UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Estadística e Investigación Operativa Aplicadas y Calidad

Métodos multivariantes dispersos aplicados al análisis de respuesta inmunitaria

Trabajo Fin de Máster

Máster Universitario en Ingeniería de Análisis de Datos, Mejora de Procesos y Toma de Decisiones

AUTOR/A: Martin Contreras, Nerea

Tutor/a: Conchado Peiró, Andrea

Director/a Experimental: FERNANDEZ-MURGA CHAVANNE, MARIA LEONOR

CURSO ACADÉMICO: 2022/2023

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MÉTODOS MULTIVARIANTES DISPERSOS APLICADOS AL ANÁLISIS DE RESPUESTA INMUNITARIA

Trabajo Fin de Máster

Máster Universitario en Ingeniería de Análisis de Datos, Mejora de Procesos y Toma de Decisiones

Dpto. de Estadística e Investigación Operativa Aplicadas y Calidad

---

CURSO ACADÉMICO 2022-2023

AUTORA: Martín Contreras, Nerea

DIRECTORA: Conchado Peiró, Andrea

DIRECTORA EXPERIMENTAL: Fernández-Murga Chavanne, Mª Leono

RESUMEN

La respuesta inmunitaria es un proceso complejo regulado en el organismo humano, mediante el cual el sistema inmunológico reconoce y combate patógenos. En el contexto de la pandemia de la COVID19, se ha llevado a cabo una investigación intensiva para comprender y evaluar la respuesta inmunitaria generada por las vacunas contra el SARS-CoV-2.

A medida que se ha comprobado la eficacia de las vacunas tipo RNAm para prevenir y reducir los síntomas de la COVID19, surge la incertidumbre de cómo evaluar la respuesta inmune en grupos de pacientes oncológicos en tratamiento activo, quienes no participaron en los ensayos clínicos de desarrollo de las vacunas Además, se necesitan respuestas a preguntas relacionadas con la duración de la inmunidad, los efectos de diferentes tratamientos en la respuesta inmunitaria y el impacto del tipo de cáncer.

Explorar la relación entre células del sistema inmune y la producción de anticuerpos ayudará a comprender cómo se modula la respuesta inmunitaria inducida por la vacunación por RNAm contra la COVID19 con cáncer. Se espera identificar posibles correlaciones entre los tipos celulares del sistema inmune y la magnitud de la respuesta inmunitaria humoral generada por la vacuna en estos pacientes.

Estos estudios presentan desafíos metodológicos significativos, ya que involucran un número reducido de pacientes que contienen una gran cantidad de variables relacionadas con la respuesta inmunitaria. El objetivo de este trabajo es abordar estos desafíos y llevar a cabo un análisis exhaustivo de la respuesta inmunitaria y la producción de anticuerpos en pacientes con cáncer en tratamiento activo tras recibir la segunda dosis de la vacuna anti-COVID19.

Con el objetivo de lograr esto, se aplicarán diferentes métodos multivariantes, entre los que se incluyen enfoques dispersos, debido a la naturaleza de los datos. Estos métodos, mediante la implementación de técnicas de penalización o regularización, permitirán realizar una selección de variables, lo cual es especialmente relevante para el estudio. De esta manera, se busca demostrar la utilidad de los métodos multivariantes dispersos y evidenciar cómo estos ofrecen una solución efectiva al problema clínico planteado.

Palabras clave: Penalización; Análisis de componentes principales dispersos; PLS disperso; Respuesta inmunitaria.

RESUM

La resposta immunitària és un procés complex regulat en l'organisme humà, mitjançant el qual el sistema immunològic reconeix i combat patògens. En el context de la pandèmia de la COVID19, s'ha dut a terme una investigació intensiva per a comprendre i avaluar la resposta immunitària generada per les vacunes contra el SARS-CoV-2.

A mesura que s'ha comprovat l'eficàcia de les vacunes tipus RNAm per a previndre i reduir els símptomes de la COVID19, sorgeix la incertesa de com avaluar la resposta immune en grups de pacients oncològics en tractament actiu, els qui no van participar en els assajos clínics de desenvolupament de les vacunes A més, es necessiten respostes a preguntes relacionades amb la duració de la immunitat, els efectes de diferents tractaments en la resposta immunitària i l'impacte del tipus de càncer.

Explorar la relació entre cèl·lules del sistema immune i la producció d'anticossos ajudarà a comprendre com es modula a resposta immunitària induïda per la vacunació per RNAm contra la COVID19 en pacients amb càncer. S'espera identificar possibles correlacions entre els tipus cel·lulars del sistema immune i la magnitud de la resposta immunitària humoral generada per la vacuna en aquests pacients.

Aquests estudis presenten desafiaments metodològics significatius, ja que involucren un nombre reduït de pacients que contenen una gran quantitat de variables relacionades amb la resposta immunitària. L'objectiu d'aquest treball és abordar aquests desafiaments i dur a terme una anàlisi exhaustiva de la resposta immunitària i la producció d'anticossos en pacients amb càncer en tractament actiu després de rebre la segona dosi de la vacuna anti-COVID19.

Amb l'objectiu d'aconseguir això, s'aplicaran diferents mètodes multivariants, entre els quals s'inclouen enfocaments dispersos, a causa de la naturalesa de les dades. Aquests mètodes, mitjançant la implementació de tècniques de penalització o regularització, permetran realitzar una selecció de variables, la qual és especialment rellevant per a l'estudi. D'aquesta manera, es busca demostrar la utilitat dels mètodes multivariants dispersos i evidenciar com aquests ofereixen una solució efectiva al problema clínic plantejat.

Paraules clau: Penalització; Anàlisi de components principals dispersos; PLS dispers; Resposta immunitària.

ABSTRACT

The immune response is a complex and regulated process in the human organism, whereby the immune system recognizes and fights pathogens. In the context of the COVID19 pandemic, intensive researches have been conducted to understand and evaluate the immune response generated by SARS-CoV-2 vaccines.

While the effectiveness of RNAm vaccines in preventing and minimizing COVID19 symptoms has been proven, there is growing uncertainty about how to evaluate the immune response in actively treated oncology patient groups who did not participate in the vaccine development clinical trials. Additionally, there is a need for answer to questions regarding the duration of immunity, the effects of different treatments on the immune response, and the impact of cancer type.

Exploring the relationship between immune system cells and antibody production will help to understand how the immune response induced by mRNA vaccination against COVID19 is modulated. It is expected to identify possible correlations between the cell types of the immune system and the magnitude of the immune response generated by the vaccine in these patients.

These studies present significant methodological challenges, as they involve a small number of patients containing many variables related to the immune response. The aim of this work is to address these challenges and conduct a comprehensive analysis of the immune response and antibody production in cancer patients on active treatment after receiving the second dose of Anti-COVID19 vaccine.

In order to achieve this, different multivariate methods will be applied, including sparse approaches, due to the nature of the data. These methods, through the implementation of penalization or regularization techniques, will allow the selection of variables, which is particularly relevant to the study. In this way, the aim is to demonstrate the utility of sparse multivariate methods and to highlight how they provide an effective solution to the clinical problem at hand.

Keyword: Penalty; Sparse principal component analysis; Sparse PLS; Immune response.

AGRADECIMIENTOS

Quisiera expresar mi profundo agradecimiento a todas las personas que han hecho posible la realización de este Trabajo. Su apoyo y contribuciones han sido fundamentales en cada etapa de este proceso.

En primer lugar, agradecer a mi directora de TFM, Andrea, por su guía, orientación y consejos a lo largo de todo el proyecto. También quiero agradecer a Leonor por su colaboración y por brindarme acceso a los recursos necesarios para llevar a cabo mi estudio. No puedo dejar de mencionar a mi familia y amigos, quienes han ofrecido su apoyo incondicional durante toda esta etapa.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

1.1. COVID19 y Vacunación

La pandemia de la COVID19 ha provocado una enorme pérdida de vidas humanas en todo el mundo y ha planteado un desafío sin precedentes para la salud pública. Según la Organización Mundial de la Salud (OMS) a fecha de Julio 2023, se han registrados más de 767,7 millones de casos de coronavirus (SARS-CoV-2), con alrededor de 7 millones de muertes relacionadas con la enfermedad (WHO, s.f.).

Como medio para frenar la propagación del virus, se han desarrollado vacunas en las que varias empresas farmacéuticas estuvieron involucradas. Entre ellas se encuentran las vacunas desarrolladas por Pfizer, Moderna, AstraZeneca, entre otras. La OMS reconoce que hasta la fecha actual se han administrado más de 13.461 millones de vacunas en todo el mundo.

Sin embargo, en los estudios clínicos realizados durante el desarrollo de las vacunas, algunos grupos de personas vulnerables de la población, como los pacientes oncológicos en tratamiento activo, entre otros, fueron excluidos. Esto ha generado incertidumbre sobre cómo evaluar la respuesta inmunitaria de estos pacientes, lo que ha llevado a la necesidad de investigar los estudios sobre la inmunidad celular y humoral en esta población.

1.2. Motivación

La preocupación sobre la vacunación y la disminución de anticuerpos a lo largo del tiempo es un tema de gran importancia entre muchas personas, especialmente para aquellos que pertenecen a grupos de mayor vulnerabilidad. Entre estos grupos, se encuentran los pacientes oncológicos en tratamiento activo, quienes enfrentan preocupaciones adicionales debido a su estado de salud comprometido.

Actualmente, existe una escasez de estudios que investiguen la relación entre la producción de anticuerpos contra la COVID19 y la respuesta inmunitaria en general, así como una falta de investigaciones que incluyan datos de pacientes oncológicos en estos estudios.

Por lo tanto, es de vital importancia llevar a cabo una investigación que analice la respuesta de los pacientes oncológicos en tratamiento activo frente a la vacuna de ARNm contra la COVID19 consiguiendo colaborar con el objetivo ODS de Salud y Bienestar. Asimismo, resulta interesante evaluar si la terapia recibida tiene algún efecto en su respuesta inmunológica, dado que los tratamientos debilitan el sistema inmunológico, lo que aumenta su susceptibilidad a infecciones y enfermedades. Además, se busca

investigar si existen diferencias en la respuesta inmunitaria dependiendo del tipo de cáncer que presenten los pacientes.

Por otro lado, en los últimos años, el uso de métodos de regularización/penalización en los modelos predictivos ha ganado relevancia con el objetivo de facilitar la interpretación de los datos. Sin embargo, el uso de estos métodos aún no está muy extendido debido a la falta de aplicabilidad práctica, ya que existen numerosos métodos, pero su utilidad real aún es desconocida. En este trabajo se busca demostrar la utilidad de los métodos multivariantes dispersos basados en la propuesta de (Lê Cao, Boitar, & Besse, 2011). Asimismo, se pretende demostrar cómo estos métodos ofrecen una solución al problema clínico planteado contribuyendo al objetivo ODS número 9, al emplear métodos novedosos.

Por último, es importante mencionar que los resultados de este trabajo se han presentado parcialmente en el congreso de Mathematical Modelling in Engineering and Human Behaviour 2023 (MME&HB2023) en la publicación (Conchado, et al., 2023).

1.3. Objetivos

1.3.1. Principal

El objetivo principal de este trabajo es comprobar si los pacientes oncológicos en tratamiento activo pueden desarrollar inmunidad frente al virus SARS-CoV-2 mediante la vacunación, tratando de investigar y comprender la respuesta inmune considerando sus características individuales y el tipo de terapia recibido. Se pretende analizar la generación de anticuerpos teniendo en cuenta que los pacientes con cáncer pueden presentar una respuesta inmune reducida debido a su propia enfermedad y los tratamientos utilizados. Es importante evaluar la eficacia de la vacunación determinando si produce una respuesta inmune adecuada en términos de generación de anticuerpos y respuesta celular.

1.3.2. Secundarios

Por otro lado, durante el desarrollo del estudio, se abordarán objetivos secundarios para complementar el objetivo principal.

1. **Examinar de forma descriptiva las Respuestas Inmunitarias en distintos pacientes oncológicos, considerando sus características individuales:**

   El objetivo es examinar la respuesta inmune que se desarrollan en pacientes oncológicos con diferentes tipos de cáncer y tratamientos. El enfoque se centra en investigar la relación entre la dosis recibida con la edad, sexo y el tipo de cáncer que padece, con la generación de anticuerpos, los niveles de respuesta celular.

   Para lograr estos objetivos, se llevará a cabo un estudio descriptivo que identifique patrones y relaciones entre la respuesta inmune que ayuden a comprender cómo la dosis administrada y las distintas características del paciente influyen en la respuesta inmune observada.

2. **Estudiar la relación entre las medidas de respuesta inmunitaria:**

   Por otro lado, se trata de modelar y examinar la relación entre distintas medidas de respuesta inmune en pacientes oncológicos, centrándose específicamente en la respuesta de anticuerpos creada contra la COVID19 tras la vacunación, y su vínculo con la respuesta celular y los niveles de los parámetros Hematológicos.

   Es importante tener en cuenta que la respuesta de anticuerpos generada por la vacuna puede tener una duración limitada, lo que implica que se requieran dosis adicionales para mantener la protección a largo plazo. Por ello, se busca estudiar esta relación para identificar posibles interacciones y efectos que la vacuna pueda tener en el resto de la respuesta inmune.

3. **Determinar la generación de anticuerpos al virus SARS-CoV-2 en función de las características del paciente y de las medidas de respuesta inmunitaria:**

   Se trata de desarrollar un modelo que permita comprender y predecir la generación de anticuerpos al virus SARS-CoV-2 en pacientes oncológicos, considerando las características individuales del paciente, la dosis administrada y otras medidas alternativas de respuesta inmune.

   El objetivo es desarrollar un modelo discriminante que pueda clasificar correctamente a los pacientes en función de su capacidad para generar anticuerpos frente al virus. Este modelo permitirá identificar los factores más influyentes en la generación de anticuerpos y proporcionará una herramienta para realizar predicciones sobre la respuesta inmune de los pacientes.

1.4. Estructura del trabajo

El trabajo se iniciará con una sólida explicación del marco teórico. En primer lugar, se bridará una explicación detallada del funcionamiento de la vacunación contra en COVID19, centrándose en las vacunas de ARNm y como estimulan la respuesta inmune de organismo. Además, se describirán los distintos métodos empleados en el estudio, tanto los métodos clásicos de análisis multivariante como los métodos dispersos utilizados para analizar los datos obtenidos.

A continuación, se expondrán los resultados obtenidos mediante la aplicación de estos métodos en el estudio abordando los objetivos generales planteados previamente, así como los objetivos secundarios. Esta sección proporcionará una visión clara y detallada de los hallazgos y su relevancia en el contexto de la investigación.

Finalmente, se elaborará una conclusión que abordará la utilidad de los métodos utilizados en el estudio y se hará un resumen de los resultados obtenidos. Se destacará la importancia de estos hallazgos y se discutirán las implicaciones prácticas que podrían derivarse de ellos. Además, se reconocerán las limitaciones del estudio y se ofrecerán recomendaciones para futuras investigaciones relacionadas.

2. MARCO TEÓRICO

2.1. Vacunación Anti-COVID19

La aparición en diciembre de 2019 del nuevo virus conocido como Coranvirus (SARS-CoV-2) tuvo consecuencias devastadoras a nivel mundial. Aunque se implantaron medidas de control como el uso de mascarillas, el distanciamiento físico y el aislamiento, estas acciones no fueron suficientes para frenar la propagación del virus. Ante esta situación, se intensificaron los esfuerzos en el desarrollo de vacunas con el objetivo de reducir la propagación y la mortalidad asociada al virus (Polack, et al., 2020).

La respuesta inmunitaria inducida por las vacunas contra el SARS-CoV-2 es un tema de gran relevancia en la lucha contra la COVID19. Entre las diferentes tecnologías en el desarrollo de vacunas, surgieron las vacunas basadas en Ácido Ribonucleico Mensajero (ARNm) como una prometedora opción. Estas vacunas aprovechan la capacidad del ARNm para proporcionar instrucciones a las células del cuerpo humano y estimular una respuesta inmunitaria específica contra el virus. Una vez administrada, las células del cuerpo producen esta proteína, lo que desencadena una respuesta de defensa del sistema inmunológico (Su, et al., 2022).

A medida que se ha avanzado en la comprensión de la eficacia de las vacunas basadas en ARNm, también se ha reconocido la importancia de evaluar su respuesta inmunitaria en grupos de pacientes especiales, como aquellos con diagnóstico de cáncer y en tratamiento activo (Baden, et al., 2021).

Los pacientes con enfermedades neoplásicas [1]representan una población particularmente vulnerable, con un mayor riesgo de adquirir infecciones por el virus SARS-CoV-2 y sufrir un curso más grave de la enfermedad. Sin embargo, hasta ahora, la eficacia de las vacunas en estos pacientes ha sido un tema de debate debido a su exclusión de los estudios clínicos iniciales que llevaron a la aprobación de las vacunas basadas en ARNm.

Investigaciones previas han sugerido que la respuesta inmunitaria generada por las vacunas en pacientes con cáncer sólido puede ser disminuida en comparación con individuos sanos. Además, se ha observado que aquellos sometidos a tratamientos como quimioterapia e inmunoterapia pueden presentar niveles más bajos de anticuerpos en comparación con aquellos que reciben otros tipos de terapias (Su, et al., 2022).

Asimismo, se ha observado que después de unos meses los niveles de anticuerpos contra el virus disminuyen significativamente en pacientes con cáncer y también en personas sin cáncer. Una de las soluciones para combatir esto se trata de las vacunas de refuerzo que pueden restaurar la respuesta de anticuerpos reducida (Waldhorn, et al., 2021).

[1] Se refiere a la presencia de tumores o neoplasias en el organismo, que pueden ser benignos o malignos, y se caracterizan por un crecimiento celular anormal y descontrolado.

2.2. Respuesta inmune

El sistema inmune (SI), formado por células y moléculas, es responsable de la respuesta ante la introducción de sustancias extrañas en el cuerpo. El objetivo principal del SI es defender al organismo contra microorganismos infecciosos y responder a sustancias extrañas y a productos dañados de las propias células o células malignas (Abbas, Lichtman, & Pillai, 2020).

El SI no se encuentra en un órgano específico, sino que es una red de tejidos, células y moléculas que trabajan en conjunto para formar una respuesta unificada conocida como respuesta inmunitaria (Castellanos-Bueno, 2020). La respuesta inmunitaria se divide en dos: inmunidad innata e inmunidad adquirida. La inmunidad innata se trata de la primera línea de defensa del cuerpo, compuesta por barreras físicas y respuestas celulares. La respuesta inmunitaria adquirida se refiere a un sistema específico de respuesta celular y humoral que desarrolla el individuo a lo largo de su vida (Abbas, Lichtman, & Pillai, 2007).

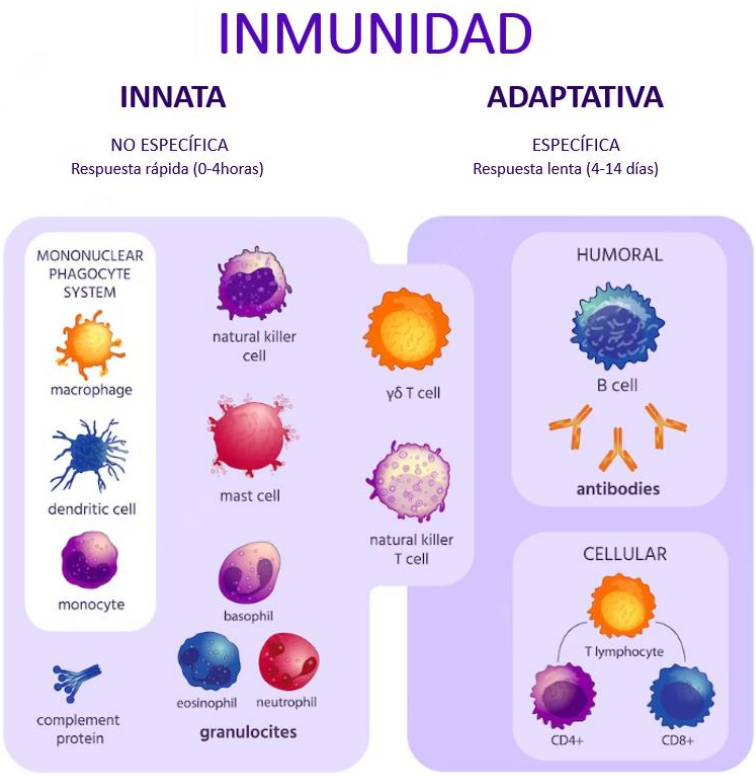

Figura 1. Células del sistema inmunitario innato y adaptativo

La Figura 1 muestra las distintas células del SI, donde se aprecia que la inmunidad innata es una respuesta rápida y uniforme que se activa ante microorganismos y células dañadas y que la adaptativa es una respuesta específica y altamente especializada que se desarrolla después de que el sistema inmunitario haya sido expuesto a un antígeno particular.

Además, las respuestas inmunitarias adaptativas se pueden identificar dos componentes, inmunidad humoral y celular. La inmunidad humoral se basa en la producción de anticuerpos por parte de los

linfocitos B, los cuales reconocen y neutralizan los microorganismos y los marcan para su eliminación. Por otro lado, la inmunidad celular es llevada a cabo por los linfocitos T y se encarga de destruir los microorganismos que se encuentran dentro de las células o los fagocitos (Hoebe, Janssen, & Beutler, 2004). Además, mediante los receptores de células B y T se obtiene información valiosa sobre como el sistema inmunológico ha interactuado con antígenos y ha respondido a diferentes estados de salud y enfermedad (Weber, y otros, 2022).

Asimismo, los anticuerpos, también conocidos como inmunoglobulinas, se consideran proteínas clave en el funcionamiento de la respuesta inmunitaria adaptativa. En función de su estructura y clase desempeñan diferentes funciones, por ejemplo, la IgM e IgG que ayudan a la eliminación de los agentes invasores, como bacterias, facilitando su destrucción (Instituto Europeo & Instituto Europeo de Salud y Bienestar Social, 2020).

Los anticuerpos no eliminan directamente las bacterias, sino que requieren la ayuda del complemento. Tras la primera vacunación, se generan anticuerpos de amplia diversidad, pero con baja especificidad. Sin embargo, con las sucesivas inmunizaciones, la especificidad y afinidad de los anticuerpos aumentan, lo que implica que no solo hay más anticuerpos, sino que también son de mejor calidad (Instituto Europeo & Instituto Europeo de Salud y Bienestar Social, 2020).

