Universidad de Valladolid

Facultad de Ciencias

Trabajo Fin de Grado

Grado en Estadística

Aplicaciones del método PLS en Quimiometría

Autor: Laura Vaquer Rivera

Tutor: Luis Ángel García Escudero

RESUMEN

El método de Partial Least Squares (PLS) es un método de análisis multivariante muy aplicado al analizar conjuntos de datos en dimensión alta. La Quimiometría es uno de los campos donde esta metodología es claramente de gran aplicabilidad debido a la gran abundancia de datos generada por la instrumentalización analítica. Con este método podemos crear procedimientos matemáticos que son capaces de predecir valores que no son directamente calculables. El propósito de este TFG es revisar, tanto a nivel metodológico como de software, la aplicación de la técnica PLS en este entorno comprobando su aplicabilidad al compararlo con otros métodos multivariantes.

ABSTRACT

The Partial Least Squares (PLS) method is a multivariate analysis method widely applied when analyzing high-dimensional data sets. Chemometrics is one of the fields where this methodology is clearly highly applicable due to the flood of data generated by analytical instrumentation. With this method we can create mathematical procedures that are capable of predicting values that are not directly calculable. The purpose of this TFG is to review, both at a methodological and software level, the application of the PLS technique in this environment, checking its applicability when comparing it with other multivariate methods.

Índice general

Capítulo 1

Introducción

1.1 Quimiometría

La Quimiometría es una ciencia orientada a extraer información de experimentos químicos mediante la aplicación de aproximaciones cuantitativas/estadísticas. Se pretende que la toma de decisiones en Química se base mayoritariamente en criterios dependientes del análisis e interpretación de datos ("data-driven") y no en aspectos subjetivos y que no sean perfectamente cuantificables. La Quimiometría proporciona herramientas y técnicas a ser tenidas en cuenta en la planificación de los experimentos químicos y en el análisis de los resultados obtenidos en los mismos.

En 1975, la ICS (International Chemometrics Society) definió la Quimiometría como "…la disciplina química que utiliza métodos matemáticos y estadísticos para diseñar o seleccionar procedimientos de medida y experimentos óptimos, y para proporcionar la máxima información química mediante el análisis de datos químicos". El nombre de Quimiometría fue dado por Svante Wold en 1971, quien definió esta área como "el arte de extraer la información relevante químicamente a partir de datos producidos en los experimentos químicos". S. Wold desarrolló también el método PLS del que trataremos en este TFG.

La Quimiometría ayuda a resolver complejos problemas usando técnicas tanto descriptivas como predictivas. Es razonable pensar que si somos capaces de *describir* y modelizar las características de los sistemas químicos conoceremos también mejor las relaciones y estructura de dichos sistemas. Por otro lado, las técnicas predictivas tienen como finalidad *predecir* nuevas propiedades o comportamientos de interés que no son fácilmente determinables de forma directa.

La Quimiometría se aplica de manera rutinaria en la Química y algunos de los problemas típicos que ha conseguido solucionar esta ciencia son:

- Reconocimiento de ciertas propiedades o actividad de cierto componente químico.
- Clasificación de las muestras, determinación de la concentración de un compuesto en una mixtura compleja.
- Predicción de una propiedad o actividad de un componente químico.
- Evaluación del estado de un proceso.

Aunque no se discute que incluso los primeros experimentos analíticos químicos ya se podían considerar que formaban parte del campo de la Quimiometría, puesto que raramente se solía omitir una breve descripción numérica de los resultados obtenidos en los mismos, su inicio más formal se reconoce con la llegada de los primeros ordenadores para uso científico en los años 70. De hecho, la Quimiometría podría considerarse también parte de un campo más amplio como es la Quimioinformática.

Existen tres revistas dedicadas a esta ciencia activas desde los años 80 hasta el día de hoy y que continúan cubriendo los avances metodológicos más significativos: *Journal of*

Chemometrics, *Chemometrics and Intelligent Laboratory Systems* y *Journal of Chemical Information and Modeling*.

El libro "Chemometrics: A textbook", publicado por D. L. Maasart en 1988 fue durante mucho tiempo considerado la biblia para las personas dedicadas a este campo.

Una característica de los datos recogidos y analizados en Química es el hecho de que pueden ser complejos puesto que suelen incluir un número bastante grande de variables a analizar. De hecho, la mayoría de las características medidas son multivariantes por naturaleza, como pueden ser los datos de espectroscopía infrarroja y ultravioleta (vistas como medidas de interacción de la radiación con la materia) los cuales a menudo producen miles de mediciones por observación.

De esta forma, la parte quizás más relevante de la Quimiometría suele ser la aplicación de técnicas de Análisis de Datos Multivariantes. La estructura multivariante de este tipo de datos nos conduce a utilizar, por ejemplo, técnicas como el Análisis de Componentes Principales (PCA) o los Mínimos Cuadrados Parciales (PLS). Esto es porque los datos son de alta dimensionalidad y porque existe una estructura de dependencia fuerte entre variables y, a menudo, relaciones lineales subyacentes bastante fuertes en estas variables. Se ha visto que los métodos PCA y PLS son muy efectivos para modelar empíricamente datos químicos explotando dichas relaciones internas entre variables y proporcionando herramientas alternativas de análisis al ser combinadas con otras técnicas estadísticas como puede ser la Regresión, el Análisis Cluster o Análisis Discriminante.

Aunque técnicas de aprendizaje no supervisado, como el Análisis Cluster, son también típicamente aplicadas en Quimiometría, en este TFG nos centraremos exclusivamente en técnicas de aprendizaje supervisado y muy especialmente en el método de Partial Least Squares (PLS). En este campo, dichas técnicas de aprendizaje supervisado son comúnmente denominadas como técnicas de "calibración".

1.2 Calibración

Como ya se ha comentado, una de las tareas fundamentales en Quimiometría es modelar una variable respuesta partiendo de una o varias variables predictoras, que resulten más fácilmente medibles. En este trabajo, a las variables respuesta las denominamos variables-y y a las variables predictoras variables-x. En ocasiones, la relación entre estas variables es determinista y bien conocida, pero en la mayoría de los casos de interés esta relación no la conocemos, por lo que usaremos técnicas multivariantes para derivar modelos predictivos que nos ayuden a comprender la relación subyacente o a poder realizar predicciones razonablemente precisas.

Estos modelos pueden ser lineales o no lineales; se pueden formular mediante una ecuación o a través de algoritmos más complejos como puedan ser las redes neuronales. En este trabajo nos centramos en modelos lineales de la forma:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_j x_j + \cdots + b_m x_m + e$$

donde b_0 es el intercept, los coeficientes b_1, ..., b_m son los coeficientes de regresión, m es el número de variables explicativas o predictoras y e es un término de error aleatorio.

Al estimar los valores de los coeficientes b's desconocidos se podría pensar que la regresión mínimo-cuadrática típica (Ordinary Least Squares, OLS) es la única técnica que

necesitaríamos aplicar bajo hipótesis de linealidad. No obstante, veremos que otros métodos estudiados en este trabajo como PLS (Partial Least Squares) o PCR (Principal Component Regression) van a resultar más convenientes.

En Quimiometría, el OLS es raramente utilizado ya que los conjuntos de datos en este campo, como ya se ha comentado, suelen tener un gran número de variables que a su vez presentan una correlación alta entre ellas y esto crea gran variabilidad o inestabilidad en las estimaciones de los coeficientes de regresión lineal. Este problema es típicamente conocido como el problema de la "multicolinealidad''. Es por ello que los métodos por excelencia a aplicar en este campo de la calibración en Quimiometría, vayan a ser más bien los métodos PLS y PCR.

Utilizar este tipo de métodos en regresión, como PLS o PCR, ofrece una serie de ventajas como es la capacidad de tratar con variables-x muy correladas o conjuntos de datos con más cantidad de variables que de observaciones. Además, PCR y PLS trabajan con componentes (variables latentes) que resultan de combinaciones lineales de las variables originales y que sirven para controlar la complejidad del modelo. Se denomina "scores'' a los valores que toman esas componentes y se denominan"loadings'' a los coeficientes utilizados en las combinaciones lineales y que describen la influencia de las variables. La representación gráfica de los "scores'' y"loadings'' puede resultar muy útil para comprender de forma visual aspectos muy interesantes en nuestros datos químicos.

Se utilizará en este trabajo consistentemente la notación T para los scores y P para los loadings, siendo esta la notación más común en Quimiometría.

A la hora de modelizar disponemos de dos tipos de conjuntos de datos, aquellos en los que se encuentran las variables-x se denominarán datos-x y los datos-y contienen a las variables-y. A su vez, estas variables se encuentran en sus propios espacios muestrales, a los cuales denominaremos como espacio-x y espacio-y. También se distinguen los scores-x y loadings-x de los scores-y y loadings-y, dependiendo de las variables que se usen para obtener dichas componentes y sus pesos.

Para comprobar si el modelo elegido describe correctamente la relación entre las variables-x y las variables-y existen diferentes medidas del error de predicción. Así, se utiliza con frecuencia el error cuadrático SSE definido como

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

donde y_i son los valores observados en la variable respuesta y $\hat{y}_i$ son los valores predichos para el total n de observaciones en nuestro conjunto de datos.

Como es bien sabido, el error SSE es un indicador bastante optimista ("aparente") del error que se comete en el proceso de calibración, ya que está basado directamente en las n observaciones que se han usado para obtener los parámetros que nos permiten obtener las predicciones. Conviene mejor utilizar estimaciones del error más realistas obtenidas usando, por ejemplo, técnicas de validación cruzada o mediante un conjunto de test independiente del que se ha usado para ajustar los parámetros. La validación cruzada o el uso de conjuntos de test independientes nos permite evitar el fenómeno del "sobrejauste'' y ayudan a estimar el error de "generalización'' que se va a incurrir al predecir nuevas observaciones.

En muchas ocasiones, el conjunto de test en el que se prueba la calibración puede resultar bastante diferente al conjunto de datos que se ha utilizado para ajustar el modelo de calibración. Nótese que, incluso, es frecuente que los aparatos de medición puedan ser hasta distintos y los experimentos realizarse en laboratorios completamente diferentes. En estos casos es interesante contar con medidas del error particularizadas a estas nuevas calibraciones. Con ese fin, si contamos con N nuevas observaciones en un entorno diferente nos puede interesar el "sesgo'' particular cometido:

$$\text{sesgo} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)$$

donde y_i son los valores reales observados e $\hat{y}_i$ sus predicciones en esos nuevos N datos. En este caso, es común considerar el error estándar de predicción (Standard Error of Prediction, SEP) definido como

$$\text{SEP} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{y}_i - \text{sesgo}^2)}.$$

Finalmente, es también frecuente considerar en la validación cruzada un esquema de "doble validación cruzada''. Parte de los datos para dicha validación cruzada se usan para determinar algún parámetro de ajuste (tuning) del procedimiento como, por ejemplo, en regresión PLS y PCR pueda ser el número total de componentes que se retienen. Posteriormente, otra parte de los datos para la validación cruzada (independientes) son usados para estudiar la eficacia de la regla obtenida basada en ese parámetro de ajuste (tuning) ya fijado.

1.3 Breve historia del método PLS

El método de Mínimos Cuadrados Parciales o Partial Least Squares (PLS) es la técnica de calibración más utilizada en Quimiometría. Herman Wold desarrolló las primeras herramientas analíticas PLS a mediados de los años 60 en Suecia. Aunque las ideas fueron en principio muy heurísticas, durante los años sucesivos H. Wold desarrolló, junto con su equipo, diferentes técnicas más formales y que dieron lugar a los procedimientos iterativos comúnmente aplicados en PLS. Algunas de las principales aplicaciones durante este desarrollo inicial del método PLS surgieron mayoritariamente del campo de la Economía y de las Ciencias Sociales.

La evolución de estos métodos en los años 80 sufrió un notable impulso de la mano de Svante Wold (hijo de Herman Wold) con la aplicación de los principios del PLS al análisis de datos químicos. Hoy en día este método PLS es utilizado de forma muy rutinaria en Quimiometría. Con el avance informático y de software se ha podido implementar dichos métodos PLS en este entorno, reflejando así el gran potencial de modelización, versatilidad y capacidad analítica de este método.

Capítulo 2

Metodología

2.1 Regresión PCR

Hemos explicado que uno de los problemas que encontramos en Quimiometría es que existe multicolinealidad o variables que tienen una fuerte relación de dependencia, es decir, la correlacción entre las variables es alta. Esta problemática reduce la efectividad del análisis introduciendo una alta inestabilidad en las estimaciones de los parámetros y las predicciones asociadas.

La Regresión en Componentes Principales (Principal Components Regression, PCR) es una técnica muy utilizada para reducir el número de variables regresoras con objetivo de manejar la multicolinealidad, es por ello que sea una técnica tan aplicada en el campo de la Quimiometría.

Para reducir la dimensionalidad este método trabaja con componentes principales resultando en combinaciones lineales de los regresores originales. La regresión en Componentes Principales es pues una combinación del Análisis de Componentes Principales (Principal Components Analysis, PCA) y regresión mínimo-cuadrática OLS.

El PCA descompone una matriz centrada X (se restan la media a las columnas de esta matriz de datos) en el producto de dos matrices, una de scores T y otra de loadings P. Para cierto número de componentes a, bastante menor que el rango de la matriz de datos X, el PCA considera la descomposición siguiente

$$X = TP^T + E,$$

siendo E una matriz de errores.

La matriz de scores T trata de recoger la máxima cantidad de "información'' posible de la matriz X entre todas las matrices ortogonales $T^TT = I$ que resultan de a combinaciones lineales de las variables originales.

Si partimos del ajuste de regresión lineal múltiple OLS dado por

$$y = Xb + e \quad (2.1)$$

siendo e un vector de errores. La idea del PCR es reemplazar en la regresión OLS a la matriz X por la matriz de scores T.

El modelo resultante es

$$y = Xb + e = (TP^T)b + e_T = Tg + e_T,$$

donde los nuevos coeficientes de regresión son $g = P^Tb$ y se cuenta con un nuevo vector de error e_T.

La información recogida por las variables originales (muy correladas) está ahora comprimida en unos pocos vectores de scores incorrelados, por lo que se puede considerar ya resuelto el problema de la multicolinealidad.

Al usar regresión OLS para estimar los coeficientes, los coeficientes serían:

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

Como consecuencia de la incorrelación de los vectores de scores T se tiene que $T^T T$ es una matriz diagonal. Por tanto, su inversa es muy fácil y, por supuesto, estable numéricamente.

Para deshacer la transformación, los coeficientes de la regresión finales para el modelo original (2.1) deben verificar que

$$P^T b_{PCR} = g.$$

La complejidad del modelo de regresión resultante puede controlarse por el número de componentes principales a seleccionado.

2.2 Regresión PLS

Alternativamente a la regresión PCR, la regresión de mínimos cuadrados parciales PLS es un método estadístico lineal que trata de encontrar relaciones lineales tras la proyección de las variables de predicción y las variables respuesta a nuevos espacios de menor dimensionalidad.

En contraposición con el análisis de regresión tipo OLS, que no reducía la dimensionalidad y utilizaba todas las variables directamente sin evitar el problema de multicolinealidad, la regresión PLS surge como una combinación de la regresión mínimo-cuadrática con ideas de reducción de dimensionalidad, del tipo a las consideradas en Análisis de Componentes Principales.

Una similitud entre PCR y PLS es que los datos-x se convierten en componentes y estas componentes son las que se relacionan con la(s) variable(s) respuesta(s).

Aunque PCR y PLS tienen esa filosofía subyacente común, normalmente PCR necesita más componentes que PLS porque no utiliza la información de la variable dependiente y para la computación de los scores. Por tanto, la diferencia entre ellos es que el método que usa el PCR para obtener las componentes latentes en el espacio-x es un tipo de aprendizaje no supervisado, tipo PCA, donde los valores que se desean predecir en el espacio-y no son utilizados. Contrariamente, se usa un método de aprendizaje supervisado en el método PLS también en esa fase.

En Análisis de Componentes Principales, la m-ésima dirección de la componente principal dada por v_m debe resolver

$$v_m = \text{argmax}_{\alpha} \text{Var}(X\alpha) \text{ con } \| \alpha \| = 1, \alpha^T S v_l = 0, l = 1, \ldots, m-1,$$

siendo S la matriz de varianzas-covarianzas muestral y v_l, para $l = 1, \ldots, m-1$, las componentes principales previamente obtenidas.

Por otro lado, la regresión OLS resuelve:

$$b = \text{argmax}_{\alpha} \text{corr}^2(y, X\alpha).$$

El método PLS busca un objetivo "compromiso", entre los dos fines que persiguen los métodos PCA y el OLS, al maximizar un criterio del tipo:

$$w_m = \text{argmax}_\alpha\{\text{corr}^2(y, X\alpha)\text{Var}(X\alpha)\} \text{ sujeto a } \| \alpha \| = 1 \text{ y que } \alpha^T S w_l = 0,$$

para $l = 1, \ldots, m-1$.

Nótese que el producto que nos encontramos en la función objetivo está claramente relacionado con $\text{Cov}^2(y, X\alpha)$ simplemente notando que para dos variables aleatorias U y V se tiene que

$$\text{Cov}^2(U, V) = \text{corr}^2(U, V)\text{Var}(U)\text{Var}(V)$$

y que $\text{Var}(y)$ no depende de α en dicha función objetivo.

A diferencia de la regresión múltiple, el PLS no presupone que los predictores sean fijos y sí que estos predictores se miden con cierto error, lo que hace que PLS sea un método más robusto ante incertidumbre de las mediciones.

El método PLS no realiza selección de variables útiles para explicar la respuesta, sino que el énfasis está exclusivamente en el desarrollo de modelos predictivos. Otros enfoques más adecuados a la selección de variables se basan, por ejemplo, en técnicas LASSO.

