Departamento de Matemáticas, Facultad de Ciencias
Universidad Autónoma de Madrid

# Mínimos cuadrados parciales (PLS) en regresión lineal múltiple y funcional

**Trabajo de fin de grado**

Grado en Matemáticas

*Autor:* David del Val

*Tutor:* Jose Ramón Berrendero

Curso 2022-2023

## Resumen

Las técnicas de reducción de dimensionalidad son herramientas extremadamente útiles a la hora de trabajar con conjuntos de datos con muchas variables. En el análisis de datos funcionales, cobran aún más importancia al permitir la proyección de datos infinito-dimensionales a espacios de dimensión finita. Uno de estos métodos con mejores resultados es el de mínimos cuadrados parciales (PLS). Sin embargo, el planteamiento de PLS es bastante diferente al de otros métodos de reducción de dimensionalidad como componentes principales (PCA) o correlaciones canónicas (CCA). Este trabajo busca introducir PLS y explorar sus propiedades desde un nuevo punto de vista. En particular, se busca que esta explicación de PLS sea fácil de entender para un lector familiarizado con estos otros métodos pero que no necesariamente conoce PLS. Asimismo, se busca consolidar resultados ya conocidos pero cuyos análisis o demostraciones no son fáciles de encontrar en la literatura.

Uno de los campos donde PLS brilla es en la estimación de modelos de regresión. En consecuencia, cubriremos los pasos necesarios para aplicar el método de reducción PLS a modelos de regresión. En particular, nos centraremos en los modelos con respuesta escalar, donde PLS es equivalente a regularizar el estimador de mínimos cuadrados. Adicionalmente, se presentarán las modificaciones necesarias para tratar datos funcionales y resultados numéricos que respaldan las propiedades teóricas obtenidas.

## Abstract

Dimensionality reduction techniques have become very useful tools when it comes to working with high-dimensionality datasets. When it comes to functional data analysis (FDA), they are even more powerful. They can be used to project data from an infinite-dimensional (functional) data space to a finite dimensional space, making it considerably easier to analyze. One of the methods used to accomplish these goals is partial least squares (PLS). However, the rationale behind it is rather different when compared to other dimensionality reduction methods such as principal components (PCA) or canonical correlations (CCA). This work intends to introduce PLS and explore its properties from a new point of view. In doing so, we hope to produce an easy-to-follow explanation for those already familiar with PCA or CCA, but not necessarily with PLS. Moreover, this document consolidates results that are already known but whose analysis or proofs are hard to find in the PLS literature.

One of the fields where PLS shines the most is adjusting linear regression models. Therefore, we will cover the steps required to adapt the PLS dimensionality reduction technique to perform linear regression. In particular, the scalar response model is of particular interest. In that case we will prove that PLS is equivalent to regularizing the ordinary least squares estimator (OLS). Moreover, the changes that PLS needs to handle functional data will be covered and numerical results will be presented to showcase the derived properties.

# Índice general

# CAPÍTULO 1
# Introducción

Dada la gran riqueza de datos disponibles en la actualidad, la gran mayoría de conjuntos de datos cuentan con muchas variables. Esta abundancia de información puede resultar muy útil para extraer comportamientos o patrones que, de otro modo, no serían distinguibles. Sin embargo, trabajar con datos con una gran dimensionalidad introduce nuevas dificultades.

Al trabajar en espacios con una dimensionalidad alta, serán necesarias muchas observaciones para poder obtener estimaciones o inferencias adecuadas. Dado que el número de observaciones necesarias crece rápidamente, en muchas ocasiones, los datos disponibles estarán demasiado dispersos. Este problema se conoce como la maldición de la dimensionalidad. Asimismo, el procesamiento de los datos es más lento y costoso cuanto mayor es el número de variables consideradas.

Para mitigar estos problemas, se emplean métodos de reducción de la dimensionalidad. Estos métodos buscan transformar los datos llegando a otro conjunto de datos cuya dimensión sea más próxima a lo que podríamos llamar su dimensión intrínsica, es decir, al mínimo número de variables necesarias para recoger toda la información presente en el conjunto de datos.

Mínimos cuadrados parciales (PLS) es uno de estos métodos. Su origen se remonta a la definición de NIPALS en Noonan y Wold (1977). No obstante, la concepción habitual de PLS es notablemente distinta. Se trata de un método que se ha simplificado con el paso del tiempo y actualmente denotamos PLS a lo que podría considerarse un caso particular del algoritmo original.

Por otro lado, hay planteamientos alternativos de PLS que concuerdan en el criterio a seguir para obtener la primera componente pero divergen a partir de ese punto. En Wegelin (2000) se puede encontrar una relación de los métodos más tradicionales mientras que De Jong (1993) introduce el método SIMPLS basado en un planteamiento diferente pero que resulta ser equivalente bajo ciertas hipótesis

Sin embargo, gran parte de la literatura que explora PLS es bastante farragosa, complicando la comprensión del método suponiendo que la mayoría de los lectores ya están familiarizados con el mismo. En el presente documento buscamos introducir PLS de forma distinta a lo que es habitual en la mayoría de artículos. En primer lugar, realizaremos un estudio de PLS comparándolo con otros métodos de reducción de la dimensionalidad más habituales como PCA y CCA. Este análisis motivará una definición formal del problema de optimización que se resuelve en PLS. Una vez hayamos definido este problema, introduciremos la herramienta que se utiliza para resolverlo: el algoritmo NIPALS.

Este planteamiento es contrario a la mayoría de los análisis de PLS como por ejemplo Rosipal y Krämer (2005) o Höskuldsson (1988). Sin embargo, nos permitirá entender el algoritmo NIPALS rápidamente al haber establecido con anterioridad la base teórica sobre la que se asienta.

La aplicación de PLS a problemas de regresión (cubierto en Höskuldsson (1988) y Wegelin (2000)) es una extensión del método de reducción de dimensionalidad. Se trata de una serie de regresiones lineales entre las componentes extraídas al aplicar la reducción de dimensionalidad. No obstante, veremos como, cuando la respuesta es escalar, se llega a un método (PLS1) mucho más simple. De hecho, PLS1 puede ser visto como un método de regularización del ajuste de una regresión por mínimos cuadrados ordinarios (ver Rosipal y Krämer (2005) o Eldén (2004)).

Por otro lado, el análisis de datos funcionales es un campo donde los métodos de reducción de la dimensionalidad son especialmente útiles. La naturaleza infinito-dimensional de los datos da lugar a problemas sobredeterminados y conjuntos de datos que presentan multicolinealidad. Nuestro objetivo será estudiar cómo se pueden adaptar las técnicas de PLS finito-dimensionales para mitigar estos problemas.

Al igual que en el caso finito-dimensional, hay varios métodos que producen resultados diferentes. Por un lado, en Preda y Saporta (2005) se introducen planteamientos similares a los métodos multivariates basados en NIPALS y, en Febrero-Bande, Galeano et al. (2017), se comparan con la regresión PCA aplicada a datos funcionales. Por otro lado, Delaigle y Hall (2012) presenta un planteamiento alternativo con propiedades distintas. En nuestro caso nos centraremos en los primeros, a fin de mantener la coherencia con las secciones anteriores.

## 1.1. Estructura del documento

El objetivo del segundo capítulo es plantear PLS partiendo de su comparación con otros métodos de reducción de dimensionalidad más conocidos. A continuación, se desarrolla el marco teórico del algoritmo NIPALS y se llega al algoritmo NIPALS para la resolución de problemas de reducción de la dimensionalidad mediante PLS.

El tercer capítulo explora la extensión del método anterior a problemas de regresión y las diversas propiedades que pueden obtenerse cuando la respuesta del modelo es escalar. De forma similar, el cuarto capítulo trata la adaptación de PLS para tratar datos funcionales, en particular, en problemas de regresión.

A fin de proporcionar ejemplos prácticos, el quinto capítulo incluye simulaciones donde PLS es puesto a prueba en diversos escenarios, tanto con datos finito-dimensionales como funcionales. Asimismo, se estudia el rendimiento de PLS en función de la estructura de autovalores de la matriz de covarianzas de la variable regresora.

Finalmente, el sexto capítulo contiene las conclusiones y comentarios finales mientras que los apéndices recogen demostraciones y algoritmos adicionales.

# CAPÍTULO 2

# Métodos lineales de reducción de dimensionalidad

A la hora de reducir la dimensión de uno o varios vectores aleatorios, es posible proceder de diferentes formas. Una primera aproximación al problema podría consistir en seleccionar solo ciertas componentes de los vectores. Este enfoque se conoce como selección de variables y puede ser efectivo en problemas en los que haya muchas características redundantes o irrelevantes.

Sin embargo, en otras situaciones es preferible no desechar componentes de estos vectores. En su lugar, se opta por técnicas que utilicen proyecciones de los vectores originales en espacios diferentes. Este primer capítulo se centra en analizar algunos de estos métodos basados en transformaciones lineales.

Es importante notar que, en lo sucesivo, se supondrá que los vectores aleatorios a analizar tienen media 0 a fin de simplificar el análisis. Con esto en mente, consideramos $X$ un vector aleatorio $M$-dimensional e $Y$ un vector aleatorio $D$-dimensional. Si queremos reducir la dimensión del espacio a $L$ coordenadas, basta seleccionar $L$ parejas de vectores: $\{(\mathbf{w}_l, \mathbf{c}_l)\}_{l=1}^{L} \subset \mathbb{R}^M \times \mathbb{R}^D$ y considerar las proyecciones de ambos vectores aleatorios a lo largo de estas direcciones:

$$X \longrightarrow \left(\mathbf{w}_1^\top X, \quad \mathbf{w}_2^\top X, \quad \ldots \quad \mathbf{w}_L^\top X\right),$$
$$Y \longrightarrow \left(\mathbf{c}_1^\top Y, \quad \mathbf{c}_2^\top Y, \quad \ldots \quad \mathbf{c}_L^\top Y\right).$$

Por tanto, los diferentes métodos lineales de reducción de dimensión se reducen a diferentes criterios para seleccionar las direcciones $\{\mathbf{w}_l\}$ y $\{\mathbf{c}_l\}$ a lo largo de las que proyectar los vectores iniciales. En lo sucesivo, denominaremos componentes a estas proyecciones. En esta sección se analizarán tres criterios diferentes para elegir estas direcciones, que dan lugar a tres de los métodos de reducción de dimensionalidad más utilizados.

En primer lugar, el análisis de componentes principales (PCA) se basa en elegir direcciones ortogonales maximizando la varianza de la proyección del vector aleatorio en cada paso. Dado un vector aleatorio $X$, la primera dirección seleccionada es la dirección $\mathbf{w}_1$ que satisface

$$\mathbf{w}_1^{\text{PCA}} = \underset{\|\mathbf{w}\|=1}{\arg\max} \operatorname{var}(\mathbf{w}^\top X),$$

donde $\|\cdot\|$ denota la norma euclídea.

Al contrario que PCA, los dos siguientes métodos se aplican a la vez a dos vectores aleatorios, produciendo parejas de direcciones. El análisis de correlaciones canónicas

(CCA) busca maximizar la correlación en vez de la varianza. Por tanto, el primer par de direcciones $(\mathbf{w}_1, \mathbf{c}_1)$ se selecciona de la siguiente forma:

$$(\mathbf{w}_1^{\text{CCA}}, \mathbf{c}_1^{\text{CCA}}) = \underset{\|\mathbf{w}\|=\|\mathbf{c}\|=1}{\arg\max}\, \text{corr}^2(\mathbf{w}^\top X, \mathbf{c}^\top Y).$$

Finalmente, mínimos cuadrados parciales (PLS) busca dos direcciones tales que se maximice la covarianza de las proyecciones, es decir, el criterio para seleccionar las dos primeras direcciones $(\mathbf{w}_1, \mathbf{c}_1)$ es

$$(\mathbf{w}_1^{\text{PLS}}, \mathbf{c}_1^{\text{PLS}}) = \underset{\|\mathbf{w}\|=\|\mathbf{c}\|=1}{\arg\max}\, \text{cov}^2(\mathbf{w}^\top X,\, \mathbf{c}^\top Y).$$

Una vez elegida la primera dirección en cada uno de los métodos es necesario añadir restricciones adicionales para definir las siguientes direcciones. En el caso de PCA y CCA, es posible obtener un problema de minimización claro para cada componente a extraer añadiendo restricciones sobre las sucesivas direcciones. Estos problemas se definen en (2.1) y (2.2).

$$\begin{aligned} \mathbf{w}_l^{\text{PCA}} = &\arg\max_{\mathbf{w}} \text{var}(\mathbf{w}^\top X) \text{ sujeto a} \\ &\|\mathbf{w}\| = 1, \\ &\mathbf{w} \perp \mathbf{w}_j^{\text{PCA}} \quad \text{para todo } j < l. \end{aligned} \tag{2.1}$$

$$\begin{aligned} (\mathbf{w}_l^{\text{CCA}}, \mathbf{c}_l^{\text{CCA}}) = &\arg\max_{(\mathbf{w},\mathbf{c})} \text{corr}^2(\mathbf{w}^\top X, \mathbf{c}^\top Y) \text{ sujeto a} \\ &\|\mathbf{w}\| = \|\mathbf{c}\| = 1, \\ &\text{corr}(\mathbf{w}^\top X, (\mathbf{w}_j^{\text{CCA}})^\top X) = 0 \quad \text{para todo } j < l, \\ &\text{corr}(\mathbf{c}^\top Y, (\mathbf{c}_j^{\text{CCA}})^\top Y) = 0 \quad \text{para todo } j < l. \end{aligned} \tag{2.2}$$

Sin embargo, en el caso de PLS, la forma de extraer componentes sucesivas no es tan clara. De hecho, si bien hasta este punto hemos introducido PLS como un solo método, puede ser considerado una familia de métodos. A la hora de extraer cada componente, todos estos métodos tienen como objetivo maximizar la covarianza pero las restricciones que se establecen entre las sucesivas componentes son diferentes.

Un planteamiento muy habitual en los métodos PLS es calcular las componentes de forma iterativa, realizando algún tipo de modificación a los datos entre cada iteración. Esta modificación se suele denominar "deflación" en la literatura. Intuitivamente, la deflación es una operación que tiene como objetivo sustraer la parte de los datos que ya puede ser explicada con los componentes ya extraídos. Con la excepción de PLS-SVD (definido en Wegelin (2000)) la deflación está presente de una forma u otra en todas las variantes de PLS más utilizados. Estos métodos se pueden clasificar en dos grandes grupos:

- Los métodos derivados del algoritmo original de NIPALS (*Non-Linear Iterative Partial Least Squares*) descrito en Noonan y Wold (1977). Estos métodos tienen

una formulación claramente iterativa y se basan en efectuar deflaciones sobre los datos directamente para obtener las componentes sucesivas. Es decir, para calcular la segunda componente se busca maximizar la covarianza de nuevo pero modificando los vectores aleatorios con los que se trabaja. En definitiva, la segunda componente componente se extraerá como

$$(\mathbf{w}_2^{\text{PLS}}, \mathbf{c}_2^{\text{PLS}}) = \underset{\|\mathbf{w}\|=\|\mathbf{c}\|=1}{\arg\max}\, \text{cov}^2(\mathbf{w}^\top X_1,\, \mathbf{c}^\top Y_1),$$

donde $X_1$ es el resultado de la de la primera deflación sobre $X$ e $Y_1$ es el resultado de la primera deflación sobre $Y$. Analizaremos estas deflaciones en la sección 2.3.

- SIMPLS[1], introducido en De Jong (1993). En este caso, la restricción toma la misma forma que en CCA, exigiendo que las sucesivas proyecciones sean incorreladas. Sin embargo, como este algoritmo está diseñado para su uso solo en problemas de regresión, no se establece ninguna restricción sobre el vector $Y$. Exploraremos la aplicación de PLS a problemas de regresión en el capítulo 3. La formulación de SIMPLS es puramente muestral y se basa en aplicar deflaciones a la matriz de covarianzas muestrales cruzadas ($\mathbf{X}^\top\mathbf{Y}$) donde $\mathbf{X}$ e $\mathbf{Y}$ son las matrices de datos de los vectores aleatorios $X$ e $Y$, respectivamente.

En el resto del presente documento nos centraremos en los métodos derivados de NIPALS. Sin embargo, en De Jong (1993), se lleva a cabo una comparación de estos frente a SIMPLS a la hora de resolver un problema de regresión y se llega a la conclusión de que son equivalentes si la respuesta es escalar.

Como resulta aparente a simple vista, hay una gran similitud entre CCA y PLS. De hecho, si expandimos la covarianza como el producto de las desviaciones típicas y la correlación, obtenemos que PLS busca componentes que maximicen

$$\text{cov}^2(\mathbf{w}^\top X, \mathbf{c}^\top Y) = \text{var}(\mathbf{w}^\top X)\, \text{corr}^2(\mathbf{w}^\top X,\, \mathbf{c}^\top Y)\, \text{var}(\mathbf{c}^\top Y).$$

Es decir, PLS tiene en cuenta tanto la información contenida en la varianza de las proyecciones de ambos vectores aleatorios como en la correlación entre ellas. Se puede entender, por tanto, como una mezcla de los criterios de CCA y PCA.

De hecho, se pueden definir métodos intermedios entre CCA y PLS considerando el problema de optimización dado por

$$\underset{\|\mathbf{w}\|=\|\mathbf{c}\|=1}{\text{máx}} \frac{\text{cov}^2(\mathbf{w}^\top X, \mathbf{c}^\top Y)}{\Big((1-\gamma_X)\,\text{var}(\mathbf{w}^\top X) + \gamma_X\Big)\Big((1-\gamma_Y)\,\text{var}(\mathbf{c}^\top Y) + \gamma_Y\Big)},$$

donde $0 \leq \gamma_X, \gamma_Y \leq 1$ pueden considerarse términos de regularización. Es inmediato ver que con $\gamma_X = 1,\ \gamma_Y = 1$ aparece el criterio de PLS y, con $\gamma_X = 0, \gamma_Y = 0$, el de CCA. Las propiedades con este planteamiento se han explorado en Vinod (1976) bajo el nombre de *canonical ridge analysis*.

[1]Los autores eligieron este acrónimo inspirados en la siguiente frase: *straightforward implementation of a statistically inspired modification of the PLS method according to the simple concept given in Table 1.* Donde la tabla 1 en De Jong (1993) resume los pasos claves del método.

## 2.1. Maximización de la varianza (PCA)

El problema de PCA se introduce en (2.1) y consiste en maximizar la varianza de $\mathbf{w}^\top X$. Como $X$ está centrado, $\mathbf{w}^\top X$ también y tenemos:

$$\text{var}(\mathbf{w}^\top X) = \mathbb{E}((\mathbf{w}^\top X)^2) = \mathbb{E}(\mathbf{w}^\top X X^\top \mathbf{w}) = \mathbf{w}^\top \Sigma_{XX} \mathbf{w},$$

donde $\Sigma_{XX}$ denota la matriz de covarianzas de $X$. Así, nuestro problema de maximización se reduce a

$$\mathbf{w}^{\text{PCA}} = \arg\max_{\mathbf{w}} \mathbf{w}^\top \Sigma_{XX} \mathbf{w} \qquad \text{sujeto a} \quad \|\mathbf{w}\| = 1.$$

Este problema puede resolverse utilizando multiplicadores de Lagrange. Para ello, se define la función auxiliar $\mathcal{L}$ como

$$\mathcal{L}(\mathbf{w}) = \mathbf{w}^\top \Sigma_{XX} \mathbf{w} - \lambda(\mathbf{w}^\top \mathbf{w} - 1)$$

y buscamos los valores en los que su gradiente se anula:

$$\frac{\partial}{\partial \mathbf{w}^\top} \mathcal{L}(\mathbf{w}) = 2\Sigma_{XX}\mathbf{w} - 2\lambda\mathbf{w} = 0 \implies \Sigma_{XX}\mathbf{w} = \lambda\mathbf{w}.$$

Por tanto, vemos que, en los extremos de la cantidad a maximizar, $\mathbf{w}$ es un autovector de $\Sigma_{XX}$. Analizando la cantidad original a maximizar en estos caso, obtenemos que $\mathbf{w}^\top \Sigma_{XX} \mathbf{w} = \lambda$ bajo la restricción $\|\mathbf{w}\| = 1$. En consecuencia, como estamos buscando su máximo, necesariamente, $\mathbf{w}$ tiene que ser un autovector asociado al autovalor máximo de $\Sigma_{XX}$.

Asimismo, los vectores que maximicen la misma cantidad pero sujetos a la restricción de ser ortogonales a las direcciones anteriores son los subsecuentes autovectores de $\Sigma_{XX}$. Por tanto, las direcciones extraídas por PCA son los autovectores de $\Sigma_{XX}$ ordenados de forma decreciente por la magnitud de sus autovalores.

## 2.2. Maximización de la correlación (CCA)

En el caso de CCA, el problema que buscamos resolver es (2.2). Al igual que para PCA, buscamos una caracterización de las direcciones de proyección como autovectores de alguna matriz. Para ello, aprovechamos que la correlación es independiente de la escala. Por tanto, podemos resolver el problema con cualquier restricción en la norma de $\mathbf{w}$ y $\mathbf{c}$ y, a posteriori, normalizar estos vectores a norma 1 si queremos recuperar la restricción original.

