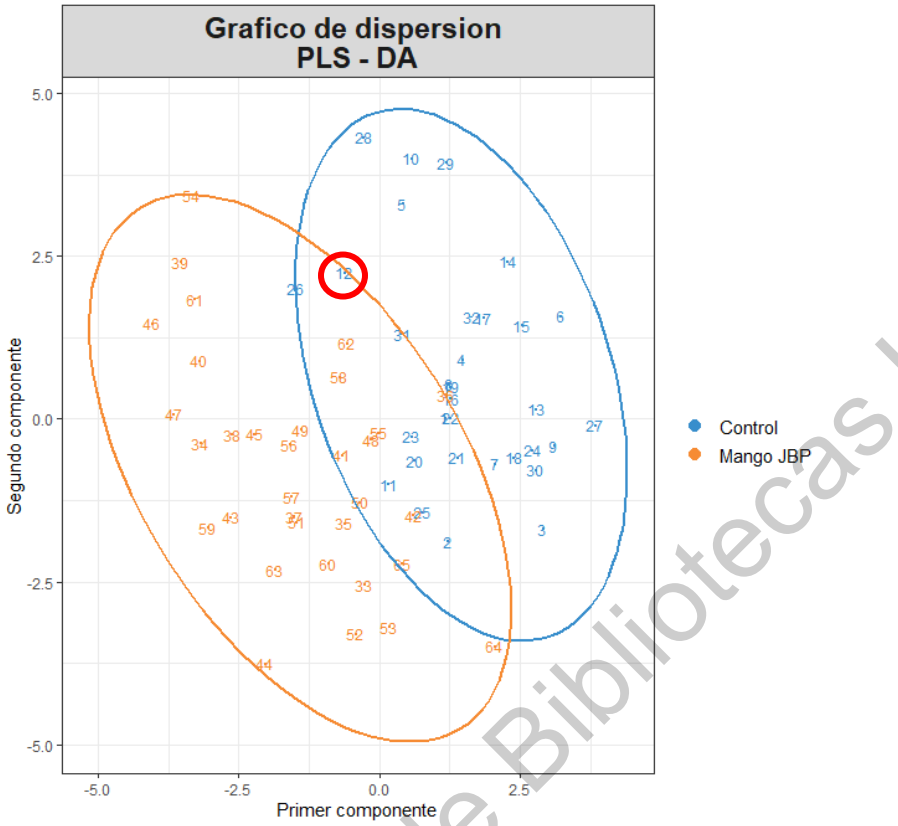

Figura 5.2.1.3 Gráfico de dispersión del modelo PLS de mixOmics. En rojo, la observación atípica del modelo de regresión logística

El dato atípico en la regresión logística no parece ser un atípico con los primeros dos componentes del modelo PLS. Se repetirá el modelo de regresión logística omitiendo este valor encontrado para investigar si tiene un efecto importante sobre los resultados obtenidos.

La regresión escalonada sugiere los siguientes modelos:

Regresión escalonada hacia adelante (Tabla 5.2.1.8)

Coeficientes	Estimaciones	AIC
Intercepto	-27.74	34.6
MIF	-0.0333	
HGBx	1.915	
PAI.I	0.0003	

Tabla 5.2.1.8 Resultado de la regresión escalonada hacia adelante, omitiendo el registro atípico

Regresión escalonada hacia atrás (Tabla 5.2.1.9)

Coeficientes	Estimaciones	AIC
Intercepto	-27.74	34.6
MIF	-0.0333	
HGBx	1.915	
PAI.I	0.0003	

Tabla 5.2.1.9 Resultado de la regresión escalonada hacia atrás, omitiendo el registro atípico

En ambas direcciones (Tabla 5.2.1.10)

Coeficientes	Estimaciones	AIC
Intercepto	-27.74	34.6
MIF	-0.0333	
HGBx	1.915	
PAI.I	0.0003	

Tabla 5.2.1.10 Resultado de la regresión escalonada en ambas direcciones, omitiendo el registro atípico

Los resultados sugieren utilizar las mismas variables para el modelo, y los resultados de la clasificación con los datos de prueba se muestran en las Tablas 5.2.1.11 y 5.2.1.12.