Una ventaja de controlar el número de componentes a es poder evitar el sobreajuste. Cuánto más complejo es un modelo, más capacidad tiene de ajustar los datos de entrenamiento, pero el error de predicción puede decrecer a medida que aumenta dicha complejidad. Un modelo lo suficientemente complejo puede ajustar prefectamente nuestro conjunto de entrenamiento con errores casi nulos entre la respuesta real y la predicha. Sin embargo, estos modelos no suelen ser útiles, ya que posiblemente no tengan buena capacidad de generalización, produciendo errores altos al intentar predecir respuestas para observaciones que no sean las del conjunto de entrenamiento con el que se ajustó el modelo.

Como ya se comentó, para manejar el error de generalización recurrimos a la validación cruzada y a la validación doblemente cruzada. La validación cruzada es la forma de remuestreo más usada en Quimiometría. El procedimiento más habitual de esta estrategia es dividir nuestro conjunto de datos en k partes de aproximadamente el mismo tamaño. Una de las partes se deja como conjunto de validación o test mientras que el resto de datos se utilizan como conjunto de entrenamiento y se modelizan aumentando la complejidad (en este caso el número a de las componentes PLS). Con este modelo se realizan predicciones al conjunto de validación y se evalúan dando lugar a un error de predicción. Realizamos el mismo procedimiento k veces, siendo cada vez uno el conjunto de validación. Una vez tenemos los k errores de predicción, el error de generalización de nuestro modelo es la media de estos errores.

El PLS es una poderosa herramienta de análisis por las mínimas exigencias en términos de tamaños de las muestras en las matrices X e Y y sobre las distribuciones de los errores implicados. Este método no precisa que los datos provengan de una distribución conocida, como la distribución normal, y se considera que no impone demasiadas restricciones al modelo.

Dentro de la regresión PLS distinguimos si la variable respuesta esta formada por una o varias variables.

En el primer caso hablaremos de regresión **PLS1**, que será cuando solo contemos con una única variable respuesta y. El algoritmo simplificado a aplicar en el caso del PLS1 es:

1) Calculamos la primera componente como aquella con máxima covarianza entre los scores-x y la variable dependiente y. En este paso, como ya se ha comentado, se busca directamente un compromiso entre máxima correlacción (OLS) y máxima varianza (PCA).

2) Quitamos la variabilidad explicada por la última componente calculada en los datos-x. Este proceso se denomina "deflación'' y corresponde a la proyección del espacio-x en un plano que sea ortogonal a la dirección de la última componente calculada. La nueva matriz resultante tiene el mismo número de variables que la original, pero con la dimensionalidad reducida en uno.

3) A partir de cada nueva matriz resultante del proceso de deflación anterior se calcula la siguiente componente buscando maximizar la covarianza.

4) Es un proceso iterativo se detiene cuando se valore que no se mejora el modelado de la respuesta y. El número final de componentes que extraigamos define la complejidad del modelo.

La regresión PLS con varias variables respuesta (matriz Y) se denomina **PLS2**. EL objetivo en PLS2 es el mismo, crear modelos de calibración para la predicción de varias variables dependientes a partir de las variables independientes. El espacio m-dimensional original para los datos-x se proyecta en un pequeño número de componentes-x y el espacio q-dimensional original para los datos-y se proyecta en un pequeño número de componentes-y. Las componentes -x e -y se relacionan por pares de máxima covarianza de sus scores.

PLS-DA (PLS Discriminant Analysis) es la variante en clasificación del método PLS cuyo objetivo es identificar a qué categoría pertenece una nueva observación. Este método lineal es muy útil para muchos tipos de datos categóricos que se obtienen en Quimiometría y, de nuevo, permite modelar directamente conjuntos de variables predictivas en alta dimensionalidad. Análogamente a la regresión PLS, la variante en clasificación trata de combatir el problema de la multicolinealidad. Al igual que pasaba con los modelos de calibración tipo PLS, este discriminante se considera una versión "supervisada" del PCA, en el sentido de que se consigue la reducción de la dimensionalidad, pero teniendo en cuenta las etiquetas de clase. Como las variables en Y pueden incluir numerosas variables, es evidente que se pueden realizar clasificaciones no binarias o multicategorías.

En este trabajo, además de probar el funcionamiento de los métodos de PLS en comparación con el OLS y el PCR, también se comparará el PLS-DA con el método de Análisis de Discriminante Lineal (LDA). Aunque los dos métodos, PLS-DA y LDA, sirven para reducir la dimensionalidad proyectándola en un espacio de menor dimensión, el LDA resulta de menor utilidad cuando se consideren numerosas variables predictoras muy correladas.

2.2.1 Aspectos matemáticos

PLS se suele explicar como un algoritmo numérico que maximiza una función objetivo bajo ciertas restricciones. La función objetivo es la covarianza entre los scores-x y los scores-y y la restricción típica suele ser la ortogonalidad de estos scores.

Para la regresión PLS2 asumimos que los datos-x y los datos-y son multivariantes y vienen dados por una matriz de dimensión $n \times m$ para los datos-x y una matriz Y de dimensión $n \times q$ para los datos-y. Asumimos que los datos de las filas de ambas matrices se obtienen de las n observaciones disponibles y que X, así, contiene la información de m variables predictoras e Y describe q variables respuesta. Si estamos ante un caso PLS1 solo tendremos una única variable respuesta y con $q = 1$.

En nuestra presentación nos centraremos en el caso más general del PLS2. Nótese que el PLS-DA sería una adaptación trivial del caso PLS2 con codificaciones razonables 0-1 de la variable categórica en las columnas de la matriz Y.

Vamos a asumir que X e Y han sido centradas, substrayendo las medias por columnas. El objetivo de la calibración sería encontrar una relación lineal de la forma:

$$Y = XB + E \quad (2.2),$$

entre las variables-x y las variables-y, siendo E una matriz de errores y B una matriz de coeficientes de regresión de tamaño $m \times q$.

En PLS2, las matrices X e Y se modelan por variables latentes de acuerdo con modelos de regresión:

$$X = TP^T + E_X \quad (2.3)$$

y

$$Y = UQ^T + E_Y, \quad (2.4)$$

donde las matrices E_X y E_Y son matrices de error. Las matrices T y U son matrices de scores; las matrices P y Q son matrices de loadings. Tanto P y Q como T y U tendrán tantas columnas como número de componentes y ese número de componentes a es menor o igual que $\min\{m, q, n\}$.

Los scores-x en T son combinaciones lineales de las variables-x y pueden ser consideradas un buen resumen de estas. Lo mismo pasa con U como resumen de las variables-y. Los vectores t_j, u_j, p_j y q_j denotan la j-ésima columna de las matrices T, U, P y Q para $j = 1, \ldots, a$.

Los scores-x y los scores-y están conectados por una relación lineal interna

$$u_j = d_j t_j + h_j,$$

$j = 1, \ldots, a$, siendo h_j el correspondiente vector de error y d_j el parámetro de regresión (simple sin intercept).

Si la relación lineal entre u_j y t_j es fuerte (es decir, los elementos de h_j son pequeños) entonces los scores-x de la primera componente PLS son buenos para predecir los scores-y, y por lo tanto para predecir los datos-y.

La relación entre los scores sería

$$U = TD + H,$$

siendo H la matriz errores y D una matriz diagonal con elementos $d_1, d_2, \ldots, d_a$. Esta ecuación da el nombre de "parcial" al método PLS porque solo se utiliza una parte de la información recogida en las componentes.

Como en PLS1 no disponemos de los scores-y, la ecuación se reduce a $y = Td + h$ y se busca maximizar la covarianza entre los scores-x e y.

Las maximizaciones están definidas mediante restricciones en los vectores de scores, que suelen ser del tipo $\| t \| = \| u \| = 1$. Los vectores de scores resultan de proyectar las matrices X e Y usando las combinaciones lineales proporcionadas por los vectores de loadings.

Tendría sentido seguir utilizando los vectores de loadings en las matrices P y Q en (2.3) y (2.4) pero por razones técnicas, utilizamos otros tipos de vectores de loadings que denotamos por w y c, siendo w el vector de loadings de las variables-x y c el vector de loadings de las variables-y, de tal forma que $t = Xw$ y $u = Yc$. El problema de maximización con restricciones inicial es entonces:

$$\max_{w,c} \text{Cov}(Xw, Yc) \quad \text{con} \ \| t \| = \| Xw \| = 1 \quad \text{y} \quad \| u \| = \| Yc \| = 1. \quad (2.5)$$

Las soluciones de este problema de maximización serán los primeros scores t_1 y u_1.

Para los siguientes vectores de scores se maximiza el mismo criterio, pero se necesitan nuevas restricciones adicionales. Estas restricciones son de ortogonalidad respecto a los anteriores vectores de scores del tipo $t_j^T t_l = 0$ y $u_j^T u_l = 0$ para $1 \leq j < l \leq a$.

Otra alternativa es requerir ortogonalidad de los vectores de loadings, lo que llevaría a scores no incorrelados, aunque esto no es lo habitual en Quimiometría donde se utiliza de forma prácticamente exclusiva la ortogonalidad de los scores.

Al tratar con la covarianza muestral, para encontrar la primera componente PLS, el problema de maximización (2.5) puede reescribirse como la maximización de

$$t^T u = (Xw)^T Yc = w^T X^T Yc. \quad (2.6)$$

Bajo las mismas restricciones de norma unitaria para los vectores en (2.5), como sucede en los desarrollos de otras técnicas estadísticas multivariantes, las soluciones de c y w se encuentran por descomposición de valor singular (SVD) de la matriz $X^T Y$. De acuerdo con esto, sobre todas las posibles direcciones de w y c, el máximo de la ecuación (2.6) se obtiene para los vectores singulares (derecha e izquierda) $w = w_1$ y $c = c_1$ correspondiendo con el máximo valor singular de $X^T Y$.

A continuación, se van a presentar diferentes algoritmos donde, en todos ellos, la primera componente se calculará de la forma estándar que acabamos de mostrar. Estos algoritmos difieren en cómo se van a calcular el resto de las componentes PLS.

Capítulo 3

Algoritmos PLS

3.1 Algoritmo Kernel

El nombre viene de usar técnicas de autovectores sobre las denominadas matrices "núcleo" que se obtienen por productos de las matrices X e Y. Recordemos que teníamos un problema de maximización (2.5) donde los óptimos w_1 y c_1 se obtenían desde la descomposición SVD de X^TY. Utilizando las propiedades del método SVD, se podría probar que las soluciones se encuentran también como

$$w_1 \text{ es el autovector del mayor autovalor de } X^TYY^TX \quad (3.1)$$

y

$$c_1 \text{ es el autovector del mayor autovalor de } Y^TXX^TY \quad (3.2).$$

Basándonos en la ecuación (2.5) ambos vectores tienen que estar normalizados, de tal forma que

$$\| Xw_1 \| = \| Yc_1 \| = 1$$

Los scores para las direcciones encontradas son las proyecciones $t_1 = Xw_1$ y $u_1 = Yc_1$, que también se toman unitarios. La variable latente p_1 se encuentra mediante regresión, es decir, se tiene

$$p_1^T = (t_1^Tt_1)^{-1}t_1^TX = t_1^TX = w_1^TX^TX.$$

Para obtener la siguiente componente, teniendo en cuenta el problema inicial (2.5) y buscando que dicho máximo se alcance en dirección ortogonal a t_1, se aplicará la deflación de la matriz X. La matriz deflactada X_1 es

$$X_1 = X - t_1p_1^T = X - t_1t_1^TX = (I - t_1t_1^T)X.$$

Se puede ver que no es necesario hacer deflación en la matriz Y porque podríamos ver que dicha deflación se hace multiplicando Y con la misma matriz $G_1 = (I - t_1t_1^T)$. Como G_1 es simétrica, con $G_1^T = G_1$, e idempotente, con $G_1G_1 = G_1$, al obtener w_2 y c_2 se obtiene el mismo resultado independientemente de que Y haya sido deflactada o no.

Las siguientes componentes PLS (t_2, p_2, y el resto de t's y p's) se obtienen del mismo algoritmo que para la primera componente pero usando la matriz deflactada X obtenida después de calcular las componentes anteriores. Los scores-y para todas las componentes se derivan de los scores-x mediante la relación:

$$u_j = Yc_j.$$

Adicionalmente, los loadings-y q_j de las a componentes se calculan mediante regresión desde (2.4) como

$$q_j^T = \left(u_j^T u_j\right)^{-1} u_j^T Y.$$

No obstante, para estimar los coeficientes finales B para el modelo en (2.2) no se necesita estimar Q. Puede verse que los coeficientes de regresión se estiman directamente como:

$$B = W(P^T W)^{-1} C^T,$$

sin usar Q, y donde la matriz C tiene por columnas a los vectores de loadings de las variables-y y la matriz W contiene por columnas a los vectores de loadings de las variables-x. Esta matriz B es la que finalmente relaciona las variables-x con las variables-y.

El algoritmo nucleo puede ser aplicado en datos univariantes (PLS1). Al igual que en PLS2, la deflación se realiza solo en la matriz X. Ahora existe solo único autovalor positivo y el correspondiente autovector es w_1.

Esta versión del algoritmo núcleo está especialmente diseñada para matrices de datos con un gran número de observaciones n. En este caso, si las dimensiones de la matriz de datos no son muy grandes, las matrices núcleo tienen dimensión notablemente más pequeña que n y la descomposición propia SVD se calcula rápido. También existe una alternativa al algoritmo núcleo para un número alto de variables-x y variables-y con un n número moderado de observaciones, que se denomina "Wide Kernel Method". La idea es multiplicar (3.1) por X por la izquierda y (3.2) por Y por la derecha. Con las relaciones $t_1 = Xw_1$ y $u_1 = Yc_1$, las nuevas matrices núcleo son XX^TYY^T y YY^TXX^T y los autovectores asociados a los autovalores más grandes serán t_1 y u_1, respectivamente.

3.2 Algoritmo NIPALS

El primer algoritmo en resolver el problema PLS fue NIPALS (Nonlinear Iterative Partial Least Squares). A pesar de que los resultados fueron útiles, había cierta confusión sobre qué estaba haciendo el algoritmo exactamente. La existencia de numerosos algoritmos ligeramente diferentes no fue de ayuda en este aspecto.

Los algoritmos Kernel y NIPALS suelen llegar a una misma solución ya que ambos usan el mismo tipo de deflación, la única diferencia entre ellos está en el cálculo de las componentes (por iteración NIPALS o como autovectores en el algoritmo Kernel) pero con resultados comparables en cuanto a precisión numérica.

A continuación, se describen todos los pasos que sigue NIPALS con la justificación de dichos pasos:

Primeramente se debe obtener la *primera componente* PLS siguiendo los pasos:

1) Inicializo u_1 con, por ejemplo, la primera columna de Y
2) $w_1 = X^T u_1 / (u_1^T u_1)$
3) $w_1 = w_1 / \| w_1 \|$
4) $t_1 = Xw_1$

5) $c_1 = Y^T t_1/(t_1^T t_1)$

6) $c_1 = c_1/\| c_1 \|$

7) $u_1^* = Y c_1$

8) $u_\Delta = u_1^* - u_1$

9) $\Delta u = \| u_\Delta \|^2$

10) Se para el proceso iterativo si $\Delta u < \epsilon$ (con ϵ un valor bastante pequeño de tolerancia) y si $\Delta u \geq \epsilon$ entonces se actualiza $u_1 = u_1^*$ y se vuelve al paso 2.

Considerando los modelos de variable latente (3.1) y (3.2) podemos ver que en el paso 2 se encontrarían los coeficientes de regresión OLS que denotamos w_1^j en el modelo de regresión

$$X = u_1^{j-1}\left(w_1^j\right)^T + e$$

siendo e un término error. Por lo tanto, esta es una regresión de los datos-x respecto al último vector de scores "potenciales" de los datos-y. Nótese que se requiere invertir la "matriz" $u_1^T u_1$ y eso es simplemente considerar $1/u_1^T u_1$. Dicho denominador se muestra en el paso 2 para hacer esas regresiones evidentes. Después de normalizar w_1 en el paso 3, los datos-x se proyectan en el paso 4 para encontrar el vector actualizado de scores de los datos-x. Después, en el paso 5, se ajustan regresiones OLS para los datos-y en relación con el vector de scores de los datos-x. Otra vez el denominador $t_1^T t_1$ se muestra para hacer dicha regresión evidente. Se normaliza c_1 a vector unitario en el paso 6 y se proyecta Y en el paso 7 para actualizar el vector de scores-y en u_1^*. Este proceso se repite hasta que los cambios en el vector determinando u_1 estabiliza.

Ya comentamos que no era fácil ver por qué este algoritmo resuelve el problema inicial (2.5) de maximizar la covarianza entre los scores-x y los scores-y. Esto se puede justificar ahora teniendo en cuenta que al inicio del paso 2, en la iteración $j + 1$, tenemos

$$w_1^{j+1} = X^T u_1^j / \left(\left(u_1^j\right)^T u_1^j\right).$$

Insertando la fórmula de u_1^j del paso 7, podemos reemplazarla de nuevo por c_1^j y los cambios en los siguientes pasos para comprobar que w_1^{j+1} es proporcional a

$$X^T Y Y^T X w_1^j \quad (3.3)$$

(esa proporcionalidad es debida a la existencia de una constante que depende de las distintas normas de sus términos). Esto demuestra que después de converger las actualizaciones para w_1 (estabilizar w_1^{j+1} en el proceso iterativo) se tiene que (3.3) implica resolver un problema de autovalores donde w_1 será el autovector de $X^T Y Y^T X$ asociado al autovalor más grande. Este era el objetivo en (3.1) del método núcleo. Análogamente, podríamos ver que $c_1^{j+1} = Y^T X X^T Y c_1$ es el mismo problema que (3.2) en el método núcleo.