Para la producción y creación de estos anticuerpos se encuentra la vacuna, diseñadas para estimular una respuesta inmunitaria específica contra un patógeno particular o sus componentes. La vacunación implica la introducción controlada en el cuerpo de componentes específicos e inactiva del patógeno, sus proteínas o fragmentos del mismo. De esta manera, estos componentes son reconocidos por el sistema inmunitario como sustancias extrañas y desencadenan una respuesta inmunitaria (Poland, Ovsyannikova, Jacobson, & Smith, 2007).

La respuesta inmunitaria desencadenada por la vacunación no solo se limita a la eliminación del patógeno presente en la vacuna, sino que también tiene un efecto duradero en el sistema inmunitario del individuo. Además, la vacunación tiene la capacidad de crear memoria inmunológica, lo que significa que el sistema inmunitario "recuerda" la exposición al patógeno y puede responder de manera más rápida y efectiva en futuros encuentros con el mismo agente infeccioso (Montoya, 2021).

2.3. Métodos Multivariantes clásicos y dispersos

2.3.1. Análisis de Componentes Principales (PCA)

El Análisis de Componentes en Principales es una técnica de aprendizaje no supervisado que tiene como objetivo analizar la posibilidad de representar adecuadamente la información de un conjunto de datos en un número menor de variables construidas como combinación lineal de las anteriores. Esta técnica permite transformar las variables originales en nuevas variables incorrelacionadas entre sí no observables directamente, pero que facilita la interpretación de los datos (Dheri, Soumen, Varinderpal, Sudeep, & Choudhary, 2019).

A continuación, se describen las principales funciones del PCA.

- Reducción de la dimensionalidad y eliminación de la correlación: Se representará la información de N observaciones y K variables de un conjunto de datos mediante un nuevo conjunto de variables latentes incorrelacionadas, menor que el original, denominadas componentes principales, siendo estas una combinación lineal de las variables originales. Estas componentes intentarán explicar la máxima variabilidad de los datos originales imponiendo ortogonalidad, donde la primera componente será la que más información recoja de los datos originales. A medida que avanza en el orden de las componentes principales, la cantidad de variabilidad explicada de cada componente disminuye gradualmente (Ringnér, 2008).
- Detección de anomalías: El PCA permite detectar observaciones atípicas que rompen la estructura de correlación y outliers severos o extremos que constituyen a una variación inusual dentro del modelo.
- Análisis exploratorio/compresión de los datos: Permite encontrar patrones en relación con las variables, así como la relación entre las distintas observaciones. Por otro lado, muestra las variables que tienen un mayor impacto en la variabilidad de los datos, ayudando a identificar aquellas variables más relevantes.

Para la creación del modelo, se parte de una matriz de datos X de dimensión NxK, donde N corresponde con el número de variables y K con las observaciones. Estos datos, son sometidos a un pretratamiento donde la media es centrada a cero y se reescala a varianza unitaria de forma que todas las variables tengan la misma longitud y media cero (Štruc & Pavešic´, 2009).

A continuación, podemos ver la descomposición de la matriz X (centrada y escala) mediante las distintas componentes principales. Esta se componente de las puntuaciones factoriales (scores) (T) y de las cargas factoriales (loadings) (P) de cada componente, además de los residuos (E) que recogen aquella variabilidad que no se ha podido explicar mediante el número de componentes seleccionadas.

$$X = t_1 p_1{}' + \cdots + t_a p_a{}' + E \quad (1)$$

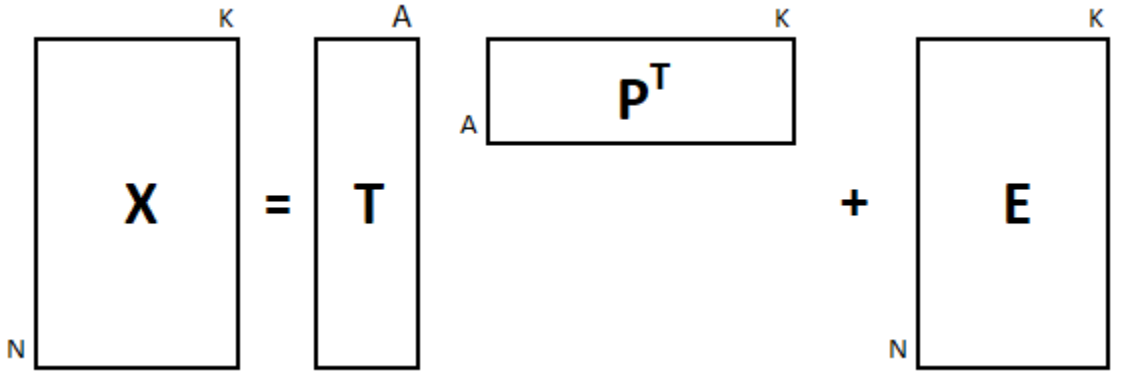

Figura 2. Estructura PCA. Decomposición de la matriz X

El cálculo de la primera componente proporciona una dirección en el espacio a lo largo de la cual se maximiza la suma de los cuadrados de todas las proyecciones de los objetos en esa dirección, donde los (scores) son los valores de las proyecciones de todos los objetos en la línea y los loadings contiene los cosenos direccionales (Paul & Johan , 2020).

Para crear la segunda componente ocurre lo mismo, se elimina la información de la primera componente $t_1 p_1'$ del conjunto de datos originales y así sucesivamente hasta obtener el número de componentes deseada. La segunda componente es una recta en el espacio ortogonal a la primera componente, formando entre las dos un plano de suma de cuadrados maximizada.

En cuanto a la validación del modelo y para poder detectar anomalías en los datos, se emplea, por un lado, el gráfico T2-Hotelling que permitir detectar observaciones que se desvían de la estructura general de los datos. El estadístico T^2 de Hotelling es una generalización multivariante del estadístico t de Student que mide la distancia multivariante entre cada observación y el centroide del conjunto de datos (Ji-Hoon, Jong-Min, Sang Wook, Dongkwon, & In-Beum, 2005). Esta se puede escribir de la siguiente manera:

T^2-Hotelling
$$T^2 = \sum_{a=1}^{A} \frac{\tau_a^2}{\lambda_a} \quad (2)$$

Por otro lado, el gráfico SPE permite detectar observaciones individuales que se desvían significativamente del modelo. Este se obtiene midiendo la distancia entre cada observación y el espacio de modelos generado por las variables principales extraídas del PCA.

SPE
$$SPE = e^T e \quad (3)$$

Finalmente, el PCA ofrece una sencilla comprensión de los datos a través los distintos gráficos. En primer lugar, el gráfico de puntuaciones factoriales (scores) que representa las proyecciones de las observaciones en el espacio de las componentes principales y, por otro lado, el gráfico de las cargas factoriales (loadings) que muestra las contribuciones de las variables originales en cada componente principal (Greenacre, et al., 2022).

2.3.2. Análisis de Componentes Principales Dispersos (sPCA)

El análisis de componentes principales, tal y como viene comentado anteriormente, se trata de una combinación lineal de todas las variables originales, sin embargo, en algunas ocasiones no todas las variables resultan relevantes, por ello que se pretende buscar métodos que produzcan cargas nulas, facilitando así la comprensión de las nuevas variables. El Análisis de Componentes Principales Sparse, tiene como finalidad conseguir que gran parte de los coeficientes de la matriz de cargas sean nulos (Rodolphe, Obozinski, & Francis, 2009).

Sparse PCA puede considerarse como un PCA modificado. Es decir, es posible realizar modificaciones en relación con la técnica principal, lo que da origen a diversos algoritmos que abordan el problema desde diferentes perspectivas y considerando distintas restricciones.

Para la aplicación del SPCA en este estudio se implementa el paquete mixOmics de R que se trata de una librería de análisis multivariante especializada en la integración de datos omics, basando el SPCA en la descomposición de valores singulares (SVD) (Huang, 2008), con una penalización "lasso" para obtener componentes principales dispersos.

El Sparse PCA utiliza la aproximación de rango bajo de la descomposición del SVD (se selecciona un subconjunto de los valores y vectores singulares más grandes y se descartan los demás) y la relación con la regresión de mínimos cuadrados para calcular las componentes y seleccionar las variables.

Cuando se resuelve utilizando el enfoque NIPALS, en cada iteración se selecciona una variable y se obtiene su vector de carga mediante una regresión local utilizando el método de mínimos cuadrados, además, se aplica una penalización "lasso" a cada vector de carga para obtener un vector de carga disperso. De esta manera, cada componente principal queda definida por una combinación lineal de variables, considerándose estas las más influyentes (Lê Cao & Welham, 2021).

En el PCA disperso, los componentes principales no están garantizados de ser ortogonales, cuando esto ocurre, la proyección de los datos se realiza en el espacio generado por el primer vector de carga y luego se ajusta la varianza explicada teniendo en cuenta la posible correlación existente entre los componentes principales. Esto significa que se considera cómo se relacionan entre sí las diferentes componentes al explicar la varianza total de los datos. A pesar de esto, cuando los datos se encuentran centrados y escalados correctamente, el vector de carga tiende a volverse ortogonal (Lê Cao & Welham, 2021).

Por último, hay que considerar que es poco probable que las variables seleccionadas en una dimensión sean seleccionadas en otra dimensión debido a la propiedad de ortogonalidad de los componentes principales.

La siguiente figura muestra el esquema del PCA Disperso donde los vectores de carga son penalizados utilizando Lasso para reducir coeficientes a cero.

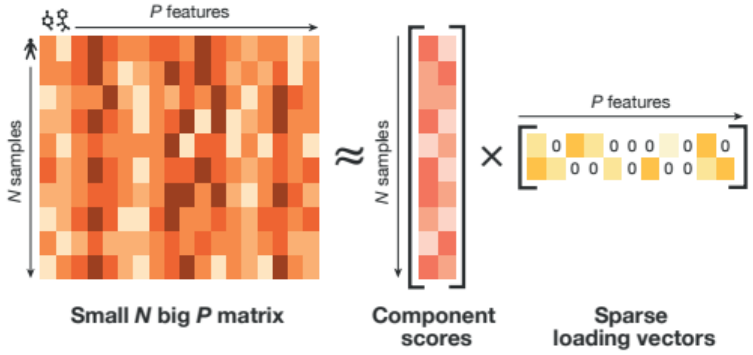

Figura 3. Representación del método SPCA

Fuente: Cao, K.-A., & Welham, Z. M (2021)

2.3.3. Regresión Lineal

Considerando el modelo de regresión tradicional de N observaciones asociadas a una variable respuesta (y_i) y a K variables predictoras. Es decir, una matriz X de dimensión (NxK) y una matriz respuesta de dimensión $(1xK)$.

Existen diferentes métodos para ajustar el modelo lineal al conjunto de datos dado, pero el más popular es el método de los mínimos cuadrados ordinario (Ordinary Least Squares, OLS). El objetivo es, en primer lugar, encontrar aquellas variables que juegan un papel importante en la predicción de la variable

respuesta, así cómo, crear un modelo que sea capaz de predecir la respuesta con datos futuros (Dastan & Adnan, 2020).

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_i \quad (4)$$

El modelo de regresión está sujeto a la ecuación **¡Error! No se encuentra el origen de la referencia.** donde β_0 y $\beta=(\beta_1, \beta_2, \ldots, \beta_p)$ se tratan de parámetros desconocidos y e_i del error cometido en la predicción. De forma matricial, siendo X una matriz centrada y con norma 1, valor Y estimado es el siguiente:

$$\hat{Y} = XB \quad (5)$$

Para poder estimar los parámetros desconocidos se emplea el método de mínimos cuadrados. Se estima β buscando minimizar la función del error cuadrático, es decir, la suma de los cuadrados entre la diferencia de las observaciones y_i y la recta de regresión mínima. La función RSS es una función de segundo grado con respecto a los parámetros y, por lo tanto, se garantiza la existencia de un valor mínimo, aunque no necesariamente es único.

$$\underset{\beta_0,\beta}{minimizar}\ RSS = \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \quad (6)$$

Sin embargo, las aproximaciones OLS no siempre resultan ser las óptimas. Aunque las estimaciones de mínimos cuadrados tienden a tener un sesgo bajo, suelen presentar una alta varianza. Por lo tanto, en ciertos casos, se puede mejorar la precisión de las predicciones al contraer algunos coeficientes hacia 0, incluso haciéndolos nulos o prácticamente nulos (Hastie, Tibshirani, & Wainwright, 2015).

Por otro lado, cuando se trabaja con un conjunto amplio de variables explicativas, es común buscar una selección más reducida de variables que capturen los efectos más significativos (Hastie, Tibshirani, & Friedman, 2009).

Es aquí, donde se plantean técnicas de técnicas de regularización o penalización, que ofrecen enfoques prometedores al reducir la dimensionalidad y simplificar los problemas. Estos métodos se aplican para abordar situaciones en las que los problemas están mal condicionados debido a la falta de unicidad en la solución o la inexistente selección de variables. Además, proporcionan un marco para abordar los problemas en los que el número de variables es significativamente mayor que el número de observaciones.

3.3.4. Regresión Ridge

La regresión Ridge trata de regularizar el sobreajuste producido por las múltiples variables predictoras correlacionadas entre sí, es decir, trata de evitar los efectos producidos por problemas de colinealidad en modelo estimados por mínimos cuadrados donde el número de variables es mayor al número de observaciones (Hoerl & Kennard, 2000).

Recordando los mínimos cuadrados de la regresión **¡Error! No se encuentra el origen de la referencia.** o btendremos una ecuación similar a excepción del término de regularización introducido en la función

objetivo, que es proporcional a la suma de los cuadrados de los coeficientes de regresión. Es decir, se define la norma ℓ2 de $\beta = (\beta 1, \dots, \beta p)$ como:

$$(||\beta||_2)^2 = \sum_{j=1}^{p} |\beta_j|^2 = \sum_{j=1}^{p} \beta_j^2 \tag{7}$$

Es decir, los coeficientes estimados por Ridge se obtienen de la siguiente manera.

$$\hat{\beta}^{ridge} = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{8}$$

$$\text{s.a. } \lambda \geq 0$$

La regresión Ridge es un método de mínimos cuadrados penalizado. Este método restringe la norma ℓ2 de los coeficientes de regresión ordinaria y para ello hace uso de un escalar λ no negativo que fuerza a los coeficientes a anularse (Kennard, 2000).

La cantidad de regularización de controla mediante un hiperparámetro λ establecido antes del ajuste del modelo. Cuanto mayor sea el valor de λ, mayor será la penalización y por tanto el valor de los coeficientes será más cercano a cero, es decir, mayor contracción de los coeficientes. En comparación con la estimación de mínimos cuadrados ordinarios, cuando λ=0 estamos en el mismo caso, sin embargo, cuando $\lambda \rightarrow \infty$, $\beta \rightarrow 0$ estamos introduciendo sesgo, pero reduciendo la varianza

La función objetico de la regresión Ridge puede ser expresada así:

$$\underset{\beta_0,\beta}{minimizar} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

$$s.a \sum_{j=1}^{p} \beta_j^2 \leq t^2 \tag{9}$$

donde t es el parámetro de penalización por complejidad.

La selección óptima de λ en la regresión Ridge, es decir, aquel valor que minimiza el error de predicción del modelo se logra mediante la técnica de validación cruzada. Es una técnica que ayuda a evaluar el rendimiento del modelo utilizando conjuntos de datos de entrenamiento y prueba aleatoriamente.

El modelo de regresión Ridge se ajusta utilizando los datos de entrenamiento y se evalúa su rendimiento utilizando los datos de prueba, calculando el error de predicción. Este proceso se repite varias veces con diferentes combinaciones de conjuntos de entrenamiento y prueba y al finalizar se promedian los errores de predicción de todas las iteraciones para obtener un resultado único. De esta manera, la validación cruzada permite seleccionar el valor de λ que produce el menor error promedio (Petrellis & Skoupil, 2023).

En la Figura 4 se muestra la representación de la estimación Ridge. Las elipses de los contornos verdes muestran función de la suma residual de los cuadrados, además en el centro se encuentra la estimación mínima cuadrática sin estar sujeta a ninguna restricción. Por otro lado, el contorno azul está representado

la restricción Ridge, $\beta_1^2 + \beta_2^2 \leq t^2$. De está manera, obtendremos la estimación Ridge, $\hat{\beta}^{ridge}$, en el punto de corte de RSS con la restricción impuesta (Hastie, Tibshirani, & Friedman, 2009).

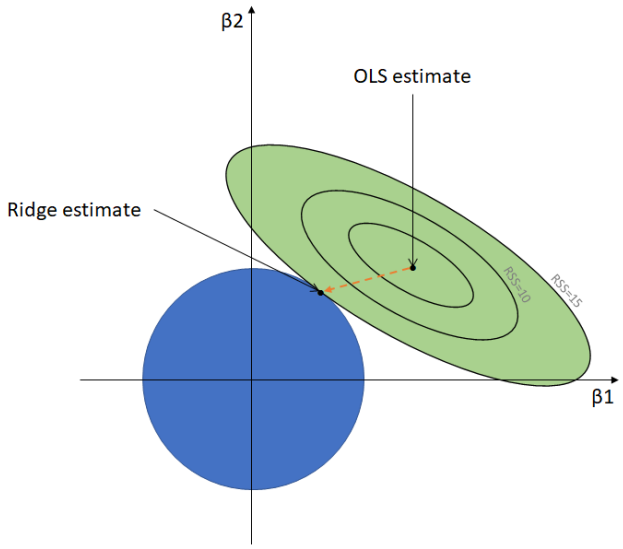

Figura 4. Estimación para la Regresión Ridge

La Regresión Ridge, sin embargo, se plantea con un inconveniente y es que contrae los coeficientes hacia cero, pero no consigue que ninguno de ellos sea nulo. Por ello, no se produce la selección de variables lo que supone un inconveniente en casos como este donde tenemos un gran número de variables predictoras.

3.3.5. Regresión Lasso (Least Absolute Shrinkage and Selection Operator)

Tras observar los beneficios e inconvenientes del método Ridge, se estudia el método Lasso donde no solo se consigue una regresión que permite estabilizar las estimaciones y predicciones, sino que permite realizar una selección de variables. Lasso es una técnica similar a Ridge de regresión regularizada, sin embargo, está sujeta a una penalización distinta que ofrece unas consecuencias importantes (Tibshirani, 1996).

La técnica Lasso aborda el problema de mínimos cuadrados mediante una restricción en la norma ℓ1 del vector de coeficientes. En otras palabras, Lasso utiliza una técnica de regularización que impone una penalización en la magnitud total de los coeficientes, reduciendo así la importancia de aquellos coeficientes que no contribuyen significativamente al modelo (Tibshirani, 1996).

Se define la norma ℓ1 de $\beta = (\beta 1, \dots, \beta p)$ como:

$$(||\beta||_1)^1 = ||\beta||_1 = \sum_{j=1}^{p} |\beta_j|^1 = \sum_{j=1}^{p} |\beta_j| \qquad (10)$$

Los coeficientes estimados por Lasso se obtienen de la siguiente manera:

$$\hat{\beta}^{Lasso} = \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = RSS + \lambda \sum_{j=1}^{p} |\beta_j| \quad (11)$$

$$s.a\ \lambda \geq 0$$

La función objetico de la regresión Lasso puede ser expresada así:

$$\underset{\beta_0,\beta}{minimizar} \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \quad (12)$$

$$s.a \sum_{j=1}^{p} |\beta_j| \leq t$$

A diferencia de la regresión Ridge que utiliza la penalización ℓ2, la regresión Lasso hace uso de la penalización ℓ1. La modificación en la función de penalización puede parecer pequeña, pero su efecto en el estimador final es significativo y dramático.

Al igual que la regresión Ridge, al penalizar los coeficientes también se contraen a cero, sin embargo, medida que aumenta el parámetro de complejidad, Lasso produce estimaciones nulas para algunos coeficientes y no nulas para otros, lo que resulta en una selección de variables continua. Este tipo de soluciones, que tienen múltiples valores iguales a cero, se conocen como soluciones dispersas (sparse). Esta forma de penalización se convierte en una especie de selección de variables continúas debido a esta característica (Hastie, Tibshirani, & Friedman, 2009).

De esta manera, Lasso reduce la variabilidad de las estimaciones al reducir los coeficientes y, al mismo tiempo, produce modelos interpretables al reducir algunos coeficientes a cero.

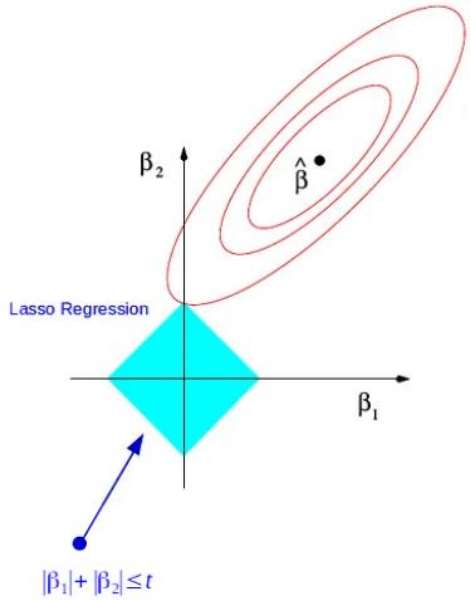

Figura 5. Estimación para la regresión Lasso

En la Figura 5 se muestra la representación de la estimación Lasso. Las elipses de los contornos verdes muestran función de la suma residual de los cuadrados, además en el centro se encuentra la estimación mínima cuadrática sin estar sujeta a ninguna restricción. Por otro lado, el contorno azul está representado la restricción Lasso, $|\beta_1| + |\beta_2| \leq t$. De esta manera, obtendremos la estimación Lasso, $\hat{\beta}^{Lasso}$, en el punto de corte de RSS con la restricción impuesta (Hastie, Tibshirani, & Friedman, 2009).

Para a selección del λ que minimiza el error de predicción de la regresión Lasso se aplica la técnica de validación cruzada tal y como se ha explicado en la regresión Ridge (Petrellis & Skoupil, 2023).

3.3.6. Regresión de Mínimos Cuadrados Parciales (PLS)

PLS es un método estadístico introducido por (Wold, Kettaneh, & Tjessem, 1996) que se ha aplicado ampliamente para en los procesos industriales como una herramienta estratégica para la mejora y optimización de procesos. Esto se debe a que los procesos industriales se caracterizan por cantidades masivas de datos altamente correlacionados (Ferrer, Aguado, Vidal-Puig, Prats, & Zarzo, 2008).