Resultados del modelo	
Precisión	0.8824
Intervalo Confianza 95%	(0.6356, 0.9854)
P – valor	0.0024
Kappa	0.7606
Sensibilidad	1
Especificidad	0.75
Valor predictivo positivo	0.8182
Valor predictivo negativo	1

Prevalencia	0.5294
Tasa de detección	0.5294
Prevalencia de detección	0.6471
Precisión balanceada	0.875

Tabla 5.2.1.11 Resultados del modelo de regresión logística, omitiendo el registro atípico

Matriz de confusión del modelo			
$n = 17$		Real	
		Control	Tratamiento
Predicción	Control	6	0
	Tratamiento	2	9

Tabla 5.2.1.12 Matriz de confusión de los datos de validación, omitiendo el registro atípico

Los resultados del nuevo modelo muestran que tiene una ligera mejora en la clasificación del grupo de control y con la correcta clasificación del grupo de tratamiento. Los residuales del nuevo modelo se distribuyen como sigue:

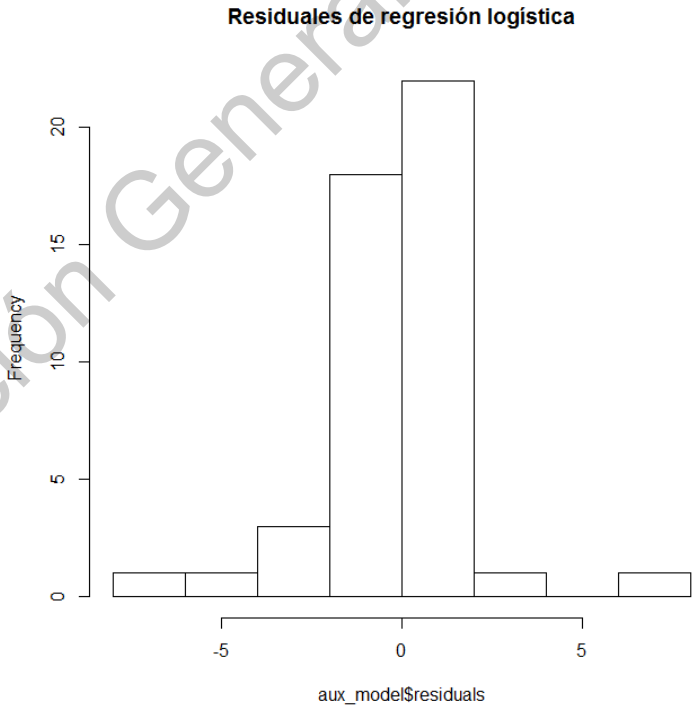

Figura 5.2.1.4 Histograma de residuales del modelo logístico sin el valor atípico

El histograma de la Figura 5.2.1.4 muestra que los residuales ahora tienen una distribución similar a una normal centrada en cero. Considerando la ligera mejora en el desempeño de clasificación, él valor atípico que se ha omitido muestra tener influencia en el comportamiento del modelo logístico. Nótese que el resultado de la regresión escalonada es la misma para las dos situaciones.

5.3 ropls

Como primer paso, se repetirán los análisis del tratamiento con el subproducto de mango para comparar la información proporcionada por este paquete y los resultados con los demás. El resultado del análisis se muestra en la Tabla 5.3.1.

R2X	R2Y	Q2	RMSEE	pR2Y	pQ2
0.0953	0.517	0.277	0.353	0.1	0.05

Tabla 5.3.1 Resultados del modelo PLS-DA con ropls

La consola de R también arroja la siguiente advertencia, dado que el modelo obtenido es de solamente un componente:

```
Warning message:

Single component model: only 'overview' and 'permutation' (in case of single
response (O)PLS(-DA)) plots available
```

El resumen que proporciona la función opls nos incluye los valores $R^2_{X_{cum}} = 0.0953$ (Varianza explicada acumulada en X), $R^2_{Y_{cum}} = 0.517$ (Varianza explicada acumulada en Y), $Q^2_{cum} = 0.277$ (Varianza predicha acumulada) y la raíz del error cuadrado medio $RMSEE = 0.353$.

El número de componentes a utilizar en el modelo se calcula automáticamente por medio de validación cruzada dentro de la función opls (en el caso de que el parámetro predI = NA, que es la opción por defecto). El criterio para decidir si un nuevo componente h se añade al modelo es el siguiente:

1. $R^2_{Y_h} \geq 1\%$
2. $Q^2_{Y_h} = 1 - \frac{PRESS_h}{RSS_{h-1}} \geq 0$ para PLS (O 5% cuando el número de muestras es menor a 100) o 1% para OPLS. La condición $Q^2_{Y_h} \geq 0$ significa que la suma de cuadrados residuales predichos ($PRESS_h$) del modelo con el nuevo componente estimado por medio de validación cruzada de 7 repeticiones es

menor que la suma de cuadrados residuales del modelo con los componentes anteriores (RSS_0 es la suma de los Y valores al cuadrado).