El algoritmo NIPALS y el algoritmo núcleo trabajan de forma diferente a la hora de hallar el *resto de componentes*, aunque a pesar de ello los resultados son bastante equivalentes. NIPALS procede mediante deflaciones de las matrices de X y de Y, tal como se indica en las siguientes líneas:

11) $p_1 = X^T t_1/(t_1^T t_1)$

12) $q_1 = Y^T u_1/(u_1^T u_1)$

13) $d_1 = u_1^T t_1/(t_1^T t_1)$

14) $X_1 = X - t_1 p_1^T$ e $Y_1 = Y - d_1 t_1 c_1^T$

Los pasos 11-13 corresponden simplemente a las regresiones OLS usando (2.2) y (2.3). El paso 14 efectúa la deflación de las matrices X e Y. Las matrices deflactadas X_1 e Y_1 se utilizan para hallar las próximas componentes PLS, siguiendo el mismo esquema proporcionado en los pasos 1 al 10.

Por último, los coeficientes de regresión B de la ecuación (2.1), que relacionan los datos-x con los datos-y, se obtienen mediante

$$B = W(P^T W)^{-1} C^T.$$

Evidentemente, las matrices W, P y C contienen los vectores w_j, p_j y c_j en sus columnas, para $j = 1, \ldots, a$.

El algoritmo NIPALS se simplifica notablemente en el caso del *PLS1* porque ya no necesitamos iteraciones dentro de la obtención de cada una de las componentes. El pseudocódigo para calcular las a componentes será:

1) Inicializamos $X_1 = X$ e iteramos los pasos 2-7 para $j = 1, \ldots, a$

2) $w_j = X_j^T y/(y^T y)$

3) $w_j = w_j/\| w_j \|$

4) $t_j = X w_j$

5) $c_j = y^T t_j/(t_j^T t_j)$

6) $p_j = X_j^T t_j/(t_j^T t_j)$

7) $X_{j+1} = X_{j+1} - t_j p_j^T$

Los coeficientes de la regresión final en el modelo $y = Xb + e_1$ se estiman como $b = W(P^T W)^{-1} c$, donde W y P recogen por columnas los vectores w_j y p_j, y c es el vector con todos los elementos c_j.

3.3 Algoritmo SIMPLS

El nombre SIMPLS deriva de "implementación directa de una modificación estadística del método PLS". La primera componente PLS de SIMPLS es idéntica a la de NIPALS o el algoritmo núcleo, siendo la obtención de las siguientes componentes ligeramente diferentes. La principal diferencia entre el algoritmo núcleo y NIPALS era el tipo de deflación mientras que en SIMPLS no hay deflación de las matrices centradas X e Y, sino que la deflación se realiza sobre la matriz de covarianzas, concretamente en la matriz de productos cruzados

$$S = X^T Y$$

entre los datos-x y los datos-y. El pseudocódigo para el algoritmo SIMPLS es:

1) Inicializo $S_0 = X^T Y$ y se iteran los siguientes pasos 2-6 para $j = 1, \ldots, a$
2) Si $j = 1, S_j = S_0$ mientras que si $j > 1$ se considera

$$S_j = S_{j-1} - P_{j-1}\left(P_{j-1}^T P_{j-1}\right)^{-1} P_{j-1}^T S_{j-1}$$

3) Se calcula w_j como el primer vector singular (por la izquierda) en la descomposión SVD de S_j
4) $w_j = w_j / \| w_j \|$
5) $t_j = X w_j$
6) $t_j = t_j / \| t_j \|$
7) $p_j = X_j^T t_j$
8) P_j se actualiza anexando por columnas los p_j anteriores de tal forma que

$$P_j = \left[p_1, p_2, .., p_{j-1}\right].$$

Los pesos resultantes w_j y los scores t_j se almacenan por columnas en las matrices W y T. La matriz W se distingue de la del resto de los algoritmos porque los pesos están directamente relacionados con la matriz X y no con las matrices deflactadas. El paso 2 tiene en cuenta la restricción de ortogonalidad de los scores t_j a todos los vectores de scores previos al utilizar S_{j-1}. El paso 3 maximiza de manera directa el problema inicial (2.5). Los scores en el paso 4 se obtienen de proyectar directamente X en la dirección óptima y los loadings del paso 5 se obtienen mediante regresión OLS.

Los coeficientes finales de la regresión son ahora

$$B = W T^T Y,$$

donde no se necesita invertir ninguna matriz en comparación con el método núcleo.

En el caso del PLS1 el algoritmo se simplifica notablemente.

3.4 Otros algoritmos

El *método O-PLS* (Ortogonal Projections to Latent Structures) pretende eliminar la "variabilidad" de X que sea ortogonal a los datos-y. Esta variabilidad ortogonal se modela con componentes añadidas para los datos-x y resulta de la descomposición $X = TP^T + T_o P_o^T + E$, donde T_o representa los scores y P_o los loadings correspondientes a la variabilidad ortogonal. Esta descomposición también puede ser considerada un modelo PLS para la matriz de datos "filtrada" $X - T_o P_o^T$. Al eliminar la variabilidad ortogonal de Y, el método O-PLS pretende maximizar simultáneamente la correlación y la covarianza entre los scores-x y los scores-y buscando una buena predicción e interpretabilidad.

El *algoritmo de autovectores* es parecido al algoritmo de núcleo aunque se trabaja de forma más directa. La idea es computar no solo el autovector asociado al autovalor más grande como en (3.1) y (3.2) sino todos los autovectores asociados a los a autovalores mayores, donde a es

el número deseado de componentes. En consecuencia, p_1,..., p_a son los vectores de loadings PLS ortogonales en el espacio-x formados por los autovectores asociados a los a autovalores mayores de $X^T Y Y^T X$. Los vectores de loadings ortogonales del espacio-y denotados por q_1,..., q_a son los autovectores de los a autovalores mayores de $Y^T X X^T Y$. Como en otras ocasiones, los scores-x y los scores-y se hallan proyectando los datos en los vectores de loadings, por ejemplo, scores-x son $t_j = X p_j$ y los scores-y serían $u_j = Y q_j$. No se aplica ningún tipo de deflación y por lo tanto los vectores de scores no son incorrelados, lo cual no soluciona el problema de maximización restringida que nos estábamos planteando. No obstante, los vectores de loadings sí que son ortogonales y esto puede ser preferible, especialmente para obtener el mismo tipo de representaciones gráficas que típicamente se hacen con otros métodos estadísticos. Utilizando los vectores de scores, los datos-x y los datos-y se pueden representar en sistemas de coordenadas ortogonales de menor dimensión, llegándose a poder distinguir estructuras de datos muy interesantes e interpretables.

PLS también tiene su variante para resolver modelos de regresión *no lineales*. En este campo del PLS no lineal, se distinguen dos aproximaciones fundamentales. Una primera aproximación consiste en cambiar la relación lineal interna por una no lineal, de tal forma que los scores-y tengan relación no lineal con los scores-x mediante el uso de funciones polinómicas, splines, bases radiales u otras. Otra posible aproximación sería que los datos-x y los datos-y originales sean transformados utilizando una función no lineal. Para lograr este propósito la teoría basada en núcleos (por ejemplo, usada en SVM) ha sido adaptada al método PLS. Con la primera aproximación los resultados pueden seguir siendo interpretables en términos de las variables originales, mientras que en la segunda aproximación se pueden obtener modelos que son poco interpretables, pero que pueden tener un excelente poder predictivo.

3.5 PLS-Robusto

Todos los algoritmos detallados previamente resolvían el problema enunciado en (2.5). Para estimar la covarianza entre los scores-x y los scores-y se ha considerado la covarianza muestral que es bien conocido que es un estimador muy poco robusto. Su falta de robustez se traduce en que unas pocas observaciones atípicas pueden afectar muy negativamente al funcionamiento del método.

Con el fin de robustificar la metodología, se ha propuesto robustecer la estimación de la covarianza y reemplazar la regresión OLS por regresión robusta.

Así, por ejemplo, con el PLS Robusto nos podríamos referir a un método que está directamente relacionado con la idea de mínimos cuadrados parciales, pero utilizando M-regresión robusta en vez de utilizar mínimos cuadrados OLS y resultando en un método denominado como "M-regresión robusta parcial". No se resuelve directamente el problema enunciado en (2.2) pero seguimos realizando regresión en los datos-y solamente con información parcial de los datos-x, dados por el modelo (2.3).

Partiendo, de nuevo, de

$$Y = Xb + e_1 = TP^T b + e_2, \quad (3.4)$$

el propósito sería estimar de forma robusta los nuevos coeficientes de regresión $g = P^T b$ en la ecuación (3.4) donde la matriz T es desconocida. Como en M-regresión, la idea es aplicar una función ρ a los residuales

$$r_i = y_i - t_i^T g$$

donde t_i son las filas de T para $i = 1, \ldots, n$. A estos residuales r_i se les aplica la función $\rho(\cdot)$ para reducir la importancia de los residuales grandes en la función objetivo, por lo que minimizamos

$$\sum \rho(y_i - t_i^T g).$$

Este problema se puede escribir como la minimización de una regresión con pesos de la forma

$$\sum_{i=1}^{n} w_i^r \, (y_i - t_i^T g)^2,$$

con pesos del tipo $w_i^r = \rho(r_i)/r_i^2$.

Además de las observaciones con valores residuales grandes, los denominados como puntos de influencia también pueden llevar a no obtener una buena estimación de los coeficientes de regresión. Esto lleva a introducir pesos adicionales para disminuir el dañino efecto de estos posibles puntos de influencia y que son observaciones atípicas en el espacio de las variables de scores en T. Los pesos robustificados resultantes, asignados a cada t_i, se denotan como w_i^t (el superíndice t en w_i^t hace referencia a que estos pesos dependen de los valores de los scores t_i).

Ambos tipos de pesos se combinan de la forma

$$w_i = w_i^r w_i^t$$

y los coeficientes de regresión g son el resultado de minimizar

$$\sum_{i=1}^{n} w_i \, (y_i - t_i^T g)^2 = \sum_{i=1}^{n} \left(\sqrt{w}_i y_i - \left(\sqrt{w}_i t_i \right)^T g \right)^2$$

Esto significa que los datos-y y los scores deben ser multiplicados por pesos apropiados $\sqrt{w}_i$ para poder aplicar los procedimientos PLS anteriormente vistos. Para poder llevar a cabo todo este proceso de determinación óptima de pesos, para conseguir robustez, hay que determinar valores iniciales de los pesos y actualizarlos utilizando un algoritmo iterativo, que no será descrito en este trabajo.

Capítulo 4

Ejemplos prácticos

4.1 Software

Toda la parte práctica ha sido realizada mediante el software R.

La librería más utilizada ha sido "chemometrics", en la cual se incluyen todos los conjuntos de datos analizados y las funciones utilizadas para realizar la regresión PLS y PCR.

Lo primero que se ha usado es la función `mvr_dcv` la cual utiliza doble validación cruzada para métodos de regresión multivariante. El modelo resultante de esta doble validación cruzada es introducido en la función `plotSEPmvr` que genera un gráfico con los valores de error por componentes de manera que se pueda llegar a la conclusión del número óptimo de componentes para ajustar el modelo. De esta manera, se han ajustado modelos con todos los algoritmos y se ha procedido a su posterior comparación.

Si se desea realizar PLS Robusto se utiliza la función `pmr_cv` y el resultado se representa con `plotSEPprm`.

A partir de ciertos datos de entrenamiento, se ajusta un modelo de regresión PLS con la función `plsr` de la librería "pls", con el cual se hallan predicciones sobre un conjunto de datos de prueba hallando un error de predicción mediante la función `defaultSummary` del paquete "caret".

En clasificación se usa el paquete "mixOmics" que ayuda a ajustar el modelo con la función `splsda`. El gráfico de scores se realiza con la función `plotIndiv` y con `plotVar` el círculo de correlación. Se utiliza validación cruzada para hallar el número de componentes óptimos con la función `tune.splsda`. Para la comparación con LDA se usa la función `lda` del paquete de R "MASS".

4.2 PLS1

4.2.1 Compuestos aromáticos policíclicos

Este ejemplo está basado en el estudio de los índices de retención de GC (cromatografía de gases) de compuestos aromáticos policíclicos (sustancias químicas que se forman durante la combustion incompleta de materia orgánica como el carbón o gasolina, así como otras sustancias orgánicas como el tabaco) los cuales dependen de 467 variables numéricas. Se disponen 209 observaciones.

Se procede a estudiar la variable respuesta:

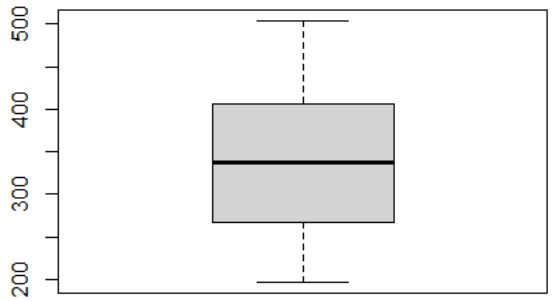

Figura 4.1: Diagrama de cajas.

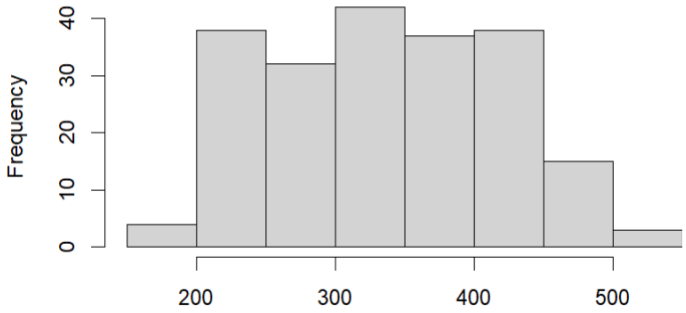

Figura 4.2: Histograma.

Se puede ver que la variable respuesta toma valores entre 200 y 500.

Se ajustan modelos de regresión PLS con tres de los algoritmos estudiados; el algoritmo núcleo, O-PLS y SIMPLS, para encontrar aquel modelo que ajuste mejor los datos.

Independientemente del algoritmo utilizado para ajustar PLS, el número de componentes ideal para ajustar este conjunto de datos son 11 componentes.

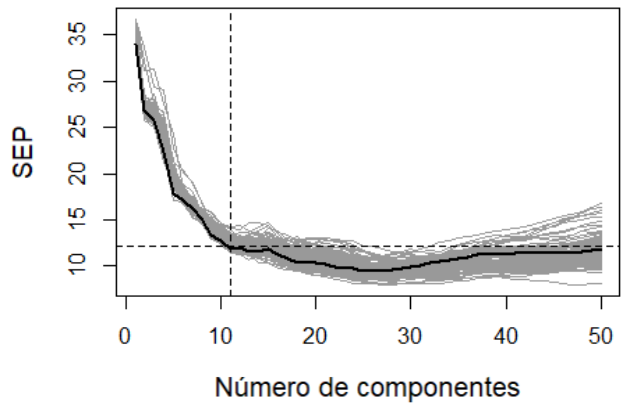

Figura 4.3: Representación del valor SEP de cada modelo con diferente número de componentes mediante validación doblemente cruzada.

	SIMPLS	Kernel	O-PLS
SEP	12.55486	12.49963	12.425558

Tabla 4.1: Estimación del error estándar de predicción para cada algoritmo con 11 componentes.

El error estándar de predicción en los tres algoritmos es muy similar, pero el menor es el modelo de regresión PLS ajustado con el algoritmo O-PLS y 11 componentes, por lo que se le tomará como modelo óptimo.

Para comprobar cómo ajusta el modelo a la variable respuesta, se realizan pruebas con los datos, dividiendo estos en un conjunto de entrenamiento y otro de prueba y ajustando el modelo óptimo hallado anteriormente.

En los siguientes gráficos se pueden ver representados los scores y las correlaciones de los dos primeros componentes:

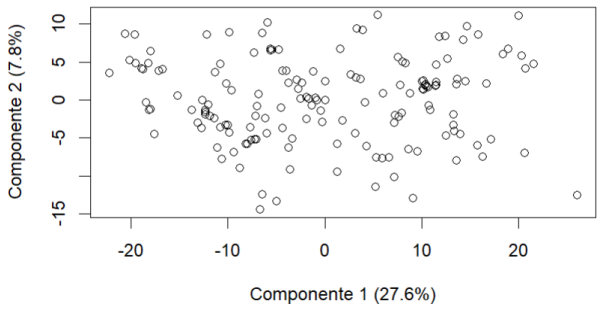

Figura 4.4: Representación de los scores de las dos primeras componentes.

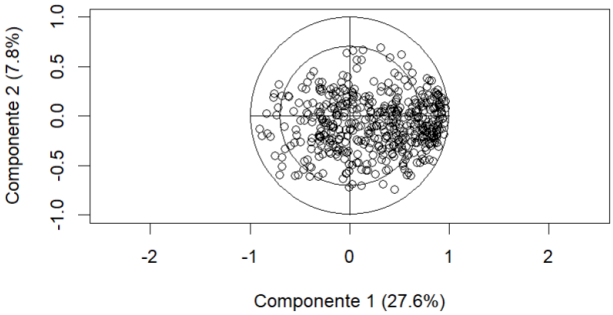

Figura 4.5: Representación de las correlaciones de las dos primeras componentes.