La regresión de Mínimos Cuadrados Parciales, PLS, es una técnica multivariante que combina algunas características tanto del Análisis de Componentes Principales como de la Regresión Lineal Múltiple (Abdi, 2010). La regresión PLS trata de encontrar un modelo de regresión proyectando tanto las variables predictoras X como las variables respuesta Y a un nuevo espacio latente incorrelacionado. Es decir, realiza una regresión de mínimos cuadrados sobre las componentes incorrelacionadas en lugar de hacerlo sobre las variables originales, de esta manera, supera los problemas de multicolinealidad que presenta la regresión lineal (Chin, 1998).

El objetivo principal de PLS es maximizar la covarianza entre las variables predictoras y la variable respuesta, al tiempo que reduce la multicolinealidad y la dimensionalidad del conjunto de datos.

A continuación, podemos ver la descomposición de la matriz X de dimensión *(NxK*) observaciones y variables predictoras y la matriz Y de dimensión *(MxK*) variables respuesta y observaciones:

$$X = TP' + E \tag{13}$$

$$Y = TC' + F \tag{14}$$

Donde (T) son las puntuaciones factoriales del espacio X, la matriz T se estima como combinaciones lineales de las variables X con coeficientes de peso W que vemos a continuación y permite resumir las variables X que están correlacionadas con las variables respuesta. (Geladi & Kowalski, 1986)

$$T = XW^* = W(P^TW)^{-1} \tag{15}$$

Los scores del espacio Y se denomina (U) como vemos a continuación en la Figura 6 y se estimas como combinación lineal de las variables Y con peso C.

La matriz P, por otro lado, se refiere a las cargas factoriales del espacio X y la matriz C a las cargas factoriales del espacio Y. Por último, los residuos (E) y (F) que recogen el error cometido (Rosipal & Trejo, 2001).

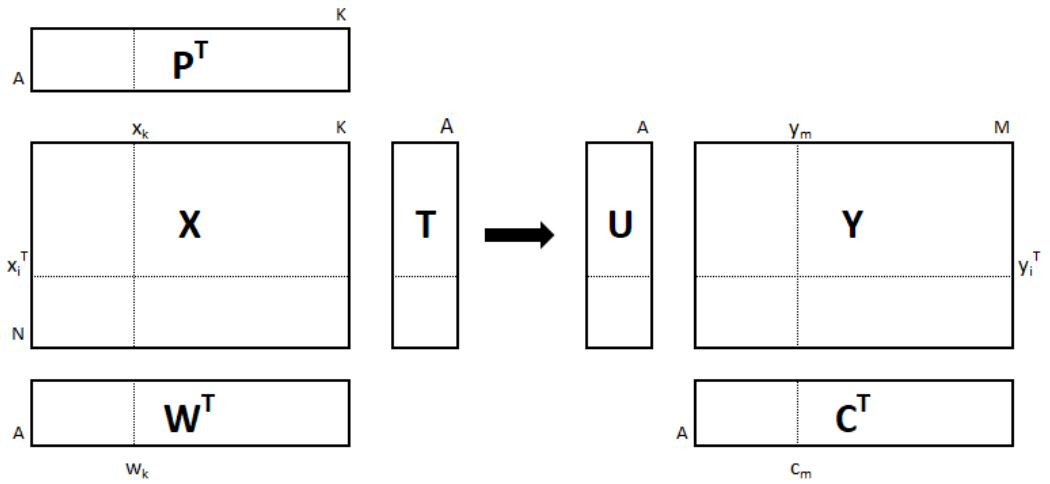

Figura 6. Estructura PLS

Por tanto, podemos decir la primera componente PLS representa la dirección en el espacio de las variables predictoras que explica la mayor variabilidad posible en la variable respuesta. Para obtener más componentes se debe usar como base la primera componente. Cada componente adicional captura la variabilidad restante que no ha sido explicada por las componentes anteriores.

3.3.7. Regresión de Mínimos Cuadrados Parciales Dispersos (sPLS)

La regresión de Mínimos Cuadrados Parciales Dispersos fue pensada para crear simultáneamente una selección de variables tanto en el espacio de las X como en el de las Y, mediante la penalización de Lasso en el vector de las cargas factoriales (loadings) (Lê Cao, Rossouw, Robert-Granié, & Besse, 2008).

Para la introducción del concepto sparsity se han propuesto dos enfoques distintos. Uno es el de (Lee, Lee, Lee, & Pawitan) basado en una modificación del algoritmo NIPALS. El otro, propuesto por (Lê Cao, Rossouw, Robert-Granié, & Besse, 2008), está basado en la introducción de penalizaciones en la descomposición SVD de la matriz de covarianzas. Este último enfoque es el implementado en mixOmics y es el que se aplica en estos análisis.

Similar a como ocurre en el SPCA, las variables menos relevantes de cada vector de loadings se les pone un peso de cero, lo que hace que las componentes sean calculadas únicamente con las variables con peso no nulo, es decir, las más relevantes (Lê Cao & Welham, 2021). El peso cero se asignan de forma óptima con el Laso, garantizando al mismo tiempo que se maximiza la covarianza entre los conjuntos de datos.

Para la selección del número de componentes óptimo, el paquete de MixOmics ofrece el criterio Q mediante la validación cruzada. El Q^2 evalúa la capacidad de un modelo para predecir la respuesta de un conjunto de. Se calcula como el coeficiente de determinación (R^2) entre los valores reales y los valores predichos para los datos de prueba, y se ajusta por el número de componentes principales o dimensiones del modelo.

Por último, para seleccionar el número de variables en cada componente se emplean diversos métodos de precisión mediante validación cruzada. Para este estudio, se aplica el Error Absoluto Medio (MAE) que

representa el promedio de las diferencias absolutas entre los valores reales y los valores predichos por el modelo (Chai & Draxler, 2014).

$$MAE = \frac{1}{N} x \sum_{t=1}^{n} |y_i - \hat{y}_i| \quad (16)$$

3.3.8. Regresión de Mínimos Cuadrados Parciales Discriminante (PLS-DA)

El PLS-DA es una extensión del PLS que se utiliza para análisis de clasificación y discriminación, cuando la variable respuesta es categórica y se busca clasificar o discriminar entre diferentes clases o categorías. Es decir, se basa en la aplicación del modelo PLS ofreciendo una alternativa atractiva al análisis discriminante lineal clásico (LDA) (Fordellone, Bellincontro, & Mencarelli, 2020).

El PLS-DA aprovecha los beneficios del PLS para manejar variables predictoras altamente correlacionadas y reducir la dimensionalidad del conjunto de datos. Al mismo tiempo, utiliza el análisis discriminante clásico para realizar la clasificación y discriminación de las diferentes categorías de la variable respuesta. La combinación de estos dos enfoques permite obtener una representación de las variables predictoras que se correlaciona con la variable respuesta categórica, lo que facilita la clasificación precisa de nuevas observaciones (Rocke & David, 2002).

La estructura del PLS-DA es igual que las del PLS, pero se debe crear un clasificador sobre la variable respuesta tal y como se ve a continuación.

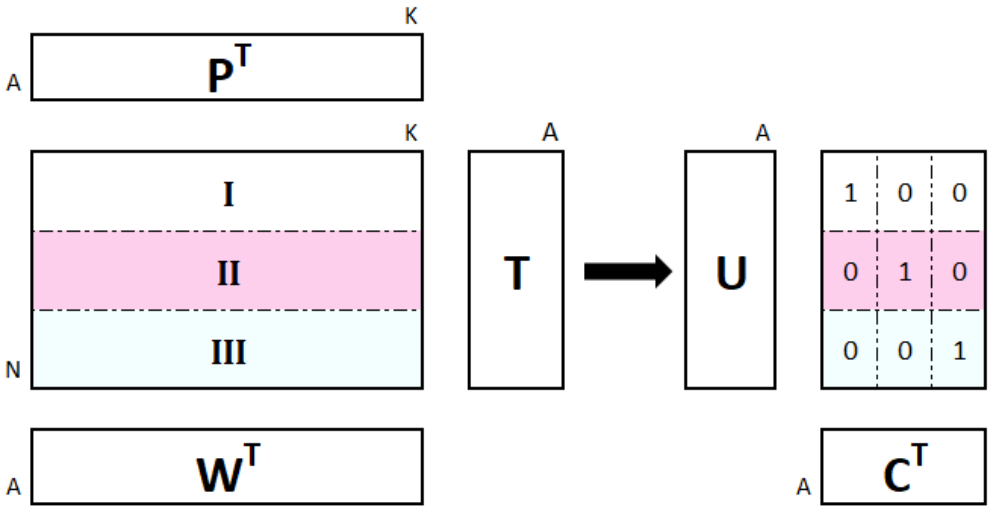

Figura 7. Estructura PLS-DA

El PLS-DA comprende y predice la pertenencia de las observaciones de la matriz X en las diferentes clases que representa la matriz Y. En la matriz Y, se asignan $1s$ en cada columna a los individuos asociados a las clases de dicha columna y $0s$ al resto de individuos. Una vez establecidas las etiquetas, se construye el modelo utilizando el mismo procedimiento y estructura que el PLS convencional. El objetivo es encontrar una relación óptima entre las variables predictoras en la matriz X y las categorías representadas en la matriz Y para lograr una clasificación precisa.

3.3.9. Regresión de Mínimos Cuadrados Parciales Discriminante (sPLS-DA)

El PLS-DA maximiza la covarianza entre las combinaciones lineales de las variables X y las combinaciones lineales de las variables dummy Y, sin embargo, cuando se quieren trabajar únicamente con las variables más significativas se debe acudir a un método disperso, en este caso, sPLS-DA (Lê Cao, Boitar, & Besse, 2011).

Para el sPLS-DA se han propuesto dos enfoques, el primero que integra el Spls en un modelo lineal generalizado (GLM) según (Chung & Keles, 2010). El segundo, el que se aplica en MixOmics y por tanto en este trabajo, lo propuso (Lê Cao, Boitar, & Besse, 2011) y está basado en la selección y clasificación en un solo paso.

sPLS-DA permite seleccionar y clasificar variables en un solo paso. Es un caso especial de PLS disperso, en el que la penalización Lasso sólo se aplica al vector de carga a asociado al conjunto de datos X (Lê Cao & Welham, 2021).

La selección del número de componentes se mediante validación cruzada donde se mide la dificultad de clasificación con el aumento de componentes, con el error de clasificación. Este error representa la proporción de observaciones clasificadas incorrectamente en comparación con el total de observaciones. El error de clasificación se calcula dividiendo el número de observaciones clasificadas incorrectamente entre el número total de observaciones. Luego, se multiplica por 100 para obtener el error de clasificación como un porcentaje (Dougherty, Sima, Hua, Hanczar, & Braga-Neto, 2010).

Por último, para la selecciona del número de variables de cada componente se emplea nuevamente la técnica de validación cruzada y se observa en este caso también el error de clasificación.

3. DESCRIPCIÓN DEL PROBLEMA

3.1. Datos y relaciones del estudio

La base de datos utilizada en este estudio contiene información relevante de pacientes oncológicos en tratamiento activo. Estos datos, obtenidos de pacientes reales, incluyen información detallada sobre su respuesta inmune mediante diversos indicadores celulares. Además, se registra si la información se recopiló antes de que el paciente recibiera cualquier dosis de la vacuna o después de la administración de la segunda dosis contra el COVID19. Para facilitar la comprensión de las variables en la base de datos, se pueden identificar cuatro grupos distintivos.

1. El primer grupo incluye características relacionadas con el paciente como la **Edad**, **Sexo**, **Cáncer**, **Tratamiento** recibido contra el mismo**, Dosis** (sin vacunar o segunda dosis) y el **Tipo de vacuna** administrada.
2. El segundo bloque se refiere a la medición de los niveles de anticuerpos IgG específicos contra el virus SARS-CoV-2 (Anti_SARS_CoV_2).
3. Por otro lado, se encuentran los parámetros Hematológicos del paciente: **Leucocitos**, **Neutrófilos**, **Linfocitos**, **Monocitos**, **Eosinófilos**, **Basófilos**, **Trombocitos** y **Hemoglobina**.
4. Por último, se encuentra el bloque de respuesta celular, el cual es el conjunto más amplio de datos en la base. Incluye variables como **CD3**, **CD4**, **CD8**, **CD3_CD56**, **CD3 memoria central**, **etc**. Por cada tipo de variable se registra el valor Total, el porcentaje y el porcentaje de APOPTOSIS [2] .

   Sin embargo, para este estudio no se consideran ni el porcentaje de Apoptosis ni el valor Total, ya que no son el objetivo de interés del estudio. Además, se han excluido las siguientes variables del estudio debido a la variabilidad prácticamente nula en los datos; *CD4 GD, CD3 CD4 gd naive, CD3 CD4 gd memoria central, CD3 CD4 gd memoria periferica, CD3 CD4 gd TEMRA*.

Inicialmente, la base de datos contiene 94 observaciones y 219 variables, sin embargo, tras excluir las variables mencionadas se reduce a 80 el número columnas.

En el anexo se proporciona una explicación más detallada de cada uno de los parámetros Hematológicos y de las variables de respuesta célula. Asimismo, se muestra la codificación de las variables de respuesta celular que se emplea a lo largo del trabajo para una comprensión más clara de los gráficos.

[2] La apoptosis es un proceso de muerte celular programada y controlada que desempeña un papel importante en el desarrollo y mantenimiento de los tejidos.

4. RESULTADOS

4.1. Estudio de la Respuesta inmunitaria en pacientes oncológicos

Para abordar el estudio de las distintas respuestas inmunes de los pacientes, se lleva a cabo un enfoque metodológico basado en los tres objetivos específicos del trabajo.

La primera sección de los resultados se centra en examinar la relación entre los anticuerpos generados contra el virus SARS-CoV-2 y las características individuales de los pacientes, como la edad, el sexo, el tipo de cáncer y el tratamiento recibido. Este análisis se realiza mediante la utilización de diversos gráficos descriptivos y se complementa con una Regresión Lineal para identificar posibles asociaciones entre estas variables.

En la segunda sección, se aborda el estudio de los niveles de los parámetros Hematológicos, que proporcionan información valiosa sobre la respuesta inmune del paciente. Para ello, se aplica un Análisis de Componentes Principales (PCA) que permite identificar los patrones principales de variación en los diferentes componentes celulares.

Por último, se explora la relación de la respuesta celular mediante el uso de un PCA y de un PCA Disperso ya que ayuda a obtener una representación más precisa y significativa de las variables que contribuyen de manera más relevante a la respuesta celular.

4.1.1. Estudio de los anticuerpos al virus SARS-CoV-2 en función del paciente

Para llevar a cabo el análisis de la relación entre los anticuerpos generados por el paciente y sus características individuales, se ha realizado un análisis exploratorio que represente la distribución de los niveles de anticuerpos en función de la dosis administrada. Además, con el fin de añadir más información visual y contextual, se emplean puntos de colores de datos característicos que ayudan a identificar posibles patrones o tendencias. Cada punto de datos estará coloreado de acuerdo con las características específicas del paciente, como la edad, el sexo, el tipo de cáncer y el tratamiento recibido.

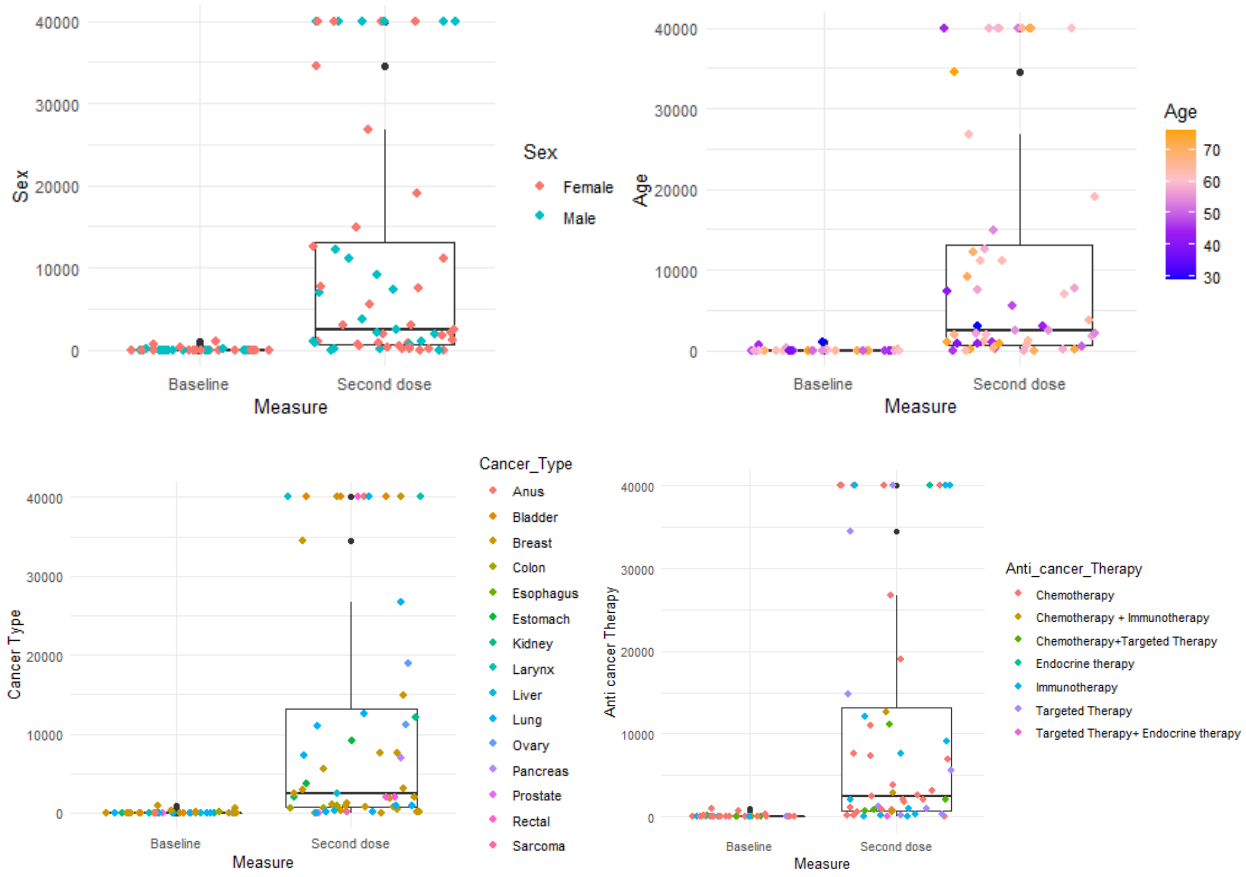

Figura 8. BoxPlot de los anticuerpos al virus SARS-CoV-2 en función de la medida y las características del paciente (Sexo, Edad, Tipo de Cáncer y Tratamiento contra el cáncer)

Estos gráficos proporcionan una representación visual de los niveles de anticuerpos en pacientes antes y después de la segunda dosis de la vacuna del COVID19.

En primer lugar, se evidencia la relevancia de la vacuna en la generación de anticuerpos. Se observa que cuando los pacientes reciben la vacuna contra la COVID19 tiene un nivel alto de anticuerpos específicos contra el SARS-CoV-2. Sin embargo, cuando aún no han recibido la vacuna se observa la ausencia de una respuesta inmunológica protectora frente a este virus.

Al analizar las características del paciente en relación con el nivel de anticuerpos, se observa que ninguna de ellas parece predominar en los niveles altos o bajos de estos. Es decir, no se identifica ninguna tendencia en cuanto al sexo del paciente que indique niveles más altos de anticuerpos. Además, no parece haber una asociación entre la edad del paciente y un nivel más elevado de anticuerpos. Lo mismo ocurre al considerar el tipo de cáncer que presenta el paciente y el tratamiento activo recibido. Con una primera y pequeña exploración descriptiva, no se encuentran correlaciones significativas entre estas variables y los niveles de anticuerpos

Para poder estudiar en profundidad la relación de las características del paciente con los anticuerpos, a continuación, se realiza una **Regresión Lineal**, permitiendo encontrar aquellas variables más significativas con respecto a los anticuerpos.

Las características del paciente se han codificado como variables categóricas, excepto, la edad, por ello, es necesario crear tantas variables dummy como categorías. En primer lugar, se ajusta un modelo de regresión lineal con todas las variables, pero se obtienen un resultado en el que ninguna de las variables del paciente resulta significativa en cuanto al número de anticuerpos.

Tabla 1. Resultado de la Regresión Lineal

Regresión Lineal			
	Coeficiente	**Error_Estandar**	**Valor_p**
(Intercept)	13597.62	15950.46	0.397
Age	84.24	175.56	0.633
Cancer_TypeBladder	-56.37	11088.34	0.996
Cancer_TypeBreast	-15786.75	10304.42	0.130
Cancer_TypeColon	-17497.94	11419.27	0.130
Cancer_TypeEsophagus	-14689.11	16828.32	0.386
Cancer_TypeEstomach	-15635.33	11552.56	0.180
Cancer_TypeKidney	-17879.30	11877.86	0.137
Cancer_TypeLarynx	926.66	12510.50	0.941
Cancer_TypeLiver	-4189.88	11800.02	0.724
Cancer_TypeLung	-14664.97	9821.39	0.140
Cancer_TypeOvary	-9589.65	11392.74	0.403
Cancer_TypePancreas	-15255.07	12637.96	0.231
Cancer_TypeProstate	-21627.90	15927.20	0.179
Cancer_TypeRectal	-6295.48	11431.83	0.584
Cancer_TypeSarcoma	-19571.27	15147.67	0.200
Anti_cancer_TherapyChemotherapy + Immunotherapy	1246.59	6810.40	0.855
Anti_cancer_TherapyChemotherapy+Targeted Therapy	-3359.03	6521.75	0.608
Anti_cancer_TherapyEndocrine therapy	36207.95	13220.02	0.008
Anti_cancer_TherapyImmunotherapy	2049.11	4654.74	0.661
Anti_cancer_TherapyTargeted Therapy	4503.67	4417.38	0.311
Anti_cancer_TherapyTargeted Therapy+ Endocrine therapy	-2696.90	12715.82	0.833

Por ese motivo, se realiza una selección tanto *Forward,* introduciendo la variable que resulte más significativo*,* como *Backward*, eliminando la variable menos significativa, en busca de las variables más significativas con el mejor resultado.

Tabla 2. Resultado de la Regresión Lineal tras selección Backward

Regresión Lineal Backward			
	Coeficiente	**Error_Estandar**	**Valor_p**
(Intercept)	-5649100	7227055	0.436
Age	209613	121296	0.087

Tras la selección Backward se obtiene un $\mathbf{R^2}$ de **0.07204**, es decir, se consigue explicar un 7% de los anticuerpos. Por otro lado, la variable más influyente en la generación de anticuerpos se trata de la edad

del paciente, sin embargo, esta no llega a ser significativa para poder explicar la cantidad de anticuerpos que crean los pacientes. Para ver su influencia se muestra el siguiente gráfico.