A fin de simplificar el problema, vamos a cambiar la restricción de la norma por $\mathbf{w}^\top \Sigma_{XX} \mathbf{w} = \mathbf{c}^\top \Sigma_{YY} \mathbf{c} = 1$. Este cambio resulta ser particularmente beneficioso porque permite que la condición de norma unidad tenga una forma muy similar a la de ortogonalidad entre las componentes. De hecho, desarrollando la correlación entre componentes, obtenemos

$$\begin{aligned} \text{corr}(\mathbf{w}^\top X, (\mathbf{w}_j^{\text{CCA}})^\top X) = 0 &\iff \text{cov}(\mathbf{w}^\top X, (\mathbf{w}_j^{\text{CCA}})^\top X) = 0 \iff \\ &\iff \mathbb{E}(\mathbf{w}^\top X X^\top \mathbf{w}_j^{\text{CCA}}) = 0 \iff \\ &\iff \mathbf{w}^\top \Sigma_{XX} \mathbf{w}_j^{\text{CCA}} = 0 \end{aligned}$$

y, de forma análoga, se puede llegar a

$$\operatorname{corr}(\mathbf{c}^\top Y, (\mathbf{c}_j^{\text{CCA}})^\top Y) = 0 \iff \mathbf{c}^\top \Sigma_{YY} \mathbf{c}_j^{\text{CCA}} = 0.$$

Por otra parte, utilizando la definición de la correlación en términos de covarianzas y varianzas podemos reescribir la expresión a maximizar. Además, aprovechando que los vectores aleatorios están centrados, podemos obtener

$$\begin{aligned}\operatorname{corr}(\mathbf{w}^\top X, \mathbf{c}^\top Y) &= \frac{\operatorname{cov}(\mathbf{w}^\top X, \mathbf{c}^\top Y)}{\sqrt{\operatorname{var}(\mathbf{w}^\top X)}\sqrt{\operatorname{var}(\mathbf{c}^\top Y)}} = \\ &= \frac{\mathbb{E}((\mathbf{w}^\top X Y^\top \mathbf{c}))}{\sqrt{\mathbb{E}((\mathbf{w}^\top X)^2)}\sqrt{\mathbb{E}((\mathbf{c}^\top Y)^2)}} = \\ &= \frac{\mathbf{w}^\top \Sigma_{XY} \mathbf{c}}{\sqrt{\mathbf{w}^\top \Sigma_{XX} \mathbf{w}}\sqrt{\mathbf{c}^\top \Sigma_{YY} \mathbf{c}}} = \mathbf{w}^\top \Sigma_{XY} \mathbf{c},\end{aligned}$$

donde el último paso se verifica ya que el denominador será siempre 1 como consecuencia de la nueva normalización.

A la luz de estas identidades, es posible replantear el problema (2.2) cambiando la restricción de norma 1 y aplicando el cambio de variable

$$\hat{\mathbf{w}} = \Sigma_{XX}^{1/2} \mathbf{w}, \quad \hat{\mathbf{c}} = \Sigma_{YY}^{1/2} \mathbf{c}. \tag{2.3}$$

Así, se llega al problema

$$\begin{aligned}(\hat{\mathbf{w}}_l^{\text{CCA}}, \hat{\mathbf{c}}_l^{\text{CCA}}) = \underset{(\hat{\mathbf{w}}, \hat{\mathbf{c}})}{\arg\max} (\hat{\mathbf{w}}^\top \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2} \hat{\mathbf{c}}^\top)^2 \text{ sujeto a} \\ \hat{\mathbf{w}}^\top \hat{\mathbf{w}} = \hat{\mathbf{c}}^\top \hat{\mathbf{c}} = 1, \\ \hat{\mathbf{w}} \perp \hat{\mathbf{w}}_j^{\text{CCA}} \quad \text{para todo } j < l, \\ \hat{\mathbf{c}} \perp \hat{\mathbf{c}}_j^{\text{CCA}} \quad \text{para todo } j < l.\end{aligned} \tag{2.4}$$

Como se puede apreciar, hemos obtenido un problema de maximización con restricciones notablemente más sencillas. De hecho, siguiendo un procedimiento similar al de PCA, se puede ver que las soluciones son los autovectores de cierta matriz.

**Proposición 2.1.** *Las soluciones al problema de optimización* (2.4) *se caracterizan del siguiente modo:*

*(a)* $\hat{\mathbf{w}}_l^{CCA}$ *es un autovector de norma unidad asociado al* $l$*-ésismo autovalor (en orden de magnitud decreciente) de la matriz* $\Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX} \Sigma_{XX}^{-1/2}$.

*(b)* $\hat{\mathbf{c}}_l^{CCA}$ *es un autovector de norma unidad asociado al* $l$*-ésismo autovalor (en orden de magnitud decreciente) de la matriz* $\Sigma_{YY}^{-1/2} \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_{YY}^{-1/2}$.

*Demostración.*

Si definimos la función $h(\hat{\mathbf{w}}, \hat{\mathbf{c}}) = \hat{\mathbf{w}}^\top \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2} \hat{\mathbf{c}}$ y $f(t) = t^2$, la función a maximizar es $f(h(\hat{\mathbf{w}}, \hat{\mathbf{c}}))$. Como $f$ es una función par, creciente en $(0, +\infty)$ y decreciente en $(-\infty, 0)$, tenemos que

$$\max_{\Omega} f(h(\hat{\mathbf{w}}, \hat{\mathbf{c}})) = \left( \max \left( \max_{\Omega} h(\hat{\mathbf{w}}, \hat{\mathbf{c}}), \min_{\Omega} h(\hat{\mathbf{w}}, \hat{\mathbf{c}}) \right) \right)^2$$

para cualquier conjunto de argumentos (pares $(\mathbf{w}, \mathbf{c})$) $\Omega$. Por tanto, el problema de maximizar $f \circ h$ se reduce a maximizar y minimizar $h$ en las restricciones dadas. Podemos resolver ambos problemas a la vez mediante los multiplicadores de Lagrange.

Es decir, utilizamos los multiplicadores de Lagrange para calcular los extremos de $h$ sujeta a $\hat{\mathbf{w}}^\top \hat{\mathbf{w}} = \hat{\mathbf{c}}^\top \hat{\mathbf{c}} = 1$. Para ello, definimos la función auxiliar $\mathcal{L}(\hat{\mathbf{w}}, \hat{\mathbf{c}})$:

$$\mathcal{L}(\hat{\mathbf{w}}, \hat{\mathbf{c}}) = \hat{\mathbf{w}}^\top \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2} \hat{\mathbf{c}} - \lambda_w (\hat{\mathbf{w}}^\top \hat{\mathbf{w}} - 1) - \lambda_c (\hat{\mathbf{c}}^\top \hat{\mathbf{c}} - 1).$$

Calculando los gradientes e igualando a 0 llegamos a las dos identidades siguientes

$$\Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2} \hat{\mathbf{c}} = \lambda_w \hat{\mathbf{w}}, \qquad \Sigma_{YY}^{-1/2} \Sigma_{YX} \Sigma_{XX}^{-1/2} \hat{\mathbf{w}} = \lambda_c \hat{\mathbf{c}} \tag{2.5}$$

y juntando ambas identidades deducimos que los vectores $\hat{\mathbf{w}}$ y $\hat{\mathbf{c}}$ tienen que satisfacer las siguientes ecuaciones de autovectores:

$$\begin{aligned} \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX} \Sigma_{XX}^{-1/2} \hat{\mathbf{w}} &= \lambda_w \lambda_c \hat{\mathbf{w}}, \\ \Sigma_{YY}^{-1/2} \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_{YY}^{-1/2} \hat{\mathbf{c}} &= \lambda_w \lambda_c \hat{\mathbf{c}}. \end{aligned} \tag{2.6}$$

Despejando $\hat{\mathbf{w}}$ del miembro derecho de (2.6) y sustituyendo en la restricción $\hat{\mathbf{w}}^\top \hat{\mathbf{w}} = 1$, se obtiene

$$\begin{aligned} 1 = \hat{\mathbf{w}}^\top \hat{\mathbf{w}} &= \hat{\mathbf{w}}^\top \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX} \Sigma_{XX}^{-1/2} \hat{\mathbf{w}} \, \frac{1}{\lambda_w \lambda_c} \\ &\overset{(2.5)}{=} \hat{\mathbf{w}}^\top \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2} \, \hat{\mathbf{c}} \, \lambda_c \, \frac{1}{\lambda_w \lambda_c}. \end{aligned}$$

A continuación, despejando $\lambda_w$, se llega a

$$\lambda_w = \hat{\mathbf{w}}^\top \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2} \, \hat{\mathbf{c}}.$$

Realizando el mismo cálculo con las expresiones análogas para $\hat{\mathbf{c}}$, se puede comprobar que

$$\hat{\mathbf{w}}^\top \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2} \, \hat{\mathbf{c}} = \lambda_c = \lambda_w.$$

Por tanto, hemos obtenido que $h(\hat{\mathbf{w}}, \hat{\mathbf{c}}) = \lambda_c = \lambda_w$ cuando $(\hat{\mathbf{w}}, \hat{\mathbf{c}})$ es un candidato a máximo o mínimo. Podemos utilizar esta identidad para calcular el valor de $f \circ h$ en su máximo:

$$f(h(\hat{\mathbf{w}}, \hat{\mathbf{c}})) = (h(\hat{\mathbf{w}}, \hat{\mathbf{c}}))^2 = \lambda_c \lambda_w = \lambda,$$

donde $\lambda$ es el autovalor de cualquiera de las dos matrices de la expresión (2.6). Por tanto, los vectores que solucionan el problema de maximizar la correlación al cuadrado son los autovectores asociados al autovalor dominante de dichas matrices.

Asimismo, como las dos matrices de (2.6) son simétricas, sus sucesivos autovectores serán ortogonales. En consecuencia, las siguientes componentes de CCA se obtienen eligiendo como direcciones de proyección los sucesivos autovectores de dichas matrices (ordenados de forma decreciente por el autovalor asociado). □

**Proposición 2.2.** *Las soluciones al problema de optimización de CCA original (2.2) se caracterizan del siguiente modo:*

*(a)* $\hat{\mathbf{w}}_l^{CCA}$ *es un autovector de norma unidad asociado al* $l$*-ésismo autovalor (en orden de magnitud decreciente) de la matriz* $\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$.

*(b)* $\hat{\mathbf{c}}_l^{CCA}$ *es un autovector de norma unidad asociado al* $l$*-ésismo autovalor (en orden de magnitud decreciente) de la matriz* $\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1/2}\Sigma_{XY}$.

*Demostración.*

Es una consecuencia directa de la proposición 2.1. Basta deshacer el cambio de variable. □

Es importante notar que los pasos claves que nos llevan a esta caracterización son el cambio en la normalización junto con el cambio de variable en (2.3). Estos dos pasos nos llevan a un problema donde todas las restricciones se pueden escribir solo en términos de las direcciones de proyección (tras el cambio de variable).

## 2.3. Maximización de la covarianza (PLS)

Como hemos indicado al principio del capítulo, el criterio de optimización asociado a PLS es el de maximizar la covarianza al proyectar sobre las direcciones elegidas. Este criterio es suficiente para obtener la primera componente. Sin embargo, aún no hemos definido formalmente las componentes sucesivas. Imitando la restricción de CCA para las siguientes componentes, es razonable considerar el problema de optimización (2.7) que definimos a continuación. Aunque este no es el problema que acaba resolviendo NIPALS, su análisis es muy ilustrativo.

$$\begin{aligned}(\mathbf{w}_l^{\text{PLS}}, \mathbf{c}_l^{\text{PLS}}) = \underset{(\mathbf{w},\mathbf{c})}{\arg\max} \left(\operatorname{cov}(\mathbf{w}^\top X, \mathbf{c}^\top Y)\right)^2 \text{ sujeto a} \\ \|\mathbf{w}\| = \|\mathbf{c}\| = 1 \\ \operatorname{cov}(\mathbf{w}^\top X, (\mathbf{w}_j^{\text{PLS}})^\top X) = 0 \quad \forall j < l \\ \operatorname{cov}(\mathbf{c}^\top Y, (\mathbf{c}_j^{\text{PLS}})^\top Y) = 0 \quad \forall j < l\end{aligned} \tag{2.7}$$

Podemos ahora intentar resolver este problema con métodos similares a los ya utilizados para los PCA y CCA. Observando las restricciones, es inmediato ver que, como las restricciones de ortogonalidad se enuncian sobre las proyecciones en vez de las direcciones $(\mathbf{w}, \mathbf{c})$, no podemos maximizar esta cantidad como maximizamos la varianza en el caso de PCA.

Sin embargo, salta a la vista que este problema es muy parecido a CCA por lo que es natural plantearse resolverlo del mismo modo. No obstante, en el argumento de CCA aprovechamos que la correlación es independiente de la escala, un hecho que ya no es cierto para la covarianza. Esta propiedad era la que permitía que las tres restricciones pudiesen expresarse como simples productos escalares con respecto al mismo producto interno. Ante la imposibilidad de proceder de este modo, nos quedamos con una restricción con respecto al producto interno usual ($\|\mathbf{w}\| = \|\mathbf{c}\| = 1$) y dos restricciones de ortogonalidad que se pueden reescribir como:

$$\mathbf{w}^\top \Sigma_{XX} \mathbf{w}_j^{\text{PLS}} = 0 \quad \text{y} \quad \mathbf{c}^\top \Sigma_{YY} \mathbf{c}_j^{\text{PLS}} = 0.$$

Es decir, nos hemos encontrado con un problema de optimización en el que no hemos podido escribir todas las restricciones sobre los vectores $\mathbf{w}, \mathbf{c}$ como simples productos internos. Por tanto, los métodos vistos anteriormente para PLS y CCA no pueden aplicarse.

Ante esta situación, los métodos basados en NIPALS optan por modificar ligeramente el problema y resolver otro problema mucho más sencillo, con resultados muy favorables en la práctica. A fin de simplificar la notación, en el resto de esta sección omitimos el superíndice $^{\text{PLS}}$ de las direcciones de proyección.

Para obtener la primera componente, simplemente se maximiza la covarianza. Denotando $X_0 = X,\ Y_0 = Y$:

$$(\mathbf{w}_1, \mathbf{c}_1) = \underset{\|\mathbf{w}\|=1, \|\mathbf{c}\|=1}{\arg\max}\ \operatorname{cov}^2(\mathbf{w}^\top X_0, \mathbf{c}^\top Y_0). \tag{2.8}$$

Para simplificar la notación, se introducen en (2.9) unas nuevas variables aleatorias que corresponden a las proyecciones a lo largo de las direcciones elegidas, es decir, a las componentes. Dado que estas variables se han extraído a partir de los datos, en buena parte de la literatura (por ejemplo Noonan y Wold (1977), Wold (1975), Wold (1980), Wegelin (2000)), se hace referencia a ellas como variables latentes extraídas.

$$\tau_1 = \mathbf{w}_1^\top X_0 \quad \upsilon_1 = \mathbf{c}^\top Y_0. \tag{2.9}$$

A continuación podemos construir unos nuevos vectores aleatorios aplicando una deflación sobre $X$ e $Y$. En particular, buscamos unos nuevos vectores $X_1$ e $Y_1$ que cumplan:

$$\begin{aligned} &X_1 \text{ es tal que } \operatorname{cov}(\mathbf{w}^\top X_1, \tau_1) = 0 \quad \forall \mathbf{w} \in \mathbb{R}^M, \\ &Y_1 \text{ es tal que } \operatorname{cov}(\mathbf{c}^\top Y_1, \upsilon_1) = 0 \quad \forall \mathbf{c} \in \mathbb{R}^D. \end{aligned} \tag{2.10}$$

Si los nuevos vectores de datos cumplen estas restricciones, podremos obtener las siguientes componentes resolviendo el mismo problema de optimización (2.8) pero con

los nuevos vectores de datos, y habremos garantizado que las proyecciones extraídas en sucesivas iteraciones sean incorreladas.

Una posibilidad para asegurar que se cumpla (2.10) es definir $X_1$ como el residuo de proyectar $X_0$ ortogonalmente sobre $\tau_1$. De forma análoga, se define $Y_1$ como el residuo de proyectar $Y_0$ ortogonalmente sobre $\upsilon_1$. Entonces, podemos obtener $X_1$ e $Y_1$ como

$$X_1 = X_0 - \frac{\mathbb{E}(X_0\tau_1)}{\mathrm{var}(\tau_1)}\tau_1, \quad Y_1 = Y_0 - \frac{\mathbb{E}(Y_0\upsilon_1)}{\mathrm{var}(\upsilon_1)}\upsilon_1.$$

Un sencillo cálculo es suficiente para verificar que se cumple (2.10). Por simetría, basta comprobarlo para $X_1$. Aprovechando que todas las variables aleatorias involucradas están centradas, obtenemos

$$\begin{aligned}\mathrm{cov}(\mathbf{w}^\top X_1, \tau_1) = \mathbb{E}(\mathbf{w}^\top X_1\tau_1) = \mathbf{w}^\top\mathbb{E}(X_1\tau_1) = \mathbf{w}^\top\mathbb{E}\left(\left(X - \frac{\mathbb{E}(X\tau_1)}{\mathrm{var}(\tau_1)}\tau_1\right)\tau_1\right) = \\ = \mathbf{w}^\top\left(\mathbb{E}(X\tau_1) - \mathbb{E}(X\tau_1)\frac{\mathrm{var}(\tau_1)}{\mathrm{var}(\tau_1)}\right) = 0, \qquad \forall\mathbf{w}\in\mathbb{R}^M.\end{aligned}$$

Definiendo las siguientes direcciones como las que resuelven el mismo problema de maximización pero con los nuevos vectores aleatorios, se obtiene el siguiente procedimiento iterativo:

$$(\mathbf{w}_l, \mathbf{c}_l) = \underset{\|\mathbf{w}\|=1,\|\mathbf{c}\|=1}{\arg\max}\ \mathrm{cov}^2(\mathbf{w}^\top X_{l-1}, \mathbf{c}^\top Y_{l-1}),$$

donde $l \geq 1$ y

- $\tau_l = \mathbf{w}_l^\top X_{l-1}$,
- $\upsilon_l = \mathbf{c}_l^\top Y_{l-1}$,
- $X_l = X_{l-1} - \frac{\mathbb{E}(X_{l-1}\tau_l)}{\mathrm{var}(\tau_l)}\tau_l$,
- $Y_l = Y_{l-1} - \frac{\mathbb{E}(Y_{l-1}\upsilon_l)}{\mathrm{var}(\upsilon_l)}\upsilon_l$,
- $X_0 = X,\ Y_0 = Y$ (datos iniciales).

Cabe mencionar que las parejas de direcciones $(\mathbf{w}_l, \mathbf{c}_l)$ obtenidas con este método no tienen el mismo significado que las definidas en el caso de PCA o CCA. En este caso, se están definiendo direcciones tales que las proyecciones de los datos deflactados a lo largo de ellas resulten en las componentes buscadas. Sin embargo, dada una dirección, proyectar los datos originales a lo largo de ella no es necesariamente equivalente a proyectar los datos deflactados. Esta discrepancia será tratada en detalle y resuelta en la sección 2.4.2.

Para cerrar esta sección, presentamos el problema de autovalores correspondiente a maximizar la covarianza.

**Proposición 2.3.** *Las soluciones al problema de optimización*

$$(\mathbf{w}, \mathbf{c}) = \underset{\|\mathbf{w}\|=1, \|\mathbf{c}\|=1}{\arg\max} \ \mathrm{cov}^2(\mathbf{w}^\top X, \mathbf{c}^\top Y),$$

*se caracterizan del siguiente modo:*

*(a)* $\mathbf{w}$ *es un autovector de norma unidad asociado al autovector dominante de la matriz* $\Sigma_{XY}\Sigma_{YX}$.

*(b)* $\mathbf{c}$ *es un autovector de norma unidad asociado al autovector dominante de la matriz* $\Sigma_{YX}\Sigma_{XY}$.

*(c) Se cumple que* $\mathbf{c} \propto \Sigma_{YX}\mathbf{w}$ *y* $\mathbf{w} \propto \Sigma_{XY}\mathbf{c}$.

*Demostración.*
Se sigue de un argumento análogo al desarrollado para CCA. □

## 2.4. NIPALS

NIPALS (Nonlinear Iterative Partial Least Squares) en su versión original fue desarrollado en la década de los 70. Este algoritmo se desarrolló para ajustar modelos que relacionaban linealmente variables latentes construidas como combinaciones lineales de las variables observables. Sin embargo, actualmente y fuera de ciertas áreas como la economía o la quimiometría, se denomina NIPALS a una especialización de este algoritmo. El algoritmo original se puede encontrar en Noonan y Wold (1977); y en Wegelin (2000) se presenta una comparación entre la versión original de NIPALS y la que se utiliza para el cálculo de PLS.

Por el contrario, en el presente trabajo, vamos a abordar el análisis de NIPALS centrándonos en el concepto más contemporáneo del algoritmo. De hecho, con frecuencia en la literatura reciente, se denomina NIPALS solo al algoritmo utilizado para calcular PLS, utilizando ambos términos casi de forma intercambiable como se puede observar en Rosipal y Krämer (2005) o Höskuldsson (1988). En nuestro caso, utilizaremos el término NIPALS para referirnos a una familia de algoritmos que permite resolver problemas de reducción de dimensionalidad (veremos que también se pueden aplicar a problemas de regresión). En particular, existen versiones de NIPALS para resolver PLS, PCA y CCA. Estas están cubiertas en Wegelin (2000) y Geladi y Kowalski (1986).

El cálculo de CCA mediante NIPALS no se aborda en esta sección. Sin embargo, se puede encontrar la versión de NIPALS correspondiente a maximizar la correlación en el apéndice B.2. El análisis correspondiente se puede encontrar en Wegelin (2000) y Lyttkens (1972).

Finalmente, antes de comenzar a analizar especializaciones de NIPALS para los distintos métodos de reducción de la dimensionalidad, vamos a proporcionar una visión de alto nivel del algoritmo. NIPALS es un algoritmo iterativo que realiza deflaciones sucesivas de los datos entre cada iteración para calcular varias componentes del

problema de reducción de dimensionalidad considerado. Denotaremos $L$ al número de componentes a extraer.

El algoritmo está compuesto por un bucle que se repite $L$ veces. En el interior de ese bucle se llevan a cabo tres bloques de acciones:

1. Calcular las direcciones de proyección para los datos deflactados. Este paso se traduce en resolver el problema de autovalores asociado a la técnica de reducción de la dimensionalidad.
2. Calcular las componentes proyectando los datos deflactados a lo largo de las direcciones halladas.
3. Realizar una deflación de las matrices de datos mediante su proyección sobre un subespacio ortogonal a los componentes ya extraídos y volver al paso 1.