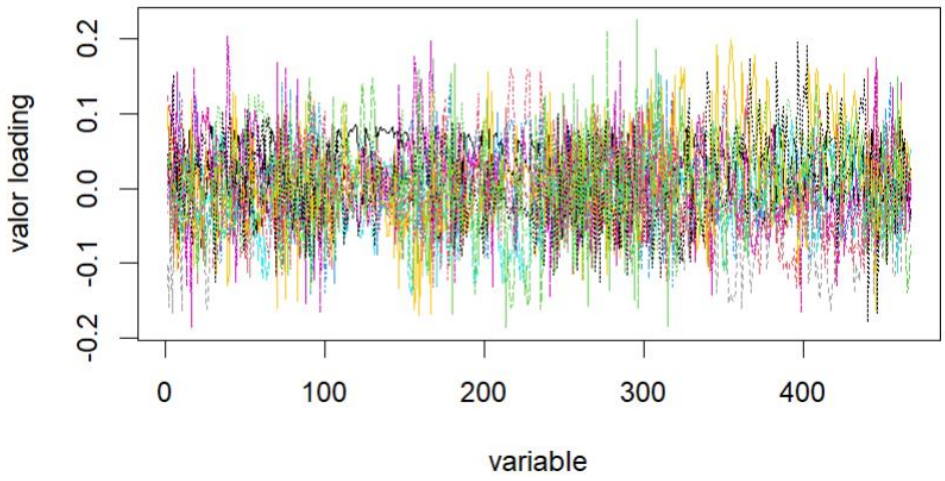

Figura 4.6: Representación de los loadings de las 11 componentes.

Se realizan predicciones y se comparan con los datos observados:

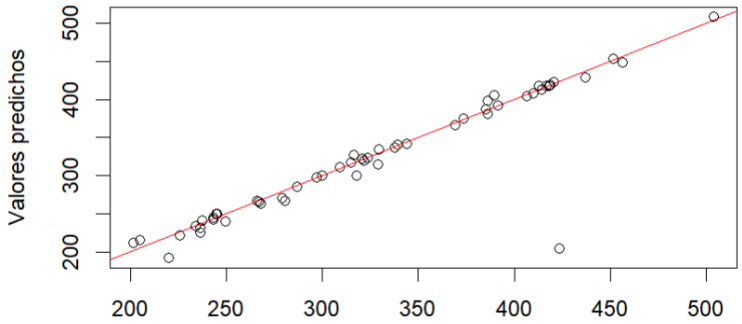

Figura 4.7: Valores observados contra valores predichos.

Se puede observar que hay un valor atípico, el cual se encuentra alejado de la línea roja. La presencia de este valor hace que el error estándar de predicción sea alto, 305.1692.

Se procede a utilizar PLS Robusto ya que la robustez hace que el análisis no se vea afectado por estos valores atípicos. Utilizando validación cruzada 10-fold se comparan los valores de SEP con los valores obtenidos cuando el 20% de los residuales más grandes en valor absoluto han sido eliminados.

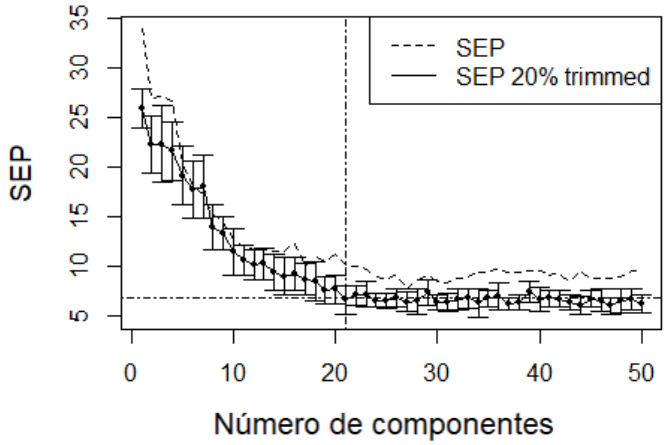

Figura 4.8: Representación del valor SEP de cada modelo con diferente número de componenentes.

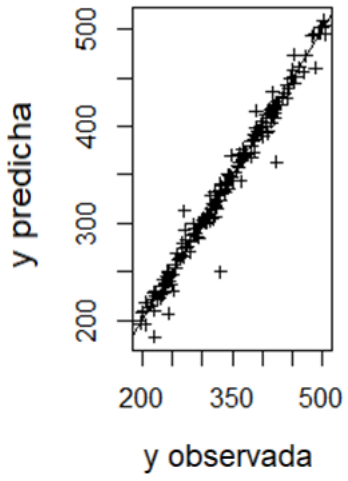

Figura 4.9: Valores reales contra predichos.

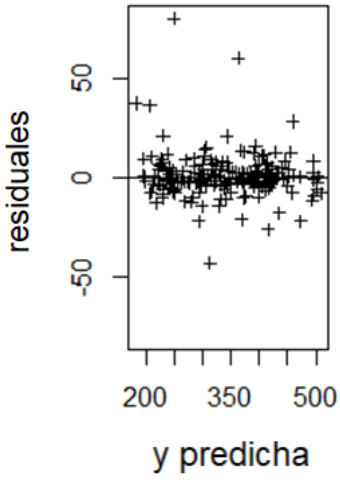

Figura 4.10: Valores predichos contra residuales.

El método de PLS Robusto hace mejorar las predicciones, el error estándar de predicción dismunuye siendo de 6.031733.

El modelo óptimo es con 21 componentes. El peso de alguna de las 209 observaciones ha disminuido drásticamente al construir el modelo robusto PLS, lo que hace que aparezca con residuales grandes. Esta disminución de pesos provoca cierta curvatura por el ajuste de valores altos y bajos de y.

Se procede a comparar el ajuste que hace el método PLS con el que haría PCR

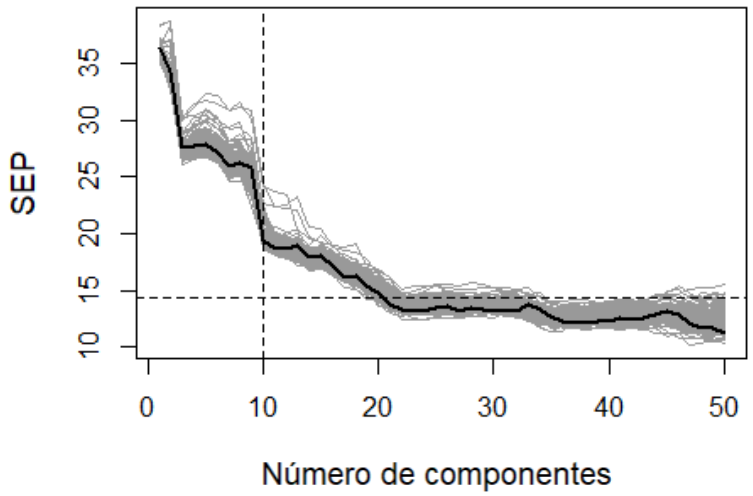

Figura 4.11: Representación del valor SEP de cada modelo con diferente número de componentes mediante validación doblemente cruzada.

El número óptimo de componentes es 10, ajustando este modelo y tras realizar predicciones se comparan con los valores observados.

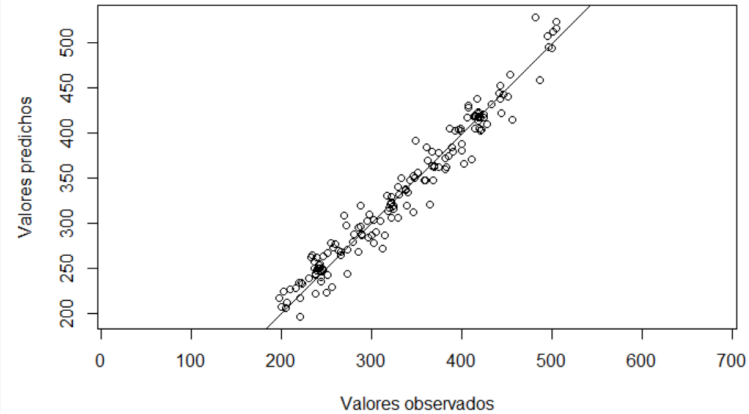

Figura 4.12: Valores observados contra valores predichos.

El error estándar de predicción es de 20.02531, siendo este mayor que con el método PLS. Como conclusión, en este caso el método PLS es mejor que PCR, pero debido a los valores atípicos que se encuentran, el menor error de predicción resulta obtenido con el método PLS-Robusto.

4.2.2 Datos de cenizas

Nos encontramos ante un conjunto de 99 muestras de ceniza originarias de diferentes biomasas medidas en 9 variables. Se han añadido variables transformadas logarítmicamente. La variable respuesta es SOT (Softening Temperature).

Las variables explicativas son todas numéricas : P205, Si02, Fe203, Al203, Ca0, Mg0, Na20, K20, Log(P205), Log(Si02), Log(Fe203), Log(Al203), Log(Ca0), Log(Mg0), Log(Na20), Log(K20).

En los siguientes gráficos se detalla el comportamiento de la variable respuesta:

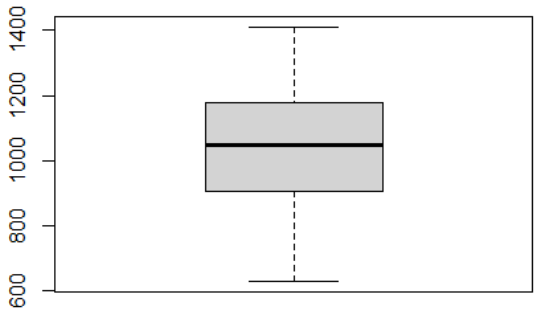

Figura 4.13: Diagrama de cajas.

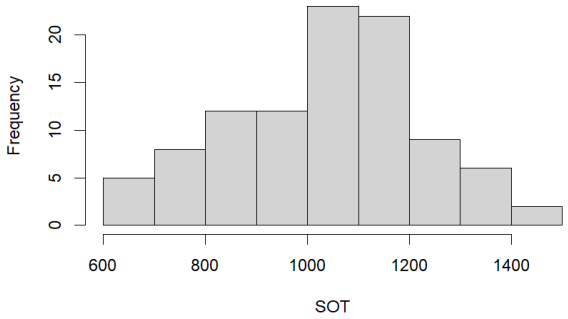

Figura 4.14: Histograma.

Se puede ver que puede tomar valores entre 600 y 1500 aunque lo mas común es un valor entre 1000-1200.

Para responder a la pregunta de si se ajustarán mejor las variables sometidas a la transformación logarítmica, se ha probado regresión PLS ajustada con el algoritmo SIMPLS.

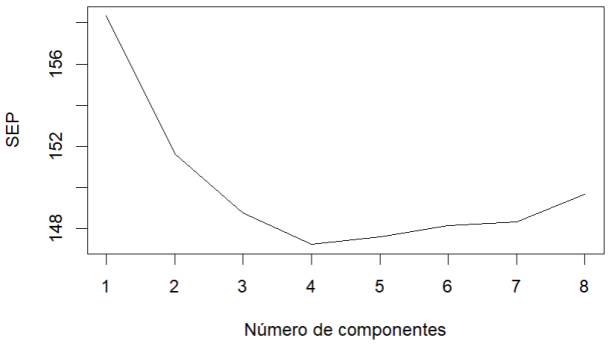

Figura 4.15: SEP para cada número de componentes ajustado con variables sin transformar.

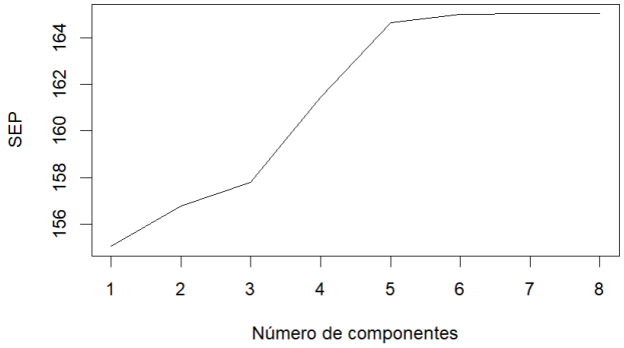

Figura 4.16: SEP para cada número de componentes ajustado con variables transformadas.

	SIMPLS	Kernel	O-PLS
SEP	147.2416	145.771	146.7714

Tabla 4.2: Estimación del error estándar de predicción para cada algoritmo con 4 componentes.

Aplicando regresión PLS sobre el conjunto de datos, el modelo óptimo sería con 4 componentes independientemente del algoritmo utilizado. Los errores de los tres algoritmos son muy parecidos aunque es el algoritmo Kernel aquel que proporciona un error menor, por lo que a partir de ahora se tomará regresión PLS ajustada con algoritmo Kernel y 4 componentes como modelo óptimo.

Para comprobar cómo ajusta el modelo creado a la variable respuesta, se realizan pruebas con los datos, dividiendo estos en un conjunto de entrenamiento y otro de prueba y ajustando el modelo óptimo hallado anteriormente.

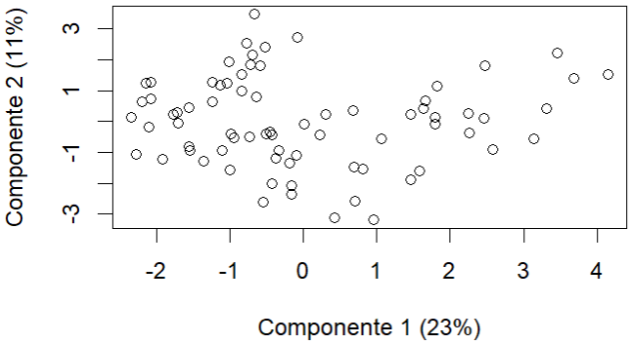

Figura 4.17: Representación de los scores de las dos primeras componentes.

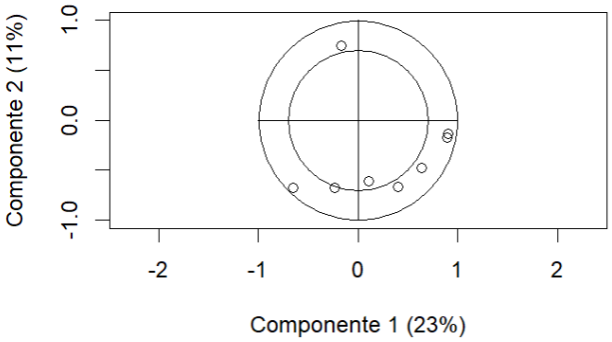

Figura 4.18: Representación de las correlaciones de las dos primeras componentes.

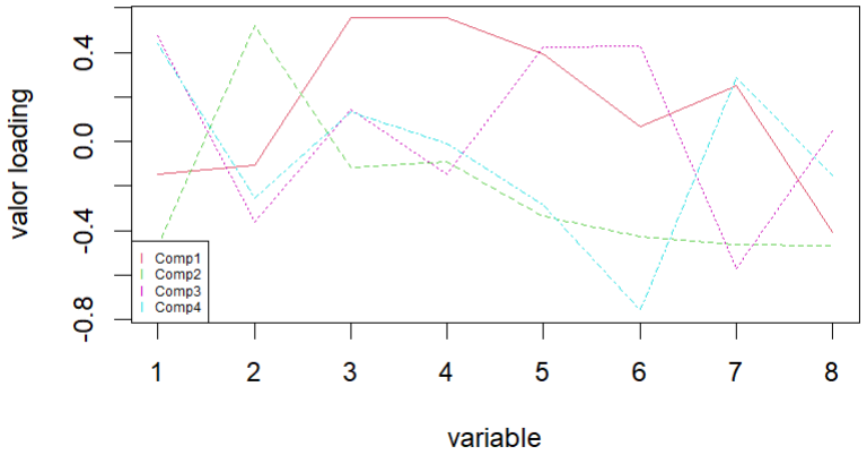

Figura 4.19: Representación de los loadings de las 4 componentes.

Se realizan predicciones y se comparan con los datos observados:

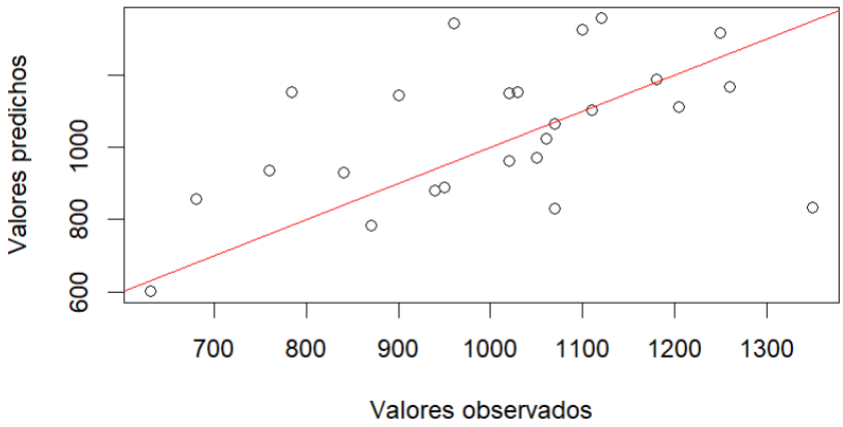

Figura 4.20: Valores observados contra valores predichos.

Se puede ver que el modelo elegido no ajusta especialmente bien debido a la gran cantidad de valores atípicos. El error estándar de predicción es de 189.2749. Una posible solución podría ser aplicar PLS-Robusto.

Utilizando validación cruzada 10-folds, se comparan los valores de SEP con los valores obtenidos cuando el 20% de los residuales más grandes en valor absoluto han sido eliminados.

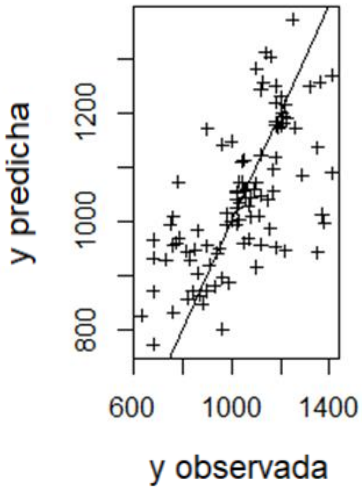

Figura 4.21: Valores reales contra predichos.

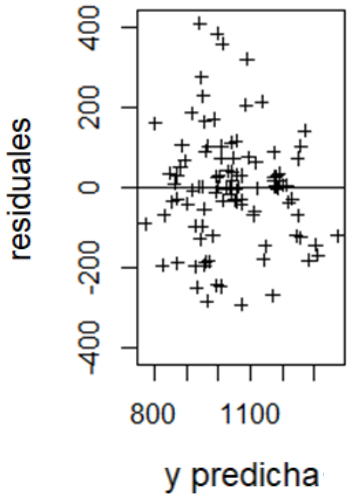

Figura 4.22: Valores predichos contra residuales.