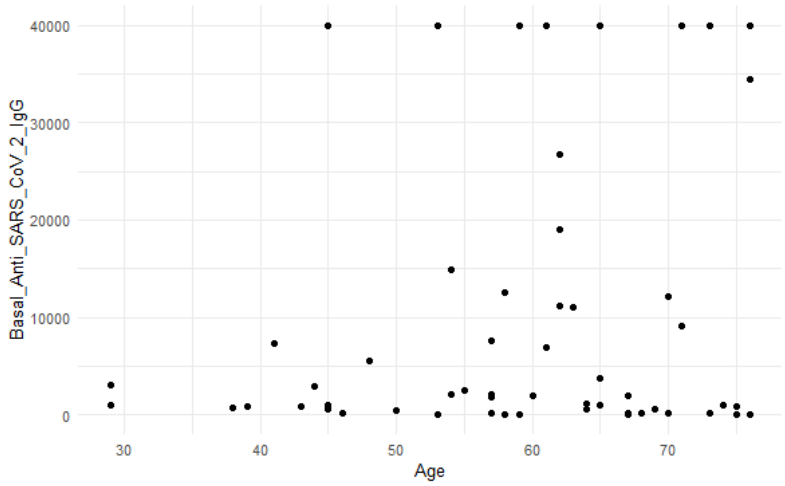

Figura 9. Distribución de los anticuerpos en función de la edad

Atendiendo a la Figura 9 que muestra el nivel de anticuerpos en función de la edad del paciente, no parece haber una diferencia significativa entre los pacientes más jóvenes y los pacientes más adultos. Por ello, en estos datos no parece que la edad pueda afectar en el aumento o descenso de anticuerpos.

Sin embargo, a fin de mejorar este resultado se decide descartar las observaciones con un nivel de anticuerpos nulo y trabajar únicamente con las restantes. Nuevamente se ajusta el modelo de regresión lineal con el total de variables, que ahora son 47, y vuelve a ser necesario aplicar un método de selección siendo el método Backward el que da el resultado óptimo.

Tabla 3. Resultado de la Regresión Lineal tras selección Backward (2)

Regresion Lineal Backward			
	Coeficiente	**Error_Estandar**	**Valor_p**
(Intercept)	40000.000	12261.087	0.003
Cancer_TypeBladder	-4030.876	14834.444	0.788
Cancer_TypeBreast	-35992.564	12716.550	0.008
Cancer_TypeColon	-36713.835	14697.106	0.018
Cancer_TypeEstomach	-36588.857	15379.690	0.024
Cancer_TypeKidney	-38970.314	16420.573	0.024
Cancer_TypeLarynx	0.000	17339.796	1.000
Cancer_TypeLiver	-18769.950	15016.704	0.221
Cancer_TypeLung	-28948.256	13373.251	0.038
Cancer_TypeOvary	-21371.310	15474.909	0.177
Cancer_TypePancreas	-33037.400	17339.796	0.066
Cancer_TypeRectal	-22939.474	14375.507	0.121
Anti_cancer_TherapyChemotherapy + Immunotherapy	-763.915	8035.608	0.925
Anti_cancer_TherapyChemotherapy+Targeted Therapy	-7051.579	7475.673	0.353
Anti_cancer_TherapyEndocrine therapy	35992.564	12716.550	0.008
Anti_cancer_TherapyImmunotherapy	6046.314	6643.329	0.370

Anti_cancer_TherapyTargeted Therapy	11003.579	5894.763	0.072

Tras realizar la selección Bacward se obtiene una **R^2** de **0.3836** donde esta vez se obtienen algunas variables significativas. Entre los distintos cánceres que padecen los pacientes resultan significativos los siguientes: Cáncer de Mama, Colon, Estomago, Riñón y de pulmón. Asimismo, la terapia endocrina resulta la más significativa en cuanto a la diferencia en los niveles de anticuerpos contra el virus del COVID19.

4.1.2. Estudio los parámetros Hematológicos

Una vez examinada la relación entre anticuerpos y perfil del paciente, se profundiza en la relación entre la respuesta inmune y los parámetros Hematológicos que se reflejan en el Hemograma, en línea con el objetivo del estudio.

Mediante el Análisis de Componentes principales se aborda el estudio de la estructura de la inmunidad en función de los parámetros Hematológicos.

En este análisis se han incluido los contadores: Leucocitos, Neutrófilos, Linfocitos, Monocitos, Eosinófilos, Basófilos, Trombocitos y Hemoglobina. Estos se incluyen habitualmente en los análisis de sangre. Se trata de buscar dimensiones subyacentes a este conjunto de contadores, para posteriormente relacionarlas con la respuesta inmune.

Previamente se observan posibles observaciones extremas y anomalías en los datos mediante los gráficos SPE y T^2-Hotelling que se muestran en la Figura 10. En ambos gráficos se pueden apreciar dos líneas que representan el percentil 95 y 99 de los datos, que cuanto más lejos esté una observación de estos umbrales, más anómala será la observación. En base a los resultados no se consideran observaciones atípicas ya que la posición relativa de aquellas más alejadas o difiere lo suficiente del resto de los datos en ninguno de los dos gráficos.

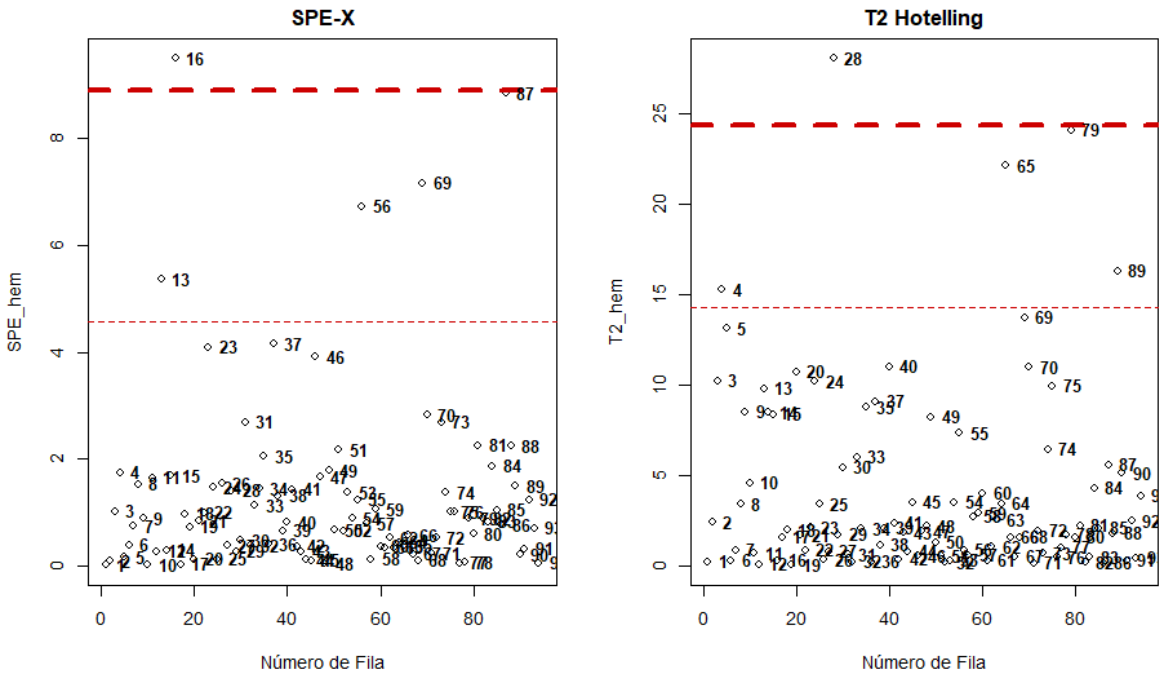

Figura 10. SPE Y T^2de Hotelling (Parámetros Hematológicos)

A continuación, para poder interpretar el PCA, en primer lugar, se observa la variabilidad explicada por cada una de las componentes. Para ello se debe tener en cuenta que se trabaja con un total de 8 variables y 94 observaciones, por tanto, por cada componente se establece el criterio de no sacar más de la mitad del número de componentes totales, es decir, 4.

El resultado está mostrado en el siguiente gráfico, donde el eje X representa el número de componentes y el eje Y, la varianza acumulada.

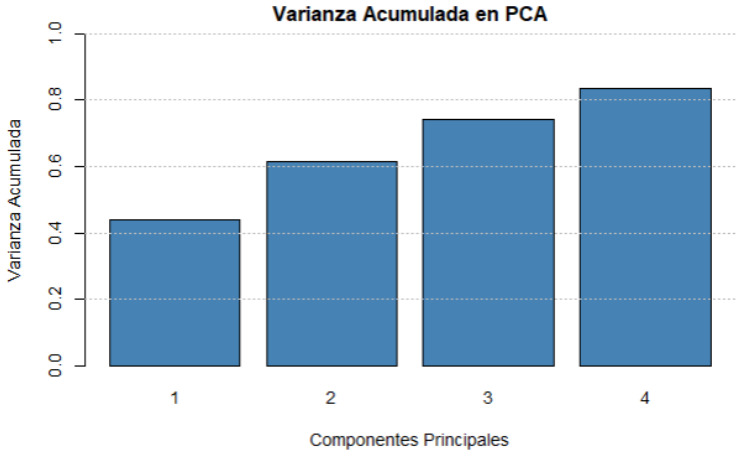

Figura 11. Variabilidad Explicada por las Componentes Principales. Parámetros Hematológicos

Al analizar la variabilidad explicada por las diferentes componentes, se puede observar que la primera componente captura el 34% de la variabilidad total de los datos, sin embargo, a medida que se aumenta

el número de componentes la variabilidad explicada aumenta, hasta conseguir explicar un 83% de los datos originales reduciendo la dimensionalidad hasta 4 componentes.

Por otro lado, es interesante observar un resumen de las distintas variables a lo largo de las componentes. Para ello, la Figura 12 muestra un gráfico de barras que proporciona la contribución de las distintas variables en cada una de las componentes.

En el gráfico se observa como algunas variables no están apenas explicas por ninguna de las dos primeras componentes, que son aquellas que más variabilidad recoge. Sin embargo, fijándonos en la variable *Hemoglobina* se puede apreciar como esta parece estar muy explicada por la tercera componente, lo que puede indicar un comportamiento diferenciado al resto de los parámetros Hematológicos. Asimismo, los *Linfocitos* también parecen mostrar una diferencia en la cuarta componente principal, a pesar de que estas no lleguen a explicar ni un 10% de los datos variabilidad total de los datos.

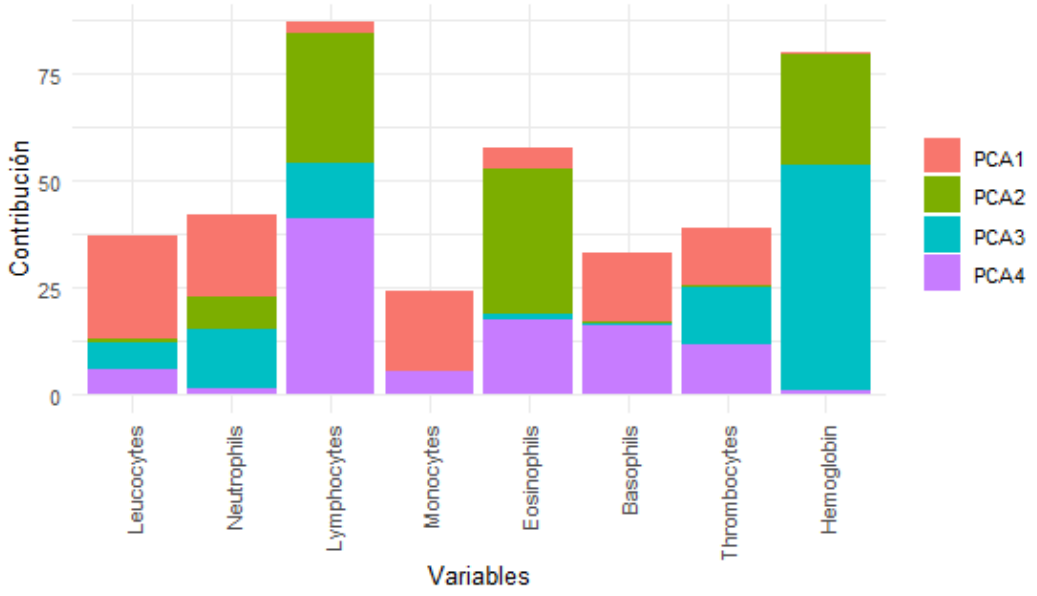

Figura 12. Resumen de los parámetros Hematológicos en el PCA

Para facilitar la interpretación de las componentes extraídas se muestran a continuación los siguientes gráficos de cargas factoriales (loadings). En el gráfico de la primera componente frente a la segunda destaca principalmente, la correlación positiva entre las variables coloreadas en rojo con mayor peso la primera componente; *Leucocitos*, *Neutrófilos*, *Trombocitos*, *Monocitos* y *Basófilos*. Por otro lado, se puede apreciar una ligera correlación entre los *Linfocitos* y los *Eosinófilos* y algo más ligera con la *Hemoglobina*, coloreados en verde, con un mayor peso a lo largo de la segunda componente.

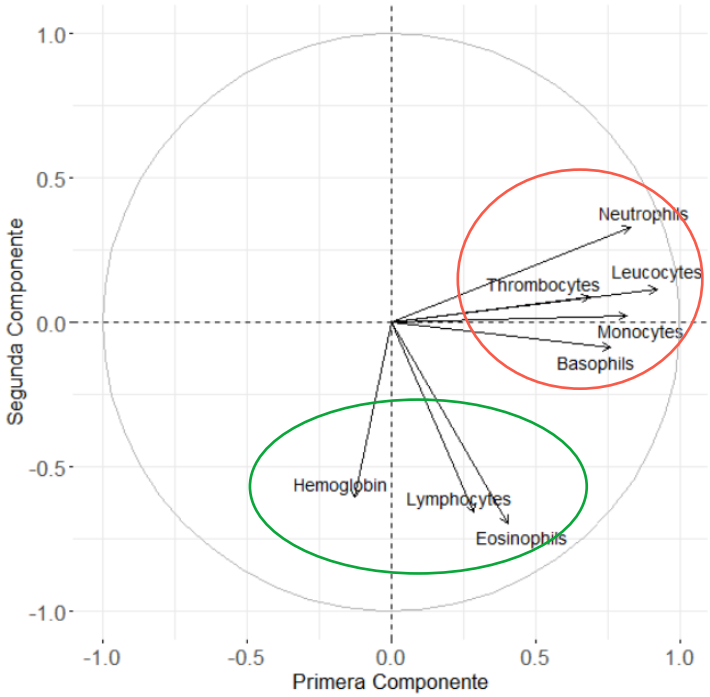

Figura 13. Gráfico de Loadings 1ªy 2ª Componente (Parámetros Hematológicos)

Sin embargo, si observamos los siguientes gráficos donde se encuentran los pesos de las variables en la tercera y cuarta componente, se muestra lo que se apreciaba en la Figura 12, como la *Hemoglobina* y los *Linfocitos* muestran un comportamiento diferenciado al resto de las variables.

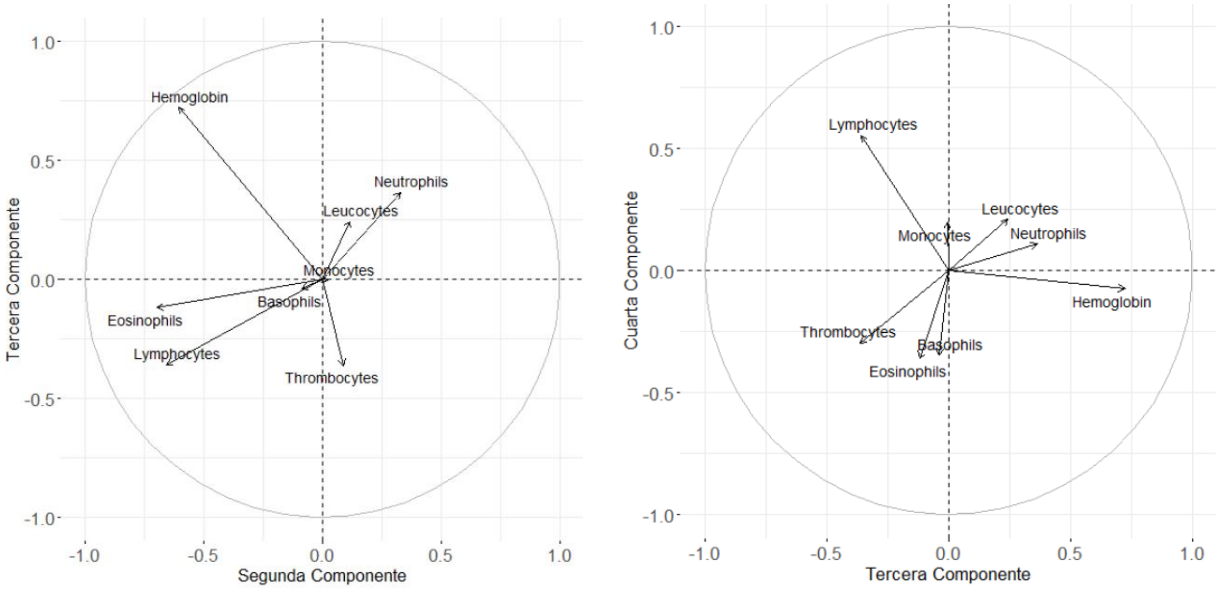

Figura 14. Gráfico de Loadings 2º, 3ªy 4ª Componente (Parámetros Hematológicos)

Asimismo, en la Figura 15 que muestra la correlación entre los distintos parámetros Hematológicos donde se puede apreciar lo indicado previamente el gráfico de las cargas factoriales (loadings) de la primera y segunda componente, es decir, la relación positiva de la *Hemoglobina*, los *Linfocitos* y los *Eosinófilos*. Por

otro lado, notar que la relación positiva más fuerte entre las células se encuentra entre los *Leucocitos* y los *Monocitos*.

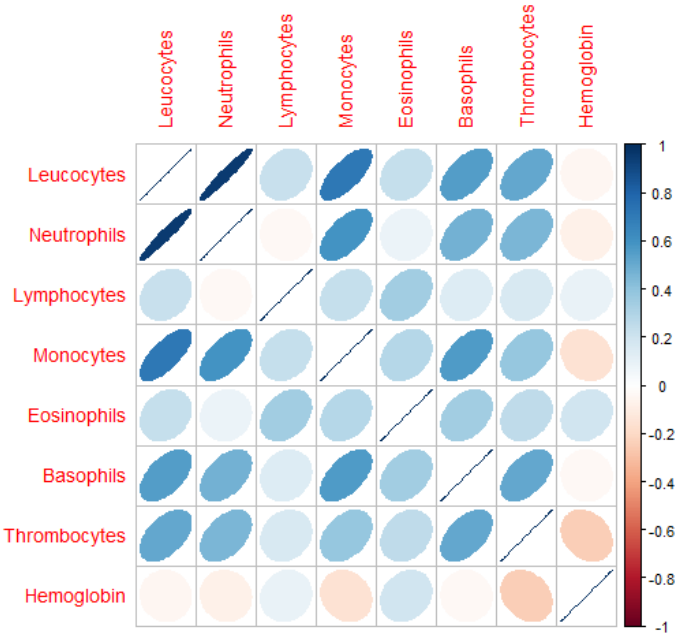

Figura 15. Correlación entre los parámetros Hematológicos

Los siguientes gráficos presentados representan los scores de la primera y segunda componente, los cuales se han coloreado en función de la dosis administrada, el tipo de cáncer y el tratamiento. En cada uno de ellos, se estudia la posible influencia de estas características.

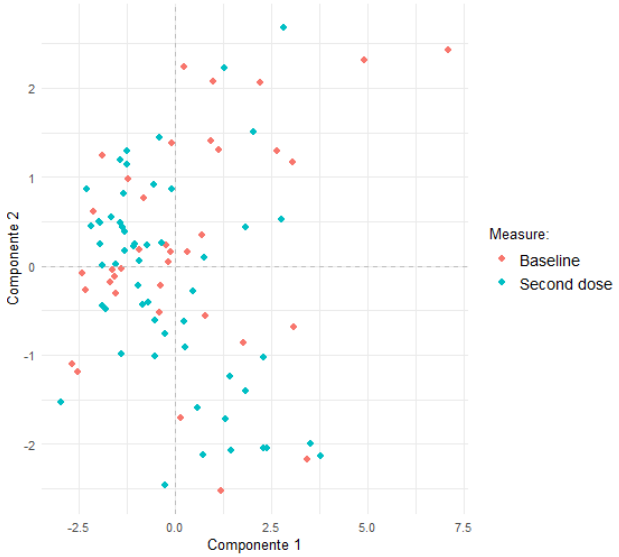

Figura 16. Gráfico de Scores en función de la dosis (Parámetros Hematológicos)

En el caso de la dosis, no parece influyente que el paciente haya recibido la segunda vacuna del SARS-CoV-2 o que no esté vacunado, ya que ninguno de los dos grupos parece tener mayor peso en un conjunto de variables.

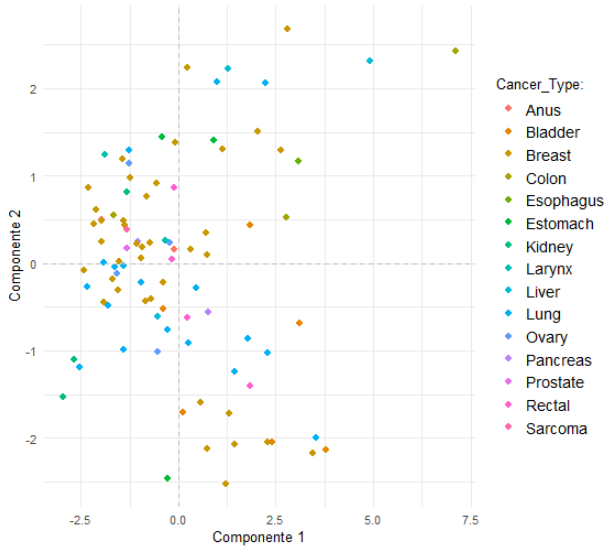

Figura 17. Gráfico de Scores en función del tipo de Cáncer (Parámetros Hematológicos)

Cuando se dibuja el mismo gráfico, pero en función al tipo de cáncer que tiene el paciente, tampoco parece influyente en ningún grupo de los parámetros Hematológicos.

Sin embargo, en el caso del tratamiento que están recibiendo los pacientes se puede apreciar que aquellos que reciben inmunoterapia parecen tener mayores valores de Hemoglobina, Linfocitos y Eosinófilos, recordando el peso de estos en la segunda componente, Figura 13.