A la vista de esta estructura, el desarrollo teórico de la sección anterior es muy significativo. La relación de los tres métodos de reducción de dimensionalidad con problemas de autovalores es clave para su resolución mediante NIPALS. Asimismo, en esta sección, pasamos del análisis poblacional al muestral. Denotaremos como $N$ al número de observaciones y las matrices de datos correspondientes a los vectores aleatorios $X$ e $Y$ serán $\mathbf{X} \in \mathbb{R}^{N\times M}$ e $\mathbf{Y} \in \mathbb{R}^{N\times D}$.

### 2.4.1. NIPALS aplicado a PCA

En primer lugar, introducimos una versión ligeramente simplificada pero suficiente para calcular las componentes principales (algoritmo 1). Se trata de un algoritmo iterativo que encuentra una componente principal por cada iteración. Las direcciones de proyección ($\mathbf{w}_l$) se definen en la línea 4. La línea 5 calcula las proyecciones ($\mathbf{t}_l$, componentes principales) y la línea 6 realiza una deflación sobre la matriz de datos $\mathbf{X} \in \mathbb{R}^{N\times M}$.

**Algoritmo 1** NIPALS simplificado para el cálculo de PCA

| | |
|---|---|
| **Entrada:** | $\mathbf{X}$ la matriz de datos y $L$ el número de componentes principales a extraer. |
| **Salida:** | $L$ direcciones de proyección:$\{\mathbf{w}_l\}_{l=1}^{L}$.<br>$L$ componentes principales: $\{\mathbf{t}_l\}_{l=1}^{L}$. |

1: $\mathbf{X}_0 \leftarrow \mathbf{X}$
2: $l \leftarrow 1$
3: **while** $l < L$ **do**
4: $\quad \mathbf{w}_l \leftarrow$ autovector dominante unitario de $\mathbf{X}_{l-1}^{\top}\mathbf{X}_{l-1}$
5: $\quad \mathbf{t}_l \leftarrow \mathbf{X}_{l-1}\mathbf{w}_l$ ▷ Componente principal
6: $\quad \mathbf{X}_l \leftarrow \left(\mathbf{I} - \frac{\mathbf{t}_l\mathbf{t}_l^{\top}}{\mathbf{t}_l^{\top}\mathbf{t}_l}\right)\mathbf{X}_{l-1}$ ▷ Deflación
7: $\quad l \leftarrow l + 1$
8: **end while**

Esta deflación permite calcular la siguiente componente principal resolviendo el mismo problema (maximizar $\mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w}$). Definimos $\mathbf{X}_l$ como la matriz de datos obtenida tras la $l$-ésima deflación.

La deflación presente en la línea 6 se puede entender como una proyección de $\mathbf{X}_{l-1}$ sobre el subespacio ortogonal a $\mathbf{t}_l$. Para demostrar que este algoritmo resulta en la resolución del problema (2.1), enunciamos las dos siguientes proposiciones.

**Proposición 2.4.** *Dado un vector $\mathbf{w} \in \mathbb{R}^M$ cualquiera y $0 < l \leq L$, si definimos $\mathbf{X}_l$ y $\mathbf{w}_l$ como en el algoritmo 1, es posible descomponer $\mathbf{w}$ como $\mathbf{w} = \mathbf{w}^\perp + \mathbf{w}^\parallel$ donde $\mathbf{w}^\perp \perp \mathbf{w}_l$ y se verifica que $\mathbf{X}_l\mathbf{w} = \mathbf{X}_{l-1}\mathbf{w}^\perp$.*

*Demostración.* Se incluye la demostración en el apéndice A (página 39). □

**Proposición 2.5.** *Dada $0 < l \leq L$ y un vector $\mathbf{w} \in \mathbb{R}^M$, podemos descomponer $\mathbf{w} = \mathbf{w}^\perp + \mathbf{w}^\parallel$ donde $\mathbf{w}^\parallel \in \operatorname{span}\{\mathbf{w}_1, \dots \mathbf{w}_l\}$ y donde $\mathbf{w}^\perp \in \operatorname{span}\{\mathbf{w}_1, \dots \mathbf{w}_l\}^\perp$. Entonces $\mathbf{X}_l\mathbf{w} = \mathbf{X}\mathbf{w}^\perp$.*

*Demostración.*
Se sigue de aplicar la proposición anterior iterativamente sobre $\mathbf{X}_l\mathbf{w}$. □

Como la cantidad a maximizar en la $l$-ésima iteración es $\mathbf{w}^\top \mathbf{X}_l^\top \mathbf{X}_l \mathbf{w}$ sujeto a $\|\mathbf{w}\| = 1$, por la proposición anterior, este valor se alcanza cuando $\mathbf{w}$ es ortogonal a las direcciones elegidas en las iteraciones anteriores. Además, el hecho de que $\mathbf{w}$ sea ortogonal a las direcciones anteriores es suficiente para que $\mathbf{w}^\top \mathbf{X}_l^\top \mathbf{X}_l \mathbf{w} = \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w}$. Por tanto, maximizar $\mathbf{w}^\top \mathbf{X}_l^\top \mathbf{X}_l \mathbf{w}$ es equivalente a maximizar $\mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w}$ sujeto a la condición de ortogonalidad buscada. Así, hemos demostrado que este algoritmo es equivalente al cálculo de PCA.

**Algoritmo 2** NIPALS-PCA

**Entrada:** $\mathbf{X}$ la matriz de datos y $L$ el número de componentes principales a extraer.
**Salida:** $L$ direcciones de proyección:$\{\mathbf{w}_l\}_{l=1}^L$.
$L$ componentes principales: $\{\mathbf{t}_l\}_{l=1}^L$.
$L$ *loadings* : $\{\mathbf{p}_l\}_{l=1}^L$.

1: $\mathbf{X}_0 \leftarrow \mathbf{X}$
2: $l \leftarrow 1$
3: **while** $l < L$ **do**
4: $\quad \mathbf{w}_l \leftarrow$ autovector dominante de norma unidad de $\mathbf{X}_{l-1}^\top \mathbf{X}_{l-1}$
5: $\quad \mathbf{t}_l \leftarrow \mathbf{X}_{l-1}\mathbf{w}_l$ ▷ *Scores*
6: $\quad \mathbf{p}_l \leftarrow \mathbf{X}_{l-1}^\top \mathbf{t}_l / (\mathbf{t}_l^\top \mathbf{t}_l)$ ▷ *Loadings*
7: $\quad \mathbf{X}_l \leftarrow \mathbf{X}_{l-1} - \mathbf{t}_l \mathbf{p}_l^\top$ ▷ *Deflación*
8: $\quad l \leftarrow l + 1$
9: **end while**

Para concluir este análisis de la aplicación de NIPALS al cálculo de PCA, presentamos su versión más habitual en el algoritmo 2. Tradicionalmente, el cálculo del

autovalor dominante (línea 4) se lleva a cabo mediante el algoritmo de la potencia aunque esta particularidad no afecta al resultado final. Asimismo, como resultado del algoritmo se construyen las siguientes matrices

$$\mathbf{W}_L = (\mathbf{w}_1, \dots, \mathbf{w}_L), \qquad \mathbf{T}_L = (\mathbf{t}_1, \dots, \mathbf{t}_L), \qquad \mathbf{P}_L = (\mathbf{p}_1, \dots, \mathbf{p}_L).$$

La introducción de los vectores $\{\mathbf{p}_l\}_{l=1}^{L}$ (denominados habitualmente *loadings*) permite obtener una expresión para $\mathbf{X}_l$ como $\mathbf{X}_l = \mathbf{X} - \mathbf{T}_l\mathbf{P}_l^\top$. Esta expresión puede ser interpretada como una forma de aproximar la matriz de datos original. En particular, tenemos que $\mathbf{X} = \mathbf{T}_L\mathbf{P}_L^\top + \mathbf{E}_L$ donde $\mathbf{E}_L$ es un error que decrece según aumenta $L$.

Aprovechamos esta oportunidad para introducir la nomenclatura con la que se suele designar a los distintos vectores involucrados en NIPALS:

- *Weights*($\mathbf{w}_l$). Son las direcciones que resuelven el problema de autovectores en cada iteración. Es importante tener en cuenta que no son necesariamente las direcciones a lo largo de las que hay que proyectar los datos originales para obtener los componentes (sí lo son en PCA). Se trata de las direcciones sobre las que hay que proyectar los datos deflactados para obtener las componentes.
- *Scores*($\mathbf{t}_l$). Son las proyecciones de los datos deflactados a lo largo de los *weights*. Equivalentemente, se trata de las componentes que se buscaba extraer.
- *Loadings*($\mathbf{p}_l$). Se trata de vectores que permiten calcular una aproximación de rango 1 de $\mathbf{X}_l$ como $\mathbf{t}_l\mathbf{p}_l^\top$.

### 2.4.2. NIPALS aplicado a PLS (NIPALS-Modo A)

En esta sección introducimos el algoritmo NIPALS correspondiente al cálculo de PLS (algoritmo 3). Se trata de un algoritmo muy similar al correspondiente a PCA cambiando la matriz sobre la que se calcula el autovalor dominante y trabajando con dos bloques de variables (X e Y) en vez de uno. Esta variación es la versión más extendida de NIPALS. Se suele denominar NIPALS Modo A siguiendo la nomenclatura original en Noonan y Wold (1977) o PLSCanonical (Pedregosa et al. (2011)).

Un primer vistazo sugiere que el análisis de este algoritmo se puede llevar a cabo de forma similar al anterior. Sin embargo, es sencillo probar con un contraejemplo que la propiedad análoga a la proposición 2.4 no se cumple en este caso.

No obstante, los vectores calculados en el algoritmo, son análogos a los calculados en el algoritmo 2 pero repetidos para cada bloque (véase algoritmo 3). Asimismo, al igual que en el caso anterior, los vectores obtenidos por el algoritmo se agrupan en las siguientes matrices:

$$\begin{aligned} \mathbf{W} &= (\mathbf{w}_1, \dots, \mathbf{w}_L) \in \mathbb{R}^{M\times L} & \mathbf{C} &= (\mathbf{c}_1, \dots, \mathbf{c}_L) \in \mathbb{R}^{D\times L} \\ \mathbf{T} &= (\mathbf{t}_1, \dots, \mathbf{t}_L) \in \mathbb{R}^{N\times L} & \mathbf{U} &= (\mathbf{u}_1, \dots, \mathbf{u}_L) \in \mathbb{R}^{N\times L} \\ \mathbf{P} &= (\mathbf{p}_1, \dots, \mathbf{p}_L) \in \mathbb{R}^{M\times L} & \mathbf{Q} &= (\mathbf{q}_1, \dots, \mathbf{q}_L) \in \mathbb{R}^{D\times L}. \end{aligned}$$

El problema de autovalores que se busca resolver en este caso es el enunciado en la proposición 2.3. Además, se aprovecha la proporcionalidad entre $\mathbf{c}$ y $\mathbf{Y}^\top\mathbf{X}\mathbf{w}$ para obtener $\mathbf{c}$ sin resolver su problema de autovectores asociado.

**Algoritmo 3** NIPALS-Modo A

**Entrada:** $\mathbf{X}$, $\mathbf{Y}$ las matrices de datos y $L$ el número de componentes a extraer.
**Salida:** *weights* :$\{\mathbf{w}_l\}_{l=1}^L$, $\{\mathbf{c}_l\}_{l=1}^L$.
*scores* :$\{\mathbf{t}_l\}_{l=1}^L$, $\{\mathbf{u}_l\}_{l=1}^L$.
*loadings* :$\{\mathbf{p}_l\}_{l=1}^L$, $\{\mathbf{q}_l\}_{l=1}^L$.

1: $\mathbf{X}_0 \leftarrow \mathbf{X}, \quad \mathbf{Y}_0 \leftarrow \mathbf{Y}$
2: $l \leftarrow 1$
3: **while** $l < L$ **do**
4: $\quad \mathbf{w}_l \leftarrow$ autovector dominante de norma unidad de $\mathbf{X}_{l-1}^\top \mathbf{Y}_{l-1} \mathbf{Y}_{l-1}^\top \mathbf{X}_{l-1}$
5: $\quad \mathbf{c}_l \leftarrow \mathbf{Y}_{l-1}^\top \mathbf{X}_{l-1}\, \mathbf{w}_l / \|\mathbf{Y}_{l-1}^\top \mathbf{X}_{l-1}\, \mathbf{w}_l\|$ ▷ *Weights* de $\mathbf{Y}$
6: $\quad \mathbf{t}_l \leftarrow \mathbf{X}_{l-1}\mathbf{w}_l$ ▷ *Scores* de $\mathbf{X}$
7: $\quad \mathbf{u}_l \leftarrow \mathbf{Y}_{l-1}\mathbf{c}_l$ ▷ *Scores* de $\mathbf{Y}$
8: $\quad \mathbf{p}_l \leftarrow \mathbf{X}_{l-1}^\top \mathbf{t}_l / (\mathbf{t}_l^\top \mathbf{t}_l)$ ▷ *Loadings* de $\mathbf{X}$
9: $\quad \mathbf{q}_l \leftarrow \mathbf{Y}_{l-1}^\top \mathbf{u}_l / (\mathbf{u}_l^\top \mathbf{u}_l)$ ▷ *Loadings* de $\mathbf{Y}$
10: $\quad \mathbf{X}_l \leftarrow \mathbf{X}_{l-1} - \mathbf{t}_l \mathbf{p}_l^\top$ ▷ Deflación de $\mathbf{X}$
11: $\quad \mathbf{Y}_l \leftarrow \mathbf{Y}_{l-1} - \mathbf{u}_l \mathbf{q}_l^\top$ ▷ Deflación de $\mathbf{Y}$
12: $\quad l \leftarrow l + 1$
13: **end while**

Por otro lado, un rápido vistazo a la deflación llevada a cabo en las líneas 10 y 11 nos permite reescribirla de dos formas distintas en (2.11). En primer lugar, podemos ver que la deflación se corresponde con la planteada en la sección 2.3 sustituyendo la esperanza y la varianza por sus análogos muestrales y simplificando el factor $N$. Por otro lado, también podemos ver que las deflaciones se reducen a proyectar sobre el espacio ortogonal a la componente extraída en el último paso:

$$\begin{aligned} \mathbf{X}_l &= \mathbf{X}_{l-1} - \mathbf{t}_l \frac{\mathbf{t}_l^\top \mathbf{X}_{l-1}}{\mathbf{t}_l^\top \mathbf{t}_l} = \left(\mathbf{I} - \frac{\mathbf{t}_l \mathbf{t}_l^\top}{\mathbf{t}_l^\top \mathbf{t}_l}\right) \mathbf{X}_{l-1}, \\ \mathbf{Y}_l &= \mathbf{Y}_{l-1} - \mathbf{u}_l \frac{\mathbf{u}_l^\top \mathbf{Y}_{l-1}}{\mathbf{u}_l^\top \mathbf{u}_l} = \left(\mathbf{I} - \frac{\mathbf{u}_l \mathbf{u}_l^\top}{\mathbf{u}_l^\top \mathbf{u}_l}\right) \mathbf{Y}_{l-1}. \end{aligned} \tag{2.11}$$

Estas deflaciones nos garantizan que las siguientes componentes (al ser combinaciones lineales de las columnas de $\mathbf{X}_l$) sean perpendiculares a las anteriores. Por tanto, se verifica la restricción de ortogonalidad entre componentes sucesivas exigida en (2.7). Además, como en cada paso se maximiza la covarianza, es de esperar que obtengamos soluciones similares a las del problema (2.7).

Sin embargo, para poder afirmar que estamos resolviendo ese problema siguiendo un razonamiento análogo al de PCA, necesitaríamos que $\mathbf{X}_l\mathbf{w} = \mathbf{X}_{l-1}\mathbf{w}$ cuando $\mathbf{X}_l\mathbf{w}$

es ortogonal a todas las componentes extraídas. Esta propiedad sí se cumple en el caso de PCA pero no se cumple en general para PLS.

En la siguiente proposición se recogen las principales propiedades de los vectores extraídos por el algoritmo.

**Proposición 2.6.** *Los vectores extraídos por el algoritmo 3 cumplen:*

1. *Los vectores de pesos de cada iteración son perpendiculares, es decir, si $j > i$, $\mathbf{w}_j \perp \mathbf{w}_i$ y $\mathbf{c}_j \perp \mathbf{c}_i$. Por tanto, matricialmente, se cumple que*

$$\mathbf{W}_L^\top \mathbf{W}_L = \mathbf{I}, \qquad \mathbf{C}_L^\top \mathbf{C}_L = \mathbf{I}.$$

2. *Las scores extraídas en cada iteración son perpendiculares a las anteriores, es decir, si $j > i$, $\mathbf{t}_j \perp \mathbf{t}_i$ y $\mathbf{u}_j \perp \mathbf{u}_i$. Por tanto, matricialmente, se cumple que $\mathbf{T}_L^\top \mathbf{T}_L$ y $\mathbf{U}_L^\top \mathbf{U}_L$ son matrices diagonales.*

3. *En cada iteración se maximiza la covarianza de las proyecciones, es decir, se cumple que*

$$\operatorname{cov}^2(\mathbf{X}_l \mathbf{w}_l, \mathbf{Y}_l \mathbf{c}_l) = \max_{\|\mathbf{r}\|=\|\mathbf{s}\|=1} \operatorname{cov}^2(\mathbf{X}_l \mathbf{r}, \mathbf{Y}_l \mathbf{s}).$$

4. *Los vectores extraídos son solución a los siguiente problemas de autovalores:*

$$\begin{aligned} \mathbf{X}_{l-1}^\top \mathbf{Y}_{l-1} \mathbf{Y}_{l-1}^\top \mathbf{X}_{l-1} \mathbf{w}_l &= \lambda \mathbf{w}_l, & \mathbf{Y}_{l-1}^\top \mathbf{X}_{l-1} \mathbf{X}_{l-1}^\top \mathbf{Y}_{l-1} \mathbf{c}_l &= \lambda \mathbf{c}_l, \\ \mathbf{X}_{l-1} \mathbf{X}_{l-1}^\top \mathbf{Y}_{l-1} \mathbf{Y}_{l-1}^\top \mathbf{t}_l &= \lambda \mathbf{t}_l, & \mathbf{Y}_{l-1} \mathbf{Y}_{l-1}^\top \mathbf{X}_{l-1} \mathbf{X}_{l-1}^\top \mathbf{u}_l &= \lambda \mathbf{u}_l. \end{aligned}$$

5. *La deflación tiene las siguientes consecuencias:*

$$\begin{aligned} \mathbf{X}_L &= \mathbf{X} - \mathbf{T}_L \mathbf{P}_L^\top, & \mathbf{Y}_L &= \mathbf{Y} - \mathbf{U}_L \mathbf{Q}_L^\top, \\ \mathbf{X}_L &= \prod_{l=k}^{L} \left( \mathbf{I} - \frac{\mathbf{t}_l \mathbf{t}_l^\top}{\mathbf{t}_l^\top \mathbf{t}_l} \right) \mathbf{X}_{k-1}, & \mathbf{Y}_L &= \prod_{l=k}^{L} \left( \mathbf{I} - \frac{\mathbf{u}_l \mathbf{u}_l^\top}{\mathbf{u}_l^\top \mathbf{u}_l} \right) \mathbf{Y}_{k-1}, \quad k < L. \end{aligned}$$

6. *Los loadings se pueden expresar como*

$$\mathbf{P}_L = \mathbf{X}^\top \mathbf{T}_L (\mathbf{D}_L^x)^{-2}, \quad \mathbf{Q}_L = \mathbf{Y}^\top \mathbf{U}_L (\mathbf{D}_L^y)^{-2},$$

*donde* $\mathbf{D}_L^x = \operatorname{diag}(\|\mathbf{t}_1\|, \dots, \|\mathbf{t}_L\|)$ *y* $\mathbf{D}_L^y = \operatorname{diag}(\|\mathbf{u}_1\|, \dots, \|\mathbf{u}_L\|)$.

*Demostración.* Los puntos dos y tres ya se justificaron en el análisis poblacional. Se incluye una demostración completa en el apéndice A (página 40). □

Llegados a este punto, es importante recordar el objetivo de un método de reducción de dimensionalidad tal y como se enunció al principio de este capítulo. Buscamos obtener unas direcciones a lo largo de las cuales proyectar los datos. Sin embargo, aún no hemos obtenido una expresión para estas direcciones. Si bien los vectores $\{\mathbf{w}_l\}$ se pueden ver como direcciones de proyección, los datos que se proyectan sobre $\mathbf{w}_l$ no

corresponden a la matriz de datos original $\mathbf{X}$ sino a la matriz de datos con la que trabaja el algoritmo en la $l$-ésima iteración.

Sin embargo, sí es posible encontrar las direcciones buscadas ya que el espacio generado por las columnas de $\mathbf{X}_l$ es un subespacio del generado por las columnas de $\mathbf{X}$. En particular, si definimos $\pi$ como la proyección sobre $(\text{span}\{\mathbf{t}_1, \dots \mathbf{t}_l\})^\perp$, debido a las deflaciones sucesivas, se cumple que

$$\text{Col}(\mathbf{X}_l) = \pi(\text{Col}(\mathbf{X})) \subseteq \text{Col}(\mathbf{X}),$$

donde $\text{Col}(\mathbf{A})$ denota el espacio generado por las columnas de $\mathbf{A}$. Es decir, si escribimos la matriz por columnas como $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_M)$, se cumple que $\text{Col}(\mathbf{A}) = \text{span}\{\mathbf{a}_1, \dots, \mathbf{a}_M\}$

Presentamos las matrices de proyección con respecto a los datos originales en la siguiente proposición. La importancia de este resultado es mayúscula ya que nos permite llevar a cabo la proyección de nuevos datos sin tener que repetir todo el algoritmo.