En las gráficas podemos ver que los valores predichos y los reales se ajustan mejor que antes y el error de predicción ha bajado respecto al modelo anterior. Antes este error era de 145.7714 y ahora de 125.4899.

Se quiere comparar el ajuste que hace el método PLS con el que haría PCR.

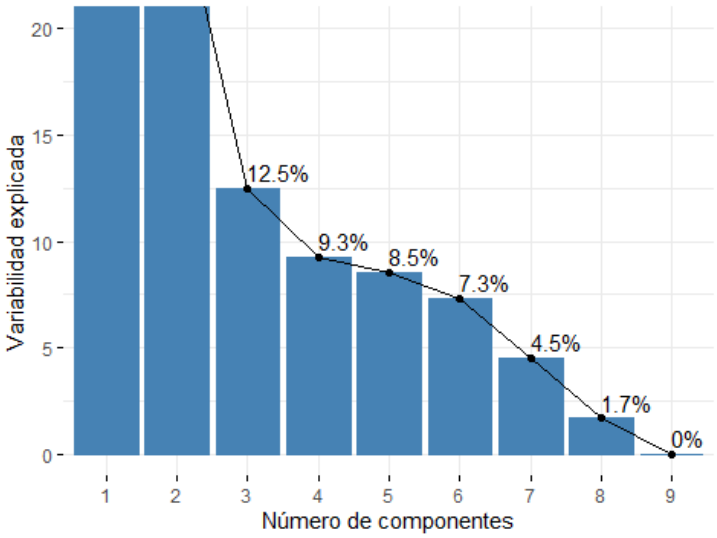

Figura 4.23: Scree plot.

El número óptimo de componentes es 3, ajustando este modelo y tras realizar predicciones se comparan con los valores observados.

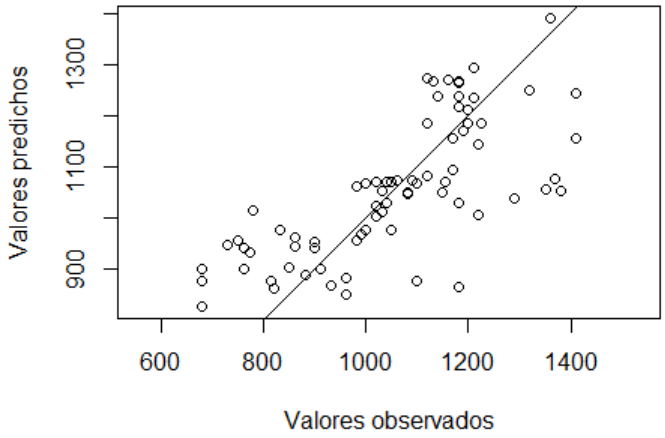

Figura 4.24: Valores observados contra valores predichos.

Se puede apreciar que el ajuste no es muy bueno y que el error de predicción es de 155.8926.

Al comparar el modelo PCR con el modelo de regresión PLS ajustado con el algoritmo Kernel y 4 componentes, cuyo error era de 155.771, se puede comprobar que PLS ajusta mejor que PCR. Sin embargo, anteriormente se vió que existían muchos valores atípicos y el mejor método para ajustar estos datos es PLS-Robusto, ofreciendo el menor error siendo este de 125.4899.

4.3 PLS2

4.3.1 Datos de cereales

El conjunto de datos a estudiar se conforma por una lista con 3 elementos y 15 observaciones en cada uno. Así, X será una matriz de 15 filas y 145 columnas, Y una matriz de 15 filas y 6 columnas y la matriz Ysc es la matriz Y escalada la cual contiene también 15 filas y 6 columnas.

Los datos X son 145 espectros infrarrojos (medida de la interacción de la radiación infrarroja con la materia) y los datos Y son 6 propiedades químicas: Heating value, C, H, N, Starch, Ash.

Se estudian las variables respuesta sin escalar y su distribución:

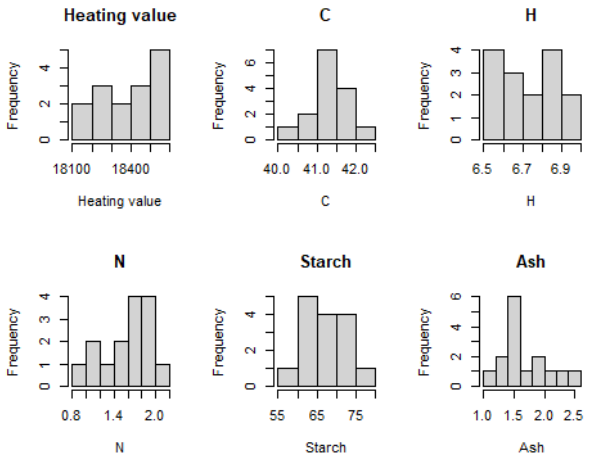

Figura 4.25: Histogramas.

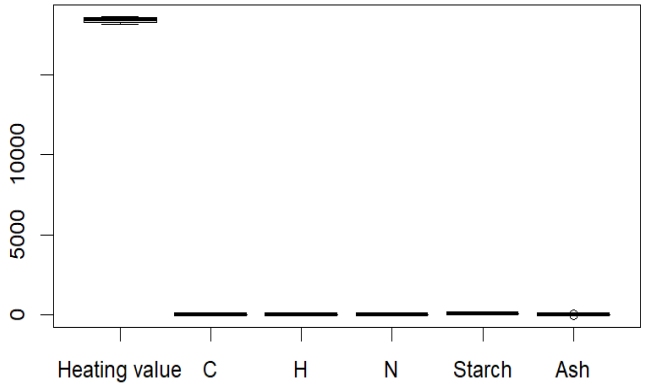

Figura 4.26: Diagramas de cajas.

Se puede ver que la que la variable “Heating value” toma valores grandes en comparación con las demás.

Se estudian las variables respuestas escaladas y su distribución:

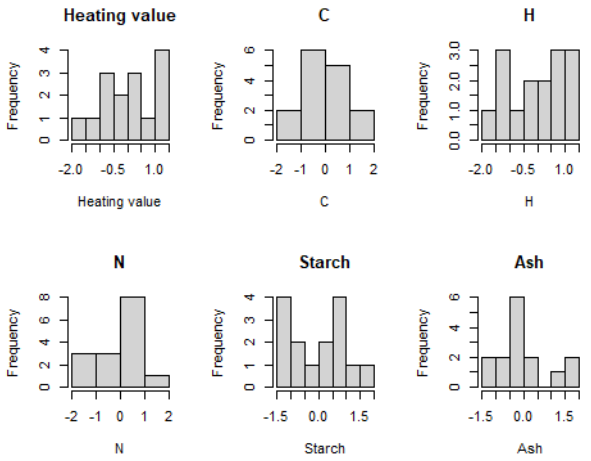

Figura 4.27: Histogramas.

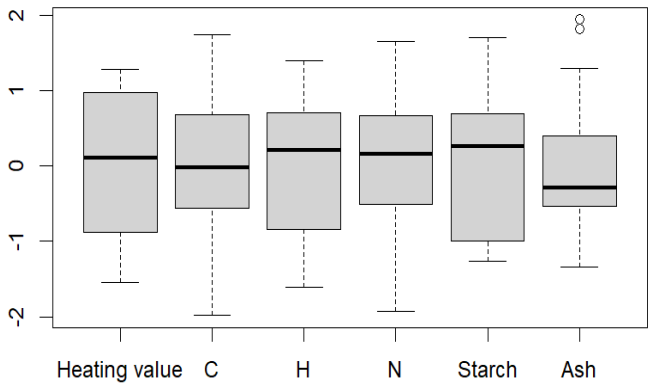

Figura 5.28: Diagramas de cajas.

Para ver si es mejor utilizar datos escalados o no escalados se realiza regresión PLS ajustada con el algoritmo O-PLS.

Se calcula la media de errores de cada variable por componente para poder alcanzar el modelo que mejor ajuste a todas las variables respuesta.

Ajuste del algoritmo O-PLS con datos escalados

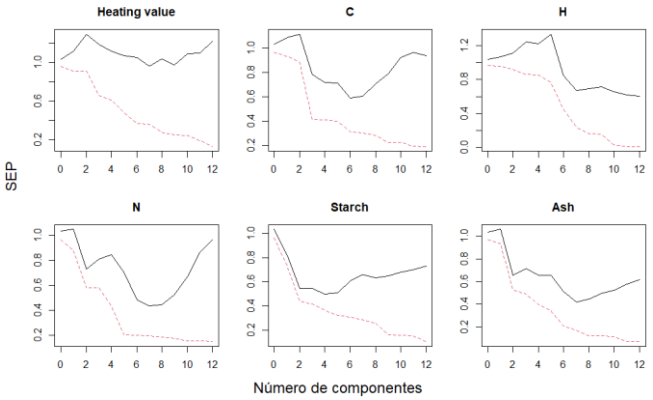

Figura 4.29: Representación por variable del valor SEP con diferente número de componentes mediante validación doblemente cruzada.

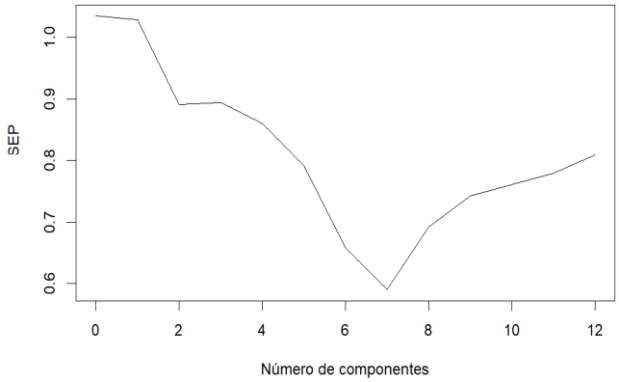

Figura 4.30: Representación de la media del valor SEP de cada variable con diferente número de componentes.

El mejor modelo ajustado con el algoritmo O-PLS y la variable respuesta escalada es aquel con 2 componentes. El error de predicción es de 0.8911413.

Ajuste del algoritmo O-PLS con datos no escalados

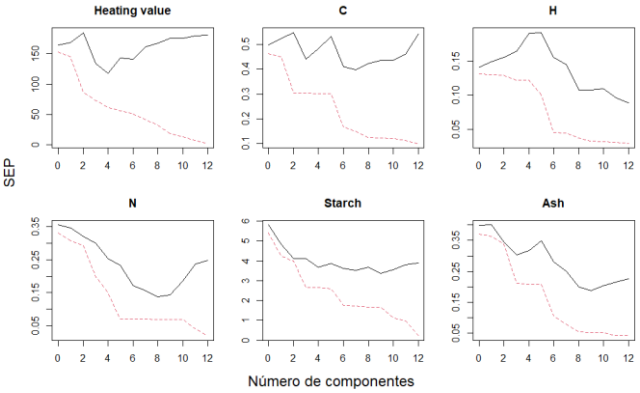

Figura 4.31: Representación por variable del valor SEP con diferente número de componentes mediante validación doblemente cruzada.

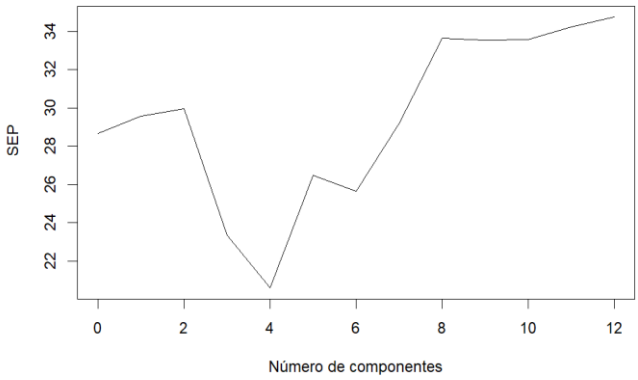

Figura 4.32: Representación de la media del valor SEP de cada variable con diferente número de componentes.

El mejor modelo ajustado con el algoritmo O-PLS y la variable respuesta no escalada es aquel con 4 componentes y su error de predicción es de 20.58871.

PLS trabaja mejor con las respuestas escaladas, se llega a la misma conclusión si se comparan el resto de algoritmos, por lo que a partir de ahora se utilizarán esas variables respuesta.

Ajuste del algoritmo Kernel

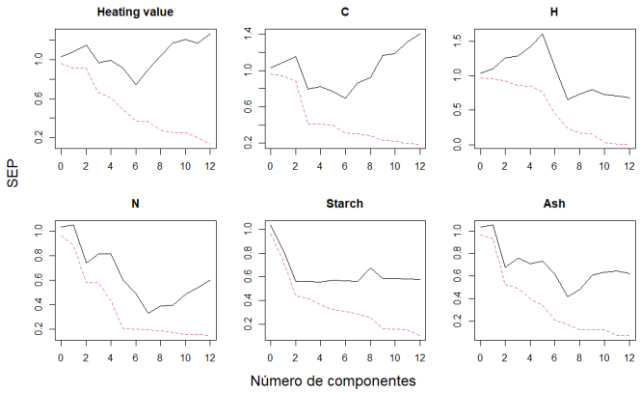

Figura 4.33: Representación por variable del valor SEP con diferente número de componentes mediante validación doblemente cruzada.

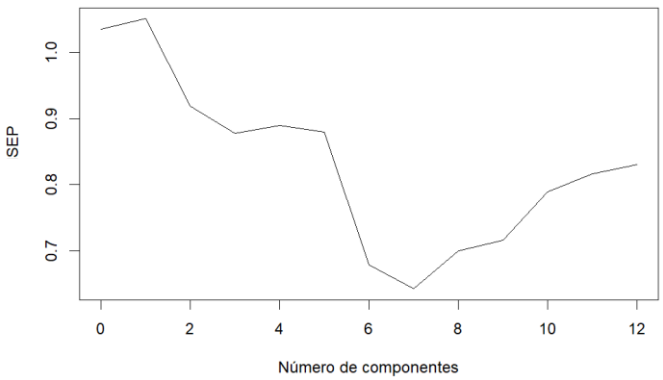

Figura 4.34: Representación de la media del valor SEP de cada variable con diferente número de componentes.

El mejor modelo ajustado con el algoritmo Kernel es aquel con 3 componentes.

Ajuste del algoritmo SIMPLS

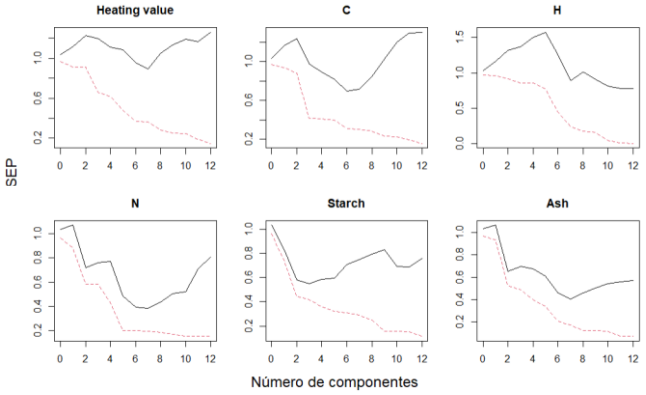

Figura 4.35: Representación por variable del valor SEP con diferente número de componentes mediante validación doblemente cruzada.

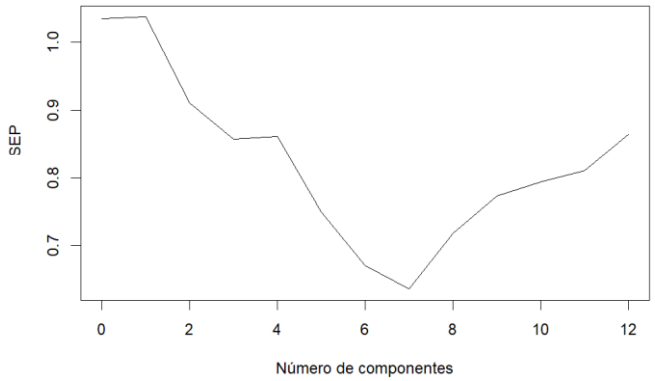

Figura 4.36: Representación de la media del valor SEP de cada variable con diferente número de componentes.

El mejor modelo con el algoritmo SIMPLS es aquel con 3 componentes.

	SIMPLS	Kernel	O-PLS
SEP	0.8568746	0.8781026	0.8911413

Tabla 4.3: Estimación del error estándar de predicción para cada algoritmo.

Por lo tanto, se ajusta un modelo 2PLS mediante el algoritmo SIMPLS con 3 componentes, ya que es el modelo que menor SEP ofrece.

Para comprobar cómo ajusta el modelo obtenido a la variable respuesta, se realizan pruebas con los datos, dividiendo estos en un conjunto de entrenamiento y otro de prueba y ajustando el modelo óptimo hallado anteriormente.

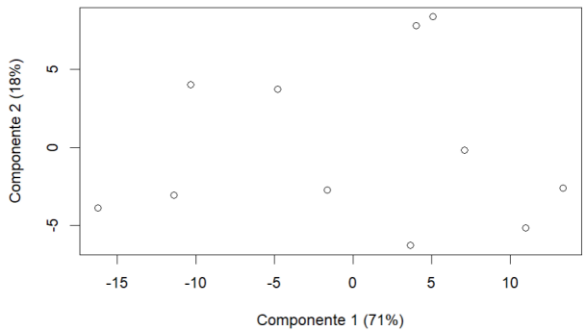

Figura 4.37: Representación de los scores de las dos primeras componentes.