El valor $Q^2_{cum} = 1 - \prod_{h=1}^{r}(1 - Q^2_Y)$ es el desempeño en la predicción del modelo. Se tiene que $Q^2_{Y_{cum}} \in [0,1]$, y mientras más grande sea el valor, mejor el desempeño.

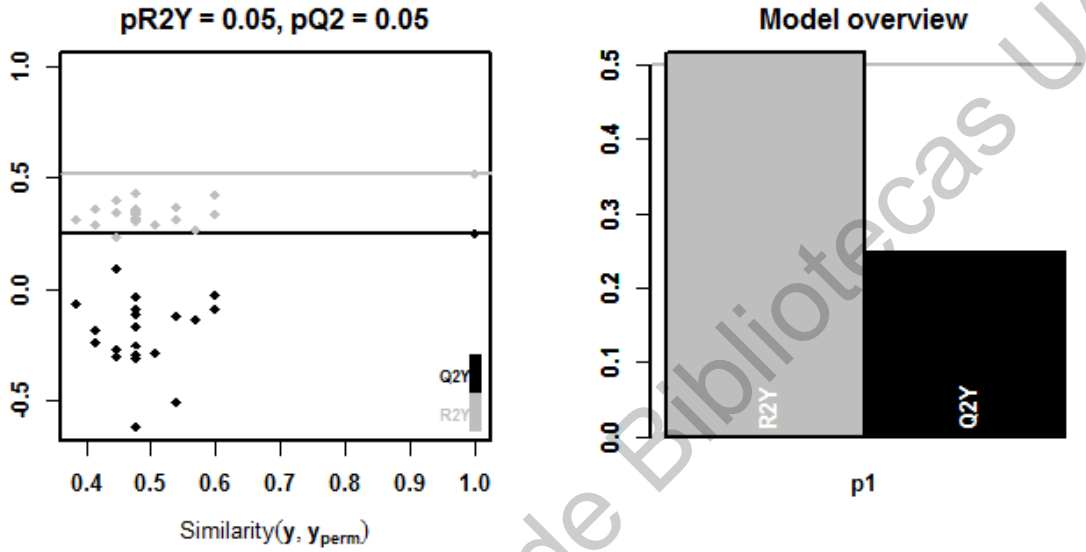

Figura 5.3.1 Gráficos del modelo PLS-DA. Izquierda: diagnóstico de significancia. Los valores R^2_Y y Q^2_Y se comparan con los valores correspondientes obtenidos al permutar de manera aleatoria la respuesta Y. Derecha: Gráfico de inercia.

El gráfico de la izquierda en la Figura 5.3.1 es el diagnóstico de *significancia*: los valores R^2 y Q^2 del modelo se comparan con los valores correspondientes obtenidos después de permutaciones aleatorias de las respuestas Y. Este método de inferencia permite obtener un cálculo de significancia alternativo, que permite no depender de la suposición de normalidad de los grupos. El test de permutación genera grupos (en este caso, individuos con tratamiento e individuos sin tratamiento) de manera aleatoria bajo el supuesto de nulidad de efecto del tratamiento, obteniendo para cada par de grupos un estadístico de significancia (Efron, Hastie, 2016). Los p valores que se muestran en la parte superior del gráfico indican la proporción de veces en las que, habiendo permutado los valores de Y (Y_{perm}) se obtienen valores más grandes para $Q^2_{Y_{perm}}$ que los reportados originalmente. La importancia de estos análisis es identificar si los valores de Q^2_Y que se obtuvieron fueron obtenidos por azar, y evitar el overfitting. Por default, el

número de permutaciones que se realizan son 20, pero se pueden modificar con el parámetro **permI**.

El gráfico de la derecha en la Figura 5.3.1 es un vistazo general de los valores de R^2 y Q^2 obtenidos por el modelo con distintos componentes. Dado que en este ejemplo se tiene 1 como número óptimo de componentes, no se puede apreciar la diferencia entre componentes. Los resultados gráficos de opls incluyen también un gráfico para diagnóstico de residuales y un gráfico de los *scores* del modelo.

El paquete opls también proporciona los coeficientes vip, con la función **getVipVn**. Utilizando esta función, las 5 variables más importantes se muestran en la Tabla 5.3.2.