**Proposición 2.7.** *Dadas las definiciones del algoritmo 3, se definen las matrices:*

$$\mathbf{R}_L^x = \mathbf{W}_L(\mathbf{P}_L^\top \mathbf{W}_L)^{-1}, \quad \mathbf{R}_L^y = \mathbf{C}_L(\mathbf{Q}_L^\top \mathbf{C}_L)^{-1}.$$

*Estas matrices cumplen que* $\mathbf{T}_L = \mathbf{X}\mathbf{R}_L^x$ *y* $\mathbf{U}_L = \mathbf{Y}\mathbf{R}_L^y$. *Por tanto, sus columnas contienen las direcciones de proyección para obtener las componentes calculadas en NIPALS. Con frecuencia, se denomina a estas matrices "rotations".*

*Demostración.* En el anexo A (página 43). □

Para concluir la sección, es importante remarcar que la versión más extendida de NIPALS utiliza el algoritmo de la potencia para calcular $\mathbf{w}_l$ (línea 4). Se puede encontrar un análisis de la aplicación del algoritmo de la potencia en Wegelin (2000). Este algoritmo se basa en la idea de que multiplicar repetidamente un vector cualquiera por la misma matriz resultará en un autovector asociado al autovalor dominante de la matriz.

El algoritmo de la potencia es una buena opción en NIPALS ya que permite obtener un autovector dominante rápidamente. No obstante, no se trata en absoluto de la única opción. Se pueden encontrar opciones alternativas que recurren, por ejemplo, a SVD.

Por completitud, se ha incluido la versión más habitual de NIPALS basada en el algoritmo de la potencia en el apéndice B.1.

# CAPÍTULO 3
# PLS aplicado a problemas de regresión

Un problema de regresión con respuesta escalar consiste encontrar un vector $\hat{\boldsymbol{\beta}}$ que ajuste el modelo muestral

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.1}$$

de forma que $\boldsymbol{\epsilon}$ (el residuo) sea pequeño en algún sentido. Como no vamos a llevar a cabo un análisis estadístico de las propiedades asintóticas de este modelo, no son necesarias algunas de las suposiciones habituales como $\mathbb{E}(\boldsymbol{\epsilon}|X) = 0$ o la homocedasticidad. Sin embargo, al igual que en el resto del documento, sí supondremos que tanto $\mathbf{X}$ como $\mathbf{y}$ son las realizaciones de un vector aleatorio o variable aleatoria centrada.

Por ejemplo, un posible vector $\hat{\boldsymbol{\beta}}$ que ajusta este modelo es el estimador de mínimos cuadrados. Dado que las variables se encuentran centradas, se puede ver claramente que este estimador equivale a minimizar la varianza de $\boldsymbol{\epsilon}$. A continuación lo definimos formalmente e incluimos una posible expresión para su cálculo.

**Definición 3.1.** Dado un modelo de regresión lineal con la estructura de (3.1), el estimador de mínimos cuadrados (OLS) se define como:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} := \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|.$$

**Proposición 3.1.** *El estimador de mínimos cuadrados tiene una fórmula cerrada. En particular, el estimador de mínimos cuadrados para el modelo* (3.1) *puede escribirse como:*

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}.$$

A lo largo de este capítulo, veremos como la aplicación de PLS a problemas de regresión nos brinda otro posible $\hat{\boldsymbol{\beta}}$. Al igual que ocurre al aplicar otros métodos de reducción de la dimensionalidad como PCA, el uso de PLS resulta en un estimador capaz de, por ejemplo, mitigar problemas de multicolinearidad o reducir la varianza de las predicciones.

PLS también puede ser aplicado a problemas de regresión multivariante, aunque pierde algunas de sus propiedades. En estos casos, el modelo que se quiere ajustar es

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E},$$

donde la matriz $\mathbf{E}$ denota ahora al residuo y el vector respuesta se ha convertido en la matriz respuesta $\mathbf{Y}$.

En la inmensa mayoría de la literatura se denomina PLS1 a la aplicación de PLS a la regresión múltiple y PLS2 a la aplicación de PLS a la regresión multivariante. En este capítulo primero introduciremos PLS2 como una modificación del método de PLS que resuelve el problema de reducción de la dimensionalidad. A continuación, presentaremos PLS1 como un caso particular de PLS2 y estudiaremos sus diferencias. Finalmente, veremos que PLS1 se puede entender como un método de regularización.

## 3.1. PLS2

La aplicación de PLS a los problemas de regresión se basa en suponer la existencia de una relación lineal entre las componentes extraídas en cada iteración. En términos poblacionales, recordando la nomenclatura introducida en la sección 2.3, suponemos que existe la siguiente relación lineal $\upsilon_l = d_l\tau_l + \epsilon_l$ donde $\epsilon_l \perp \tau_l$ y $d_l \in \mathbb{R}$.

En términos muestrales, suponemos que existen $\{d_l\}_{i=l}^{L}$ tales que $\mathbf{u}_l = d_l\mathbf{t}_l + \mathbf{h}_l$ para $l = 1, \dots, L$, donde $\mathbf{h}_l$ denota un residuo. Para calcular los escalares $d_l$, se pueden ajustar los $L$ modelos de regresión lineal por mínimos cuadrados. Así, se obtiene $d_l = \frac{\mathbf{t}_l^\top \mathbf{u}_l}{\mathbf{t}_l^\top \mathbf{t}_l}$.

Como estamos suponiendo que existe una relación lineal subyacente, es de esperar que $\|\mathbf{h}_l\|$ sea pequeño y, con la expresión anterior para $d_l$, hemos obtenido una forma de estimar $\mathbf{u}_l$.

Por tanto, podemos sustituir las apariciones de $\mathbf{u}_l$ en el algoritmo NIPALS-Modo A (algoritmo 3) por esta aproximación. Vemos que este vector aparece en las lineas 9 y 11 y afecta a la definición de los vectores $\mathbf{q}_l$ y a la deflación de la matriz $\mathbf{Y}$. Estas expresiones son las siguientes:

$$\mathbf{q}_l \leftarrow \mathbf{Y}_{l-1}^\top \mathbf{u}_l / (\mathbf{u}_l^\top \mathbf{u}_l) \qquad \mathbf{Y}_l \leftarrow \mathbf{Y}_{l-1} - \mathbf{u}_l \mathbf{q}_l^\top.$$

Si sustituimos directamente la expresión $\mathbf{u}_l = d_l\mathbf{t}_l$,

$$\mathbf{q}_l \leftarrow d_l \mathbf{Y}_{l-1}^\top \mathbf{t}_l / (d_l^2 \mathbf{t}_l^\top \mathbf{t}_l) \qquad \mathbf{Y}_l \leftarrow \mathbf{Y}_{l-1} - d_l \mathbf{t}_l \mathbf{q}_l^\top.$$

Si, además, absorbemos el factor $d_l$ de la deflación en la expresión de $\mathbf{q}_l$ llegamos a una definición de $\mathbf{q}_l$ distinta pero que resulta en un algoritmo más simple. En definitiva, se obtiene

$$\tilde{\mathbf{q}}_l \leftarrow \mathbf{Y}_{l-1}^\top \mathbf{t}_l / (\mathbf{t}_l^\top \mathbf{t}_l) \qquad \mathbf{Y}_l \leftarrow \mathbf{Y}_{l-1} - \mathbf{t}_l \tilde{\mathbf{q}}_l^\top.$$

En algunas fuentes como Rosipal y Krämer (2005) o Höskuldsson (1988) se lleva a cabo una normalización distinta del vector $\mathbf{c}_l$ en el algoritmo NIPALS de forma que este coincida con $\tilde{\mathbf{q}}_l$. Sin embargo, en el presente documento se ha optado por emplear esta notación alternativa para mantener la coherencia con el análisis de PLS como método de reducción de la dimensionalidad. Asimismo, el algoritmo presentado en esta sección (algoritmo 4) obvia el cálculo de los vectores que no resultan ser necesarios para el cálculo de la matriz de regresión.

**Algoritmo 4** NIPALS-PLS2

**Entrada:** $\mathbf{X}$, $\mathbf{Y}$ las matrices de datos y $L$ el número de componentes a extraer.
**Salida:** *weights* :$\{\mathbf{w}_l\}_{l=1}^{L}$.
*scores* :$\{\mathbf{t}_l\}_{l=1}^{L}$.
*loadings* :$\{\mathbf{p}_l\}_{l=1}^{L}$, $\{\tilde{\mathbf{q}}_l\}_{l=1}^{L}$.

1: $\mathbf{X}_0 \leftarrow \mathbf{X}, \quad \mathbf{Y}_0 \leftarrow \mathbf{Y}$
2: $l \leftarrow 1$
3: **while** $l < L$ **do**
4: $\quad \mathbf{w}_l \leftarrow$ autovector dominante de norma unidad de $\mathbf{X}_{l-1}^\top \mathbf{Y}_{l-1} \mathbf{Y}_{l-1}^\top \mathbf{X}_{l-1}$
5: $\quad \mathbf{t}_l \leftarrow \mathbf{X}_{l-1}\mathbf{w}_l$ ▷ *Scores* de $\mathbf{X}$
6: $\quad \mathbf{p}_l \leftarrow \mathbf{X}_{l-1}^\top \mathbf{t}_l / (\mathbf{t}_l^\top \mathbf{t}_l)$ ▷ *Loadings* de $\mathbf{X}$
7: $\quad \tilde{\mathbf{q}}_l \leftarrow \mathbf{Y}_{l-1}^\top \mathbf{t}_l / (\mathbf{t}_l^\top \mathbf{t}_l)$ ▷ *Loadings* de $\mathbf{Y}$
8: $\quad \mathbf{X}_l \leftarrow \mathbf{X}_{l-1} - \mathbf{t}_l \mathbf{p}_l^\top$ ▷ Deflación de $\mathbf{X}$
9: $\quad \mathbf{Y}_l \leftarrow \mathbf{Y}_{l-1} - \mathbf{t}_l \tilde{\mathbf{q}}_l^\top$ ▷ Deflación de $\mathbf{Y}$
10: $\quad l \leftarrow l + 1$
11: **end while**

Esta simplificación tiene como consecuencia adicional que la deflación en el bloque Y podría omitirse como enunciamos a continuación. Sin embargo, hemos optado por incluirla en el algoritmo 4 ya que facilita su análisis y comparativa con los algoritmos anteriores.

**Lema 3.2.** *La deflación llevada a cabo en la línea 9 del algoritmo 4 puede eliminarse sin modificar el resultado del algoritmo.*

*Demostración.* En el apéndice A (página 43). □

Como consecuencia directa de este lema, al desaparecer la deflación sobre $\mathbf{Y}$, el número de componentes que podemos extraer ya no está limitado por la cantidad de columnas de $\mathbf{Y}$. Sin estos cambios, la matriz de datos $\mathbf{Y}$ sería 0 como mucho en $D$ iteraciones (en cada iteración se sustraía una aproximación de rango uno). Esta propiedad es tremendamente útil ya que, de lo contrario, la regresión basada en PLS sería de muy poco uso cuando contamos con una única variable respuesta.

Estos cambios en el algoritmo hacen que algunas de las identidades derivadas para NIPALS-Modo A ya no se cumplan. Sin embargo, las que solo involucran a matrices del bloque X ($\mathbf{W}_L, \mathbf{T}_L, \mathbf{Q}_L$) siguen siendo válidas ya que en su demostración no se utiliza la definición de $\mathbf{Q}_L$ o la fórmula de deflación de $\mathbf{Y}$.

Como consecuencia de la deflación de $Y$ en la línea 9 del algoritmo 4, se verifica $\mathbf{Y} = \mathbf{T}_L \tilde{\mathbf{Q}}_L^\top + \tilde{\mathbf{F}}_L$ donde $\tilde{\mathbf{F}}_L$ denota el residuo. Como la expresión para $\mathbf{T}_L$ obtenida en la proposición 2.7 sigue siendo válida, se llega a la siguiente identidad

$$\mathbf{Y} = (\mathbf{X}\mathbf{W}_L(\mathbf{P}_L^\top \mathbf{W}_L)^{-1})\tilde{\mathbf{Q}}_L^\top + \tilde{\mathbf{F}}_L.$$

Por tanto, hemos encontrado una posible expresión de la matriz de regresión. A continuación, recogemos en una proposición las expresiones más habituales de la misma.

**Proposición 3.2.** *En la regresión PLS se busca una matriz* $\hat{\mathbf{B}}$ *tal que* $\mathbf{Y} = \mathbf{X}\hat{\mathbf{B}} + \mathbf{E}$ *donde* $\mathbf{E} = \mathbf{Y} - \hat{\mathbf{Y}}$ *es una matriz de residuos. Si se consideran* $L$ *componentes,* $\hat{\mathbf{B}}_L$ *es el resultado de las siguientes expresiones equivalentes:*

$$\hat{\mathbf{B}}_L = \mathbf{W}_L(\mathbf{P}_L^\top \mathbf{W}_L)^{-1}\tilde{\mathbf{Q}}_L^\top,$$

$$\hat{\mathbf{B}}_L = \mathbf{W}_L(\mathbf{W}_L^\top \mathbf{X}^\top \mathbf{X}\mathbf{W}_L)^{-1}\mathbf{W}_L^\top \mathbf{X}^\top \mathbf{Y}. \tag{3.2}$$

*Además, si definimos* $\tilde{\mathbf{T}}_L$ *y* $\tilde{\mathbf{P}}_L$ *tales que* $\tilde{\mathbf{T}}_L\tilde{\mathbf{P}}_L^\top = \mathbf{T}_L\mathbf{P}_L^\top$ *y las columnas de* $\tilde{\mathbf{T}}$ *tienen norma unidad, la siguiente expresión también es equivalente*

$$\hat{\mathbf{B}}_L = \mathbf{W}_L(\tilde{\mathbf{P}}_L^\top \mathbf{W}_L)^{-1}\tilde{\mathbf{T}}_L^\top \mathbf{Y}.$$

*Demostración.* En el apéndice A (página 44). □

Por completitud, se incluye en el anexo B.3 el pseudocódigo que se puede encontrar en la mayoría de las fuentes para el método PLS2, así como una comparativa con el aquí presentado.

## 3.2. PLS1

Dado que PLS1 puede verse como un caso particular de PLS2, las identidades derivadas en la sección anterior siguen siendo válidas. En lo sucesivo, denotaremos por $\mathbf{y}$ al vector de datos correspondiente a la variable respuesta en vez de $\mathbf{Y}$.

Como ya adelantábamos en la introducción de este capítulo, PLS1 es relevante como caso particular de PLS2 porque, en este caso, el problema se simplifica notablemente. Ya no es necesario calcular el autovector dominante de la matriz de covarianzas cruzadas. La siguiente proposición nos brinda una forma de calcular $\mathbf{w}_l$ directamente.

**Proposición 3.3.** *Dada una matriz* $M \times M$ *de la forma* $\mathbf{A} = \mathbf{a}\mathbf{a}^\top$*, donde* $\mathbf{a} \in \mathbb{R}^M$*,* $\mathbf{a}/\|\mathbf{a}\|$ *es un autovalor dominante unitario de* $\mathbf{A}$*.*

*Demostración.* En el apéndice A (página 45). □

Aplicando esta proposición y omitiendo la parte del algoritmo correspondiente a la deflación de $\mathbf{y}$, se obtiene el algoritmo 5. Es importante notar que, si bien hemos optado por no normalizar las componentes del bloque $X$, algunas fuentes como Phatak y Hoog (2002) normalizan todos los vectores. Se ha optado por no normalizar dichos vectores para mantener la consistencia con secciones anteriores. Sin embargo, como ilustra la última parte de la proposición 3.2, que las columnas de la matriz $\mathbf{T}$ tengan norma unidad puede ser beneficioso para ciertos análisis.

**Algoritmo 5** NIPALS-PLS1

**Entrada:** $\mathbf{X}$, $\mathbf{y}$ las matrices de datos y $L$ el número de componentes a extraer.
**Salida:** *weights* :$\{\mathbf{w}_l\}_{l=1}^{L}$.
*scores* :$\{\mathbf{t}_l\}_{l=1}^{L}$.
*loadings* :$\{\mathbf{p}_l\}_{l=1}^{L}$.

$\mathbf{X}_0 \leftarrow \mathbf{X}, \quad \mathbf{y}_0 \leftarrow \mathbf{y}$
$l \leftarrow 1$
**while** $l < L$ **do**
  $\mathbf{w}_l \leftarrow \mathbf{X}_{l-1}^{\top}\mathbf{y}/\|\mathbf{X}_{l-1}^{\top}\mathbf{y}\|$ ▷ *Weights* de $\mathbf{X}$
  $\mathbf{t}_l \leftarrow \mathbf{X}_{l-1}\mathbf{w}_l$ ▷ *Scores* de $\mathbf{X}$
  $\mathbf{p}_l \leftarrow \mathbf{X}_{l-1}^{\top}\mathbf{t}_l/(\mathbf{t}_l^{\top}\mathbf{t}_l)$ ▷ *Loadings* de $\mathbf{X}$
  $\mathbf{X}_l \leftarrow \mathbf{X}_{l-1} - \mathbf{t}_l\mathbf{p}_l^{\top}$ ▷ Deflación de $\mathbf{X}$
  $l \leftarrow l + 1$
**end while**

## 3.3. PLS1 como método de regularización

En esta sección veremos como el algoritmo de PLS1 es equivalente a aplicar una regularización al estimador de mínimos cuadrados. En particular, demostraremos que hay una equivalencia con el estimador de mínimos cuadrados restringido a un determinado subespacio. El objetivo de esta restricción es reducir la varianza del estimador aunque esto implique introducir un sesgo. Si se consigue reducir la varianza lo suficiente, esperamos obtener un estimador con un error cuadrático medio menor.

A continuación, se definen los espacios de Krylov. Aunque pueda parecer una definición artificial, estos espacios aparecen de forma natural en algunos métodos de optimización iterativos. En particular, uno de estos métodos es el método del gradiente conjugado. Tal y como recoge Wright, Nocedal et al. (1999, p. 108), este algoritmo resuelve problemas de optimización de forma iterativa explorando espacios de Krylov de orden creciente.

**Definición 3.3.** Dada $\mathbf{A} \in \mathbb{R}^{M\times M}$ una matriz simétrica definida positiva y $\mathbf{b} \in \mathbb{R}^M \setminus \{\mathbf{0}\}$, el espacio de Krylov de orden $L$ definido por $\mathbf{A}$ y $\mathbf{b}$ es:

$$\mathcal{K}_L(\mathbf{A}, \mathbf{b}) := \operatorname{span}\left\{\mathbf{b}, \mathbf{A}\mathbf{b}, \dots \mathbf{A}^{L-1}\mathbf{b}\right\}.$$

Antes de proceder con la explicación de las propiedades que caracterizan a estos espacios, incluimos un resultado que motiva la restricción del estimador de mínimos cuadrados a espacios de Krylov.

**Proposición 3.4.** *Sea* $\mathbf{A}$ *una matriz simétrica* $M \times M$ *tal que* $\det(\mathbf{A}) \neq 0$. *Entonces existe un polinomio* $P$ *de grado* $m-1$ *tal que* $\mathbf{A}P(\mathbf{A}) = \mathbf{I}$, *donde* $m$ *es el número de autovalores distintos de* $\mathbf{A}$. *En particular,* $\mathbf{A}^{-1} = a_0 + a_1\mathbf{A} + \cdots + a_{m-1}\mathbf{A}^{m-1}$ *para ciertos coeficientes* $a_j \in \mathbb{R}$.

*Demostración.* En el apéndice A (página 45). □

Esta proposición nos muestra que podemos expresar la matriz inversa como un polinomio. Por tanto, una posible aproximación a esta inversa sería dicho polinomio truncado hasta un cierto grado.

Si recordamos la fórmula para el estimador de mínimos cuadrados (proposición 3.1) vemos que $\hat{\beta}_{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. Por tanto, se podría obtener una aproximación al estimador trabajando con combinaciones lineales de potencias de $\mathbf{A} = \mathbf{X}^\top\mathbf{X}$ multiplicadas por $\mathbf{b} = \mathbf{X}^\top\mathbf{y}$. Esto es exactamente lo que ocurre al restringir el estimador al espacio de Krylov.

**Proposición 3.5.** *El estimador OLS restringido al espacio* $\mathcal{K}_L(\mathbf{X}^\top\mathbf{X}, \mathbf{X}^\top\mathbf{y})$ *tiene la expresión*

$$\hat{\beta} = \mathbf{R}_L(\mathbf{R}_L^\top\mathbf{X}^\top\mathbf{R}_L)^{-1}\mathbf{R}_L\mathbf{X}^\top\mathbf{y}, \tag{3.3}$$

*donde* $\mathbf{R}_L$ *es una matriz cuyas columnas forman una base ortonormal del espacio de Krylov* $\mathcal{K}_L(\mathbf{X}^\top\mathbf{X}, \mathbf{X}^\top\mathbf{y})$.

*Demostración.* En el apéndice A (página 46). □

Para concluir esta sección, hemos de probar que el estimador OLS restringido es equivalente al estimador de PLS1. Sin embargo, comparando las expresiones (3.2) y (3.3), es inmediato ver que basta demostrar que las columnas de la matriz $\mathbf{W}_L$ obtenida en PLS1 forman una base del espacio de Krylov.

**Proposición 3.6.** *Las columnas de la matriz* $\mathbf{W}_L$ *generada por NIPALS-PLS1 forman una base ortonormal del espacio de Krylov* $\mathcal{K}_L(\mathbf{X}^\top\mathbf{X}, \mathbf{X}^\top\mathbf{y})$.

*Demostración.* En el apéndice A (página 46). □

Así, podemos concluir que PLS1 es equivalente a la restricción de OLS al espacio de Krylov $\mathcal{K}_L(\mathbf{X}^\top\mathbf{X}, \mathbf{X}^\top\mathbf{y})$. Es decir, PLS1 puede ser considerado una técnica de regularización del estimador de mínimos cuadrados habitual.