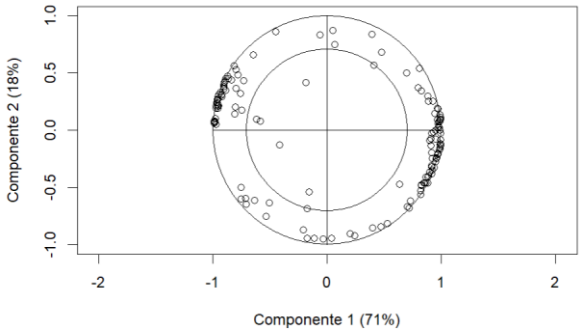

Figura 4.38: Representación de las correlaciones de las dos primeras componentes.

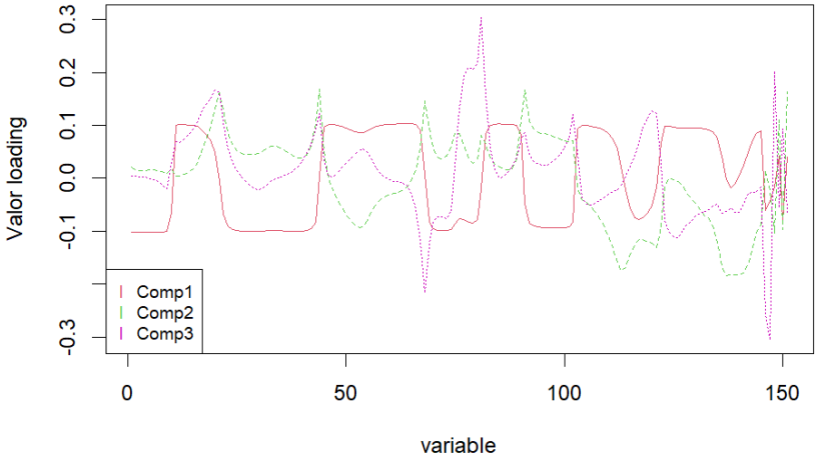

Figura 4.39: Representación de los loadings de las 3 componentes.

Se realizan predicciones y se comparan con los valores observados.

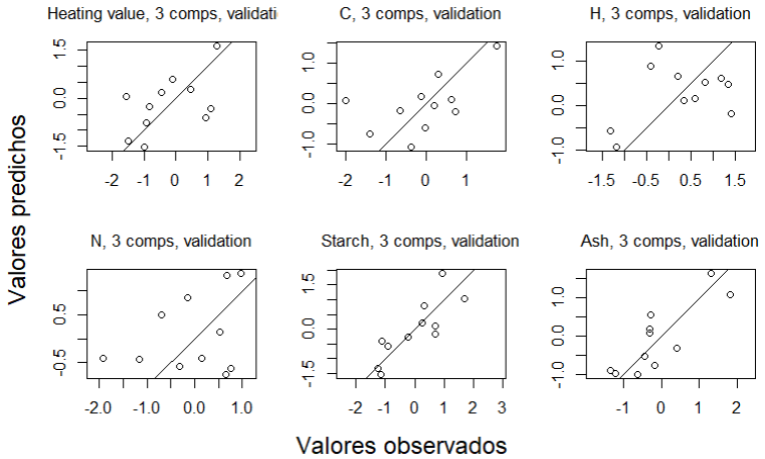

Figura 4.39: Valores observados contra valores predichos de cada variable respuesta.

	Heating value	C	H	N	Starch	Ash
SEP	1.195508	0.7566593	1.137042	0.9139621	0.6233376	0.6912288

Tabla 4.4: Estimación del valor SEP para la predicción de cada variable respuesta.

El modelo obtiene de media un error de 0.8862895.

Se compara el ajuste que hace el método PLS con el que haría PCR.

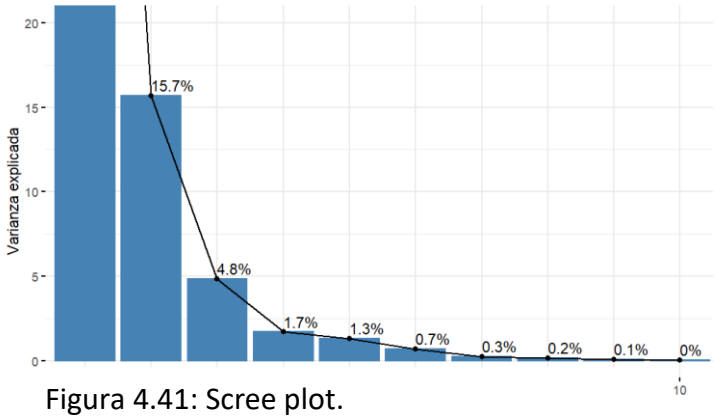

Figura 4.41: Scree plot.

Se ajusta regresión PCR con 3 componentes.

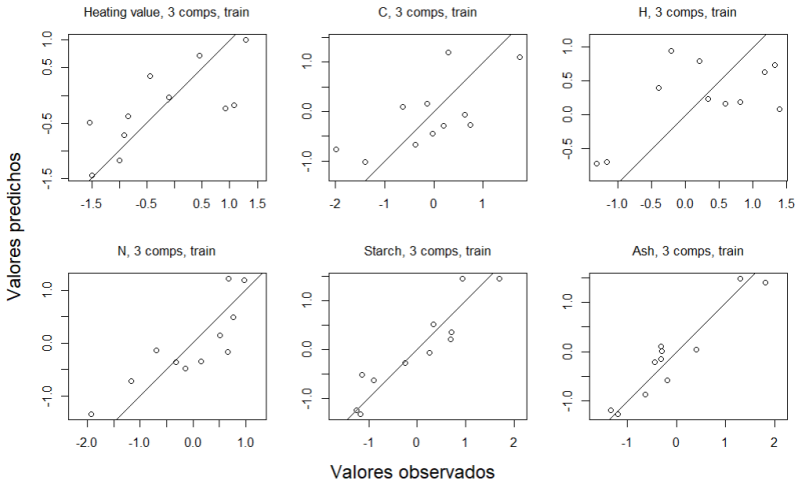

Figura 4.42: Valores observados contra valores predichos de cada variable respuesta.

	Heating value	C	H	N	Starch	Ash
SEP	1.379559	0.9967431	1.347803	0.5289272	0.5593389	0.7168927

Tabla 4.5: Estimación del valor SEP para la predicción de cada variable respuesta.

De media el error de esta regresión es de 0.921544, este valor es mayor que el modelo óptimo obtenido con regresión PLS, que era el algoritmo SIMPLS con 3 componentes, el cual resultaba en un error de 0.8862895.

Nos podemos fijar también que en el ajuste del modelo PCR, aparte de tener mayor error medio, la única variable de la que ofrece un error menor es N.

4.3.2 Radiación infrarroja cercana

Para 166 purés de fermentación alcohólica de diferentes materias primas (centeno, trigo y maíz) se disponen 235 variables (X) que contienen las primeras derivadas de los valores de absorción de radiación infrarroja cercana (NIR) a 1115-2285 mm. Encontramos dos variables respuesta, la concentracion de glucosa y de etanol en g/L.

Se procede a estudiar las variables respuesta y su distribución:

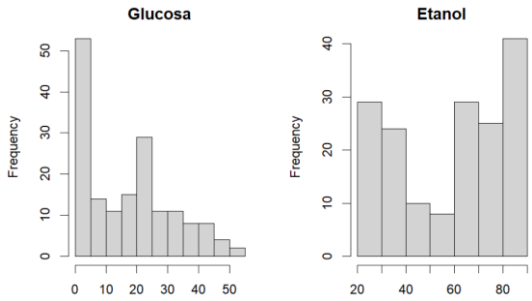

Figura 4.43: Histogramas.

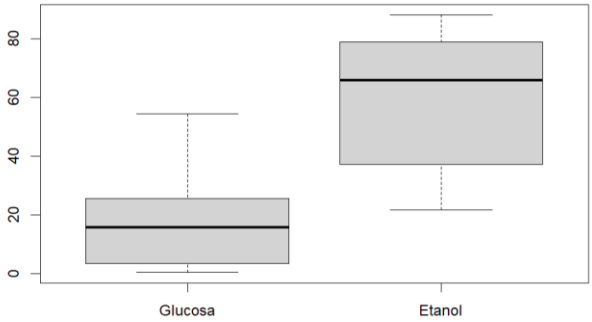

Figura 4.44: Diagramas de cajas.

La variable Etanol toma valores más grandes que la variable Glucosa.

Se busca el algoritmo óptimo para ajustar un modelo de regresión PLS con el menor error.

Tras ajustar los tres algoritmos, lo óptimo sería tomar 4 componentes para realizar el mejor ajuste.

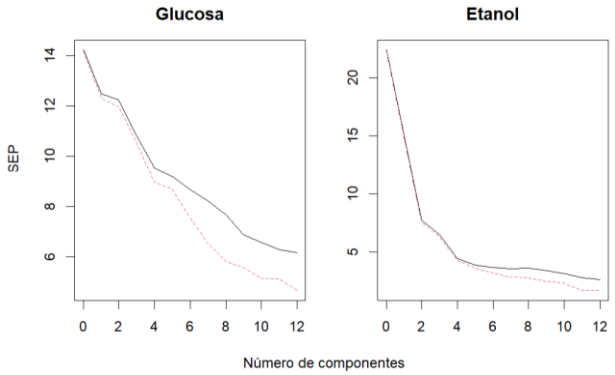

Figura 4.45: Representación por variable del valor SEP con diferente número de componentes mediante validación doblemente cruzada.

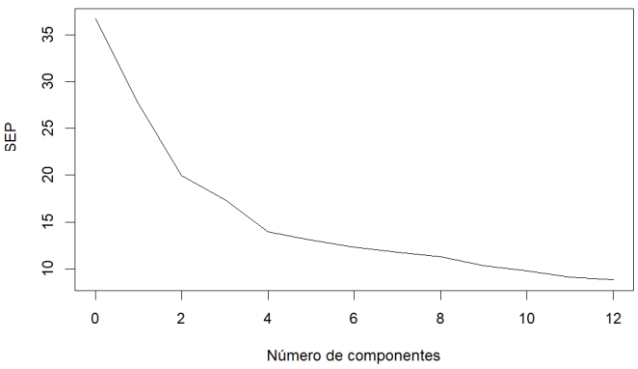

Figura 4.46: Representación de la media del valor SEP de cada variable con diferente número de componentes.

	SIMPLS	Kernel	O-PLS
SEP	6.973879	7.034282	6.97219

Tabla 4.6: Estimación del error estándar de predicción para cada algoritmo con 4 componentes.

Se ajusta un modelo 2PLS con el algoritmo O-PLS con 4 componentes, ya que es el modelo que menor SEP ofrece aunque no exista mucha variación del error entre algoritmos.

Para comprobar cómo ajusta el modelo a la variable respuesta, se realizan pruebas con los datos, dividiendo estos en un conjunto de entrenamiento y otro de prueba y ajustando el modelo óptimo hallado anteriormente.

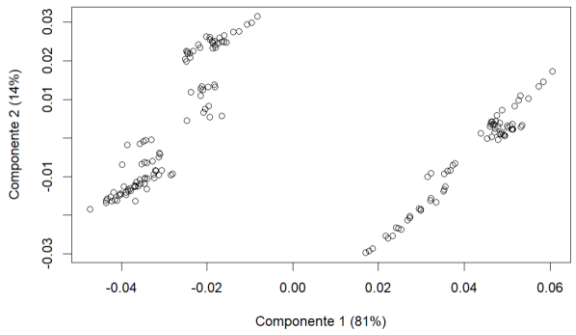

Figura 4.47: Representación de los scores de las dos primeras componentes.

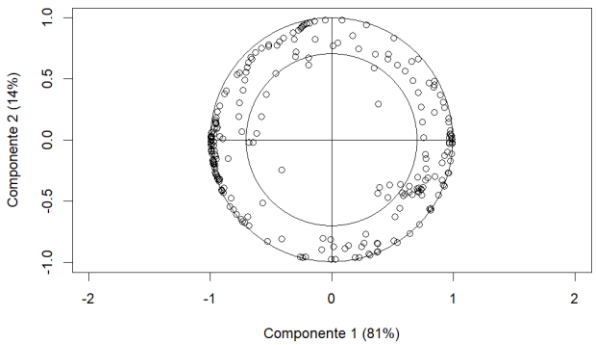

Figura 4.48: Representación de las correlaciones de las dos primeras componentes.

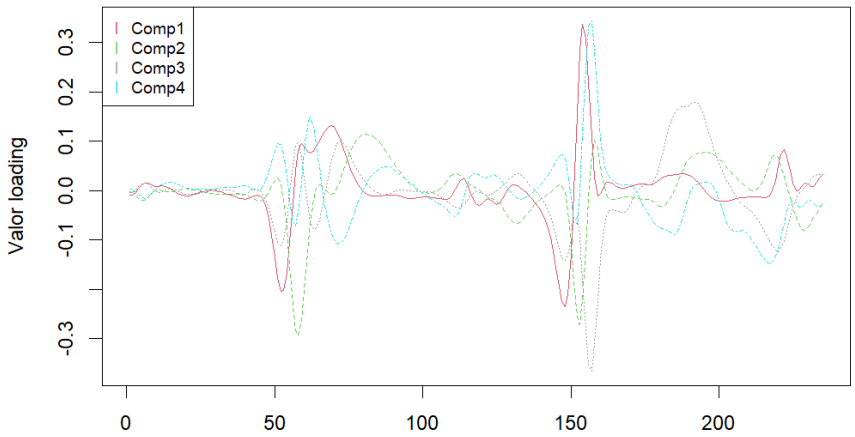

Figura 4.49: Representación de los loadings de las 4 componentes.

Se realizan predicciones y se comparan con los valores observados.

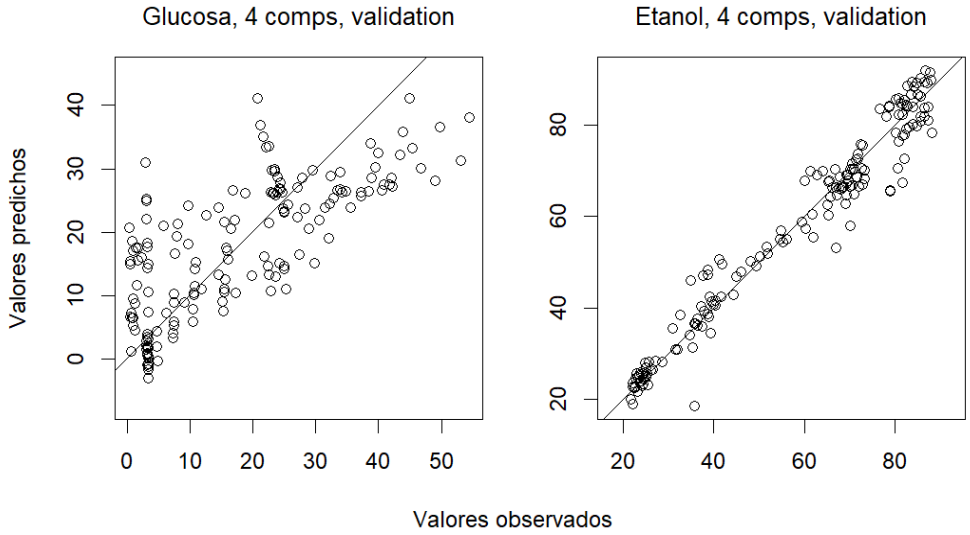

Figura 4.50: Valores observados contra valores predichos de cada variable.

	Glucosa	Etanol
SEP	8.977	4.28

Tabla 4.7: Estimación del valor SEP para la predicción de cada variable respuesta.

Tanto en el gráfico de reales contra predichos como en el error calculado vemos que este modelo predice mejor Etanol que Glucosa.

Se compara el ajuste que hace el método PLS con el que haría PCR.

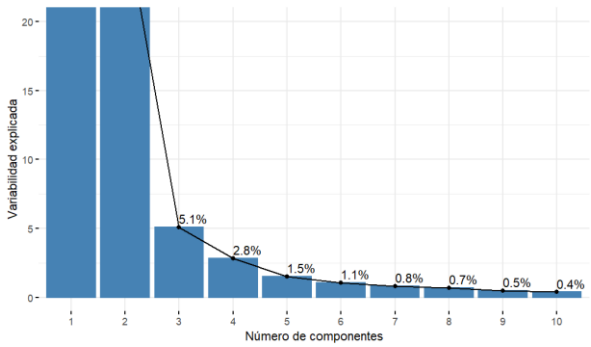

Figura 4.51: Scree plot.

El modelo PCR óptimo sería con 3 componentes.

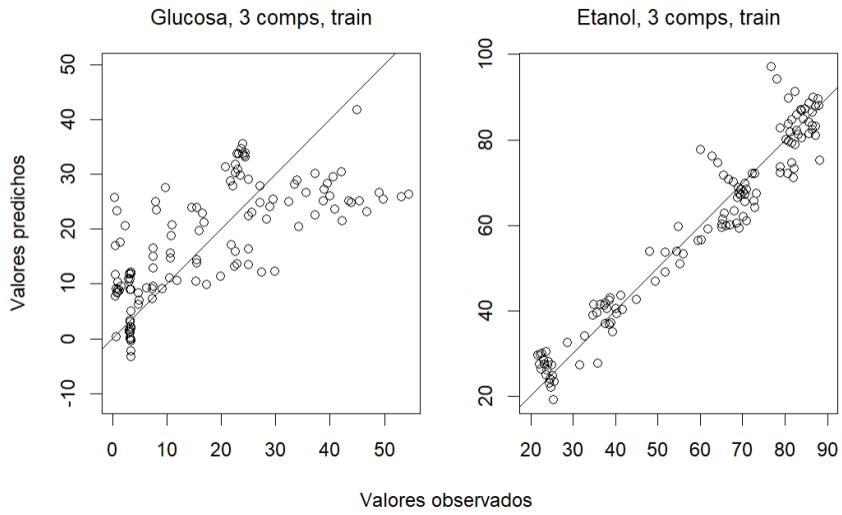

Figura 4.52: Valores observados contra valores predichos de cada variable.