Por otro lado, aquellos pacientes que reciben Quimioterapia tienen menores niveles en todo el conjunto de las células, posiblemente debido al nivel de agresividad de este tratamiento para el paciente.

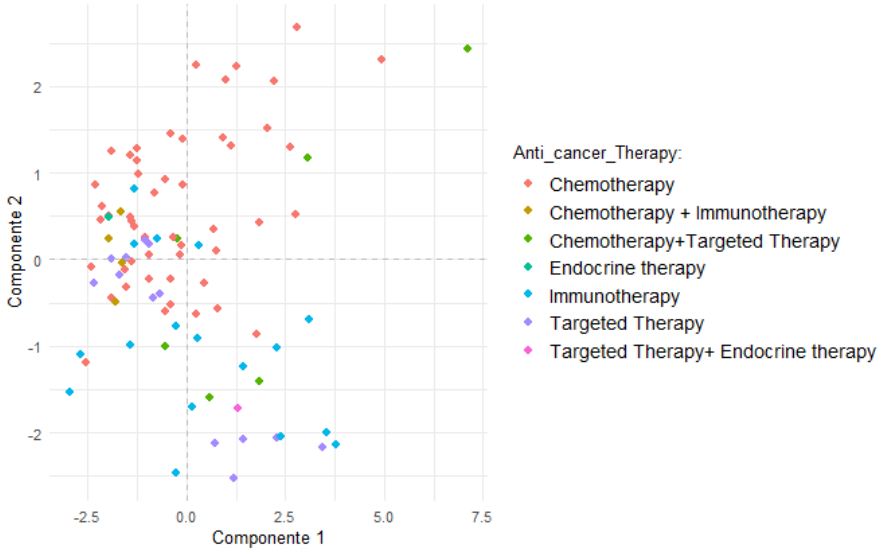

Figura 18. Gráfico de Scores en función del tratamiento aplicado (Parámetros Hematológicos)

4.1.3. Estudio de la Respuesta Celular

Para abordar el estudio de la estructura de la respuesta celular se emplea, en primer lugar, el **Análisis de componentes principales (PCA)**, ya que nos permite explorar los datos, encontrar patrones y relaciones entre las distintas variables.

Antes de sumergirnos en los detalles del PCA, es esencial explorar el impacto de las observaciones extremas y anomalías en los datos, ya que estas pueden distorsionar la estructura y afectar negativamente en los resultados de las técnicas empleadas. Para ello, se comienza aplicando el gráfico SPE que permite detectar posibles observaciones atípicos que rompen la estructura de correlación, y el gráfico T^2-Hotelling para detectar las observaciones extremas. Sin embargo, en los resultados que muestra la Figura 19 no se aprecia ninguna observación muy distante al resto como para considerarse anomalía.

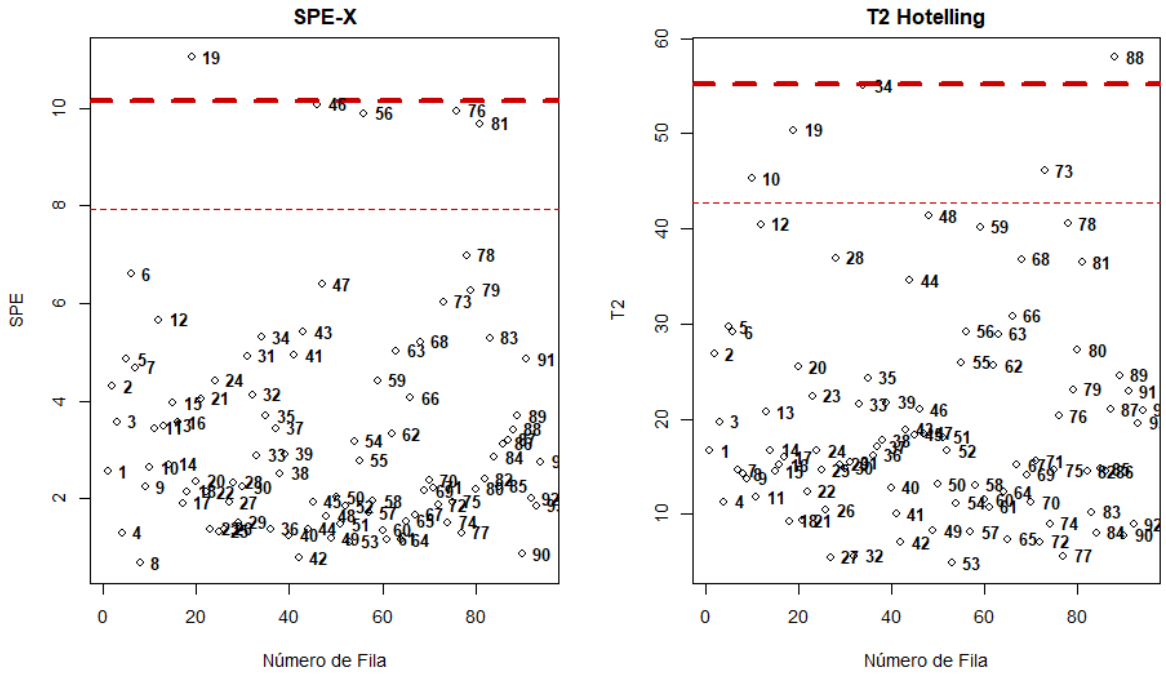

Figura 19. Gráfico SPE y T2 de Hotelling (Respuesta Celular)

Una vez observadas aquellas posibles observaciones extremas que podrían afectar negativamente en el estudio de los datos, se aplica PCA para reducir la dimensionalidad del conjunto de variables, formado por 67 variables.

En primer lugar, se examina la variabilidad explicada por las distintas componentes extraídas para decidir el número de componentes retenidas. La primera componente explica más que un 22% de la variabilidad total. Teniendo en cuenta que se están analizando un total de 67 variables no es sorprendente notar que serán necesarias más componentes para poder explicar bien los datos. Sin embargo, para este estudio se considera suficientes 4 componentes ya que permite explicar más de un 50% de la variabilidad total de los datos originales.

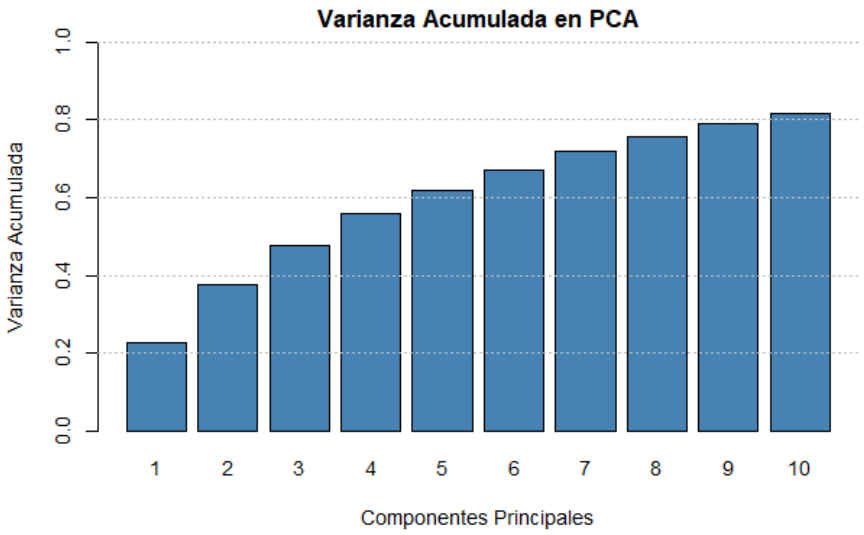

Figura 20. Variabilidad Explicada por las Componentes Principales. Respuesta Celular

A continuación, para estudiar la influencia de las distintas variables a lo largo de las componentes se muestra la contribución de estas a lo largo de las componentes. Esto es útil para comprender qué variables tienen mayor influencia en la formación de las componentes y cómo contribuyen a la variabilidad total del conjunto de datos.

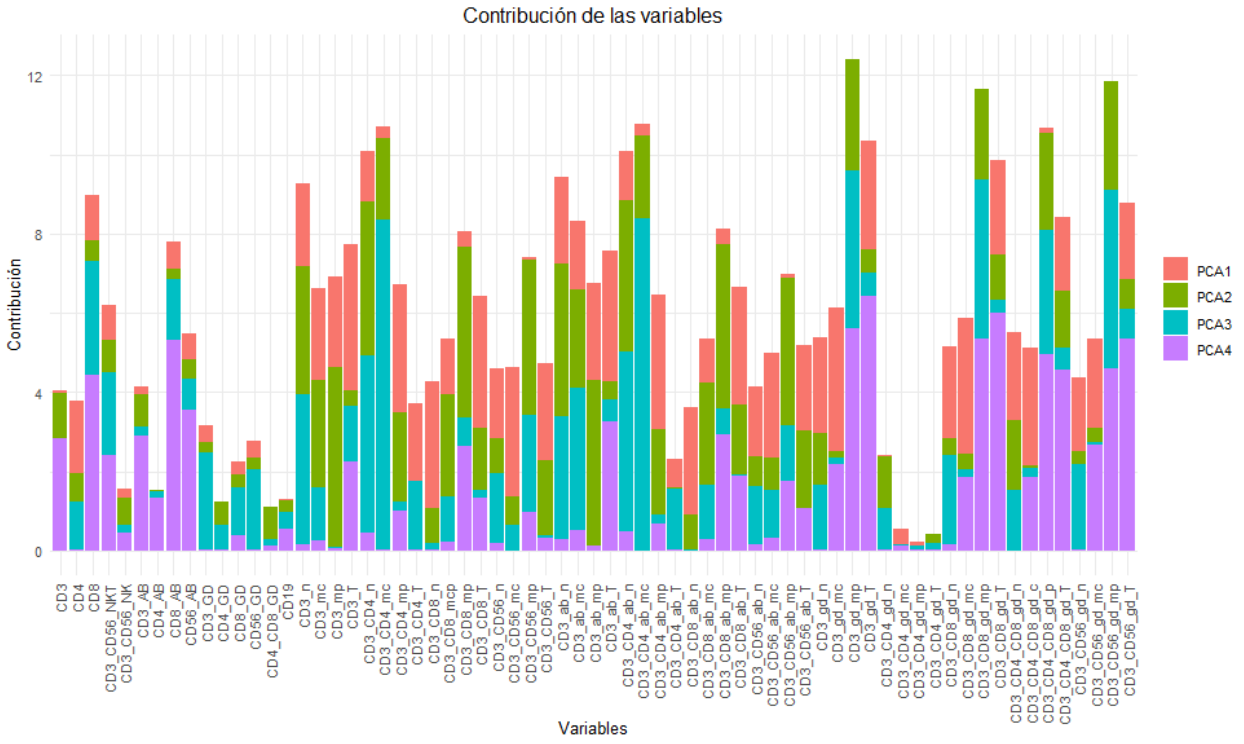

Figura 21. Resumen de las variables de Respuesta Celular en el PCA

El gráfico muestra como algunas variables como *CD3_CD56_NKT, CD19, CD3, CD4_GD_MC*… tienen muy caja contribución en todas las variables, mostrando su poca significancia en las componentes con mayor peso.

Para poder interpretar de manera más clara lo que sugieren la primera y la segunda componente y la contribución de las variables, ya que el gráfico anterior dificulta su interpretación, emplea tanto el gráfico de puntuaciones factoriales o scores que muestra las puntuaciones factoriales como el gráfico de loadings o cargas factoriales, donde podremos ver la relación entre las distintas variables y observaciones.

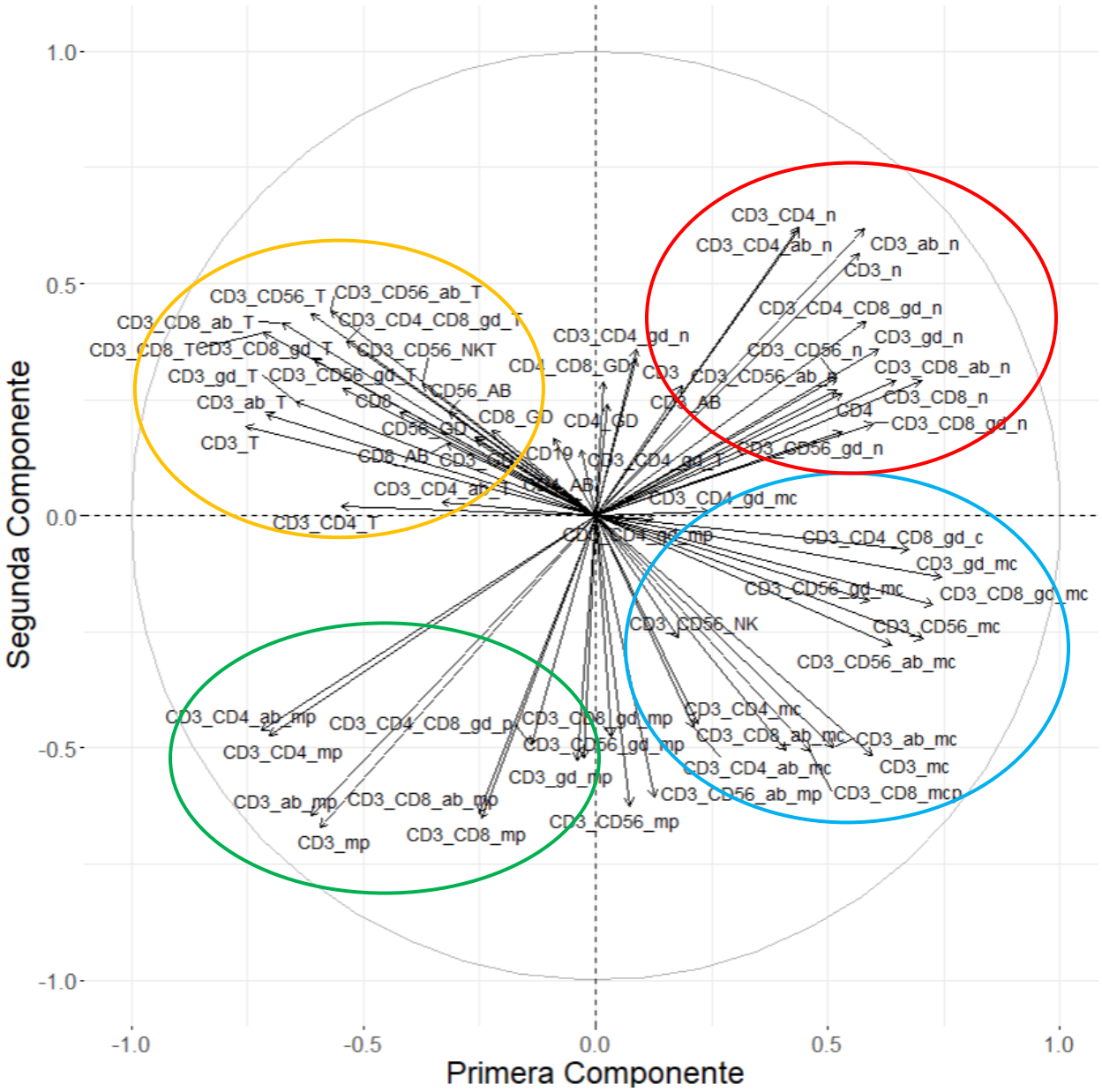

Figura 22. Gráfico de Loadings de la 1ª vs la 2ª Componente (R. Celular)

El gráfico anterior muestra las cargas de las variables a lo largo de las dos primeras componentes principales, es decir, está mostrando la importancia de cada variable en cada componente. Asimismo, este gráfico ayuda a identificar y encontrar patrones de correlación, tanto positiva como negativa, entre las variables originales. De esta manera, las variables que tienen vectores que apuntan en la misma dirección tienen una correlación positiva, mientras que las variables que tienen vectores que apuntan en direcciones opuestas tienen una correlación negativa.

El gráfico, muestra la división por grupos donde se observa en color azul, aquellas variables que representan la *memoria central*, en color verde las variables con *memoria periférica*, en rojo las variables *Naive* y por último en Naranja las variables que representan *TEMRA*.

Por esto mismo, se aprecia como cada grupo de variables está positivamente correlacionadas entre ellas y como en el caso de las variables con "memoria central" parecen presentar una correlación negativa con las variables que contienen *TEMRA*. Asimismo, las variables *Naive* parecen presentar también una correlación positiva entre ellas, pero negativa con las variables de "memoria periférica".

Para comprobar que estamos en lo cierto, se realiza una matriz de correlación por grupos de variables. De esta manera, el primer gráfico muestra tal y como esperábamos la correlación entre las variables memoria central y *Temra* y el segundo gráfico la correlación entre las variables memoria periférica y *Naive*.

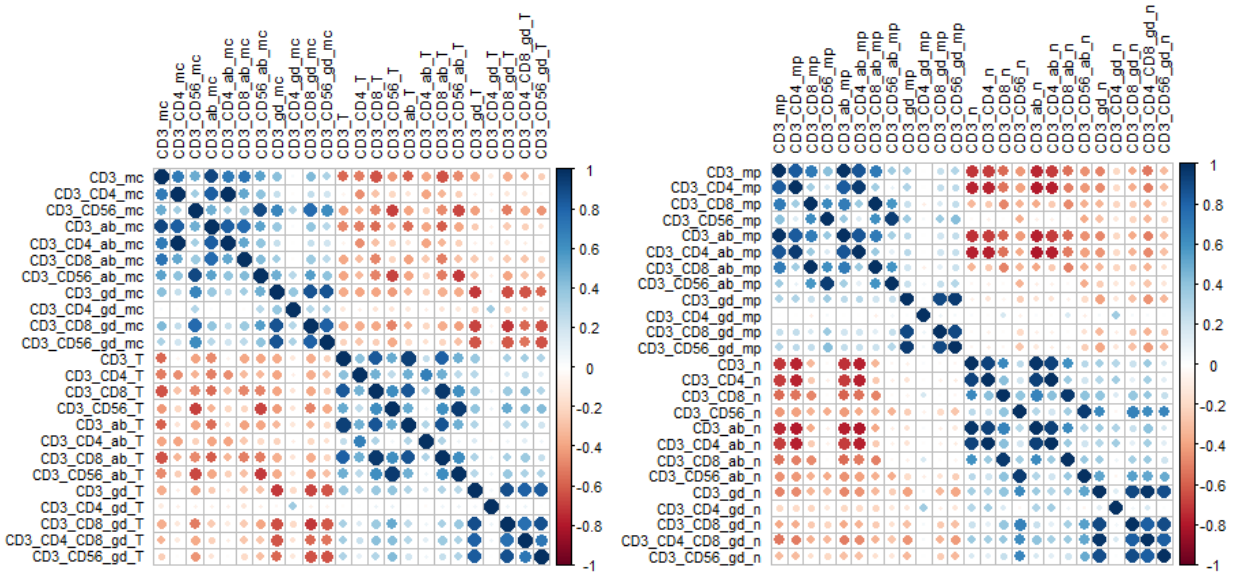

Figura 23. Matriz de correlación entre "Memoria Central" y "Naive""Memoria Periférica" y "Temra"

Además, se han generado gráficos que muestran las puntuaciones factoriales de la primera y segunda componente, nuevamente coloreados según la dosis administrada, el tipo de cáncer y el tratamiento correspondiente. Tal y como ocurría con la inmunidad de los parámetros Hematológicos, la respuesta celular tampoco detecta ninguna diferencia significativa con relación a estas características.

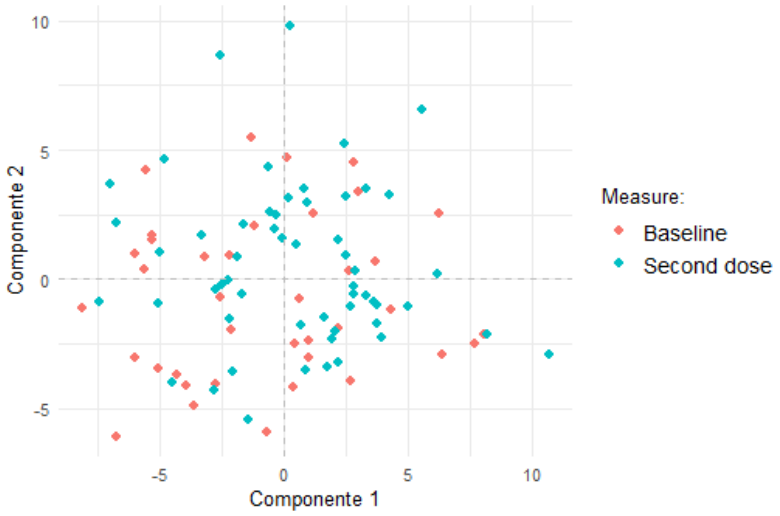

Figura 24. Gráfico de Scores en función de la dosis (Respuesta Celular)

En el caso de la dosis aplicada al paciente, no se aprecia ninguna diferencia en cuanto a un grupo de respuesta celular ya que los puntos de colores se ven dispersos entre todos ellos sin tener un mayor peso en un conjunto de variables.

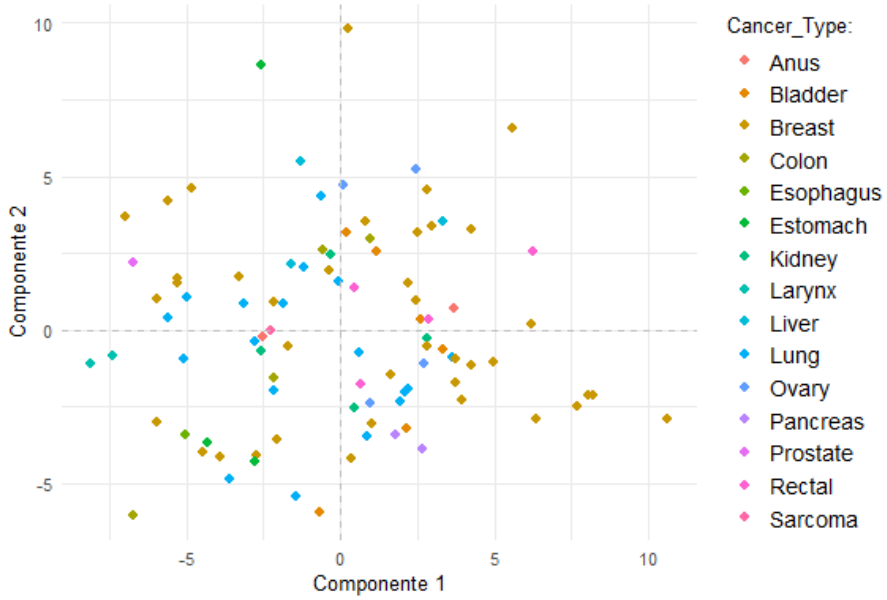

Figura 25. Gráfico de Scores en función del tipo de Cáncer (Respuesta Celular)

En relación con el tipo de cáncer que afecta al paciente, al igual que en el gráfico previo, no se observa una mayor puntuación en ningún grupo de respuesta celular. En otras palabras, no se evidencia una asociación significativa entre el tipo de cáncer y la respuesta celular.