# CAPÍTULO 4
# PLS aplicado a datos funcionales

En este capítulo exploraremos la aplicación de PLS al análisis de datos funcionales. En particular nos centraremos en su aplicación a la resolución de modelos de regresión lineal con respuesta escalar. No obstante, gran parte del análisis que se llevará a cabo puede extenderse a modelos con respuesta vectorial. Asimismo, por simplicidad, supondremos que las funciones involucradas están definidas en el intervalo $[0, 1]$. Sin embargo los argumentos presentados serían también válidos para cualquier otro intervalo.

## 4.1. Modelo de regresión funcional

Para el análisis de datos funcionales, en Kokoszka y Reimherr (2017) se proponen varios modelos de regresión dependiendo de la naturaleza de la variable respuesta. En nuestro caso, utilizamos el modelo de regresión con respuesta escalar, que tiene la siguiente expresión poblacional:

$$Y = \int_0^1 \beta(t)X(t)dt + \varepsilon, \tag{4.1}$$

donde $X(t)$ es un proceso estocástico en $L^2$, $Y$ es la variable aleatoria respuesta y $\varepsilon$ es ruido aleatorio independiente de $X(t)$. El ajuste del modelo se traduce a encontrar una función $\widehat{\beta}(t) \in L^2$.

Sin embargo, el ajuste de este modelo no es tan sencillo como podría parecer a primera vista. Una posibilidad sería discretizar las funciones e intentar ajustar el modelo finito-dimensional resultante. Esto es, si optamos por discretizar las funciones en los puntos $\{t_m\}_{m=1}^M$ y contamos con $N$ observaciones, se llega al siguiente modelo multivariante (en su versión muestral)

$$y_n = \sum_{m=1}^{M} w(t_m)X_n(t_m)\beta(t_m) + \epsilon_n, \quad n = 1, \dots N,$$

donde $w(t_m)$ representa el peso de la cuadratura de integración en el punto $m$-ésimo de la rejilla.

Es inmediato observar que contamos con $N$ ecuaciones y $M$ incógnitas. Por tanto, si la rejilla es fina, podemos llegar a una situación en la que hay múltiples soluciones. Este resultado es particularmente preocupante ya que sería de esperar que el uso de

rejillas más finas a la hora de discretizar el problema llevase a resultados más cercanos a la función buscada.

Por otro lado, se podría considerar realizar un ajuste por mínimos cuadrados como se suele llevar a cabo en el caso finito-dimensional. En la proposición 3.1 vimos que el estimador de mínimos cuadrados de un modelo de regresión múltiple era $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$. Sin embargo, el operador de covarianzas, análogo a la matriz $\mathbf{X}^\top \mathbf{X}$, no es invertible en el caso funcional.

Ante esta situación, una de las soluciones más habituales es sustituir las funciones en (4.1) por su expansión en alguna base $\mathcal{B} = \{\phi_k\}_{k=1}^K$. El uso de una base finita implica la imposibilidad de expresar cualquier función en la misma. Sin embargo, esta misma propiedad será la que nos permitirá obtener soluciones a nuestro problema de regresión. Asimismo, esta aproximación al problema tiene la ventaja de que, incluso cuando $K$ es pequeño, la expresión de las observaciones en la base es una función tan suave como los elementos de la base.

Denotaremos $\{x_k\}_{k=1}^K$ a los coeficientes de $X$ en la base $\mathcal{B}$ (son variables aleatorias). Similarmente, denotamos $\{b_k\}_{k=1}^K$ a los coeficientes de $\beta$. Es decir, contamos con las siguientes identidades:

$$X(t) = \sum_{k=1}^{K} x_k \phi_k(t), \qquad \beta(t) = \sum_{k=1}^{K} b_k \phi_k(t).$$

Para reescribir el modelo (4.1), es muy útil considerar la matriz de productos internos $\mathbf{G} = (g_{i,j})_{1 \leq i,j \leq K}$ donde $g_{i,j} = \int_0^1 \phi_i(t)\phi_j(t)dt$. Utilizando esta matriz, podemos reescribir el modelo como:

$$\begin{aligned} Y &= \int_0^1 \left( \sum_{i=1}^{K} x_i \phi_i(t) \sum_{j=1}^{K} b_j \phi_j(t) dt \right) + \varepsilon = \\ &= \sum_{1 \leq i,j \leq K} x_i b_j \int_0^1 \phi_i(t)\phi_j(t)dt + \varepsilon = \sum_{1 \leq i,j \leq K} x_i b_j g_{i,j} + \varepsilon = \\ &= \tilde{X}^\top \mathbf{G}\mathbf{b} + \varepsilon, \end{aligned}$$

donde $\tilde{X}$ denota el vector aleatorio compuesto por las variables aleatorias $x_i$ y $\mathbf{b}$ denota el vector de coeficientes de $\beta$. Observando este modelo, vemos que hemos llegado a un modelo de regresión múltiple en dimensión finita que sí sabemos ajustar.

Con frecuencia, se opta por utilizar una base ortonormal. De este modo, la matriz $\mathbf{G}$ obtenida en este desarrollo es la identidad y se simplifica aún más el problema. Por el contrario, en determinadas circunstancias, puede ser beneficioso expresar los datos en una base diferente a la del coeficiente ($\beta$). En estos casos, se sustituye la matriz de productos internos $\mathbf{G}$ por la matriz de productos cruzados entre los elementos de las bases.

Por otra parte, si bien la expansión en bases nos permite resolver el problema de regresión, introduce otro problema: la elección de la base. Una de las elecciones más habituales es una base genérica como la base de Fourier o B-Splines. Sin embargo, no contamos con ninguna garantía de que la información útil para el problema de

regresión se encuentre recogida en los primeros coeficientes de la expansión en estas bases.

Para mitigar este riesgo, a menudo se utiliza la base de los componentes principales. De forma análoga al caso finito-dimensional, esta base es aquella tal que las proyecciones de $X$ a lo largo de sus elementos maximizan la varianza. Asimismo, es bien conocido (ver Ramsay y Silverman (2013)) que esta base está compuesta por las autofunciones del operador de covarianza.

Sin embargo, aunque la expansión en la base de componentes principales sea muy útil a la hora de recoger la información en $X$ en los primeros componentes, tampoco tenemos la seguridad de que la información que relaciona $X$ con $Y$ se encuentre en estos primeros componentes. Por el contrario, la definición de bases obtenidas con el criterio PLS podría ser útil en estas circunstancias ya que, al maximizar la covarianza entre $X$ e $Y$, es de esperar que la información necesaria para la regresión se encuentre en las primeras componentes.

Finalmente, también es de vital importancia la elección del número de funciones de la base ($K$). Este parámetro puede interpretarse como un parámetro de regularización ya que restringe el espacio en el que se consideran posibles soluciones ($\beta$) del problema.

## 4.2. Maximización de la covarianza funcional

Al igual que en el caso finito-dimensional, hay varios métodos para maximizar la covarianza. En Preda y Saporta (2005) y Febrero-Bande, Galeano et al. (2017) se estudian métodos similares al algoritmo NIPALS en su versión de regresión. Por otra parte, en Delaigle y Hall (2012) se propone un método alternativo (APLS) y se lleva a cabo un análisis teórico de sus propiedades asintóticas.

Con el objetivo de mantener un claro paralelismo con los métodos ya estudiados, se ha optado por explorar los métodos introducidos en Preda y Saporta (2005), pero limitados al caso en el que la variable $Y$ es escalar. La mayor diferencia de estos con respecto al método finito-dimensional es el uso del producto interno en $L^2$, que da lugar a la norma $L^2$, definida como

$$\|f\|_{L^2} = \left(\int_0^1 (f(t))^2\, dt\right).$$

Utilizando esta norma, los pesos de PLS para cada iteración se pueden definir como

$$w_l = \underset{w\in L^2, \|w\|_{L^2}=1}{\arg\max} \left(\operatorname{cov}\left(\int_0^1 X_l(t)w(t)\, dt, Y_l\right)\right)^2. \tag{4.2}$$

Por otro lado, las componentes se extraen ahora como proyecciones con el producto interno de $L^2$, es decir,

$$t_l = \int_0^1 X_{l-1}(t)w(t)\, dt.$$

Finalmente, la deflación al final de cada paso viene dada por

$$X_l(t) = X_{l-1}(t) - p_l(t)t_l, \qquad Y_l(t) = Y_{l-1}(t) - q_l(t)t_l,$$

donde $p_l(t)$ y $q_l(t)$ se pueden definir como el resultado del ajuste de la regresión por mínimos cuadrados de $X_{l-1}$ e $Y_{l-1}$ sobre $t_l$. Esto es, $X_l$ es el residuo del ajuste por mínimos cuadrados del modelo de regresión $X_{l-1} = p_l t_l + \varepsilon_l^x$, donde $p_l$ es el parámetro a ajustar. Similarmente, $Y_l$ es el residuo correspondiente al modelo $Y_{l-1} = q_l t_l + \varepsilon_l^y$, donde $q_l$ es el parámetro a ajustar.

### 4.2.1. Comparación con el método multivariante

Si comparamos estas definiciones con las expresiones análogas en NIPALS, vemos que la única diferencia radica en el uso del producto interno de $L^2$. En consecuencia, es de esperar que podamos obtener un algoritmo muy similar.

Una posible aproximación al problema (4.2) es discretizar $X(t)$ en una rejilla y aplicar el método multivariante que ya hemos analizado. Por otro lado, también es posible utilizar un desarrollo en bases. En tal caso, por la linealidad de la integral (y de la esperanza), tal y como ocurría con la regresión, llegamos al método multivariante que ya hemos estudiado.

Efectuar dicho desarrollo en bases finitas plantea de nuevo la dificultad de elegir la base. Sin embargo, se podría optar por llevar a cabo el cálculo de la base PLS utilizando una base genérica (Fourier o B-Splines, por ejemplo) con muchas funciones. Esto nos proporcionaría una base con menos funciones pero con la capacidad de recoger la información útil para el problema de regresión en las primeras componentes de la misma. En este sentido, PLS se puede utilizar de nuevo como un método de reducción de la dimensionalidad que lleve al ajuste de modelos de regresión con menos parámetros.

### 4.2.2. Análisis en términos funcionales

En contraposición a la sección anterior, buscamos ahora desarrollar el método iterativo planteado en (4.2) pero manteniéndonos en el terreno funcional.

Para buscar las funciones $\{w_l\}_{l=1}^L$ que maximicen la covarianza funcional, es necesario definir una serie de operadores que resultan análogos a las matrices de covarianzas cruzadas en el caso finito-dimensional. En Preda y Saporta (2005), definen los operadores para el caso general en el que la variable respuesta es un vector aleatorio. A continuación, incluimos las definiciones simplificadas para el caso escalar

**Definición 4.1.** Dado un proceso estocástico $X(t)$ en $L^2$ y una variable aleatoria escalar $Y$, se definen los operadores:

$$C_{YX} : L^2 \longrightarrow \mathbb{R} \qquad\qquad C_{XY} : \mathbb{R} \longrightarrow L^2$$
$$f \longrightarrow \int_0^1 \mathbb{E}(X(s)Y) f(s)\, ds \qquad\qquad x \longrightarrow f(t) = \mathbb{E}(X(t)Y)x.$$

Asimismo, se definen por composición los operadores $U_X : L^2 \to L^2$ y $U_Y : \mathbb{R} \to \mathbb{R}$ como

$$U_X = C_{XY} \circ C_{YX}, \qquad U_Y = C_{YX} \circ C_{XY}.$$

Centramos nuestra atención en el operador $U_X$ ya que este se corresponde con la matriz $\mathbf{X}^\top \mathbf{y}\mathbf{y}^\top \mathbf{X}$ y, en el caso multivariante, las $\mathbf{w}_l$ se obtenían como autovectores de dicha matriz.

**Proposición 4.1.** *La función de covarianzas $c(t) = \operatorname{cov}(X(t), Y)$ es una autofunción asociada al autovalor dominante del operador $U_X$.*

*Demostración.*

Por la definición anterior, se puede escribir la siguiente identidad para el operador $U_X$:

$$U_X(f)(t) = \mathbb{E}(X(t)Y) \int_0^1 \mathbb{E}(X(s)Y) f(s)\, ds.$$

A partir de esta expresión, podemos acotar la norma $L^2$ de la imagen por el operador de una función genérica $f \in L^2$ aplicando la desigualdad de Cauchy Schwarz:

$$\begin{aligned}
\|U_X(f)\|_{L^2}^2 &= \int_0^1 \left( \mathbb{E}(X(t)Y) \int_0^1 \mathbb{E}(X(s)Y) f(s)\, ds \right)^2 dt = \\
&= \left( \int_0^1 \mathbb{E}(X(s)Y) f(s)\, ds \right)^2 \int_0^1 \left( \mathbb{E}(X(t)Y) \right)^2 dt \leq \\
&\leq \|E(X(\cdot)Y)\|_{L^2}^2 \|f(\cdot)\|_{L^2}^2 \|E(X(\cdot)Y)\|_{L^2}^2 = \\
&= \|E(X(\cdot)Y)\|_{L^2}^4 \|f(\cdot)\|_{L^2}^2.
\end{aligned}$$

Así, $\|U_X(f)\|_{L^2} \leq \|E(X(\cdot)Y)\|_{L^2}^2 \|f(\cdot)\|_{L^2}$ y, por tanto, los autovalores del operador están acotados por $\|E(X(\cdot)Y)\|_{L^2}^2$. Por otro lado, recordando que las variables aleatorias están centradas, tenemos que $c(t) = \operatorname{cov}(X(t), Y) = \mathbb{E}(X(t)Y)$ y podemos comprobar que esta es una autofunción del operador con un autovalor que alcanza la cota:

$$\begin{aligned}
U_X(c)(t) &= \mathbb{E}(X(t)Y) \int_0^1 \mathbb{E}(X(s)Y)\mathbb{E}(X(s)Y)\, ds = \\
&= \|\mathbb{E}(X(\cdot)Y)\|_{L^2}^2 \mathbb{E}(X(t)Y) = \|\mathbb{E}(X(\cdot)Y)\|_{L^2}^2 c(t).
\end{aligned}$$

□

Por otro lado, queremos probar que las funciones que maximizan la covarianza son autofunciones del operador $U_X$.

**Proposición 4.2.** *Consideramos $X(t)$ un proceso estocástico en $L^2$ e $Y$ una variable aleatoria escalar. Además, sin pérdida de generalidad, suponemos que $\mathbb{E}(X(t)) = 0\ \forall t \in [0,1]$ y $\mathbb{E}(Y) = 0$. En estas circunstancias, la función $f \in L^2([0,1])$ que maximiza*

$$\left(\operatorname{cov}\left(\int_0^1 X(t)f(t)\,dt,\ Y\right)\right)^2$$

*es una autofunción del operador $U_X$ asociada a su autovalor dominante.*

*Demostración.* En el apéndice A (página 47) □

Estas últimas dos proposiciones nos proporcionan una fórmula cerrada para $w_l$. Por tanto, podemos ya definir el método iterativo análogo al algoritmo 5. Este viene dado por las siguientes identidades:

$$\begin{aligned} w_l(t) &= \frac{\operatorname{Cov}(X_{l-1}(t), Y)}{\|\operatorname{Cov}(X_{l-1}(\cdot), Y)\|_{L^2}}, \\ t_l &= \int_0^1 X_l(t) w_l(t)\,dt, \\ p_l(t) &= \frac{\operatorname{Cov}(X_{l-1}(t), t_l)}{\operatorname{Var}(t_l)}, \\ X_l(t) &= X_{l-1}(t) - t_l p_l(t). \end{aligned}$$

De hecho, sustituyendo la varianza y covarianza por sus fórmulas muestrales, llegamos a un método completamente análogo al algoritmo 5, siendo la única diferencial el uso del producto interno en $L^2$.

Este método se puede aplicar tanto a funciones discretizadas en una rejilla como expresadas en un desarrollo en bases. En el primer caso, es necesario aproximar las integrales empleando unos pesos de integración. En el segundo caso se aprovecha la la linealidad de la expansión en bases y la integral. En particular, solo es necesario emplear las matrices de productos internos en vez de las matrices de pesos de integración. Se puede encontrar una implementación de este método en el paquete de R `fda.usc` (Febrero-Bande y Oviedo de la Fuente (2012)).

# CAPÍTULO 5
# Resultados computacionales

El objetivo de este capítulo es mostrar el rendimiento que puede ser esperado de la aplicación de PLS a problemas de regresión. Asimismo, compararemos su capacidad de resumir información en las componentes extraídas con la de PCA. En primer lugar, exploraremos el caso finito-dimensional y, a continuación, su aplicación a datos funcionales.

Hay varias formas de medir el rendimiento de la aplicación de PLS a problemas de regresión. En primer lugar, podríamos medir la capacidad del modelo de ajustarse a los datos. Sin embargo, un buen resultado en este aspecto no garantiza que el modelo sea útil a la hora de llevar a cabo predicciones para nuevas observaciones.

Por tanto, hemos optado por llevar a cabo una separación de los datos de forma que un solo un subconjunto de los mismos se utilice para calcular la matriz de regresión. Una vez el modelo ha sido ajustado, el resto de los datos se pueden utilizar para medir su rendimiento. Para reducir la dependencia de los resultados en la partición de datos elegida, se ha optado por utilizar validación cruzada, midiendo el rendimiento del modelo con distintas particiones.

Para cada partición de los datos, el rendimiento del modelo se ha determinado calculando el coeficiente de determinación (denotado habitualmente como $R^2$). Si denotamos como $\mathbf{y} = (y_1, \dots, y_N)^\top$ a las observaciones de la variable respuesta en el conjunto de test, $\hat{\mathbf{y}} = (\hat{y_1}, \dots \hat{y_N})^\top$ las predicciones del modelo para cada observación e $\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, el coeficiente de determinación se calcula como

$$R^2(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2}.$$

Haciendo la media de los valores de $R^2$ para cada partición de datos, llegamos a una medida del rendimiento con menor dependencia en la partición de datos elegida. El valor máximo que podemos obtener es 1, que indicaría una predicción perfecta de los valores. Por el contrario, predicciones arbitrariamente malas pueden dar lugar a valores de $R^2$ negativos. Esto es consecuencia de no medir el rendimiento del modelo con los mismo datos con los que se ha ajustado.

Sin embargo, un predictor que siempre predijese el valor medio de las respuestas recibiría una puntuación de 0. Por tanto es de esperar que nuestros experimentos resulten en valores entre 0 y 1. Cuanto mayor sea el valor, una mayor proporción de la varianza ha podido ser explicada por el modelo, llevando a mejores resultados. Para estas pruebas, hemos recurrido a las implementaciones proporcionadas en scikit-learn (Pedregosa et al. (2011)).

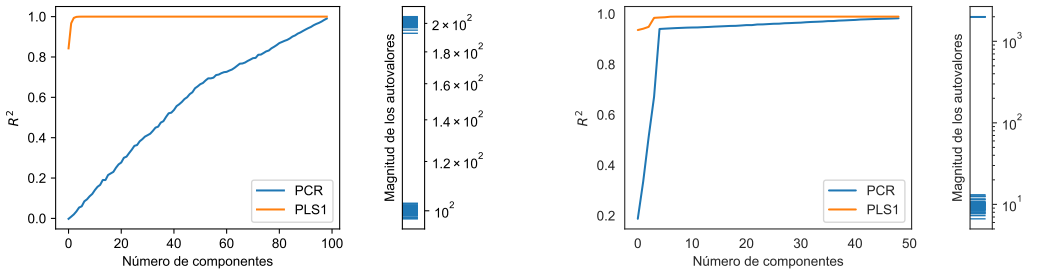
## 5.1. Influencia de la matriz de covarianzas en el rendimiento

Es bien conocido que PCA obtiene sus mejores resultados cuando la matriz de covarianza tiene unos pocos autovalores que dominan frente al resto. Por tanto, cabe preguntarse si existe un resultado parecido en el caso de PLS. En esta sección, nos centramos en la aplicación de PLS a la regresión múltiple.

Un resultado que no hemos cubierto en este documento es la equivalencia de la regresión PLS1 con el método del gradiente conjugado. No se trata de un resultado complicado ya que el algoritmo del gradiente conjugado explora espacios de Krylov crecientes en cada iteración, por lo que es de esperar que minimizar el error cuadrático con dicho algoritmo resulte en el estimador PLS. Se puede consultar la prueba en Phatak y Hoog (2002).

Por otro lado, la influencia de la distribución de los autovalores en la velocidad de convergencia del algoritmo del gradiente conjugado se ha analizado en detalle en Wright, Nocedal et al. (1999). La conclusión alcanzada es que podemos esperar muy buenos resultados cuando los autovalores estén agrupados en bloques de autovalores de magnitud similar. Hemos creado una simulación para respaldar este resultado teórico.

En esta simulación hemos buscado predecir una variable aleatoria $Y$ que es combinación lineal de una serie de variables aleatorias normales $X_1, \dots, X_M$. El vector aleatorio $X = (X_1, \dots, X_M)^\top$ ha sido modelado como una normal multivariante. De este modo, es sencillo generar una matriz de datos $\mathbf{X}$ con una matriz de covarianzas específica. Basta construir cada fila de $\mathbf{X}$ como una observación del vector normal multivariante con la matriz de covarianza deseada.




(a) Dos bloques de autovalores.

(b) Unos pocos autovalores dominantes.

Figura 5.1: Comparación de la varianza explicada entre PLS1 y PCR según la distribución de los autovalores de la matriz de covarianzas.