	Glucosa	Etanol
SEP	9.045715	4.783345

Tabla 4.8: Estimación del valor SEP para la predicción de cada variable respuesta.

Si ajustamos el modelo de regresión PCR también se comete menos error al predecir Etanol, pero ambos errores son mayores que los del modelo de regresión PLS, siendo el error medio de este de 6.982115.

4.4 PLS-DA

4.4.1 Vidrios arqueológicos

Este conjunto de datos muestra 13 medidas para 180 recipientes de vidrio arqueológicos de diferentes grupos. Se toman 13 medidas de cada recipiente, todas estas medidas son variables numéricas.

Existen cuatro tipos de cristal a los que se refiere numéricamente del 1 al 4. Para cada recipiente de cristal se indica a qué grupo pertenece.

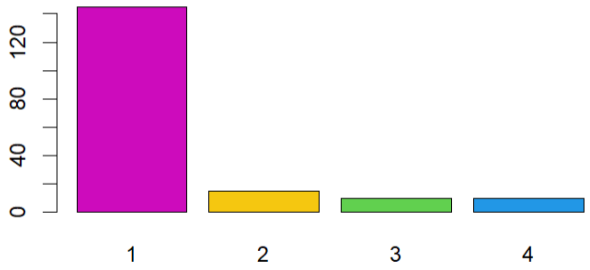

Figura 4.53: Distribución de los grupos.

Cabe recalcar, que el grupo con más observaciones es el 1, con 145. El resto de grupos son menos predominantes, 15 observaciones pertenecen al grupo 2, mientras que al grupo 3 y 4 le corresponde a cada uno 10 obervaciones.

Observamos cómo son las correlaciones entre las variables a estudiar en cado uno de los componentes. Aquellas variables que se encuentran más cercanas en el círculo están más correladas.

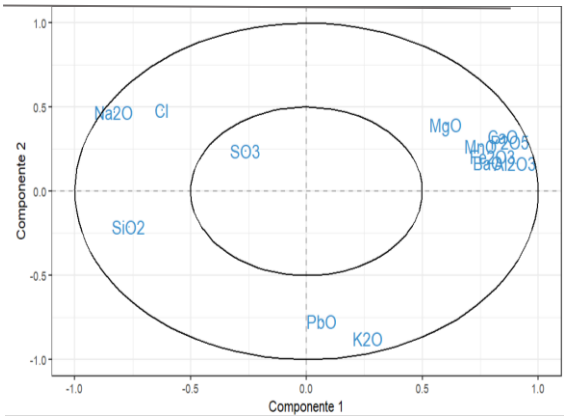

Figura 4.54: Círculo de correlación en dos componentes.

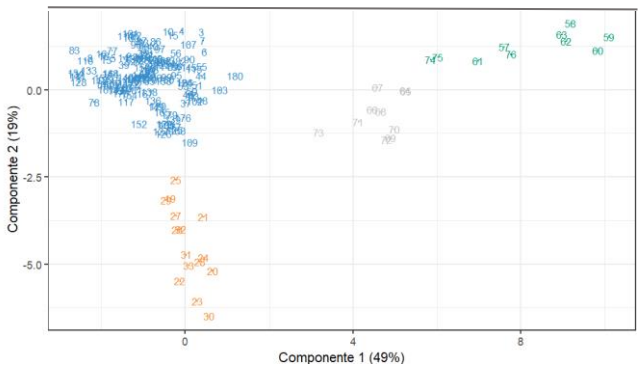

Figura 4.55: Representación de observaciones en dos componentes.

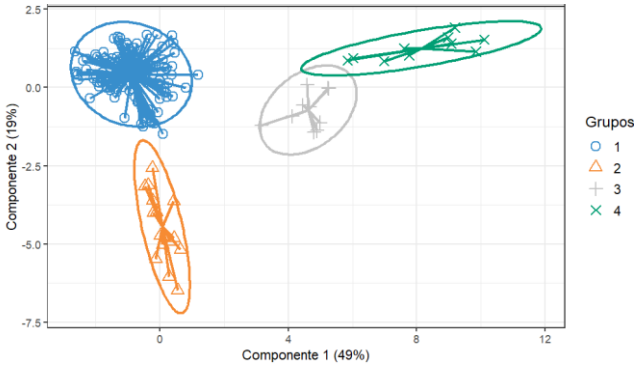

Figura 4.56: Elipsoides de confianza 95%.

Se puede ver como se diferencia cada clase de cristal gracias a la división por componentes. Se eligen 2 para poder verlo bien representado. Como se puede observer, el primer componente explica un 49% de la variabilidad mientras que el segundo explica un 19%. En total en esta representación se está explicando un 68% de la variabilidad.

En cuanto a las elipsodes podemos ver que el grupo 1 y 2 están completamente diferenciados mientras que en el 3 y 4 sus elipsoides de sobreponen dando lugar a posibles errores en la clasificación.

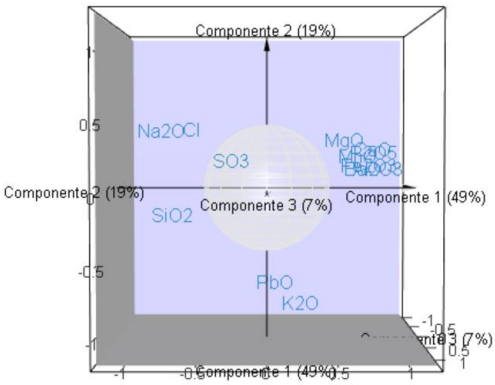

Figura 4.57: Círculo de correlación en tres componentes.

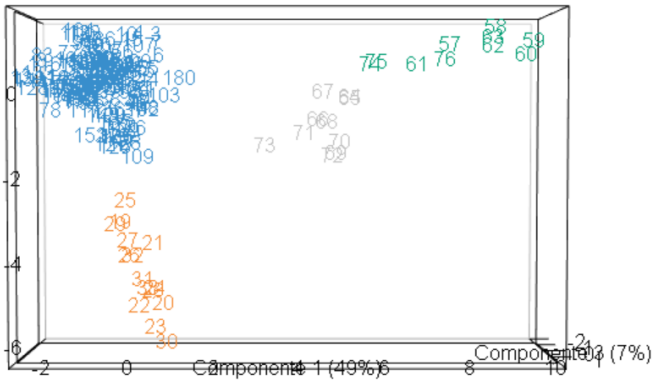

Figura 4.58: Representación de observaciones en tres componentes.

Se puede apreciar como con 3 componentes se distinguen mejor los grupos. El tercer componente explica un 7% de la variabilidad por lo que con 3 componentes la variabilidad total explicada sería de un 75%, lo que sería ya un porcentaje alto.

Se decide el número óptimo de componentes para el modelo mediante validación cruzada.

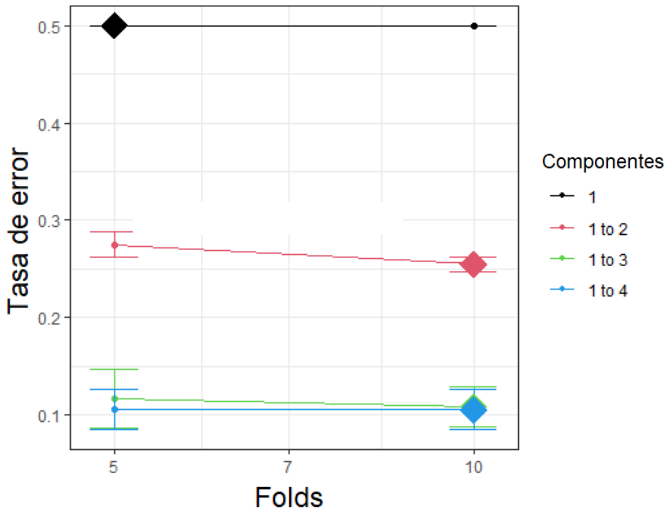

Figura 4.59: Evolución del error dependiendo del número de componentes.

Como se puede ver en el gráfico, a medida que aumenta el número de componentes disminuye el error, pero le aporta más complejidad. Tras realizar las pruebas sobre la tasa de error y ver en el gráfico que 3 es el primer número de componentes a partir del cual el error no disminuye drásticamente, el modelo óptimo sería PLS-DA con 3 componentes.

El error de generalización ajustando este modelo al conjunto de datos mediante validación cruzada es de 0.1625.

Se pretende comparar la clasificación que hace el método PLS-DA con el que haría LDA.

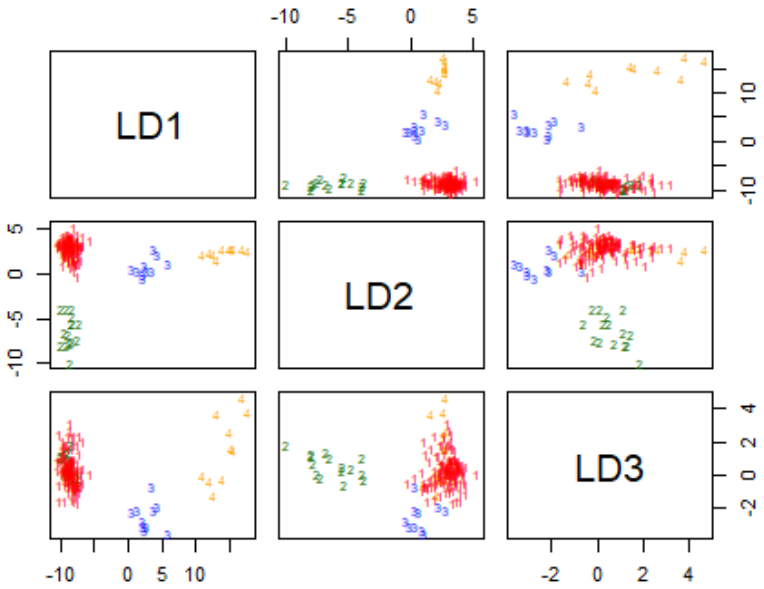

Figura 4.60: Representación del clasificador LDA.

El error de generalización con el método LDA mediante validación cruzada se estima a 0.004166667.

En este caso se puede apreciar que el método de discriminación lineal ofrece un menor error de predicción que PLS. Estos resultados parecen razonables, ya que PLS, como se mostró anteriormente, funciona mejor que otros discriminantes en conjuntos de alta dimensión y en este conjunto de datos se disponen pocas observaciones y variables.

4.4.2 Plantas de hyptis

Se disponen datos relacionados con la planta de hyptis suaveolens. El conjunto de datos está formado por 30 observaciones y 7 variables con medidas de estas plantas. Se pretende distinguir la localización de estas plantas, variable que toma valores East-high, East-low, North y South dependiendo de estas 7 variables. Se puede encontrar una variable auxiliar llamada grupo, que contiene un valor del 1 al 4 dependiendo de esta localización.

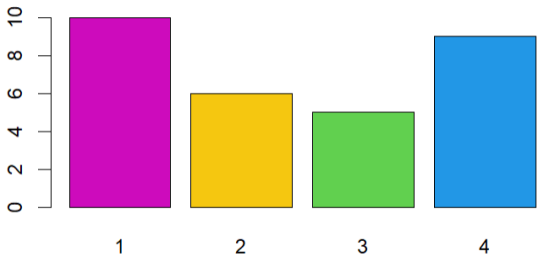

Figura 4.60: Distribución de los grupos.

Se pueden encontrar 10 observaciones pertenecientes al grupo 1, 6 pertenecientes al 2, 5 al 3 y 9 al 4.

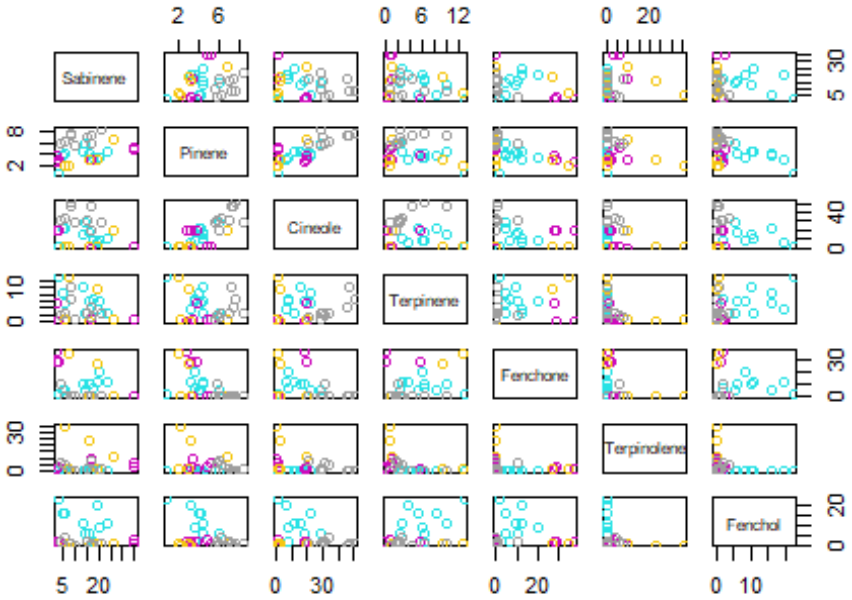

Figura 4.62: Gráfico de pares.

No se diferencian muy bien los grupos, los colores están entremezclados.

Se estudian las correlaciones entre las variables en cada uno de los componentes. Aquellas que se encuentran más cerca en el círculo estan más correladas.

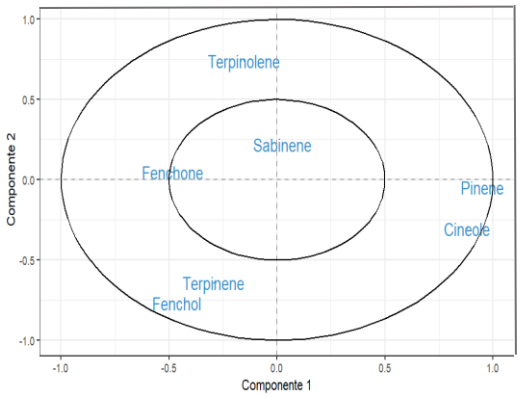

Figura 4.63: Círculo de correlación en dos componentes.

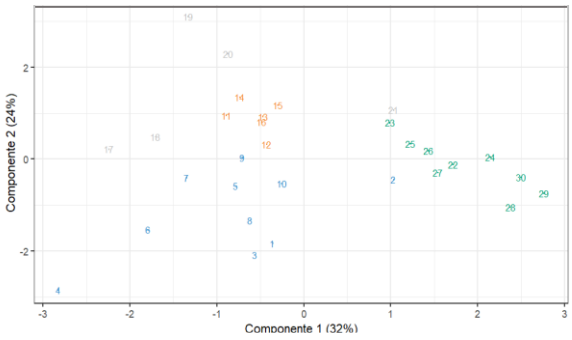

Figura 4.64: Representación de observaciones en dos componentes.

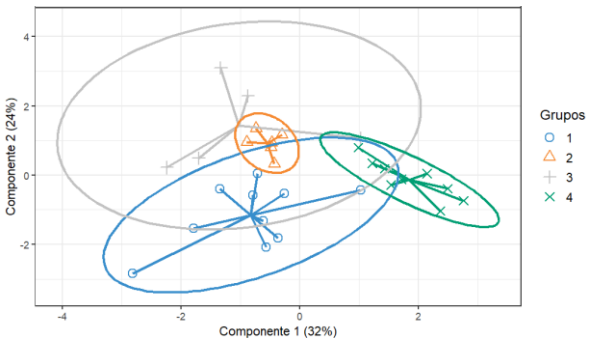

Figura 4.65: Elipsoides de confianza 95%.

Los grupos se encuentran representados muy juntos y hay posibilidades de que el modelo se pueda confundir al clasificar. Con dos componentes se explica un 56% de la variabilidad. La manera de que estos grupos se pudieran distinguir mejor sería utilizando más componentes.

Las elipsoides de cada grupo se sobreponen, incluso la elipsoide del grupo 2 esta contenida en la eliposide del grupo 3.

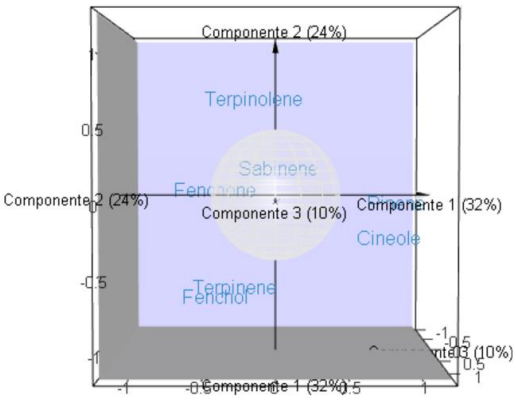

Figura 4.66: Círculo de correlación en tres componentes.

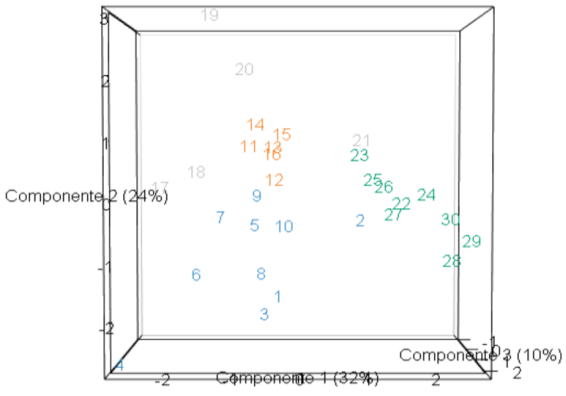

Figura 4.67: Representación de observaciones en tres componentes.

Los grupos se distinguen algo mejor pero aun así no lo hacen correctamente. El tercer componente explica un 10% de la variabilidad, por lo que con 3 componentes la variabilidad total explicada seria de un 66%.