Variable	Coeficiente vip
MIF	2.65659402
NEUTx	1.82630858
HGBx	1.80172185
G.CSF	1.70369574
PAI.I	1.67668747

Tabla 5.3.2 Coeficientes VIP del primer componente

De las 5 variables más importantes, 4 de ellas son también consideradas en el paquete mixOmics, con la adición de la variable G.CSF. Con excepción de PAI.I, los valores VIP de las demás variables en común entre los paquetes son muy similares, y las 3 variables más importantes conservan el orden de importancia.

6 Conclusiones

Con el análisis del tratamiento con el subproducto de mango, se logró obtener información de variables que tienen capacidad de discriminar a los sujetos de estudio respecto a la presencia o ausencia del tratamiento, así como ofrecer pistas respecto al efecto que produce el mismo. En este punto, se muestra en los resultados de los modelos que la variable proteínica MIF es la que muestra mayor cambio con el tratamiento de mango, y el efecto en las variables más importantes no se concentra en un solo tipo de variable, mostrando que variables hematológicas también tienen cambios con el producto.

Los resultados obtenidos por los distintos paquetes que incorporan la modelación PLS – DA muestran que esta técnica es útil para encontrar una combinación lineal

de variables que logre una buena separación entre los grupos considerados. Es importante también resaltar que el uso de esta técnica debe ser parte de una serie de pasos en el proceso de investigación, y no se debe confiar demasiado en los resultados que se puedan obtener sin realizar otros análisis. Entre los paquetes considerados, no se aprecian diferencias importantes respecto a la información entregada, salvo en la facilidad de instalación, ya que mixOmics no se encuentra en el repositorio CRAN de R, por lo que su instalación requiere un par de pasos extra, mientras que MetaboAnalystR requiere de requisitos adicionales que pueden complicar un poco su instalación.

Las relaciones de PLS con otras técnicas multivariadas como PCA, FDA y CCA en el contexto de clasificación es un ejemplo de cómo surgen diferentes tipos de análisis a partir de diferentes problemas de eigenvalores y el uso de la descomposición espectral como método común para su resolución, y describen tanto el proceso algebraico como el objetivo de optimización que se lleva a cabo. PLS – DA sirve como un punto intermedio entre FDA y PCA, pues ofrece tanto un análisis exploratorio de datos con el objetivo de la reducción de dimensionalidad en unas pocas variables latentes como también un modelo de clasificación con capacidad predictiva para variables de interés, lo que permite afrontar mejor las situaciones de alta dimensionalidad, por ejemplo, en el estudio de datos ómicos.

En esta tesis se ha discutido el uso de PLS – DA como método para encontrar una estructura entre variables independientes y variables dependientes cuando el número de variables supera el número de muestras. Las estimaciones de las variables latentes consideran combinaciones lineales de todas las variables originales. Para un trabajo posterior, se puede estudiar una variante del estudio de la técnica PLS considerando un término de regularización para forzar que algunos coeficientes de dichas combinaciones lineales sean cero. Esto puede ser útil en evitar el sobreajuste de un modelo y obtener resultados que capturen mejor las relaciones entre las matrices de datos. En un aspecto personal, el estudio de estas técnicas durante la elaboración del presente trabajo me ha permitido ampliar el abanico de posibilidades al momento de enfrentar un estudio multivariado, tanto en el ámbito académico como en el laboral.

Distintos métodos de regularización para PLS y PLS – DA se pueden encontrar en el paquete **plsVarSel**, con los que se pueden obtener resultados más eficientes en casos donde $p \gg n$, como observa Mehmood et. al. (2020). Entre los beneficios de los métodos de regularización se encuentra una mayor interpretabilidad de los modelos y un comportamiento más estable de los mismos.

7 Bibliografía

Alsberg, B., Kell, D., & Goodacre, R. (1998). Variable Selection in Discriminant Partial Least-Squares Analysis. *Analytical Chemistry*, *70*(19), 4126-4133. doi: 10.1021/ac980506o`

Anaya-Loyola, M. A., García-Marín, G., García-Gutiérrez, D. G., Castaño-Tostado, E., Reynoso-Camacho, R., López-Ramos, J. E., … Pérez-Ramírez, I. F. (2020). A mango (Mangifera indica L.) juice by-product reduces gastrointestinal and upper respiratory tract infection symptoms in children. Food Research International, 136, 109492. https://doi.org/10.1016/j.foodres.2020.109492

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal Of Chemometrics*, *17*(3), 166-173. doi: 10.1002/cem.785

Bartlett, M. (1938). Further aspects of the theory of multiple regression. Mathematical Proceedings of the Cambridge Philosophical Society, 34(1), 33-40. doi:10.1017/S0305004100019897