Dos matrices de covarianzas distintas se han considerado. En primer lugar, se ha construido una matriz cuyos autovalores se agrupen en dos clusters. En segundo lugar, se ha considerado una matriz que tenga unos pocos autovalores que claramente dominen al resto. En la figura 5.1, se presentan los resultados de estos dos escenarios. Para cada uno de ellos, incluimos la evolución del rendimiento de PCR y PLS1 según

el número de componentes. Asimismo, se incluye a la derecha de cada comparativa un gráfico mostrando la distribución de los autovalores de la matriz de covarianzas utilizada en dicha comparativa.

En la figura 5.1, se puede ver como PLS1 obtiene muy buenos resultados cuando encontramos muy pocos bloques de autovalores. En 5.1a, los autovalores se encuentran agrupados en dos bloques pero de magnitud similar. Por tanto, PCR necesita un gran número de componentes para obtener buenos resultados mientras que PLS1 es capaz de explicar prácticamente toda la varianza con los dos primeros componentes.

En el escenario mostrado en 5.1b, 5 autovalores toman valores mucho mayores a los del resto. En este caso, vemos como PCR obtiene buenos resultados una vez alcanza las 5 componentes. Similarmente, el rendimiento de PLS1 crece hasta llegar a los 5 componentes. Sin embargo, como los autovalores de mayor magnitud son similares, la progresión es mucho más rápida.

En la figura 5.2, comparamos el rendimiento de PLS1 según la distribución de los autovalores de la matriz de covarianzas. Como se puede apreciar, el rendimiento es mejor cuanto menos bloques de autovalores haya. En particular, los peores resultados se obtienen cuando la distribución de autovalores es uniforme en un intervalo lo bastante grande como para que no se comporten como un bloque pero lo bastante pequeño como para que parte de los autovalores no dominen al resto.

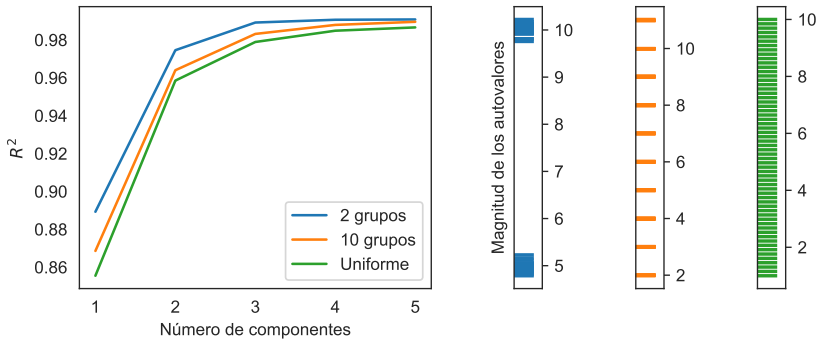


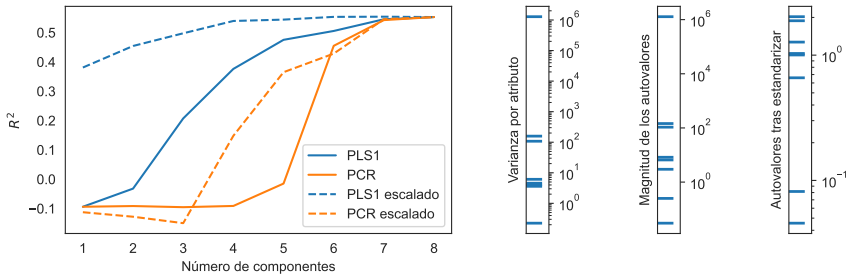

Figura 5.2: Comparación de la varianza explicada por PLS1 según la distribución de los autovalores de la matriz de covarianzas.

## 5.2. Resultados en conjuntos de datos reales

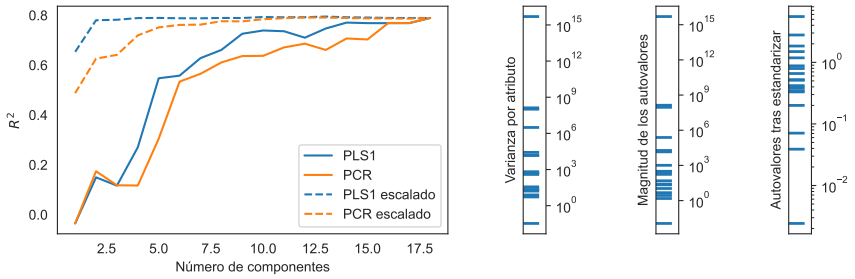
En esta sección, estudiamos los resultados obtenidos por PLS1 en conjuntos de datos reales. Compararemos su rendimiento con el de PCR y comprobaremos que, al igual que en PCR, el escalado de los datos puede jugar un papel crucial. En particular, denominamos escalado a la división de las observaciones de cada variable por su varianza, de forma que tengan varianza unidad.

Se incluyen los resultados para dos conjuntos de datos: *California Housing* (Kelley Pace y Barry (1997)) y *Life Expectancy* (Russell y Wang (2018)). En el primero de ellos, el objetivo es predecir el precio medio de las casas en California en

cada distrito utilizando atributos como la edad media de la población en el distrito o los ingresos medios. En el segundo, la variable respuesta es la esperanza de vida en un país dados valores como la población, el nivel de educación o la mortalidad infantil.




(a) Conjunto de datos *California Housing*.




(b) Conjunto de datos *Life Expectancy*.

Figura 5.3: Comparación del rendimiento de PLS1 y PCR en conjuntos de datos reales.

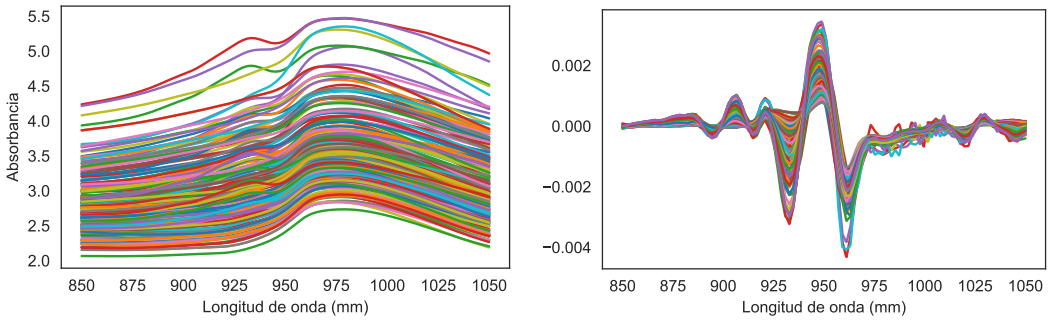
Como se puede apreciar, en ambos casos, el resultado mejora cuando se lleva a cabo un escalado de los datos. Comprobando la distribución de las varianzas de los atributos, se puede comprobar que, en ambos casos, uno de los atributos tiene una varianza mucho mayor. En estas situaciones, al igual que ocurre en PCR, el escalado permite que aumentar la importancia de los demás atributos a la hora de maximizar la covarianza (varianza en el caso de PCR).

Por otro lado, estos ejemplos muestran claramente como PLS1 alcanza mejores resultados que PCR. Observando la distribución de los autovalores, tanto antes como después de normalizar, podemos ver que PLS1 puede producir mejores resultados que PCR incluso cuando la estructura de autovalores no coincide con el mejor caso estudiado.

## 5.3. Resultados en conjuntos de datos funcionales

En esta sección presentaremos los resultados obtenidos al aplicar la regresión basada en PLS a conjuntos de datos funcionales (FPLSR). Para ello utilizaremos el conjunto de datos Tecator (descargado de http://lib.stat.cmu.edu/datasets/tecator).

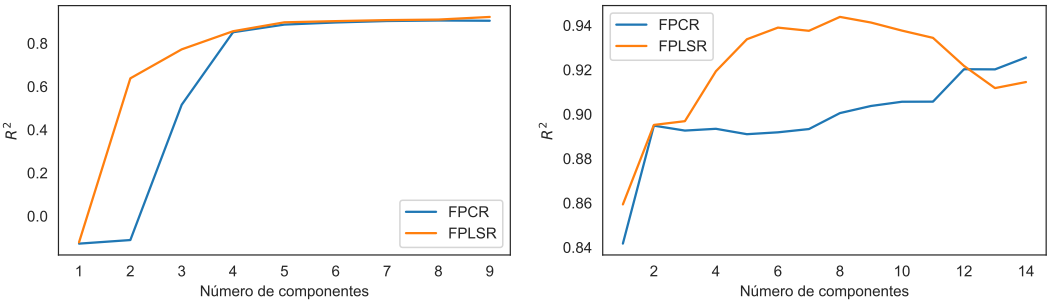
Nuestro objetivo es predecir el contenido de proteína de cada una de las muestras de carne dado su espectro de absorción o la segunda derivada del mismo. En este conjunto de datos, el uso de la segunda derivada como variable regresora suele conducir a mejores resultados. Por tanto, hemos optado por medir el rendimiento en ambos casos. En la figura 5.4 se puede ver el aspecto de los espectros de absorción y de sus segundas derivadas.




(a) Absorbancia. (b) Segundas derivadas.

Figura 5.4: Espectros de absorción de las 215 muestras de carne.

Por otro lado, en la figura 5.5, se puede ver la puntuación obtenida por FPCR y FPLSR al predecir el contenido de proteína a partir de los espectros de absorción (5.5a) y a partir de las segundas derivadas (5.5b).Como se puede apreciar, FPLSR obtiene resultados mucho mejores cuando hay pocos componentes. Según el número de componentes aumenta, los resultados se asemejan cada vez más. En 5.5b podemos apreciar como el aumento de componentes puede llevar a un sobreajuste del modelo, disminuyendo su rendimiento. Este sobreajuste aparece antes utilizando FPLSR, pero también se puede encontrar en FPCR aumentando el número de componentes.




(a) A partir de la absorbancia. (b) A partir de las segunda derivadas.

Figura 5.5: Rendimiento de FPCR y FPLSR según el número de componentes.

Esta diferencia cuando el número de componentes es bajo manifiesta la mayor capacidad del criterio PLS a la hora de extraer la información relevante para el problema de regresión de las observaciones funcionales. Según el número de componentes aumenta, ambas bases son capaces de recoger más información y esta ventaja va desapareciendo.

# CAPÍTULO 6
# Conclusiones

En este trabajo, hemos llevado a cabo un profundo análisis de PLS como método de reducción de la dimensionalidad. Entender esta aplicación de PLS en profundidad ha sido clave a la hora de estudiar la extensión del método a otros escenarios, ya sean problemas de regresión o el análisis de datos funcionales. Si bien la mayoría de los resultados presentados están presentes en multitud de artículos y estudios sobre PLS, el orden y el enfoque de los mismos ha sido cuidadosamente ajustado con el objetivo de proporcionar una visión de PLS tan clara como sea posible.

Al contrario que en la inmensa mayoría de la literatura, hemos evitado describir el algoritmo NIPALS durante más de diez páginas. Por el contrario, hemos relacionado PLS con otras técnicas más conocidas y hemos planteado el problema que NIPALS busca resolver. De este modo, el algoritmo se convierte solo en un método para alcanzar un objetivo bien establecido, facilitando su comprensión.

Por otro lado, hemos conseguido enunciar PLS2 y PLS1 como leves modificaciones del método de reducción de la dimensionalidad. Estas modificaciones han sido motivadas enunciando el modelo de relaciones lineales entre las variables latentes extraídas, un paso omitido en gran parte de los artículos. En particular, para PLS1, se han demostrado los resultados que permiten obtener un algoritmo tan simple y se ha proporcionado una prueba completa de su estrecha relación con el estimador de mínimos cuadrados habitual.

Respecto a la aplicación de PLS al análisis de datos funcionales, por limitaciones de tiempo y extensión, hemos preferido limitarnos al modelo de regresión con respuesta escalar. Esta es un área del trabajo donde habría sido posible profundizar más. El análisis presentado puede ser fácilmente extendido a modelos con respuesta multivariante o, incluso, funcional. Asimismo, muy recientemente se ha estado estudiando la incorporación de regularización a estos modelos (ver Aguilera et al. (2016)). La aplicación de PLS en estas situaciones es un tema sobre el que a día de hoy se siguen publicando artículos explorando nuevas facetas y aplicaciones.

Finalmente, el cierre del trabajo ha consistido en una serie de experimentos numéricos que muestran el rendimiento de PLS en diversos escenarios. De entre los tres grupos de experimentos llevados a cabo, el más significativo es probablemente el estudio de la influencia de la estructura de autovalores de la matriz de covarianza. Este resultado se limita a extender a PLS propiedades ya conocidas de otro método equivalente (gradiente conjugado). Sin embargo, nos aporta intuición acerca de los escenarios en los que PLS puede ser dramáticamente superior a PCA.

# Bibliografía


Aguilera, A., Aguilera-Morillo, M., y Preda, C. (2016). Penalized versions of functional PLS regression. *Chemometrics and Intelligent Laboratory Systems*, *154*, 80-92.

De Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and intelligent laboratory systems*, *18*(3), 251-263.

Delaigle, A., y Hall, P. (2012). Methodology and theory for partial least squares applied to functional data. *The Annals of Statistics*, *40*(1), 322-352.

Eldén, L. (2004). Partial least-squares vs. Lanczos bidiagonalization—I: analysis of a projection method for multiple regression. *Computational Statistics and Data Analysis*, *46*(1), 11-31.

Febrero-Bande, M., Galeano, P., y González-Manteiga, W. (2017). Functional Principal Component Regression and Functional Partial Least-squares Regression: An Overview and a Comparative Study. *International Statistical Review*, *85*(1), 61-83.

Febrero-Bande, M., y Oviedo de la Fuente, M. (2012). Statistical Computing in Functional Data Analysis: The R Package fda.usc. *Journal of Statistical Software*, *51*(4), 1-28.

Geladi, P., y Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, *185*, 1-17.

Höskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics*, *2*(3), 211-228.

Kelley Pace, R., y Barry, R. (1997). Sparse spatial autoregressions. *Statistics and Probability Letters*, *33*(3), 291-297.

Kokoszka, P., y Reimherr, M. (2017). *Introduction to Functional Data Analysis*. CRC Press.

Lyttkens, E. (1972). Regression aspects of canonical correlation. *Journal of Multivariate Analysis*, *2*(4), 418-439.

Noonan, R., y Wold, H. (1977). NIPALS Path Modelling with Latent Variables. *Scandinavian Journal of Educational Research - SCAND J EDUC RES*, *21*, 33-61.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., y Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

Phatak, A., y Hoog, F. (2002). Exploiting the connection between PLS, Lanczos methods and conjugate gradients: Alternative proofs of some properties of PLS. *Journal of Chemometrics*, *16*, 361-367.

Preda, C., y Saporta, G. (2005). PLS regression on a stochastic process [Partial Least Squares]. *Computational Statistics and Data Analysis*, *48*(1), 149-158.

Ramsay, J., y Silverman, B. (2013). *Functional Data Analysis*. Springer New York.

Rosipal, R., y Krämer, N. (2005). Overview and Recent Advances in Partial Least Squares. *Lecture Notes in Computer Science*, *3940*, 34-51.

Russell, D., y Wang, D. (2018). Life Expectancy (WHO), Version 1.

Vinod, H. (1976). Canonical ridge and econometrics of joint production. *Journal of Econometrics*, *4*(2), 147-166.

Wegelin, J. (2000). A Survey of Partial Least Squares (PLS) Methods, with Emphasis on the Two-Block Case. *Technical report*.

Wold, H. (1975). Soft Modelling by Latent Variables: The Non-Linear Iterative Partial Least Squares (NIPALS) Approach. *Journal of Applied Probability*, *12*(S1), 117-142.

Wold, H. (1980). Model Construction and Evaluation When Theoretical Knowledge Is Scarce. En *Evaluation of Econometric Models* (pp. 47-74). Academic Press.

Wright, S., Nocedal, J., et al. (1999). Numerical optimization. *Springer Science*, *35*(67-68), 7.

# APÉNDICE A
# Demostraciones

**Proposición 2.4** Dado un vector $\mathbf{w} \in \mathbb{R}^M$ cualquiera y $0 < l \leq L$, si definimos $\mathbf{X}_l$ y $\mathbf{w}_l$ como en el algoritmo 1, es posible descomponer $\mathbf{w}$ como $\mathbf{w} = \mathbf{w}^\perp + \mathbf{w}^\parallel$ donde $\mathbf{w}^\perp \perp \mathbf{w}_l$ y se verifica que $\mathbf{X}_l\mathbf{w} = \mathbf{X}_{l-1}\mathbf{w}^\perp$.

*Demostración.*

Como $\mathbf{w}_l$ es un autovector de $\mathbf{X}_{l-1}^\top\mathbf{X}_{l-1}$, se cumple que, para algún $\lambda \in \mathbb{R}$,

$$\mathbf{w}_l = \lambda\mathbf{X}_{l-1}^\top\mathbf{X}_{l-1}\mathbf{w}_l \quad \text{y, equivalentemente,} \quad \mathbf{w}_l^\top = \lambda\mathbf{w}_l^\top\mathbf{X}_{l-1}^\top\mathbf{X}_{l-1}. \tag{A.1}$$

Asimismo, como $\mathbb{R}^M = \operatorname{span}\{\mathbf{w}_l\} \oplus \operatorname{span}\{\mathbf{w}_l\}^\perp$, por las propiedades de la suma directa, tenemos una descomposición $\mathbf{w} = \mathbf{w}^\parallel + \mathbf{w}^\perp$ tal que $\mathbf{w}^\parallel \in \operatorname{span}\{\mathbf{w}_l\}$ y $\mathbf{w}^\perp \in \operatorname{span}\{\mathbf{w}_l\}^\perp$.

Por otro lado aprovechando que $\|\mathbf{w}_l\| = 1$, es inmediato ver que $\mathbf{w}^\parallel = \|\mathbf{w}^\parallel\|\mathbf{w}_l$ ya que todos los vectores en $\operatorname{span}\{\mathbf{w}_l\}$ son constantes multiplicadas por $\mathbf{w}_l$, que tiene norma unidad. En consecuencia, obtenemos la siguiente identidad:

$$\mathbf{w}_l\mathbf{w}_l^\top\mathbf{w}^\parallel = \mathbf{w}_l\mathbf{w}_l^\top(\|\mathbf{w}^\parallel\|\mathbf{w}_l) = \mathbf{w}_l\mathbf{w}_l^\top(\|\mathbf{w}^\parallel\|\mathbf{w}_l) = \mathbf{w}_l\|\mathbf{w}^\parallel\| = \mathbf{w}^\parallel. \tag{A.2}$$

Utilizando estas identidades y la identidad de deflación en la línea 6 del algoritmo 1, podemos ya obtener el resultado buscado:

$$\begin{aligned}
\mathbf{X}_l\mathbf{w} &= \left(\mathbf{I} - \frac{\mathbf{t}_l\mathbf{t}_l^\top}{\mathbf{t}_l^\top\mathbf{t}_l}\right)\mathbf{X}_{l-1}\mathbf{w} = \left(\mathbf{I} - \frac{\mathbf{X}_{l-1}\mathbf{w}_l\mathbf{w}_l^\top\mathbf{X}_{l-1}^\top}{\mathbf{w}_l^\top\mathbf{X}_{l-1}^\top\mathbf{X}_{l-1}\mathbf{w}_l}\right)\mathbf{X}_{l-1}(\mathbf{w}^\perp + \mathbf{w}^\parallel) = \\
&= \left(\mathbf{X}_{l-1}\mathbf{w}^\perp - \frac{\mathbf{X}_{l-1}\mathbf{w}_l(\mathbf{w}_l^\top\mathbf{X}_{l-1}^\top\mathbf{X}_{l-1})\mathbf{w}^\perp}{\mathbf{w}_l^\top\mathbf{X}_{l-1}^\top\mathbf{X}_{l-1}\mathbf{w}_l}\right) + \\
&+ \left(\mathbf{X}_{l-1}\mathbf{w}^\parallel - \frac{\mathbf{X}_{l-1}\mathbf{w}_l(\mathbf{w}_l^\top\mathbf{X}_{l-1}^\top\mathbf{X}_{l-1})\mathbf{w}^\parallel}{\mathbf{w}_l^\top\mathbf{X}_{l-1}^\top\mathbf{X}_{l-1}\mathbf{w}_l}\right) = \\
&\overset{(A.1)}{=} \mathbf{X}_{l-1}\mathbf{w}^\perp - \frac{\mathbf{X}_{l-1}\mathbf{w}_l(\lambda\mathbf{w}_l^\top)\mathbf{w}^\perp}{\lambda} + \mathbf{X}_{l-1}\mathbf{w}^\parallel - \frac{\mathbf{X}_{l-1}\mathbf{w}_l(\lambda\mathbf{w}_l^\top)\mathbf{w}^\parallel}{\lambda} = \\
&\overset{(A.2)}{=} \mathbf{X}_{l-1}\mathbf{w}^\perp - 0 + \mathbf{X}_{l-1}\mathbf{w}^\parallel - \mathbf{X}_{l-1}\mathbf{w}^\parallel = \\
&= \mathbf{X}_{l-1}\mathbf{w}^\perp.
\end{aligned}$$

□

**Proposición 2.6** Los vectores extraídos por el algoritmo 3 cumplen:

1. Los vectores de pesos de cada iteración son perpendiculares, es decir, si $j > i$, $\mathbf{w}_j \perp \mathbf{w}_i$ y $\mathbf{c}_j \perp \mathbf{c}_i$. Por tanto, matricialmente, se cumple que
$$\mathbf{W}_L^\top \mathbf{W}_L = \mathbf{I}, \qquad \mathbf{C}_L^\top \mathbf{C}_L = \mathbf{I}.$$
2. Las *scores* extraídas en cada iteración son perpendiculares a las anteriores, es decir, si $j > i$, $\mathbf{t}_j \perp \mathbf{t}_i$ y $\mathbf{u}_j \perp \mathbf{u}_i$. Por tanto, matricialmente, se cumple que $\mathbf{T}_L^\top \mathbf{T}_L$ y $\mathbf{U}_L^\top \mathbf{U}_L$ son matrices diagonales.
3. En cada iteración se maximiza la covarianza de las proyecciones, es decir, se cumple que
$$(\operatorname{cov}(\mathbf{X}_l \mathbf{w}_l, \mathbf{Y}_l \mathbf{c}_l))^2 = \max_{\|\mathbf{r}\|=\|\mathbf{s}\|=1} (\operatorname{cov}(\mathbf{X}_l \mathbf{r}), \mathbf{Y}_l \mathbf{s})^2.$$
4. Los vectores extraídos son solución a los siguiente problemas de autovalores
$$\begin{aligned} \mathbf{X}_{l-1}^\top \mathbf{Y}_{l-1} \mathbf{Y}_{l-1}^\top \mathbf{X}_{l-1} \mathbf{w}_l = \lambda \mathbf{w}_l, \qquad & \mathbf{Y}_{l-1}^\top \mathbf{X}_{l-1} \mathbf{X}_{l-1}^\top \mathbf{Y}_{l-1} \mathbf{c}_l = \lambda \mathbf{c}_l, \\ \mathbf{X}_{l-1} \mathbf{X}_{l-1}^\top \mathbf{Y}_{l-1} \mathbf{Y}_{l-1}^\top \mathbf{t}_l = \lambda \mathbf{t}_l, \qquad & \mathbf{Y}_{l-1} \mathbf{Y}_{l-1}^\top \mathbf{X}_{l-1} \mathbf{X}_{l-1}^\top \mathbf{u}_l = \lambda \mathbf{u}_l. \end{aligned}$$
5. La deflación tiene las siguientes consecuencias:
$$\begin{aligned} \mathbf{X}_L &= \mathbf{X} - \mathbf{T}_L \mathbf{P}_L^\top, & \mathbf{Y}_L &= \mathbf{Y} - \mathbf{U}_L \mathbf{Q}_L^\top, \\ \mathbf{X}_L &= \prod_{l=k}^{L} \left( \mathbf{I} - \frac{\mathbf{t}_l \mathbf{t}_l^\top}{\mathbf{t}_l^\top \mathbf{t}_l} \right) \mathbf{X}_{k+1}, & \mathbf{Y}_L &= \prod_{l=k}^{L} \left( \mathbf{I} - \frac{\mathbf{u}_l \mathbf{u}_l^\top}{\mathbf{u}_l^\top \mathbf{u}_l} \right) \mathbf{Y}_{k+1}, \quad k < L. \end{aligned}$$
6. Los *loadings* se pueden expresar como
$$\mathbf{P}_L = \mathbf{X}^\top \mathbf{T}_L (\mathbf{D}_L^x)^{-1} \quad \mathbf{Q}_L = \mathbf{Y}^\top \mathbf{U}_L (\mathbf{D}_L^y)^{-1},$$
donde $\mathbf{D}_L^x = \operatorname{diag}(\|\mathbf{t}_1\|, \dots, \|\mathbf{t}_L\|)$ y $\mathbf{D}_L^y = \operatorname{diag}(\|\mathbf{u}_1\|, \dots, \|\mathbf{u}_L\|)$.