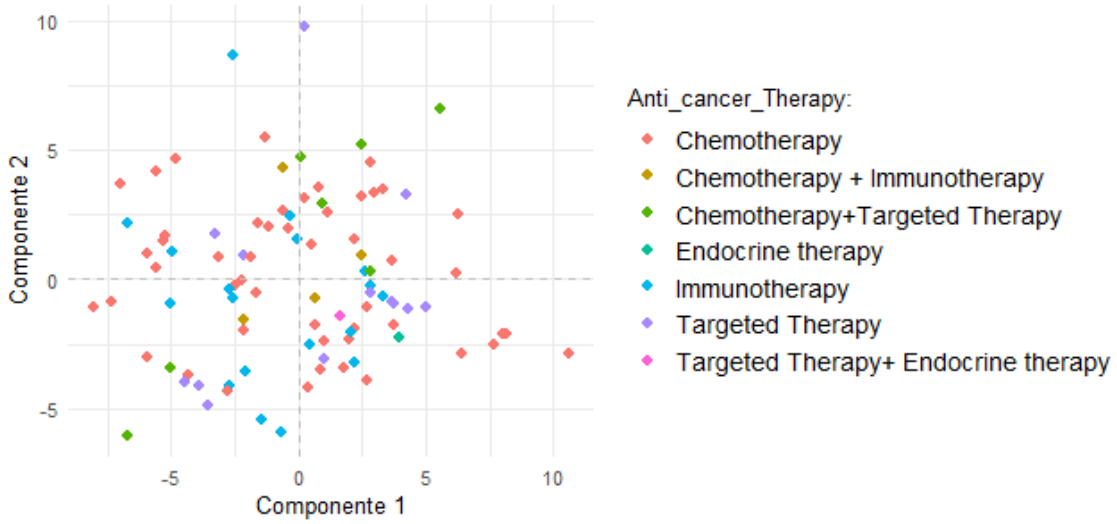

Figura 26. Gráfico de Scores en función del tratamiento aplicado (Respuesta Celular)

Finalmente, al igual que se pudo observar en las características previamente analizadas, no es posible afirmar que el tratamiento que el paciente esté recibiendo tenga un impacto en el valor de la respuesta celular.

Con el propósito de mejorar el análisis de los datos, se aplica a continuación el **Análisis de Componentes Principales Dispersos (SPCA)** debido a sus múltiples ventajas. El SPCA se enfoca en identificar las componentes principales más relevantes en conjuntos de datos dispersos, lo que permite lidiar de forma adecuada con la alta dimensionalidad y la escasez de observaciones. Al aplicar el SPCA, se espera obtener una representación más compacta y significativa de los datos, lo cual resultará en una extracción de información más precisa y útil para los análisis posteriores.

Para poder comenzar aplicando SPCA con las variables de respuesta celular es necesario seleccionar el número de componentes a evaluar mediante la variabilidad explicada por las componentes. En el PCA de respuesta celular, la Figura 20 muestra la relación entre el número de componentes y la cantidad de variabilidad explicada por cada una de ellas. Se ha observado que utilizando 4 componentes principales es posible explicar aproximadamente el 50% de la variabilidad total de los datos. Por ello, se procede a aplicar un SPCA con 4 componentes.

A continuación, en el proceso de selección del número de variables en cada componente, se utiliza la validación cruzada. Se configura una cuadrícula de valores previamente, donde la densidad de los valores es delgada al principio y se vuelve más gruesa a medida que aumenta el número de variables.

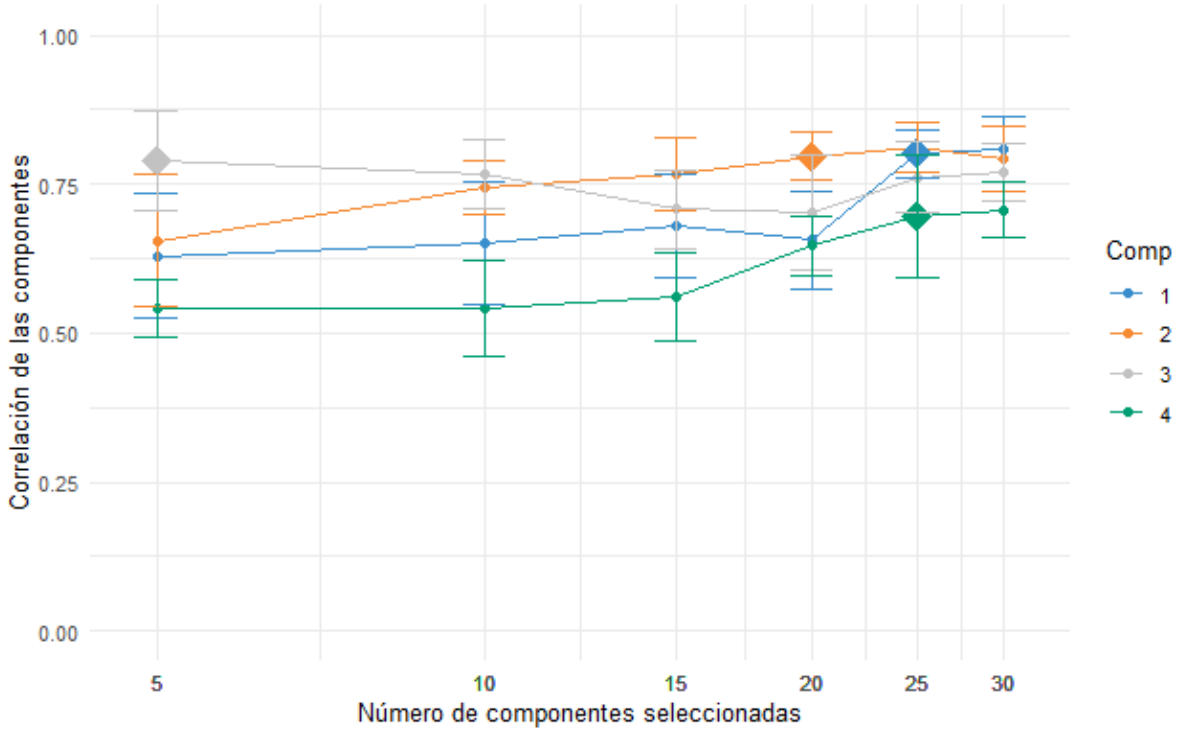

Figura 27. Ajuste del número de variables a seleccionar con SPCA

El gráfico anterior muestra, por un lado, en el eje Y la correlación media entre las componentes predichas y las componentes basadas en validación cruzada en cada una de las componentes y por otro lado, en el eje X el número de variables a seleccionar. Es decir, el gráfico está mostrando el número óptimo de variables a seleccionar para maximizar la correlación media de cada componente, siendo 25 variables en la primera componente, 20 en la segunda, 5 en la tercera y 30 variables en la segunda.

Para poder ajustar ahora el modelo SPCA se consideran en lugar de 4 componentes como se comentaba anteriormente, únicamente 3, ya que el objetivo es encontrar la respuesta celular que muestra la variación en los datos.

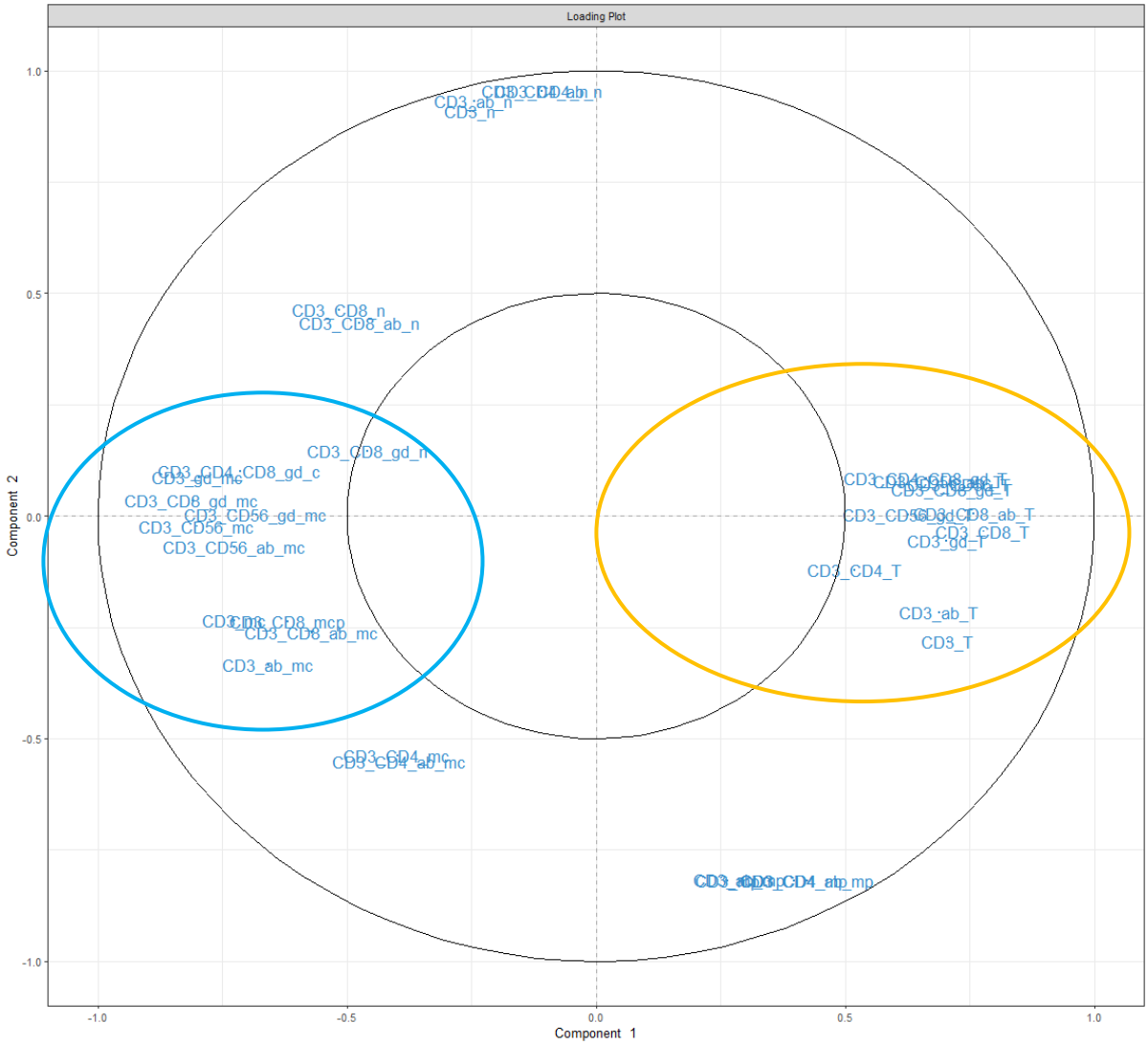

Figura 28. Gráfico de Loadings de la 1ª vs la 2ª Componente SPCA (R. Celular)

Una vez ajustado el modelo SPCA, el gráfico de Loadings de la primera frente a la segunda componente muestra aquellas variables seleccionadas como las más relevantes. A diferencia de la Figura 22 del PCA, que mostraba todas las variables, se aprecia en este gráfico como las variables más relevantes son aquellas de Memoria Periférica y de Temra, mientras que las variables relacionadas con la Memoria Central y la mayoría de las variables Naive pasan a un segundo plano. Es importante destacar que tanto las células de Memoria Periférica y Temra como las células de Memoria Central y Naive están negativamente correlacionados entre ellas, como se mostraba en la Figura 23, algo que también se puede apreciar en el gráfico SPCA. Por otro lado, a lo largo de la segunda componente se muestran como significativas algunas células Naive.

Ayudando a comprender este resultado, se muestran los pesos de carga de las variables la primera y segunda componente, ordenadas desde la menos importante, arriba, hasta la más importante, abajo.

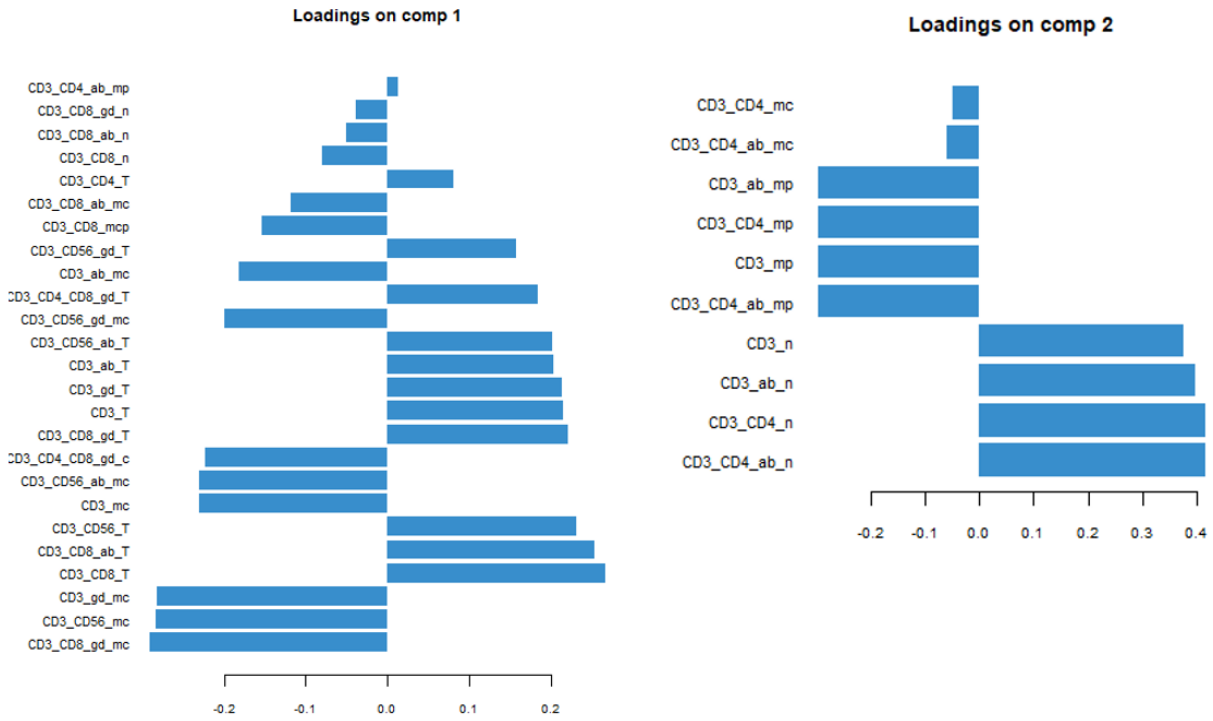

Figura 29. Peso de las variables SPCA en la componente 1 y la componente 2

Tal y como se mencionada, las variables que más peso tienen en la primera componente son las variables de Memoria Central y Temra. Asimismo, en la segunda componente las variables con mayor peso son CD3 y CD4 Naive, además de alguna de Memoria Periférica.

Por otro lado, se han generado los gráficos de las puntuaciones factoriales (scores) de la primera y segunda componente, nuevamente coloreados según la dosis administrada, el tipo de cáncer y el tratamiento correspondiente. Tal y como era de esperar y como mostraban los gráficos del PCA (Figura 24, Figura 25 y Figura 26), no se detecta ninguna diferencia significativa con relación a estas características.

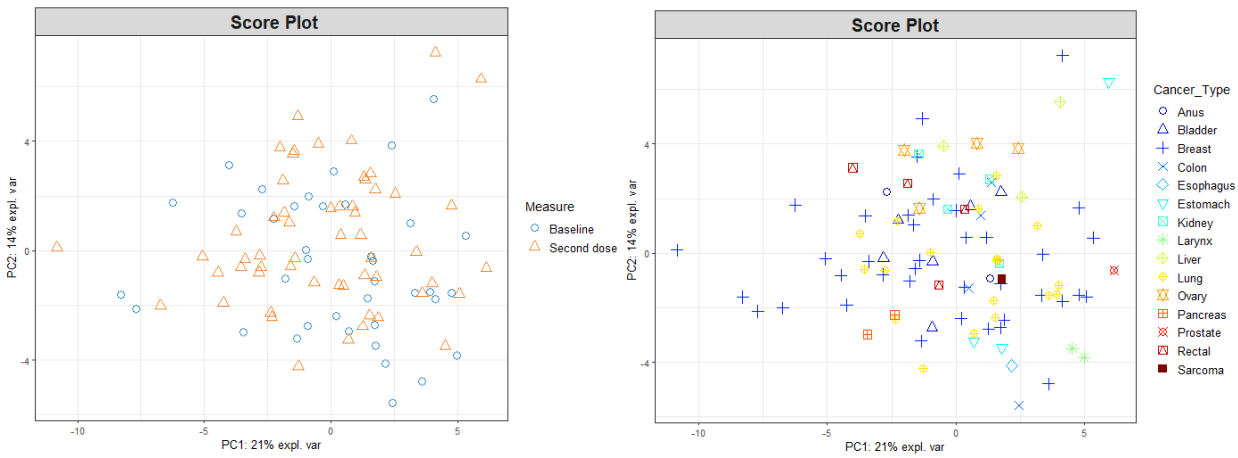

Figura 30. Gráfico de Scores en función de la dosis y del tipo de cáncer SPCA (Respuesta Celular)

Los gráficos anteriores muestran que no parece haber ninguna diferencia significativa, tanto en la dosis como en el tipo de cáncer que padezca el paciente a la hora de desarrollar inmunidad en los parámetros Hematológicos. Asimismo, en el siguiente gráfico que muestra el tratamiento que recibe el paciente, igualmente comentar que ninguno de ellos podría considerarse influyente, ya que no hay un tratamiento con mayor peso en ningún grupo de respuesta celular.

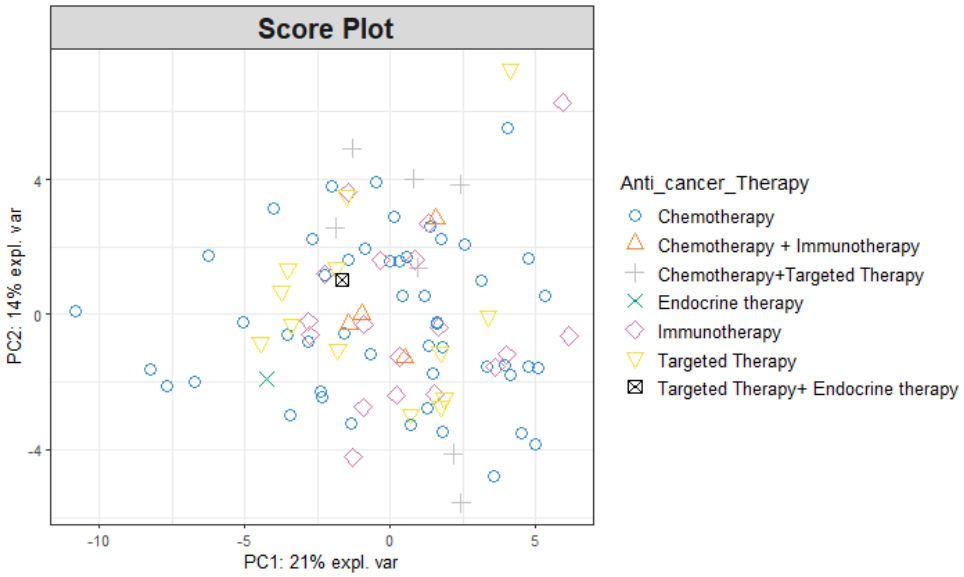

Figura 31. Gráfico de Scores en función del tratamiento aplicado SPCA (Respuesta Celular)

Por otro lado, se muestra el gráfico de las cargas factoriales de la segunda frente a la tercera componente, en la que se habían seleccionado 5 variables. En un primer instante, se había considerado una tercera componente ya que resultaba llamativo que se seleccionaran 5 variables en esta componente. Sin embargo, no parece revelar una información muy significativa más que la importancia de la agrupación de las variables CD3_gd_mp, CD3_CD56_gd_mp, CD3_CD8_gd_mp y CD3_CD4_CD8_gd_p con una correlación negativa con la variable CD3_gd_T. No obstante, el gráfico de los pesos de las variables en la tercera componente muestra como esta última variable CD3_gd_T apenas tiene peso en la tercera componente.

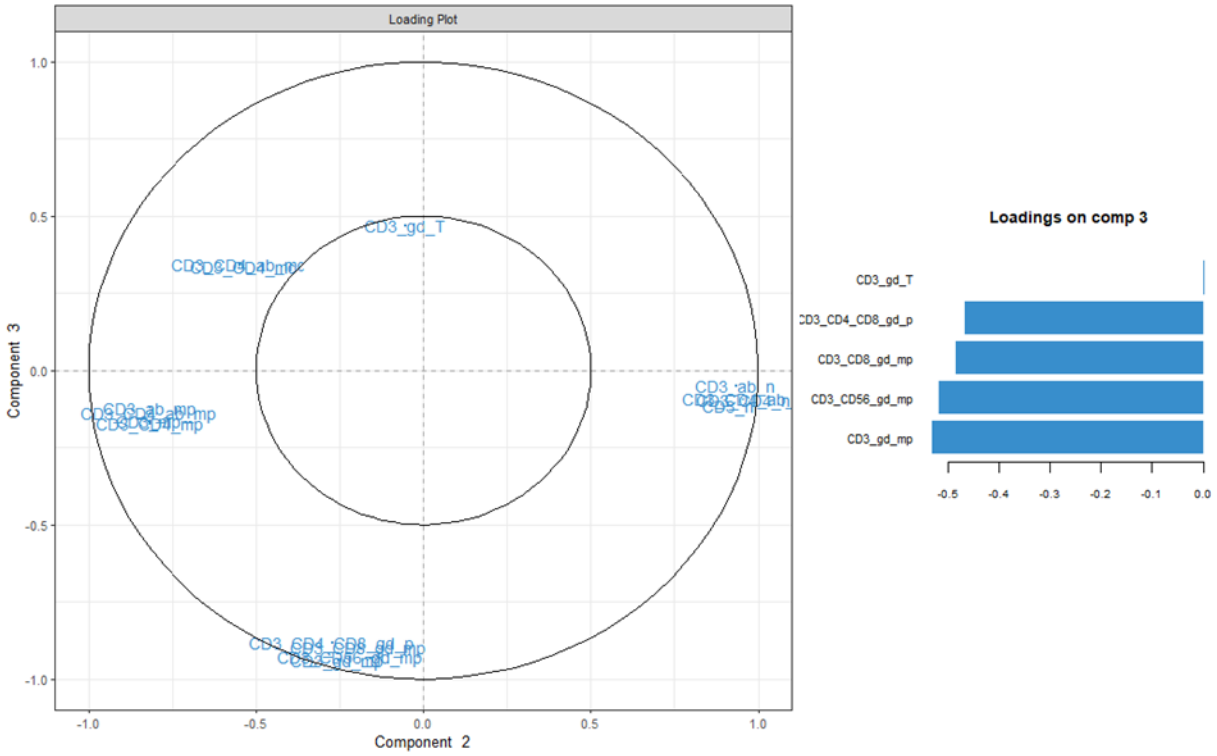

Figura 32. Gráfico de Loadings de la 2ª vs la 3ª Componente SPCA (R. Celular) (izq) Peso de las variables SPCA en la 3ª componente (drch)

4.2. Relación entre medidas de respuesta inmunitaria

En este apartado, se llevará a cabo un análisis para investigar la posible relación entre los anticuerpos generados contra el virus SARS-CoV-2 y el resto de la respuesta inmune, tanto en los parámetros Hematológicos como en la respuesta celular.