Brereton, R., & Lloyd, G. (2014). Partial least squares discriminant analysis: taking the magic away. *Journal Of Chemometrics*, *28*(4), 213-225. doi: 10.1002/cem.2609

Chiang, L. H., Russell, E. L., & Braatz, R. D. (2000). Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis. Chemometrics and Intelligent Laboratory Systems, 50(2), 243-252. https://doi.org/10.1016/s0169-7439(99)00061-1

Chun, H., & Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal Of The Royal Statistical Society: Series B (Statistical Methodology)*, *72*(1), 3-25. doi: 10.1111/j.1467-9868.2009.00723.x

Efron, B., & Hastie, T. (2016). Computer Age Statistical Inference: Algorithms, Evidence, and Data Science (Institute of Mathematical Statistics Monographs). Cambridge: Cambridge University Press. doi:10.1017/CBO9781316576533

Eriksson, L., Trygg, J., & Wold, S. (2014). A chemometrics toolbox based on projections and latent variables. *Journal Of Chemometrics*, *28*(5), 332-346. doi: 10.1002/cem.2581

Frank, I., Friedman, J. (1993), A statistical view of some chemometrics regression tools. *Technometrics*, 35: 109-148.

Geladi, P. and Kowalski, B.P. (1986) Partial least-squares regression: A tutorial. Analytica Chimica Acta, 185(1), 1-17.

Gromski, P., Muhamadali, H., Ellis, D., Xu, Y., Correa, E., Turner, M., & Goodacre, R. (2015). A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta*, *879*, 10-23. doi: 10.1016/j.aca.2015.02.012

Haenlein, M., & Kaplan, A. (2004). A Beginner's Guide to Partial Least Squares Analysis. *Understanding Statistics*, *3*(4), 283-297. doi: 10.1207/s15328031us0304_4

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). An Introduction To Statistical Learning. 1st ed. Springer.

Johnstone, I., & Titterington, D. (2009). Statistical challenges of high-dimensional data. *Philosophical Transactions Of The Royal Society A: Mathematical, Physical And Engineering Sciences*, *367*(1906), 4237-4253. doi: 10.1098/rsta.2009.0159

Lee, L., Liong, C., & Jemain, A. (2018). Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. *The Analyst*, *143*(15), 3526-3539. doi: 10.1039/c8an00599k

Liu, Y., & Rayens, W. (2007). PLS and dimension reduction for classification. *Computational Statistics*, *22*(2), 189–208. https://doi.org/10.1007/s00180-007-0039-y

Mateos-Aparicio, G. (2011). Partial Least Squares (PLS) Methods: Origins, Evolution, and Application to Social Sciences. *Communications In Statistics - Theory And Methods*, *40*(13), 2305-2317. doi: 10.1080/03610921003778225

Mehmood, T., Liland, K., Snipen, L. and Sæbø, S., 2020. A Review Of Variable Selection Methods In Partial Least Squares Regression.

Pérez-Enciso, M. & Tenenhaus, M. (2003). Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Hum Genet, 112,* 581-592. https://doi.org/10.1007/s00439-003-0921-9

Rohart, F., Gautier, B., Singh, A., & Lê Cao, K. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Computational Biology*, *13*(11), e1005752. doi: 10.1371/journal.pcbi.1005752

Rosipal R., Krämer N. (2006) Overview and Recent Advances in Partial Least Squares. In: Saunders C., Grobelnik M., Gunn S., Shawe-Taylor J. (eds) Subspace, Latent Structure and Feature Selection. SLSFS 2005. Lecture Notes in Computer Science, vol 3940. Springer, Berlin, Heidelberg

Thévenot, E. A. (2016). ropls: PCA, PLS (-DA) and OPLS (-DA) for multivariate analysis and feature selection of omics data.

Wold, H. (1973b). Nonlinear iterative partial least squares (NIPALS) modeling: some current developments. In: Krishnaiah, P. R., ed. Multivariate Analysis II. Proc. Int. Symp. Multivariate Anal. held at Wright State University, Dayton, Ohio, June 19–24, 1972. New York: Academic Press, pp. 383–407.

Wold, H. (1975). Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. Perspectives in Probability and Statistics, In Honor of MS Bartlett, 117–144.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics And Intelligent Laboratory Systems*, *58*(2), 109-130. doi: 10.1016/s0169-7439(01)00155-1

Xia, J., Psychogios, N., Young, N. and Wishart, D. (2009). MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Research, 37(Web Server), pp. W652-W660.