*Demostración.*

Demostramos cada apartado por separado. Sin embargo, no seguiremos el mismo orden en el que se han enunciado. Asimismo, debido a la simetría del algoritmo, solo demostramos las identidades correspondientes al bloque $X$.

**5.** Como resultado de la deflación, es inmediato ver que $\mathbf{X}_L = \mathbf{X} - \sum_{l=1}^{L} \mathbf{t}_l \mathbf{p}_l^\top$. Pero, por otro lado,
$$\begin{aligned} \left( \mathbf{T}_L \mathbf{P}_L^\top \right)_{i,j} &= \sum_{k=1}^{L} (\mathbf{T}_L)_{i,k} (\mathbf{P}_L)_{j,k} = \sum_{k=1}^{L} (\mathbf{t}_k)_i (\mathbf{p}_k)_j = \\ &= \sum_{k=1}^{L} (\mathbf{t}_k \mathbf{p}_k^\top)_{i,j} = \left( \sum_{k=1}^{L} \mathbf{t}_k \mathbf{p}_k^\top \right)_{i,j}. \end{aligned}$$
Así, se ha demostrado que la identidad buscada.

Por otro lado, la segunda expresión es el resultado de aplicar (2.11) iterativamente.

**2.** Se puede demostrar inductivamente. Los dos primeros vectores son ortogonales porque

$$\begin{aligned}\mathbf{t}_1^\top \mathbf{t}_2 = \mathbf{t}_1^\top \mathbf{X}_1 \mathbf{w}_1 = \mathbf{t}_1^\top \left(\mathbf{I} - \frac{\mathbf{t}_1 \mathbf{t}_1^\top}{\mathbf{t}_1^\top \mathbf{t}_1}\right) \mathbf{X}_0 \mathbf{w}_2 = \\ = (\mathbf{t}_1^\top - \mathbf{t}_1^\top)\mathbf{X}_0 \mathbf{w}_2 = 0.\end{aligned}$$

Por tanto, basta demostrar que $\{\mathbf{t}_1, \dots, \mathbf{t}_{l+1}\}$ son ortogonales dos a dos si $\{\mathbf{t}_1, \dots, \mathbf{t}_l\}$ lo son. Para ello, solo es necesario comprobar que $\mathbf{t}_{l+1}$ es ortogonal a todos los demás vectores $\mathbf{t}_j$, $j \leq l$. Si $j = l$, procedemos como en el paso anterior. Si $j < l$, aplicando las identidades obtenidas de la deflación:

$$\begin{aligned}\mathbf{t}_j^\top \mathbf{t}_{l+1} = \mathbf{t}_j^\top \mathbf{X}_l \mathbf{w}_{l+1} = \mathbf{t}_j^\top \prod_{i=l+1}^{j} \left(\mathbf{I} - \frac{\mathbf{t}_i \mathbf{t}_i^\top}{\mathbf{t}_i^\top \mathbf{t}_i}\right) \mathbf{X}_{j-1} = \\ = \mathbf{t}_j^\top \left(\mathbf{I} - \frac{\mathbf{t}_j \mathbf{t}_j^\top}{\mathbf{t}_j^\top \mathbf{t}_j}\right) \mathbf{X}_{j-1} = 0.\end{aligned}$$

La penúltima igualdad se cumple porque, debido a la ortogonalidad de las proyecciones anteriores, $\mathbf{t}_j^\top \left(\mathbf{I} - \frac{\mathbf{t}_i \mathbf{t}_i^\top}{\mathbf{t}_i^\top \mathbf{t}_i}\right) = \mathbf{t}_j$ si $j < i < l$.

**6.** Vamos a demostrar $\mathbf{P}_{L} = \mathbf{X}^\top \mathbf{T}_{L} (\mathbf{D}_L^x)^{-2}$ columna a columna. En particular, la $l$-ésima columna de esta matriz tiene la expresión

$$\begin{aligned}\mathbf{X}^\top \mathbf{t}_l (\|\mathbf{t}_l\|)^{-2} = (\mathbf{X}_{l-1} + \mathbf{T}_{l-1}\mathbf{P}_{l-1}^\top)^\top \mathbf{t}_l (\|\mathbf{t}_l\|)^{-2} = \\ = \mathbf{X}_{l-1}^\top \mathbf{t}_l (\|\mathbf{t}_l\|)^{-2} + \mathbf{P}_{l-1}\mathbf{T}_{l-1}^\top \mathbf{t}_l (\|\mathbf{t}_l\|)^{-2} = \\ = \mathbf{X}_{l-1}^\top \mathbf{t}_l / (\mathbf{t}_l^\top \mathbf{t}_l) + 0 = \mathbf{p}_l.\end{aligned}$$

En el penúltimo paso, el segundo sumando se anula por la ortogonalidad de las componentes.

**4.** Se trata de una consecuencia de la definición de la línea 4 junto con las expresiones utilizadas para calcular $\mathbf{t}_l$, $\mathbf{c}_l$ y $\mathbf{u}_l$. Por ejemplo, en el caso de $\mathbf{c}_l$, el problema de autovalores se obtiene del siguiente modo:

$$\begin{aligned}\mathbf{X}_{l-1}^\top \mathbf{Y}_{l-1} \mathbf{Y}_{l-1}^\top \mathbf{X}_{l-1} \mathbf{w}_l = \lambda \mathbf{w}_l \\ \mathbf{Y}_{l-1}^\top \mathbf{X}_{l-1} \mathbf{X}_{l-1}^\top \mathbf{Y}_{l-1} \mathbf{Y}_{l-1}^\top \mathbf{X}_{l-1} \mathbf{w}_l = \lambda \mathbf{X}_{l-1}^\top \mathbf{Y}_{l-1} \mathbf{w}_l \\ \mathbf{Y}_{l-1}^\top \mathbf{X}_{l-1} \mathbf{X}_{l-1}^\top \mathbf{Y}_{l-1} (\mathbf{Y}_{l-1}^\top \mathbf{X}_{l-1} \mathbf{w}_l) = \lambda (\mathbf{X}_{l-1}^\top \mathbf{Y}_{l-1} \mathbf{w}_l) \\ \mathbf{Y}_{l-1}^\top \mathbf{X}_{l-1} \mathbf{X}_{l-1}^\top \mathbf{Y}_{l-1} \mathbf{c}_l = \lambda \mathbf{c}_l.\end{aligned}$$

En los otros dos casos, el argumento es completamente análogo.

**1.** A partir de la definición de los problemas de autovalores, se puede deducir que se verifica $\mathbf{w}_l = \mathbf{X}_{l-1}\mathbf{u}_l / \|\mathbf{X}_{l-1}\mathbf{u}_l\|$. Para simplificar la notación, denotaremos

al escalar como $c_l$, es decir, $\mathbf{w}_l = c_l \mathbf{X}_{l-1} \mathbf{u}_l$ Utilizando esta identidad, se puede demostrar la ortogonalidad buscada:

$$\begin{aligned}\mathbf{w}_l^\top \mathbf{w}_i &= (c_l \mathbf{X}_{l-1}^\top \mathbf{u}_l)^\top \mathbf{w}_i = c_l \mathbf{u}_l^\top \mathbf{X}_{l-1} \mathbf{w}_i = \\ &= c_l \mathbf{u}_l^\top \left(\mathbf{I} - \frac{\mathbf{t}_{l-1}\mathbf{t}_{l-1}^\top}{\mathbf{t}_{l-1}^\top \mathbf{t}_{l-1}}\right) ... \left(\mathbf{I} - \frac{\mathbf{t}_i \mathbf{t}_i^\top}{\mathbf{t}_i^\top \mathbf{t}_i}\right) \mathbf{X}_{i-1} \mathbf{w}_i = \\ &= c_l \mathbf{u}_l^\top \left(\mathbf{I} - \frac{\mathbf{t}_{l-1}\mathbf{t}_{l-1}^\top}{\mathbf{t}_{l-1}^\top \mathbf{t}_{l-1}}\right) ... \underbrace{\left(\mathbf{I} - \frac{\mathbf{t}_i \mathbf{t}_i^\top}{\mathbf{t}_i^\top \mathbf{t}_i}\right) \mathbf{t}_i}_{0} = 0.\end{aligned}$$

**3.** Para simplificar la notación, en este apartado omitiremos los subíndices correspondientes a la iteración del algoritmo. Sean $\mathbf{r} \in \mathbb{R}^M$ y $\mathbf{s} \in \mathbb{R}^D$ dos vectores cualesquiera de norma unidad. Vamos a ver que la covarianza obtenida al proyectar sobre estos vectores nunca es mayor a la obtenida al proyectar sobre $\mathbf{w}$ y $\mathbf{c}$.

En primer lugar, aplicando la definición de la covarianza muestral y la descomposición SVD de $\mathbf{X}^\top \mathbf{Y}$ como $\mathbf{X}^\top \mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ , podemos ver que la cantidad a maximizar es

$$(\text{cov}(\mathbf{X}\mathbf{r}, \mathbf{Y}\mathbf{s})^2 = (\mathbf{r}^\top \mathbf{X}^\top \mathbf{Y}\mathbf{s})^2 = (\mathbf{r}^\top \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \mathbf{s})^2 .$$

Consideramos ahora los cambios de variable $\mathbf{d} = \mathbf{U}^\top \mathbf{r}$ y $\mathbf{e} = \mathbf{V}^\top \mathbf{s}$. Como las matrices $\mathbf{U}$ y $\mathbf{V}$ de la descomposición SVD son unitarias, se cumple que $\|\mathbf{d}\| = \|\mathbf{e}\| = 1$. A continuación, se aplica este cambio de variable. Asimismo, denotamos $\sigma_1$ como el mayor valor singular en $\mathbf{\Sigma}$. Aplicando la desigualdad de Cauchy-Schwartz, podemos acotar el cuadrado de la covarianza como

$$(\mathbf{r}^\top \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \mathbf{s})^2 = (\mathbf{d}^\top \mathbf{\Sigma}\mathbf{e})^2 \leq \|\mathbf{\Sigma}^\top \mathbf{d}\|\|\mathbf{e}\| = \|\mathbf{\Sigma}^\top \mathbf{d}\| \leq \sigma_1^2.$$

Ahora nos falta demostrar que $(\text{cov}(\mathbf{X}\mathbf{w}, \mathbf{Y}\mathbf{s}))^2 = \sigma_1^2$. Como sabemos que $\mathbf{w}$ es un el autovector dominante de $(\mathbf{X}^\top \mathbf{Y})(\mathbf{X}^\top \mathbf{Y})^\top$. Por tanto, $\mathbf{w} = \mathbf{u}_1$ (primera columna de $\mathbf{U}$). De nuevo, como $\mathbf{U}$ es una matriz unitaria, $\mathbf{U}^\top \mathbf{w} = \mathbf{e}_1$ (el primer vector de la base canónica). Del mismo modo, $\mathbf{V}^\top \mathbf{c} = \mathbf{e}_1$. Así, obtenemos

$$(\text{cov}(\mathbf{X}\mathbf{w}, \mathbf{Y}\mathbf{c}))^2 = (\mathbf{w}^\top \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \mathbf{c})^2 = (\mathbf{e}_1^\top \mathbf{\Sigma}\mathbf{e}_1)^2 = \sigma_1^2.$$

□

**Proposición 2.7** Dadas las definiciones del algoritmo 3, se definen las matrices:

$$\mathbf{R}_L^x = \mathbf{W}_L(\mathbf{P}_L^\top \mathbf{W}_L)^{-1} \quad \mathbf{R}_L^y = \mathbf{C}_L(\mathbf{Q}_L^\top \mathbf{C}_L)^{-1}.$$

Estas matrices cumplen que $\mathbf{T}_L = \mathbf{X}\mathbf{R}_L^x$ y $\mathbf{U}_L = \mathbf{Y}\mathbf{R}_L^y$. Por tanto, sus columnas contienen las direcciones de proyección para obtener las componentes obtenidas en NIPALS. Con frecuencia, se denomina a estas matrices *rotations*.

*Demostración.*

Se puede obtener a partir de las propiedades de la deflación presentadas en la proposición 2.6. Lo demostramos para $\mathbf{T}_L$.

$$\begin{aligned}\mathbf{X}\mathbf{R}_L^x &= \mathbf{X}\mathbf{W}_L(\mathbf{P}_L^\top \mathbf{W}_L)^{-1} = (\mathbf{T}_L\mathbf{P}_L^\top + \mathbf{X}_L)\mathbf{W}_L(\mathbf{P}_L^\top \mathbf{W}_L)^{-1} = \\ &= \mathbf{T}_L\mathbf{P}_L^\top \mathbf{W}_L(\mathbf{P}_L^\top \mathbf{W}_L)^{-1} + \mathbf{X}_L\mathbf{W}_L(\mathbf{P}_L^\top \mathbf{W}_L)^{-1} = \\ &= \mathbf{T}_L + \underbrace{\mathbf{X}_L\mathbf{W}_L}_{\mathbf{0}}(\mathbf{P}_L^\top \mathbf{W}_L)^{-1} = \mathbf{T}_L.\end{aligned}$$

Se cumple que $\mathbf{X}_L\mathbf{w}_l = \mathbf{0}$ si $l \leq L$ ya que

$$\begin{aligned}\mathbf{X}_L\mathbf{w}_l &= \left(\mathbf{I} - \frac{\mathbf{t}_L\mathbf{t}_L^\top}{\mathbf{t}_L^\top\mathbf{t}_L}\right) \ldots \left(\mathbf{I} - \frac{\mathbf{t}_l\mathbf{t}_l^\top}{\mathbf{t}_l^\top\mathbf{t}_l}\right)\mathbf{X}_l\mathbf{w}_l = \\ &= \left(\mathbf{I} - \frac{\mathbf{t}_L\mathbf{t}_L^\top}{\mathbf{t}_L^\top\mathbf{t}_L}\right) \ldots \left(\mathbf{I} - \frac{\mathbf{t}_l\mathbf{t}_l^\top}{\mathbf{t}_l^\top\mathbf{t}_l}\right)\mathbf{t}_l = \mathbf{0}.\end{aligned}$$

□

**Lema 3.2**. La deflación llevada a cabo en la línea 9 del algoritmo 4 puede eliminarse sin modificar el resultado del algoritmo.

*Demostración.*

Dejando de lado la misma deflación, vemos que $\mathbf{Y}_{l-1}$ aparece solo en la línea 4 y 7. En el caso de la linea 4, la equivalencia se debe a que la deflación en $\mathbf{X}_{l-1}$ e $\mathbf{Y}_{l-1}$ sea una proyección sobre el mismos espacio (ortogonal a $\mathbf{t}_l$). Desarrollando la matriz $\mathbf{X}_{l-1}^\top\mathbf{Y}_{l-1}$ se obtiene

$$\begin{aligned}\mathbf{X}_{l-1}^\top\mathbf{Y}_{l-1} &= \mathbf{X}^\top\left(\mathbf{I} - \frac{\mathbf{t}_1\mathbf{t}_1^\top}{\mathbf{t}_1^\top\mathbf{t}_1}\right) \ldots \left(\mathbf{I} - \frac{\mathbf{t}_{l-1}\mathbf{t}_{l-1}^\top}{\mathbf{t}_{l-1}^\top\mathbf{t}_{l-1}}\right)\left(\mathbf{I} - \frac{\mathbf{t}_{l-1}\mathbf{t}_{l-1}^\top}{\mathbf{t}_{l-1}^\top\mathbf{t}_{l-1}}\right) \ldots \left(\mathbf{I} - \frac{\mathbf{t}_1\mathbf{t}_1^\top}{\mathbf{t}_1^\top\mathbf{t}_1}\right)\mathbf{Y} = \\ &= \mathbf{X}^\top\left(\mathbf{I} - \frac{\mathbf{t}_{l-1}\mathbf{t}_{l-1}^\top}{\mathbf{t}_{l-1}^\top\mathbf{t}_{l-1}}\right) \ldots \left(\mathbf{I} - \frac{\mathbf{t}_1\mathbf{t}_1^\top}{\mathbf{t}_1^\top\mathbf{t}_1}\right)\mathbf{Y} = \mathbf{X}_{l-1}^\top\mathbf{Y}.\end{aligned}$$

Como los proyectores sobre los espacios ortogonales a $\mathbf{t}_j$ son idempotentes, la aplicación de dos proyectores se puede sustituir por la de uno solo. Por otro lado, en la línea 7, la equivalencia se debe a la ortogonalidad de los componentes.

$$\mathbf{Y}_{l-1}^\top\mathbf{t}_l = \mathbf{Y}^\top\left(\mathbf{I} - \frac{\mathbf{t}_1\mathbf{t}_1^\top}{\mathbf{t}_1^\top\mathbf{t}_1}\right) \ldots \left(\mathbf{I} - \frac{\mathbf{t}_{l-1}\mathbf{t}_{l-1}^\top}{\mathbf{t}_{l-1}^\top\mathbf{t}_{l-1}}\right)\mathbf{t}_l = \mathbf{Y}^\top\mathbf{t}_l.$$

□

**Proposición 3.2**. En la regresión PLS se busca una matriz $\hat{\mathbf{B}}$ tal que $\mathbf{Y} = \mathbf{X}\hat{\mathbf{B}} + \mathbf{E}$ donde $\mathbf{E} = \mathbf{Y} - \hat{\mathbf{Y}}$ es una matriz de residuos. Si se consideran $L$ componentes, $\hat{\mathbf{B}}_L$ es el resultado de las siguientes expresiones equivalentes:

$$\hat{\mathbf{B}}_L = \mathbf{W}_L(\mathbf{P}_L^\top \mathbf{W}_L)^{-1}\tilde{\mathbf{Q}}_L^\top. \tag{A.3}$$

$$\hat{\mathbf{B}}_L = \mathbf{W}_L(\mathbf{W}_L^\top \mathbf{X}^\top \mathbf{X}\mathbf{W}_L)^{-1}\mathbf{W}_L^\top \mathbf{X}^\top \mathbf{Y}. \tag{A.4}$$

Además, si definimos $\tilde{\mathbf{T}}_L$ y $\tilde{\mathbf{P}}_L$ tales que $\tilde{\mathbf{T}}_L\tilde{\mathbf{P}}_L^\top = \mathbf{T}_L\mathbf{P}_L^\top$ y las columnas de $\tilde{\mathbf{T}}$ tienen norma unidad, la siguiente expresión también es equivalente:

$$\hat{\mathbf{B}}_L = \mathbf{W}_L(\tilde{\mathbf{P}}_L^\top \mathbf{W}_L)^{-1}\tilde{\mathbf{T}}_L^\top \mathbf{Y}. \tag{A.5}$$

*Demostración.*

Para ver la equivalencia entre (A.3) y (A.4), necesitamos una expresión para $\mathbf{Q}_L$. Siguiendo un razonamiento análogo al del sexto apartado de la proposición 2.6, se obtiene que $\tilde{\mathbf{Q}}_L = \mathbf{Y}^\top \mathbf{T}_L(\mathbf{D}_L^x)^{-2}$. Utilizando esto:

$$\begin{aligned}
\mathbf{W}_L(\mathbf{P}_L^\top \mathbf{W}_L)^{-1}\tilde{\mathbf{Q}}_L^\top &= \mathbf{W}_L(\mathbf{P}_L^\top \mathbf{W}_L)^{-1}(\mathbf{D}_L^x)^{-2}\mathbf{T}_L^\top \mathbf{Y} = \\
&= \mathbf{W}_L(\mathbf{P}_L^\top \mathbf{W}_L)^{-1}(\mathbf{D}_L^x)^{-2}(\mathbf{X}\mathbf{W}_L(\mathbf{P}_L^\top \mathbf{W}_L)^{-1})^\top \mathbf{Y} = \\
&= \mathbf{W}_L(\mathbf{W}_L^\top \mathbf{P}_L(\mathbf{D}_L^x)^2\mathbf{P}_L^\top \mathbf{W}_L)^{-1}\mathbf{W}_L^\top \mathbf{X}^\top \mathbf{Y} = \\
&= \mathbf{W}_L(\mathbf{W}_L^\top(\mathbf{X}^\top \mathbf{T}_L(\mathbf{D}_L^x)^{-2})(\mathbf{D}_L^x)^2\mathbf{P}_L^\top \mathbf{W}_L)^{-1}\mathbf{W}_L^\top \mathbf{X}^\top \mathbf{Y} = \\
&= \mathbf{W}_L(\mathbf{W}_L^\top \mathbf{X}^\top \mathbf{T}_L\mathbf{P}_L^\top \mathbf{W}_L)^{-1}\mathbf{W}_L^\top \mathbf{X}^\top \mathbf{Y} = \\
&= \mathbf{W}_L(\mathbf{W}_L^\top \mathbf{X}^\top \mathbf{X}_L^\top \mathbf{W}_L)^{-1}\mathbf{W}_L^\top \mathbf{X}^\top \mathbf{Y}.
\end{aligned}$$

Queda demostrar la equivalencia con la expresión (A.5). Para ello, hemos de definir las matrices $\tilde{\mathbf{T}}_L$ y $\tilde{\mathbf{P}}_L$ tales que la columnas de $\tilde{\mathbf{T}}_L$ tengan norma unidad.