Inicialmente, se plantea la hipótesis de que los anticuerpos del SARS-CoV-2 podrían generar memoria en el resto de la respuesta inmune. Sin embargo, hasta el momento no se ha encontrado evidencia científica que respalde esta afirmación. Por lo tanto, se procederá a realizar un estudio para evaluar la relación entre estas variables, enfocándose únicamente en los pacientes que hayan generado anticuerpos, lo que implica reducir la muestra de estudio.

El análisis se dividirá en dos bloques principales. En el primer bloque se examinará la relación entre los anticuerpos del virus SARS-CoV-2 y los parámetros Hematológicos utilizando la Regresión de Mínimos Cuadrados Parciales (PLS) que permitirá explorar y modelar la posible asociación entre los niveles de anticuerpos y las diferentes medidas de los parámetros Hematológicos.

En el segundo bloque, se estudiará la relación entre los anticuerpos del virus SARS-CoV-2 y la respuesta celular utilizando métodos de regresión Lasso, regresión Ridge y Regresión de Mínimos Cuadrados

Parciales Dispersos (sPLS), que permitirán encontrar aquellas variables más significativas en la generación de anticuerpos.

4.2.1. Relación entre los Anticuerpos al virus SARS-CoV-2 y los Parámetros Hematológicos

Para encontrar la relación de los anticuerpos SARS-CoV-2 con los parámetros Hematológicos se procede a ajustar un modelo de Regresión de Mínimos Cuadrados Parciales (PLS), donde la variable respuesta serán los anticuerpos y las X todos los parámetros Hematológicos.

Para ajustar el modelo, en primer lugar, se crea un primer modelo de prueba con todas las componentes y se aplica validación cruzada con 10 pliegues para obtener el valor Q^2 del mismo, ya que mide la capacidad predictiva del modelo.

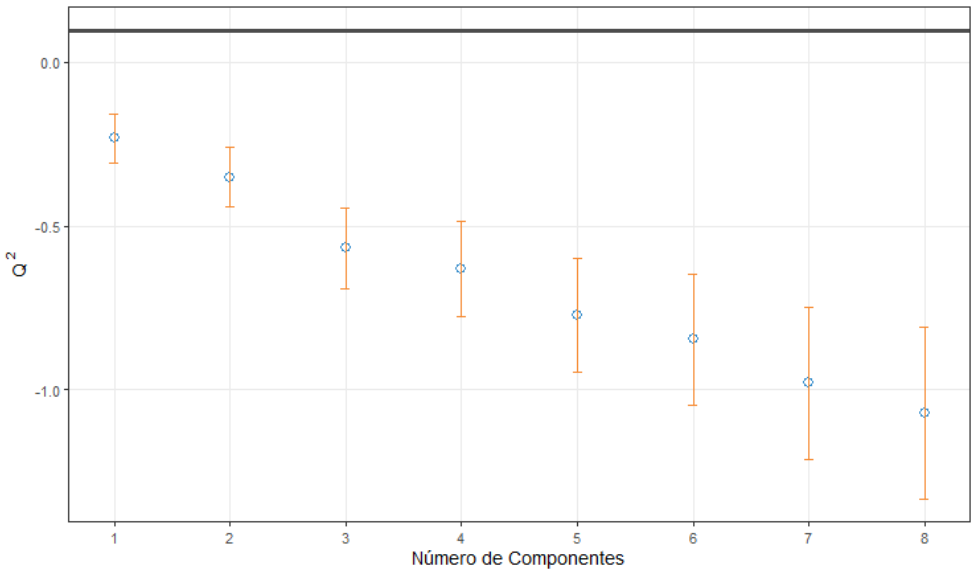

Figura 33. Selección del Nº de componetes PLS con el criterio Q^2

En el gráfico anterior se muestra en el eje de las X el número óptimo de componentes que se debe añadir al modelo teniendo en cuenta el valor Q^2. Por otro lado, la línea horizontal que se observa en la parte superior con valor 0,0975 indica el umbral en el que, por debajo de este, añadir una componente no será beneficioso para el modelo.

En este caso, el resultado indica que no se debería aplicar ninguna componente principal para ajustar el modelo, es decir, que no se debería crear modelo. Esto puede estar avisando de la poca relación de los anticuerpos del COVID con el resto de los parámetros Hematológicos. A pesar de este resultado, ya que el objetivo es profundizar más en esta relación, se decide aplicar el modelo con 2 componentes y analizar el resultado.

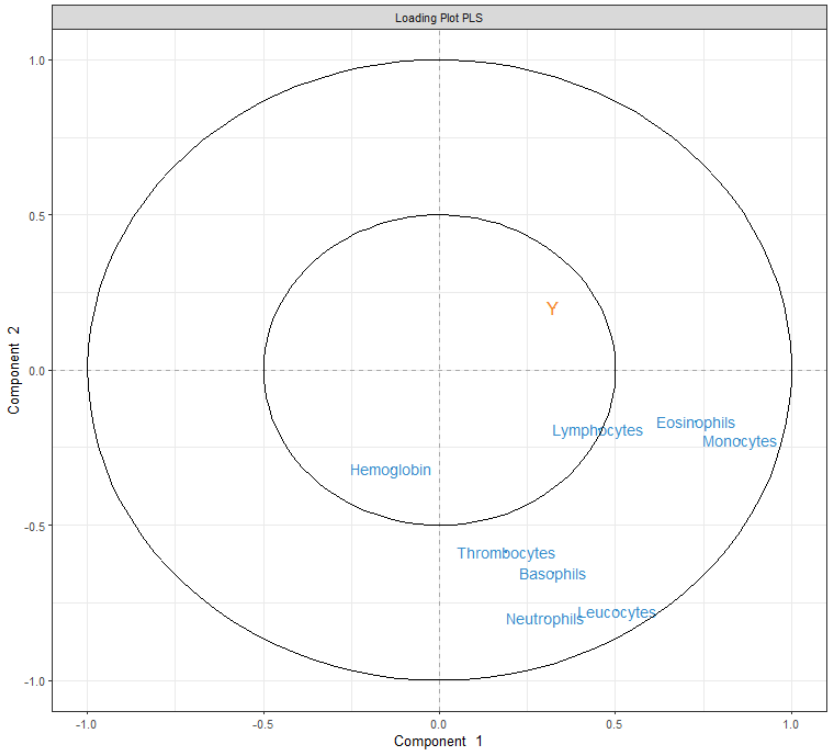

Figura 34. Gráfico de Weigthings PLS de la 1º y 2ª Componente

En primer lugar, tras ajustar el modelo con dos componentes principales se encuentra el gráfico de Weigthings. Este gráfico muestra el poco peso de los anticuerpos del COVID19 en ambas componentes principales, así como la poca relación con el resto de las células.

Por otro lado, es destacable comparar este resultado con el obtenido en el PCA de los parámetros Hematológicos donde se encontraban 2 agrupaciones de variables, Figura 13. En el PCA veíamos la relación entre la Hemoglobina, los Linfocitos y los Eosinófilos y por otro lado el resto de las células, sin embargo, ahora parece que la Hemoglobina no está tan relacionada con los Linfocitos y los Eosinófilos. La relación ahora parece encontrarse entre estos últimos, Linfocitos y Eosinófilos, con los Monocitos.

Por último, el gráfico que se encuentra a continuación muestra la proyección de los scores de los parámetros Hematológicos del Hemograma, X, como de los anticuerpos Y coloreados en función del tipo de cáncer. En ambos casos no parece haber un peso significativo del tipo de cáncer en función de los parámetros Hematológicos ni en función de la cantidad de anticuerpos.

Por tanto, teniendo en cuenta la capacidad explicativa del modelo con el criterio Q^2 que mostraba la Figura 33 y el resultado observado tanto en el gráfico de Weigthings como en el de scores, no parece que los anticuerpos a priori tengan una relación evidente con los parámetros Hematológicos.

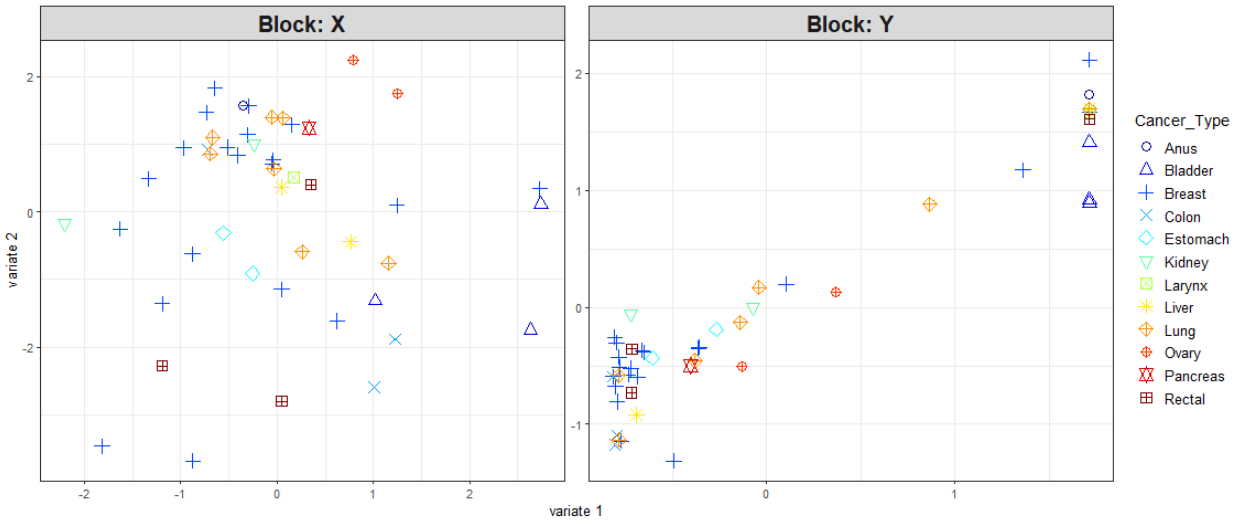

Figura 35. Score Plot PLS de los parámetros Hematológicos

4.2.2. Relación entre los Anticuerpos al virus SARS-CoV-2 y la Respuesta Celular

Para explorar la relación entre los anticuerpos del SARS-CoV-2 con la Respuesta Celular se emplean distintos métodos dispersos que ayudan a identificar aquellas variables más relevantes del modelo. Esto se aplica porque se trabaja con un número alto de variables predictoras (Respuesta Celular) frente a un número más pequeño de observaciones, ya que se trabaja únicamente con aquellos pacientes que han creado anticuerpos frente al SARS-CoV-2.

Los primeros métodos dispersos que se aplican son la regresión **Ridge** y la **Lasso**. Estos aplican una penalización a los coeficientes de las variables que ayudan a seleccionar aquellas variables más relevantes en la relación con los anticuerpos.

En primer lugar, se aplica un modelo que combina tanto la regularización ℓ1 (Lasso) como la ℓ2 (Ridge), lo que permite la selección de características y la reducción de la varianza del modelo.

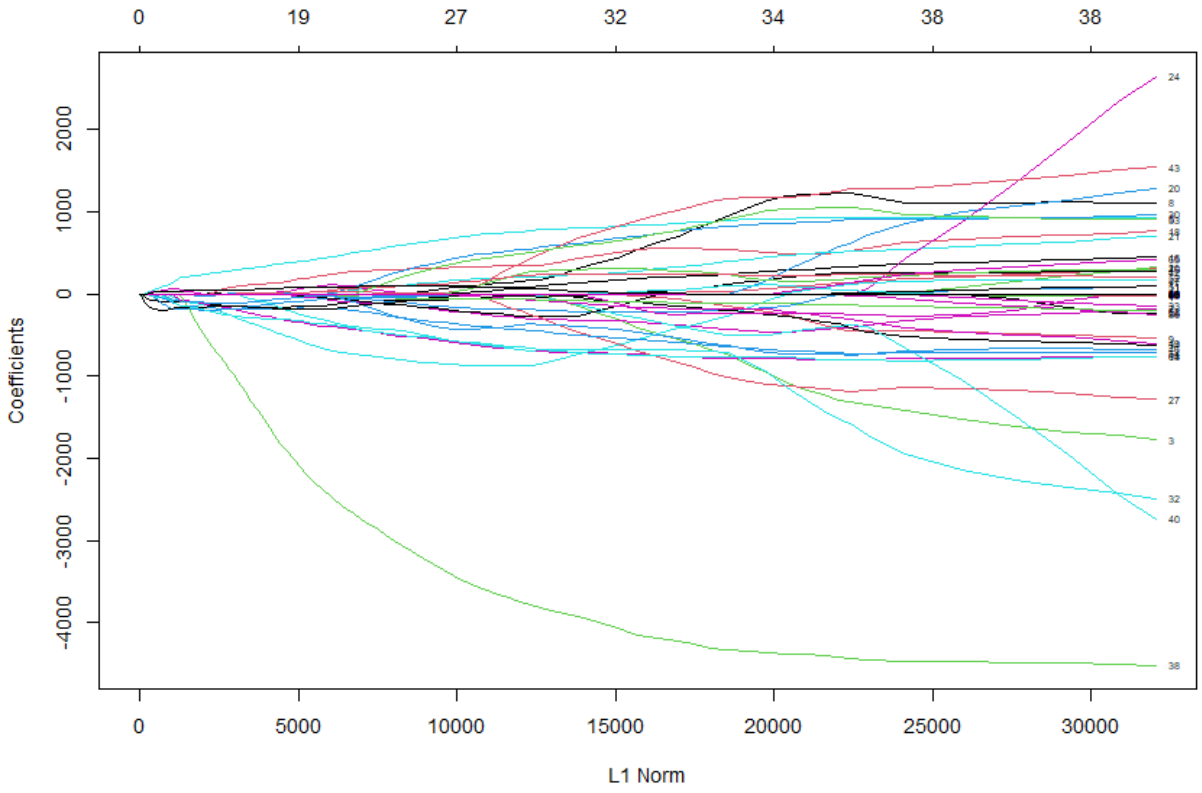

Figura 36. Gráfico de Coeficientes frente a la norma ℓ1

El gráfico se observan las variables en cada una de las curvas dibujadas de distintos colores. Este gráfico muestra la trayectoria de los coeficientes de las variables frente a la norma ℓ1 de todo el vector de coeficiente a medida que λ varía, y el eje superior muestra el número de coeficientes distintos de cero. En cuento al resultado, se observa como la variable 38 se encuentra algo alejada al resto, indicando un comportamiento algo distinguido en lo datos. Esto se debe a la variable *CD3 CD4 ab TEMRA* la cual presenta una variabilidad muy pequeña de su respuesta celular.

Para visualizar los valores del gráfico, a continuación, se muestra parte de los resultados:

Tabla 4. Coeficientes en función de la norma ℓ1

Df	%Dev	Lambda
0	0.00	5162.00
1	0.99	4927.00
2	1.90	4703.00
2	2.73	4490.00
2	3.49	4286.00
3	4.71	4091.00
4	6.43	3905.00
4	8.15	3727.00
3	9.72	3558.00

Esta tabla muestra de izquierda a derecha el número de coeficientes distintos de cero, el porcentaje de desviación nula explicada y el valor de λ.

A continuación, para comenzar a ajustar los modelos, empezando por la regresión **Ridge,** se debe seleccionar el primer lugar el número óptimo de λ. Para ello se aplica validación cruzada con 20 capas y se observa el resultado.

El siguiente gráfico muestra la curva de validación cruzada en color rojo junto a las curvas de desviación estándar a lo largo de la secuencia de λ además del error MSE cometido. Asimismo, muestra dos valores especiales a lo largo de la secuencia dibujados con puntos verticales. Empezando por la izquierda se encuentra el valor de λ que da el error de validación cruzada mínimo, mientras que el valor que se encuentra a la derecha es el valor de λ que da el modelo más regularizado. Estos valores de λ son 130875,4 para el error de validación cruzada mínimo y 5161997 el que proporciona el error más regularizado. Estos valores de λ son muy altos lo que indica que se ajustaría un modelo RIDGE con un error muy grande, dando a entender que los anticuerpos del SARS-CoV-2 no parecen tener una relación aparente con la respuesta celular.

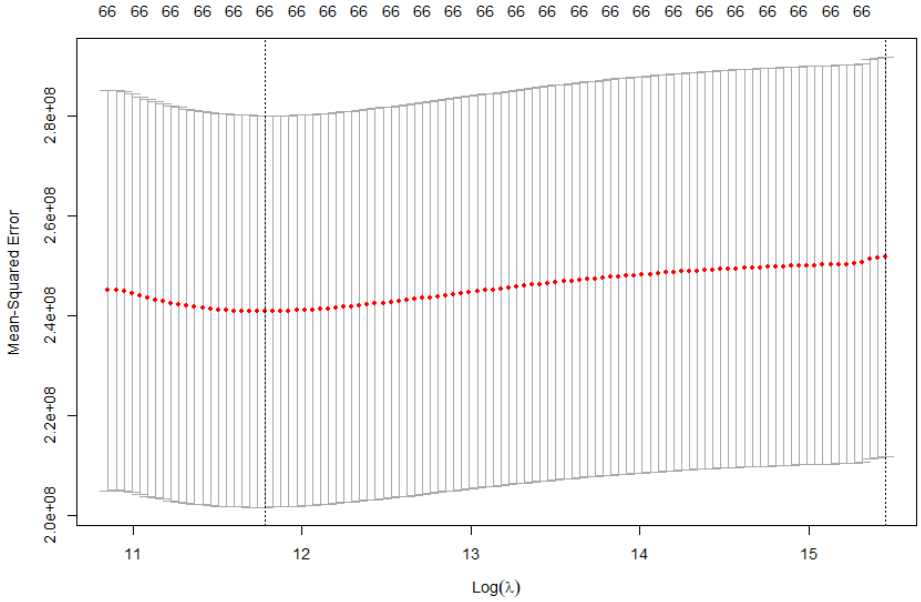

Figura 37. Gráfico de validación cruzada para la búsqueda de λ en la Regresión Ridge

Cuando se trata de ajustar el modelo Ridge con el λ "óptimo" que nos proporciona la validación cruzada no se obtiene ninguna variable significativa para la predicción de los anticuerpos del COVID, esto se debe a que aparentemente no parecen tener una relación evidente ya que el error que proporciona este es muy alto.

Por otro lado, para ajustar la regresión **Lasso**, que, al contrario del Ridge, fuerza a algunos coeficientes a cero mediante la penalización aplicada ayudando a una selección de variables, se aplica nuevamente validación cruzada con 20 pliegues para seleccionar el valor de λ.

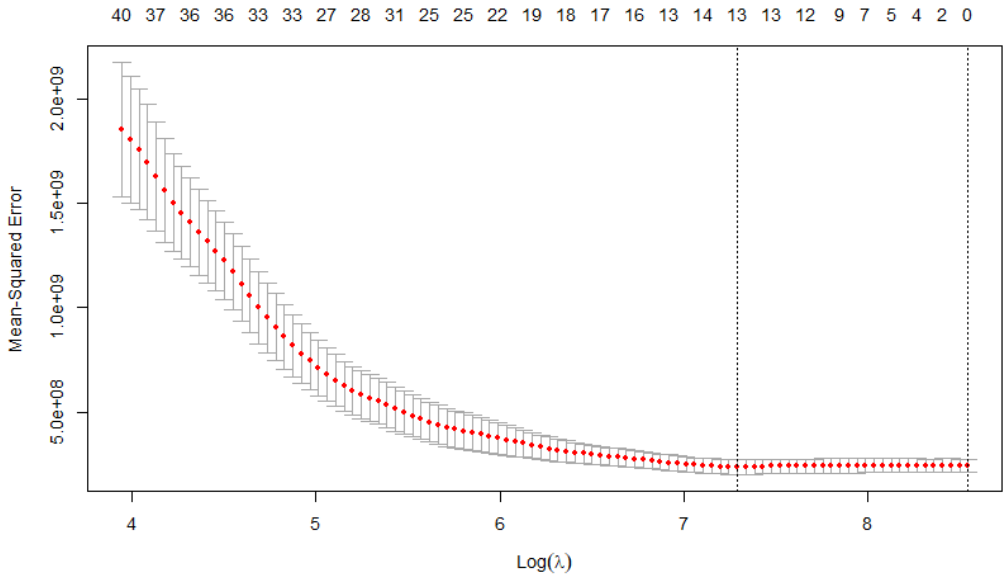

Figura 38. Gráfico de validación cruzada para la búsqueda de λ en la Regresión Lasso

En este caso al aplicar la validación cruzada se obtiene el valor de λ mínimo de 1470,155 y el valor óptimo de λ con el error más regularizado de 5161,997, nuevamente valores altísimos. Por tanto, al igual que en la regresión Ridge cuando se aplica un modelo de Regresión Lasso con el λ=5161,977 no se obtiene ninguna variable de la respuesta celular que resulte significativa.

Para tratar de mejorar el estudio de la relación de los anticuerpos del COVID con la Respuesta Celular se aplica el método de **Regresión de Mínimos Cuadrados Parciales Dispersos (sPLS)** ya que implica diversas ventajas. Este ayudará a la selección de variables automáticas ya que estaremos trabajando con un número mayor de variables que observaciones y, por otro lado, ayudar a estabilizar el modelo.