Para decidir el número óptimo de componentes que se deberían escoger se realiza validación cruzada.

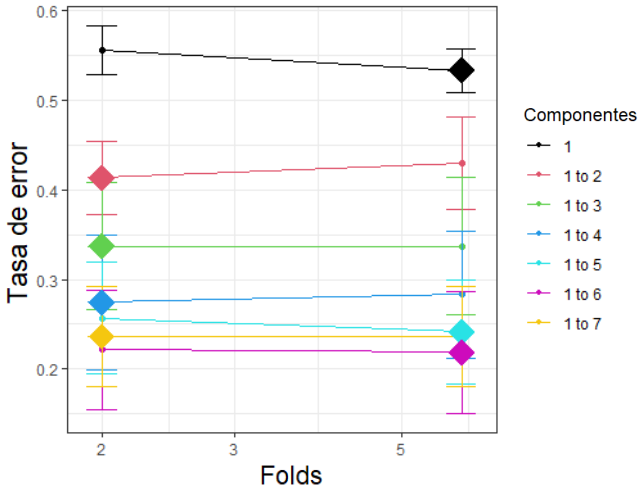

Figura 4.68: Evolución del error dependiendo del número de componentes.

Como se puede apreciar en el gráfico, a medida que aumenta el número de componentes disminuye el error, pero le aporta más complejidad. Tras realizar las pruebas sobre la tasa de error y ver en el gráfico que 5 es el primer número de componentes a partir del cual el error no disminuye drásticamente, el modelo óptimo sería PLS-DA con 5 componentes.

El error de generalización ajustando este modelo al conjunto de datos mediante validación cruzada es de 0.74166667.

Se quiere comparar la clasificación que hace el método PLS-DA con el que haría LDA.

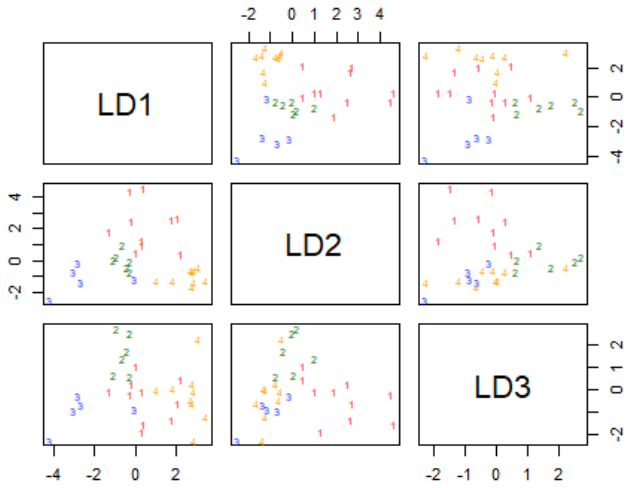

Figura 4.69: Representación del clasificador LDA.

Si se halla el error de generalización con este método mediante validación cruzada este se estima a 0.225.

Podemos ver que es un conjunto de datos difícil de clasificar, los grupos están muy entremezclados. Aún cometiendo errores altos, LDA clasifica mejor que PLS-DA. Al igual que pasaba con el anterior conjunto de datos, este conjunto no es el ideal para un clasificador como PLS-DA porque se presentan escasas observaciones y variables.

4.4.3 Masas espectrales

Nos encontramos ante una masa espectral de 600 compuestos químicos, donde 300 contienen estructura Phenyl (grupo 1) y 300 compuestos no tienen esta escructura (grupo 2). La masa espectral ha sido transformada a 658 variables conteniendo los rasgos. Los dos grupos están codificados como -1 (grupo 1) y +1 (grupo 2).

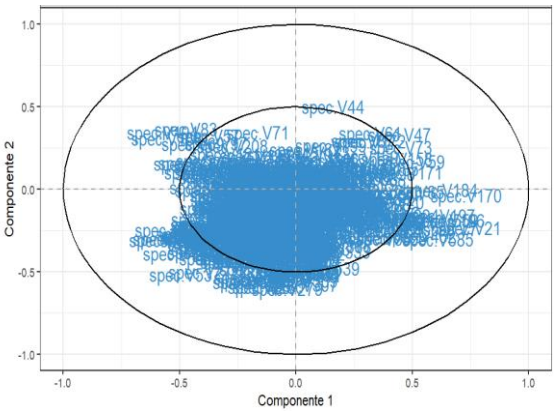

Figura 4.70: Círculo de correlación en dos componentes.

Se puede apreciar que las variables explicativas están muy correladas.

Para un primer acercamiento a los datos se ajusta un modelo PLS-DA con dos componentes.

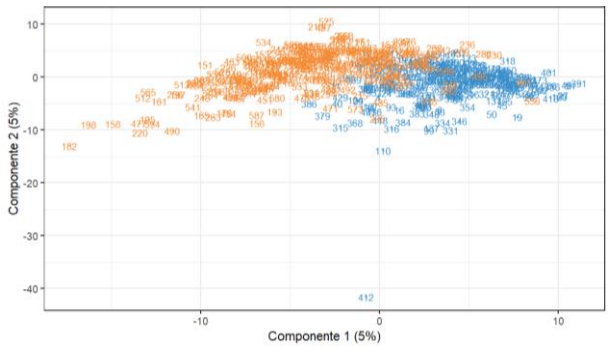

Figura 4.71: Representación de observaciones en dos componentes.

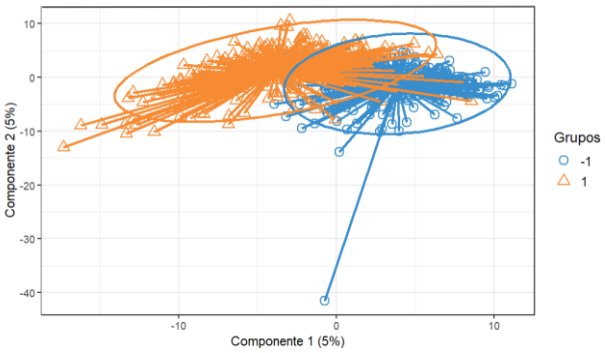

Figura 4.72: Elipsoides de confianza 95%.

Se puede apreciar que los grupos están solapados y que los primeros dos componentes explican poca variabilidad, cada uno un 5%. No se distinguen bien debido a la alta correlación entre las variables.

Los elipsoides están solapados, lo que podría ser problemático.

Se probará a ajustar un modelo con 3 componentes.

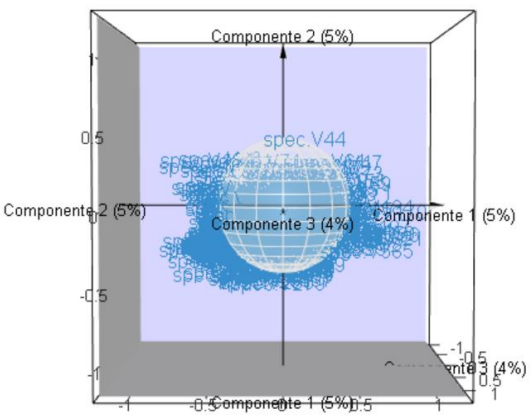

Figura 4.73: Círculo de correlación en tres componentes.

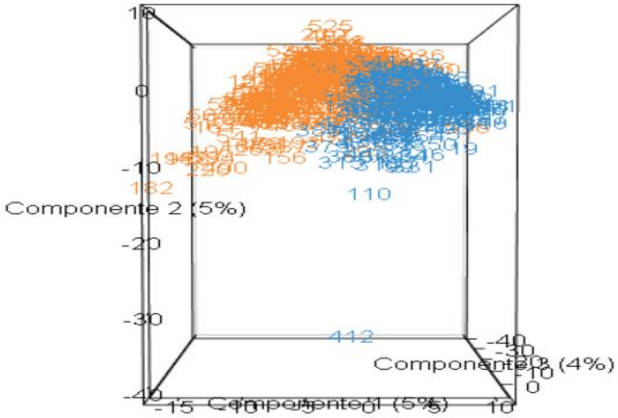

Figura 4.74: Representación de observaciones en tres componentes.

Los grupos siguen sin distinguirse. El tercer componente explica un 4% de la variabilidad. El modelo con 3 componentes explicaría un 14% de la variabilidad.

El número óptimo de componentes se decide mediante validación cruzada.

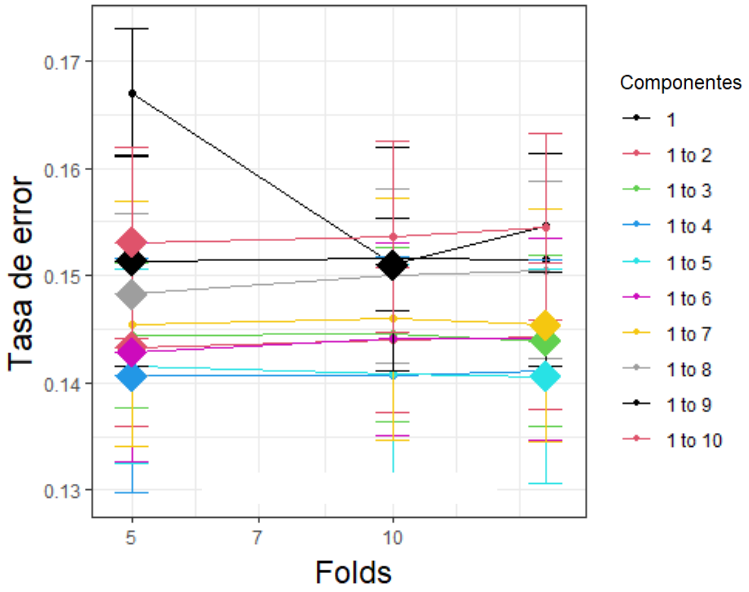

Figura 4.75: Evolución del error dependiendo del número de componentes.

En consecuencia de que los componentes expliquen poca variabilidad, una vez que se llega a 2 componentes hay poca diferencia en el error entre estos modelos. Por lo tanto, el modelo óptimo sería con 2 componentes ya que el error es parecido a con más componentes siendo el modelo menos complejo. El error de generalización ajustando este modelo al conjunto de datos mediante validación cruzada es de 0.17375.

Se pretende comparar la clasificación que hace el método PLS-DA con el que haría LDA.

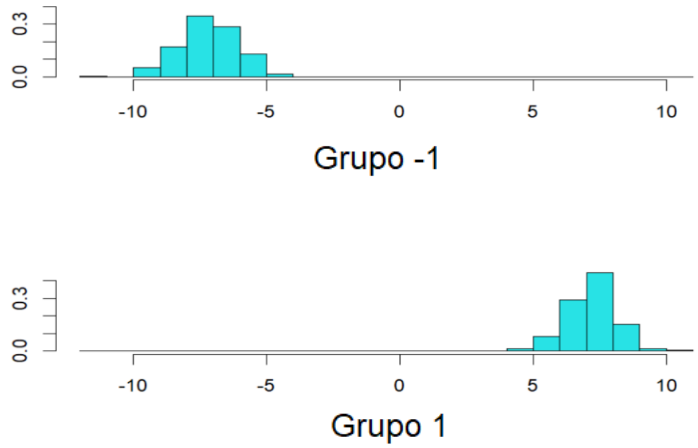

Figura 4.76: Representación del clasificador LDA.

Si se hallara el error de generalización con este método mediante validación cruzada este se estima a 0.33333.

El conjunto de datos del que se dispone es ideal para utilizar PLS-DA, ya que existe un gran volumen de variables explicativas. Se puede ver reflejado en que el error de generalización que ofrece PLS-DA con 2 componentes es menor que LDA. Se ha comprobado la existencia de alta correlación entre variables y de multicolinealidad. PLS es un método ideal para casos así ya que tiene en cuenta la multicolinealidad.

Capítulo 5

Conclusiones

La Quimiometría es un campo en el que aparecen datos de alta dimensionalidad. El PLS es uno de los métodos por excelencia para sacar valor a este tipo datos, que típicamente incluyen numerosas variables bastante correladas. De esta forma, el PLS permite realizar predicciones (calibración en terminología de Quimiometría) de menor variabilidad y llegar a errores de predicción menores.

Este método PLS ha crecido bastante en relación con la Quimiometría aunque también puede ser aplicable en otros muchos campos. En este trabajo nos hemos centrado en la aplicación del método PLS a la Quimiometría.

Para muchos métodos de análisis multivariante esta correlación fuerte entre variables es un serio problema que obliga a realizar una selección previa de las variables. Por otro lado, el PLS ajusta modelos correctamente incluso con variables correladas, viéndose así su potencial en comparación con otros métodos multivariantes para este tipo de casos tan comunes en Quimiometría. El método PLS reduce los predictores a un conjunto más pequeño de componentes incorrelados y sobre este nuevo conjunto de componentes realiza la regresión mínimo-cuadrática habitual.

El PLS, además, no precisa de rígidas suposiciones sobre los datos. En los ejemplos prácticos en esta memoria se han probado conjuntos de datos muy dispares, tanto en dimensionalidad como respecto a sus características distribucionales. Esta poca imposición de restricciones hace que este método sea sencillo de manejar. Este es un punto fuerte de este método en datos químicos, donde se dispone de datos muy complejos y donde el PLS puede ser visto como una metodología sencilla de aplicar.

En este trabajo se han probado diferentes algoritmos para ajustar la regresión PLS, como son los algoritmos Kernel, O-PLS y SIMPLS. Tras estudiar su eficacia en diferentes conjuntos de datos, se puede llegar a la conclusión de que los errores de predicción que ofrecen estos algoritmos son generalmente muy parecidos. No hay ninguno que sea claramente mejor al resto. Con los conjuntos de datos de una única variable respuesta (PLS1) se ha podido apreciar el interés de aplicar un método de regresión PLS que sea más robusto, ya que PLS habitual (no robusto) no tiene la capacidad de ajustar bien los datos cuando existen datos atípicos o puntos de influencia. En cambio, conseguimos solucionar este problema y realizar buenas predicciones si utilizamos un enfoque PLS robusto.

Otra conclusión a la que se puede llegar tras este trabajo es que PLS, aún modelando conjuntos de datos de grandes dimensiones, no suele necesitar muchas componentes para obtener un buen ajuste. El método de Regresión en Componentes Principales PCR y el método de Mínimos Cuadrados Parciales PLS, en este aspecto de reducción de la dimensionalidad, suelen ser ambos bastante razonables en Quimiometría. En los ejemplos prácticos se han comparado los errores entre ambos métodos viendo que son en general parecidos. En algunos casos el error ofrecido por PCR era incluso menor que algunos de los algoritmos PLS pero en estos conjuntos de datos siempre ha existido un algoritmo PLS que ofrecía un menor error.

	SIMPLS	Kernel	O-PLS	PCR	PLS-Robusto
Compuestos aromáticos policíclicos	12.55486	12.49963	12.42558	20.02531	**6.031733**
Datos de cenizas	147.2416	145.771	146.7714	155.8926	**125.4899**
Datos de cereales	**0.8568746**	0.871020	0.8911413	0.921544	
Radiación infrarroja cercana	6.973879	7.034282	**6.97219**	6.982115	

Tabla 6.1: Comparación de errores en regresión.

A la hora de clasificar, se han utilizado conjuntos de datos que no tienen grandes dimensiones salvo el último. Con ellos, se ha podido apreciar que en estos casos PLS-DA no clasifica tan bien como otros clasificadores multivariantes como LDA. Sin embargo, el conjunto de datos de masas espectrales tenia una dimensionalidad muy alta y muchas variables correladas, aquí se ha podido comprobar que PLS-DA consigue mejores modelos que clasificadores como LDA.

	PLS-DA	LDA
Vidrios en arqueología	0.1625	**0.004166667**
Plantas de hyptis	0.74166667	**0.225**
Masas espectrales	**0.17375**	0.333

Tabla 6.2: Comparación de errores en clasificación.

Capítulo 6

Bibliografía

[1] Gaviria, C. *Regresión por Mínimos Cuadrados Parciales PLS Aplicada a Datos Variedad Valuados*. Tesis, Universidad Nacional de Colombia, 2016.

[2] Gil, C. *Análisis de Componentes Principales (PCA).* RPubs, Junio, 2018. https://rpubs.com/Cristina_Gil/PCA

[3] Hastie, T., Tibshirani, R. y Friedman, J.. *The Elements of Statistical Learning.* Vol.3, 80-82.

[4] Le Cao, K., Dejean, S. y Abadi, A. Chapter 4 PLS - Discriminant Analysis. *mixOmix vignette*. https://mixomicsteam.github.io/Bookdown/index.html

[5] Manne, R. Analysis of two partial least squares algorithms for multivariate regression. *Chemometrics and Intelligen Laboratory Systems*, Vol. 2, 187- 197,1987.

[6] Márquez, C. *Modelo de regresión PLS.* Universidad de Sevilla. 2017.

[7] Mateos-Aparicio, G. y Caballero, J. La regresión por mínimos cuadrados parciales: orígenes y evolución. *Historia de la probabilidad y la estadística (IV)*, Vol 29, 441–447. 2009.

[8] Mevik, B., Wehrens, R. *Package 'pls'*. Cran R, August 2020.

[9] Peinado, A. *Quimiometría. Aplicación y desarrollo de técnicas quimiométricas*. Vol. 4, 57-68.

[10] Vaquerizo, R. Regresión PLS con R. *Análisis y decisión*. 2014.

[11] Varmuza, K. y Filzmoser, P. *Introduction to Multivariate Statistical Analysis in Chemometrics*. Editorial CRC Press, 2009.

[12] Varmuza, K. y Filzmoser, P. *Package 'chemometrics'*. Cran R, March 2017.

[13] Vega-Vilca, J. y Guzmán, J. Regresión PLS y PCA como solución al problema de la multicolinealidad en regresión múltiple. *Revista de matemática: Teoría y aplicaciones* 18(1):9-20, 2011.

[14] Wold, S., Sjöström, M., Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, Vol.58, 109–130, 2001.