Para normalizar las columnas, podemos definir esta matriz como $\tilde{\mathbf{T}}_L = \mathbf{T}_L(\mathbf{D}_L^x)^{-1}$. Así, para que se cumpla $\tilde{\mathbf{T}}_L\tilde{\mathbf{P}}_L^\top = \mathbf{T}_L\mathbf{P}_L^\top$, basta definir $\tilde{\mathbf{P}}_L = \mathbf{P}_L\mathbf{D}_L^x$. Con estas definiciones, podemos ver que las expresiones son equivalentes:

$$\begin{aligned}
\mathbf{W}_L(\tilde{\mathbf{P}}_L^\top \mathbf{W}_L)^{-1}\tilde{\mathbf{T}}_L^\top \mathbf{Y} &= \mathbf{W}_L(\mathbf{D}_L^x\mathbf{P}_L^\top \mathbf{W}_L)^{-1}\mathbf{D}_L^x\mathbf{T}_L^\top \mathbf{Y} = \\
&= \mathbf{W}_L(\mathbf{P}_L^\top \mathbf{W}_L)^{-1}(\mathbf{D}_L^x)^{-2}\mathbf{T}_L^\top \mathbf{Y} = \\
&= \mathbf{W}_L(\mathbf{P}_L^\top \mathbf{W}_L)^{-1}\tilde{\mathbf{Q}}_L^\top.
\end{aligned}$$

□

**Proposición 3.3.** Dada una matriz $M \times M$ de la forma $\mathbf{A} = \mathbf{a}\mathbf{a}^\top$, donde $\mathbf{a} \in \mathbb{R}^M$, $\mathbf{a}/\|\mathbf{a}\|$ es un autovalor dominante unitario de $\mathbf{A}$.

*Demostración.*

Se trata de una consecuencia directa de la desigualdad de Cauchy-Schwarz. Sea $\mathbf{v} \in \mathbb{R}^M$ tal que $\|\mathbf{v}\| = 1$. Entonces

$$\|\mathbf{a}\mathbf{a}^\top \mathbf{v}\|^2 = |\mathbf{a}^\top \mathbf{v}|^2 \|\mathbf{a}\|^2 \leq \|\mathbf{a}\|^2 \|\mathbf{v}\|^2 \|\mathbf{a}\|^2 = \|\mathbf{a}\|^4. \tag{A.6}$$

Por otro lado, considerando $\mathbf{v} = \frac{\mathbf{a}}{\|\mathbf{a}\|}$, se obtiene $\mathbf{a}\mathbf{a}^\top \frac{\mathbf{a}}{\|\mathbf{a}\|} = \mathbf{a}\|\mathbf{a}\|$. Por tanto $\mathbf{v}$ es un autovector con autovalor asociado $\|\mathbf{a}\|$. Como este alcanza la cota en (A.6), es el autovector buscado. □

**Proposición 3.4.** Sea $\mathbf{A}$ una matriz simétrica $M \times M$ tal que $\det(\mathbf{A}) \neq 0$. Entonces, existe un polinomio $P$ de grado $m-1$ tal que $\mathbf{A}P(\mathbf{A}) = \mathbf{I}$, donde $m$ es el número de autovalores distintos de $\mathbf{A}$ En particular, $\mathbf{A}^{-1} = a_0 + a_1\mathbf{A} + \cdots + a_{m-1}\mathbf{A}^{m-1}$ para ciertos coeficientes $a_j \in \mathbb{R}$.

*Demostración.*

Sea $Q^\star(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I})$ el polinomio característico de $\mathbf{A}$. En relación a este polinomio, podemos considerar $Q(x)$, el polinomio mínimo de $Q^\star$. Es decir, $Q(x)$ es el polinomio de menor grado con las mismas raíces que $Q^\star(x)$. La existencia de este polinomio está garantizada porque $\mathbf{A}$ es simétrica y todos sus autovalores son reales.

Observando la definición de $Q^\star$, es claro que $Q(\mathbf{A}) = c_0\mathbf{I} + c_1\mathbf{A} + \ldots c_m\mathbf{A}^m = 0$ ya que $Q$ cuenta con las mismas raíces. Por otro lado, como $\det(\mathbf{A}) \neq 0$, $Q(\mathbf{0}) \neq 0$ y podemos afirmar que $c_0 \neq 0$. Por tanto tenemos que

$$\begin{aligned} Q(\mathbf{A}) = c_0\mathbf{I} + c_1\mathbf{A} + \ldots c_m + \mathbf{A}^m &= 0 \\ \implies -c_0\mathbf{I} &= \mathbf{A}(c_1 + c_2\mathbf{A} + \ldots c_m + \mathbf{A}^{m-1}) \\ \implies \mathbf{I} &= \mathbf{A}\frac{-c_1 - c_2\mathbf{A} - \cdots - c_m\mathbf{A}^{m-1}}{c_0} \\ \implies \mathbf{A}^{-1} &= P(\mathbf{A}) = a_0 + a_1\mathbf{A} + \cdots + a_{m-1}\mathbf{A}^{m-1}, \end{aligned}$$

donde $a_i = -c_{i+1}/c_0$ para $i = 0, \ldots, m-1$. □

**Proposición 3.5.** El estimador OLS restringido al espacio $\mathcal{K}_L(\mathbf{X}^\top\mathbf{X}, \mathbf{X}^\top\mathbf{y})$ tiene la expresión

$$\hat{\boldsymbol{\beta}} = \mathbf{R}_L(\mathbf{R}_L^\top\mathbf{X}^\top\mathbf{R}_L)^{-1}\mathbf{R}_L\mathbf{X}^\top\mathbf{y},$$

donde $\mathbf{R}_L$ es una matriz cuyas columnas forman una base ortonormal del espacio de Krylov $\mathcal{K}_L(\mathbf{X}^\top\mathbf{X}, \mathbf{X}^\top\mathbf{y})$.

*Demostración.*

La proposición 3.1 proporciona una expresión para el estimador OLS sin restringir el espacio. Sin embargo, si tenemos una base del espacio al que se está restringiendo el estimador, estos problemas son prácticamente equivalentes.

Sea $\mathbf{R}_L$ una matriz cuyas columnas formen base de $\mathcal{K}_L(\mathbf{X}^\top\mathbf{X}, \mathbf{X}^\top\mathbf{y})$. Usando esta base, $\hat{\boldsymbol{\beta}}$ se puede expresar como $\hat{\boldsymbol{\beta}} = \mathbf{R}_L\hat{\boldsymbol{\alpha}}$ para algún $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^L$. A continuación vemos que

$$\min_{\boldsymbol{\beta}\in\mathcal{K}_L(\mathbf{A},\mathbf{b})} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\| = \min_{\boldsymbol{\alpha}\in\mathbb{R}^L} \|\mathbf{y} - \mathbf{X}\mathbf{R}_L\boldsymbol{\alpha}\|.$$

Aplicando la proposición 3.1, la segunda expresión alcanza su valor mínimo cuando

$$\hat{\boldsymbol{\alpha}} = (\mathbf{R}_L^\top\mathbf{X}^\top\mathbf{X}\mathbf{R}_L)^{-1}\mathbf{R}_L^\top\mathbf{X}^\top\mathbf{y}.$$

Por tanto, hemos obtenido la expresión buscada para $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = \mathbf{R}_L(\mathbf{R}_L^\top\mathbf{X}^\top\mathbf{X}\mathbf{R}_L)^{-1}\mathbf{R}_L^\top\mathbf{X}^\top\mathbf{y}.$$

□

**Proposición 3.6.** Las columnas de la matriz $\mathbf{W}_L$ generada por NIPALS-PLS1 forman una base ortonormal del espacio de Krylov $\mathcal{K}_L(\mathbf{X}^\top\mathbf{X}, \mathbf{X}^\top\mathbf{y})$.

*Demostración.*

Ya hemos demostrado que los vectores $\mathbf{w}_l$ son ortogonales dos a dos (proposición 2.6). Como en su definición ya se especifica que tienen que tener norma unidad, solo falta por demostrar que están contenidos en el espacio de Krylov. Esta demostración se lleva a cabo por inducción.

Observando la primera iteración del algoritmo, vemos que

$$\mathbf{w}_1 \in \operatorname{span}(\mathbf{X}^\top\mathbf{y}), \qquad \mathbf{p}_1 \in \operatorname{span}((\mathbf{X}^\top\mathbf{X})\mathbf{X}^\top\mathbf{y}).$$

A continuación, asumimos que para $l$ se cumple

$$\mathbf{w}_l \in \operatorname{span}\Big(\mathbf{X}^\top\mathbf{y}, \dots, (\mathbf{X}^\top\mathbf{X})^{l-1}\mathbf{X}^\top\mathbf{y}\Big),$$
$$\mathbf{p}_l \in \operatorname{span}\Big((\mathbf{X}^\top\mathbf{X})\mathbf{X}^\top\mathbf{y}, (\mathbf{X}^\top\mathbf{X})^{l}\mathbf{X}^\top\mathbf{y}\Big).$$

Por tanto, para ciertos escalares $c_1, c_2, c_3$ se verifica

$$\mathbf{w}_{l+1} = c_1 \mathbf{X}_l^\top \mathbf{y} = c_1(\mathbf{X}_{l-1}^\top \mathbf{y} - (\mathbf{t}_l^\top \mathbf{y})\mathbf{p}_l) = c_2 \mathbf{w}_l + c_3 \mathbf{p}_l,$$

y vemos que $\mathbf{w}_{l+1} \in \text{span}\Big((\mathbf{X}^\top\mathbf{X})\mathbf{X}^\top\mathbf{y}, (\mathbf{X}^\top\mathbf{X})^l\mathbf{X}^\top\mathbf{y}\Big) \subset \mathcal{K}_{l+1}(\mathbf{X}^\top\mathbf{X}, \mathbf{X}^\top\mathbf{y})$. □

**Proposición 4.2.** Consideramos $X(t)$ un proceso estocástico en $L^2$ e $Y$ una variable aleatoria escalar. Además, sin pérdida de generalidad, suponemos que $\mathbb{E}(X(t)) = 0 \ \forall t \in [0,1]$ y $\mathbb{E}(Y) = 0$. En estas circunstancias, la función $f \in L^2([0,1])$ que maximiza

$$\left(\text{cov}\left(\int_0^1 X(t)f(t)\,dt,\ Y\right)\right)^2$$

es una autofunción del operador $U_X$ asociada a su autovalor dominante.

*Demostración.*

Desarrollamos la expresión de la covarianza aprovechando que tanto la proyección del proceso estocástico como $Y$ están centradas.

$$\begin{aligned}
\left(\text{cov}\left(\int_0^1 X(t)f(t)\,dt,\ Y\right)\right)^2 &= \left(\mathbb{E}\left(Y\int_0^1 X(t)f(t)\,dt\right)\right)^2 = \\
&\overset{\text{Fub}}{=} \left(\int_0^1 (YX(t)f(t))\,dt\right)^2 = \\
&= \left(\int_0^1 f(t)\mathbb{E}\left(X(t)Y\right)dt\right)^2 = \\
&= \int_0^1 f(t)\mathbb{E}(X(t)Y)\int_0^1 f(s)\mathbb{E}(X(s)Y)ds\,dt = \\
&= \langle f, U_X(f)\rangle_{L^2}.
\end{aligned}$$

Como el operador $U_X$ es autoadjunto y compacto (Preda y Saporta (2005)), los resultados de maximización recogidos en Ramsay y Silverman (2013, p. 397) nos aseguran que la máxima covarianza al cuadrado se alcanza cuando $f$ es la autofunción correspondiente al autovalor dominante de $U_X$. □

# APÉNDICE B
# Algoritmos

## B.1. Formulación habitual de NIPALS-Modo A

Esta es la forma más habitual en la que se encuentra el algoritmo NIPALS usando para resolver problemas de reducción de la dimensionalidad utilizando el criterio PLS. Como se puede apreciar, la mayor diferencia es la utilización del método de la potencia como mecanismo para resolver el problema de autovalores.

**Algoritmo 6** NIPALS modo A para el cálculo de PLS

$l \leftarrow 1$
$\mathbf{X}_0 \leftarrow \mathbf{X} \qquad \mathbf{Y}_0 \leftarrow \mathbf{Y}$
**while** $l < L$ **do**
  $\mathbf{u} \leftarrow$ random vector
  **repeat**
    $\mathbf{w} \leftarrow \mathbf{X}_{l-1}^\top \mathbf{u} \;/\; \|\mathbf{X}^\top \mathbf{u}\|$
    $\mathbf{t} \leftarrow \mathbf{X}_{l-1}\mathbf{w}$
    $\mathbf{c} \leftarrow \mathbf{Y}_{l-1}^\top \mathbf{t} \;/\; \|\mathbf{Y}^\top \mathbf{t}\|$
    $\mathbf{u} \leftarrow \mathbf{Y}_{l-1}\mathbf{c}$
  **until** convergence
  $\mathbf{p}_l \leftarrow \mathbf{X}_{l-1}^\top \mathbf{t} \;/\; \|\mathbf{t}\|^2$
  $\mathbf{q}_l \leftarrow \mathbf{Y}_{l-1}^\top \mathbf{u} \;/\; \|\mathbf{u}\|^2$
  $\mathbf{w}_l \leftarrow \mathbf{w} \qquad \mathbf{t}_l \leftarrow \mathbf{t} \qquad \mathbf{c}_l \leftarrow \mathbf{c} \qquad \mathbf{u}_l \leftarrow \mathbf{u}$
  $\mathbf{X}_l \leftarrow \mathbf{X}_{l-1} - \mathbf{t}_l \mathbf{p}_l^\top$
  $\mathbf{Y}_l \leftarrow \mathbf{Y}_{l-1} - \mathbf{u}_l \mathbf{q}_l^\top$
  $l \leftarrow l + 1$
**end while**

Sin embargo, el método de la potencia se aplica de forma ligeramente distinta a la versión estándar. Aprovechando las relaciones entre los *weights* y las *scores*, se pueden calcular los cuatro vectores a la vez en el método de la potencia. Es decir, se está utilizando el método de la potencia para resolver varios problemas de autovectores a la vez, uno para cada vector considerado. Esta caracterización de los *weights* y las *scores* como autovectores ya ha sido introducida en la Proposición 2.6.

Por otro lado, el uso del método de la potencia para calcular los autovectores, es un detalle de implementación. Por ejemplo, algunas implementaciones utilizan

la descomposición SVD de la matriz de covarianzas cruzadas para encontrar dichos autovectores. No obstante, el método de la potencia es una buena opción ya que su convergencia suele ser muy rápida. Además, al calcular la descomposición SVD se obtiene mucha información que no se aprovecha ya que solo estamos interesados en el autovector asociado al autovalor dominante.

## B.2. NIPALS para el cálculo de CCA

**Algoritmo 7** NIPALS para el cálculo de PCA

$l \leftarrow 1$
$\mathbf{X}_0 \leftarrow \mathbf{X} \qquad \mathbf{Y}_0 \leftarrow \mathbf{Y}$
**while** $l < L$ **do**
  $\mathbf{u} \leftarrow$ random vector
5: **repeat**
    $\mathbf{w} \leftarrow (\mathbf{X}_{l-1}^{+})^{\top}\mathbf{u} \;/\; \|(\mathbf{X}_{l-1}^{+})^{\top}\mathbf{u}\|$
    $\mathbf{t} \leftarrow \mathbf{X}_{l-1}\mathbf{w} \;/\; \|\mathbf{X}_{l-1}\mathbf{w}\|$
    $\mathbf{c} \leftarrow (\mathbf{Y}_{l-1}^{+})^{\top}\mathbf{t} \;/\; \|(\mathbf{Y}_{l-1}^{+})^{\top}\mathbf{t}\|$
    $\mathbf{u} \leftarrow \mathbf{Y}_{l-1}\mathbf{c} \;/\; \|\mathbf{Y}_{l-1}\mathbf{c}\|$
10: **until** convergence
  $\mathbf{p}_l \leftarrow \mathbf{X}_{l-1}^{\top}\mathbf{t}$
  $\mathbf{q}_l \leftarrow \mathbf{Y}_{l-1}^{\top}\mathbf{u}$
  $\mathbf{w}_l \leftarrow \mathbf{w} \qquad \mathbf{t}_l \leftarrow \mathbf{t} \qquad \mathbf{c}_l \leftarrow \mathbf{c} \qquad \mathbf{u}_l \leftarrow \mathbf{u}$
  $\mathbf{X}_l \leftarrow \mathbf{X}_{l-1} - \mathbf{t}_l\mathbf{p}_l^{\top}$
15: $\mathbf{Y}_l \leftarrow \mathbf{Y}_{l-1} - \mathbf{u}_l\mathbf{q}_l^{\top}$
  $l \leftarrow l + 1$
**end while**

nota: $\mathbf{A}^{+}$ denota la pseudo-inversa de $\mathbf{A}$

La diferencia de esta versión de NIPALS con la correspondiente a PLS es la resolución de un problema distinto de autovalores. Este problema de autovalores se resuelve mediante el método de la potencia entre las líneas 5 y 10. Utilizando la fórmula para la pseudo-inversa en función del producto de la matriz por la traspuesta, se puede demostrar que se están resolviendo los problemas de autovalores introducidos en la proposición 2.2.

## B.3. Formulación habitual de PLS2

La gran diferencia entre este algoritmo y el algoritmo 4 es el uso del método de la potencia para el cálculo del autovector dominante de la matriz $\mathbf{X}_{l-1}^\top \mathbf{Y}_{l-1} \mathbf{Y}_{l-1}^\top \mathbf{X}_{l-1}$. Además, esta implementación aprovecha la caracterización de los *weights* y *scores* como soluciones de los problemas de autovalores enunciados en la proposición 2.6.

Sin embargo, observando las matrices involucradas en el cálculo de la matriz de regresión (proposición 3.2), vemos que estamos calculando más elementos de los necesarios para calcular la matriz de regresión. Es por ello que la versión presentada en 4 prescinde de estos vectores.

**Algoritmo 8** NIPALS-PLS2 habitual

$l \leftarrow 1$
$\mathbf{X}_0 \leftarrow \mathbf{X} \qquad \mathbf{Y}_0 \leftarrow \mathbf{Y}$
**while** $l < L$ **do**
  $\mathbf{u} \leftarrow$ random vector
  **repeat**
    $\mathbf{w} \leftarrow \mathbf{X}_{l-1}^\top \mathbf{u} \;/\; \|\mathbf{X}^\top \mathbf{u}\|$
    $\mathbf{t} \leftarrow \mathbf{X}_{l-1} \mathbf{w}$
    $\mathbf{c} \leftarrow \mathbf{Y}_{l-1}^\top \mathbf{t}$
    $\mathbf{u} \leftarrow \mathbf{Y}_{l-1} \mathbf{c} / \|\mathbf{c}\|^2$
  **until** convergence
  $\mathbf{p}_l \leftarrow \mathbf{X}_{l-1}^\top \mathbf{t} / \|\mathbf{t}\|^2$
  $\mathbf{q}_l \leftarrow \mathbf{Y}_{l-1}^\top \mathbf{t} / \|\mathbf{t}\|^2$
  $\mathbf{w}_l \leftarrow \mathbf{w} \qquad \mathbf{t}_l \leftarrow \mathbf{t} \qquad \mathbf{c}_l \leftarrow \mathbf{c} \qquad \mathbf{u}_l \leftarrow \mathbf{u}$
  $\mathbf{X}_l \leftarrow \mathbf{X}_{l-1} - \mathbf{t}_l \mathbf{p}_l^\top$
  $\mathbf{Y}_l \leftarrow \mathbf{Y}_{l-1} - \mathbf{t}_l \mathbf{q}_l^\top$
  $l \leftarrow l + 1$
**end while**