Para poder ajustar el modelo sPLS primero es necesario saber cuántas variables se deben incluir en el modelo. Para ello, se crea un modelo PLS con muchas componentes, en este caso se aplican 10, y se aplica validación cruzada de 10 pliegues para obtener el valor de la capacidad predictiva del modelo Q^2.

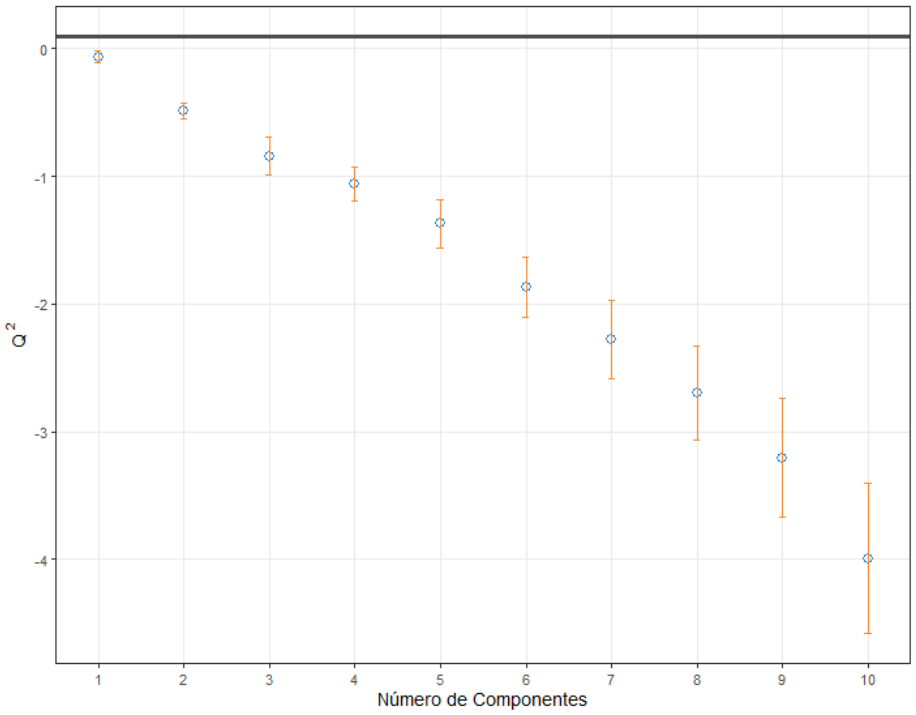

Figura 39. Selección de Número de Componetes sPLS, criterio Q2

Este gráfico muestra el valor Q^2 de cada dimensión añadida al modelo y la línea 0,0975 indica el umbral por debajo del cual añadir una dimensión puede no ser beneficiosa para mejorar la precisión en PLS. Los resultados de este gráfico indican que no se recomienda añadir ningún componente en el gráfico. Tal y como ocurría con el estudio de los parámetros Hematológicos, esto está indicando la poca relación de estas dos respuestas inmunitarias. Sin embargo, para poder trabajar con el modelo PLS y poder observar unos resultados más allá, se realiza el modelo con 2 dimensiones a pesar de que esté indicado que no se verá ningún resultado.

El siguiente paso para ajustar el modelo sPLS con dos dimensiones es buscar el número óptimo de variables a seleccionar en cada una de estas componentes, ya que el sPLS realiza una selección de variables. Para ello, se emplea el error absoluto medio MAE.

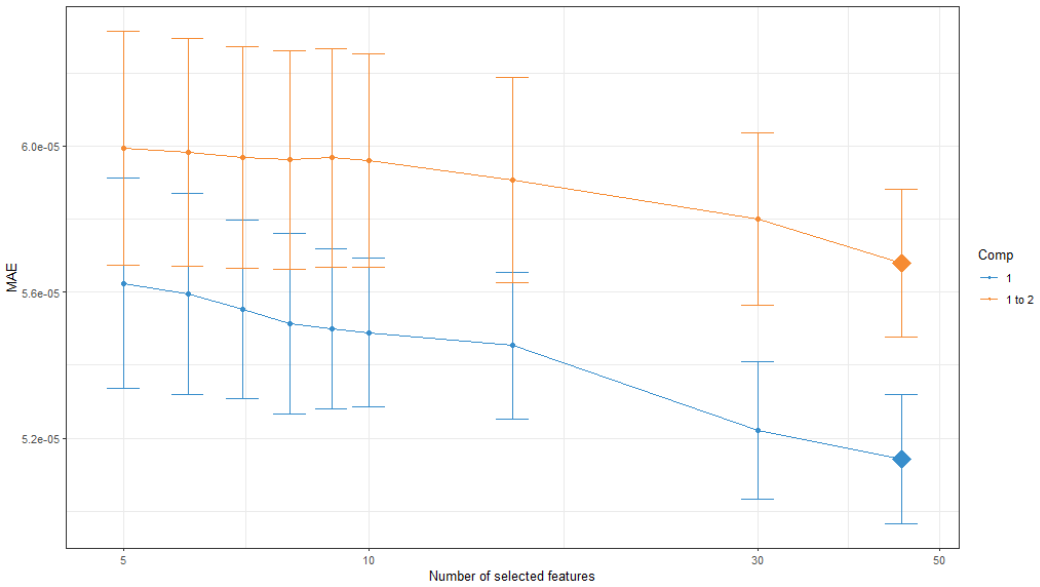

Figura 40. Criterio del Erros Absoluto para elegir el número de variables a seleccionar en sPLS

El gráfico anterior está mostrando que incluir una segunda componente amplia el erro MAE, encajando esta observación con el resultado de selección del número de componentes. Asimismo, muestra el número de variables óptimos a seleccionar en ambas compontes, siendo 45 tanto en la primera como en la segunda componente. Por tanto, se procede a ajustar un modelo sPLS con dos componentes de 45 variables en cada una de ellas ya que es lo que nos ofrece un mejor error.

Para poder analizar los resultados, se muestra a continuación el gráfico de ponderación o Weightings que muestra la estructura de correlación entre ambos espacios, X e Y, con la selección de variables del sPLS.

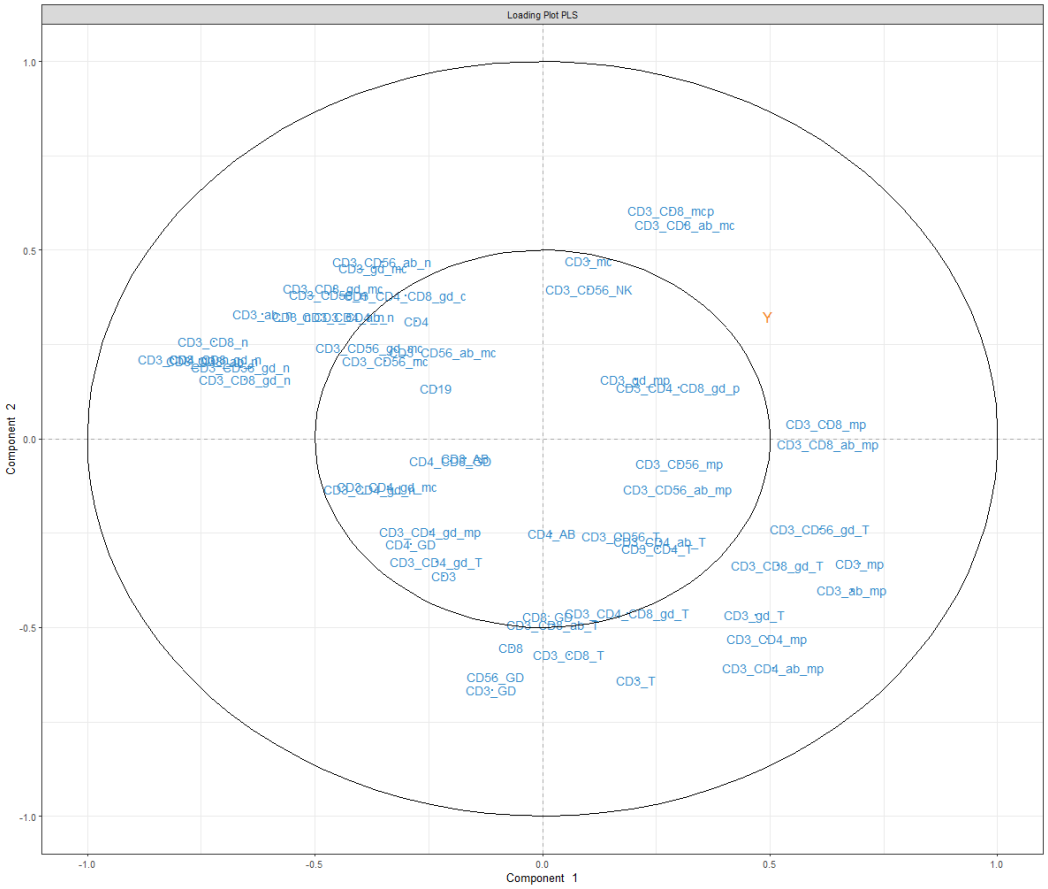

Figura 41. Gráfico de Weightings sPLS de 1º y 2º componente

El gráfico muestra como la variable que representa los anticuerpos del SARS-CoV-2 (Y) tienen un peso pequeño tanto en la primera como en la segunda componente. Esto muestra la poca relación de los anticuerpos del Covid con la respuesta celular, como venían indicando tanto el gráfico Q2 y que muestra capacidad predictiva del modelo como el error MAE.

Por otro lado, se muestran las puntuaciones factoriales tanto de la respuesta celular como de los anticuerpos Y coloreados en función del tipo de cáncer. En ambos casos no parece haber un peso significativo del tipo de cáncer en la respuesta celular y en los anticuerpos del Covid.

Por tanto, teniendo en cuenta la capacidad explicativa del modelo y el resultado observado tanto en el gráfico de Weigthings como en el de scores, no parece que los anticuerpos a priori tengan una relación evidente con la Respuesta Celular.

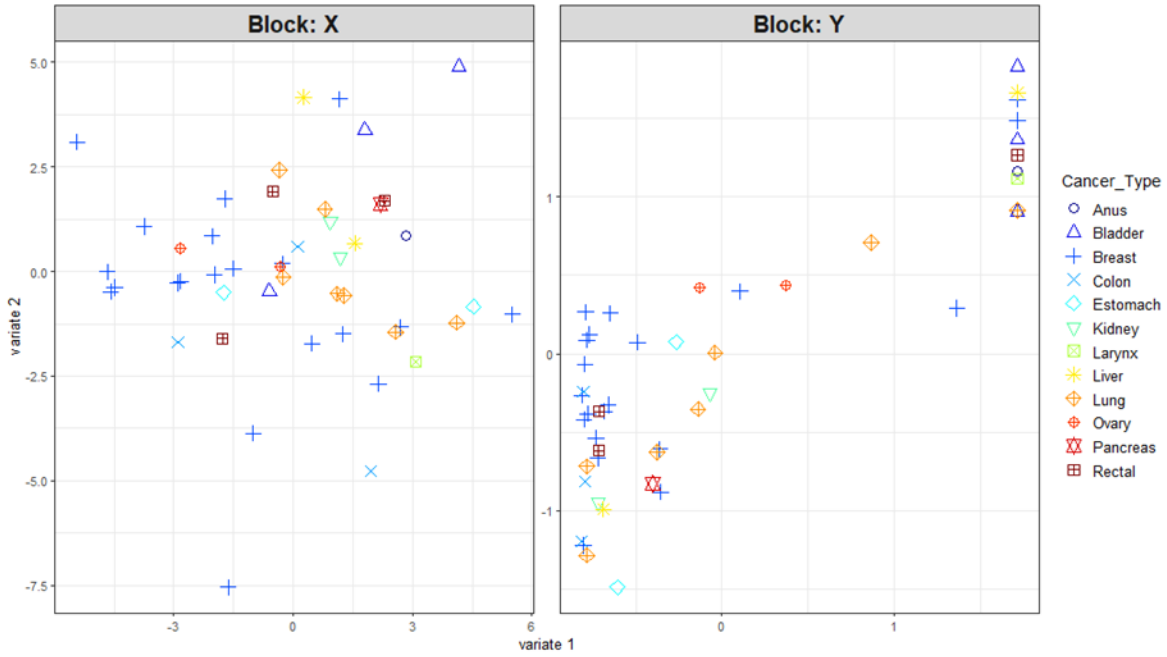

Figura 42. Gráfico de scores sPLS

4.3. Determinar la generación de anticuerpos al virus SARS-CoV-2 en función de características del paciente y de las medidas de respuesta inmunitaria

Para modelizar la generación de anticuerpos frente al SARS-CoV-2 se codifica la variable que contiene la cantidad de anticuerpos en una variable binaría que indica si los pacientes tienen anticuerpos o no. A continuación, se aplica un **Análisis Discriminante de Mínimos Cuadrados Parciales (PLS-DA)** para trata de diferenciar a los pacientes con anticuerpos de los que no los crean en función de las características del paciente (Sexo, Edad, Cáncer y Tratamiento), de los parámetros Hematológicos y de la respuesta Celular más significativa obtenida del SPCA que ha ayuda a la selección de variables.

En primer lugar, se evalúa el rendimiento del PLS-DA para clasificar a los pacientes con anticuerpos de los que no tienen mediante validación cruzada con 10 pliegues introduciendo un número alto de componentes, en este caso 20.

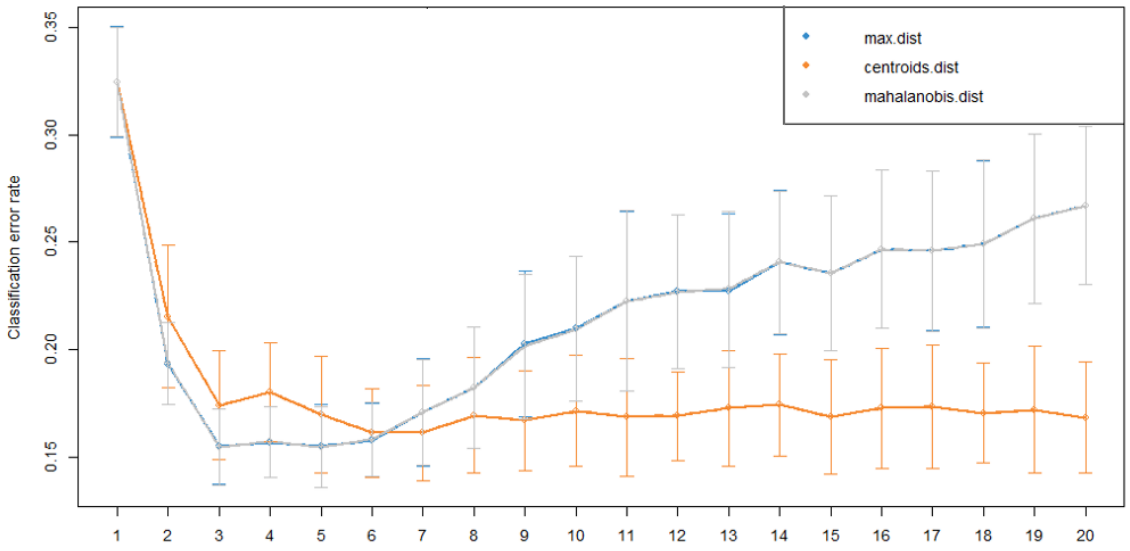

Figura 43. Ajuste del Nº de componentes PLS-DA

En cada componente se ha repetido la validación cruzada para evaluar el rendimiento de la clasificación para cada tipo de distancia de predicción: Distancia máxima, Centroides y distancia de Mahalanobis, además, las barras muestran la desviación estándar en los pliegues repetidos. El gráfico muestra que la tasa del error alcanza un mínimo a partir de 3 componentes y que en algún caso ese error vuelve a aumentar cuando se incluyen más componentes en el modelo. Por ello, se procede a aplicar un modelo PLS-DA con 3 compontes y a analizar el resultado.

Los siguientes gráficos muestras la distribución de las puntuaciones factoriales en la primera y segunda componente. En ellos no se puede apreciar una diferencia entre el grupo de pacientes que no tienen anticuerpos (0) y aquellos que si han conseguido desarrollar anticuerpos (1). Este resultado va acorde con los análisis realizados previamente ya que se ha ido obteniendo un resultado en que los anticuerpos del COVID no parecen tener relación con ninguna de las respuestas inmunes ni con las características del paciente.

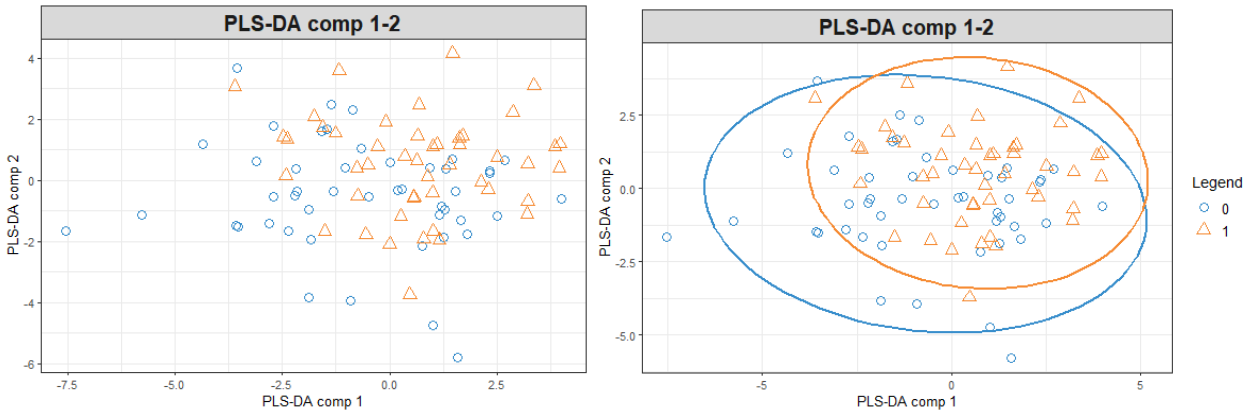

Figura 44. Gráficos de Sores del PLS-DA

Por otro lado, se muestra la correlación de las distintas variables, así como el peso de las variables en la primera y segunda componente, donde los colores indican la clase para la que una variable en concreto

se expresa al máximo. El peso de las variables en la tercera componente se muestra en el Anexo en la Figura B 1.

Observando los gráficos de la Figura 45 y la Figura 46 se podría decir que la variable más influyente en la clasificación de los pacientes con anticuerpos del COVID sería el *CD3* ya que se encuentra con puntuaciones más altas el *CD3 Naive, CD3 Memoria Periférica* y *CD3 AB Naive*. A pesar de tener poco peso y no ser muy significativo, el CD3 resulta ser lo más relevante para la clasificación de los pacientes con anticuerpos tras la obtención de la vacuna del Covid.

Por otro lado, la Figura 47 está mostrando el peso del cáncer de Colon en la segunda componente, revelando la influencia de este tipo de cáncer en los niveles más alto de anticuerpos.

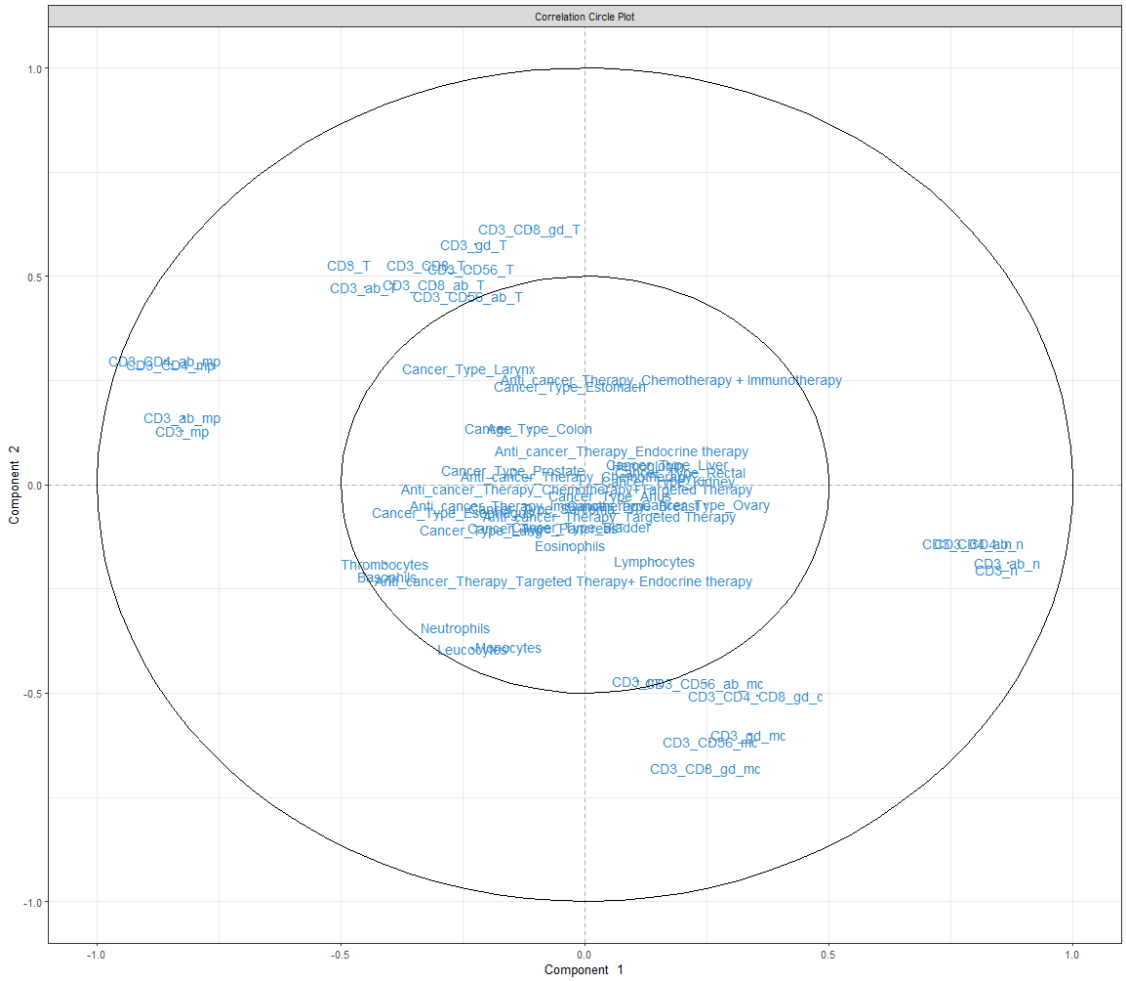

Figura 45. Gráfico de correlación de las variables del PLS-DA

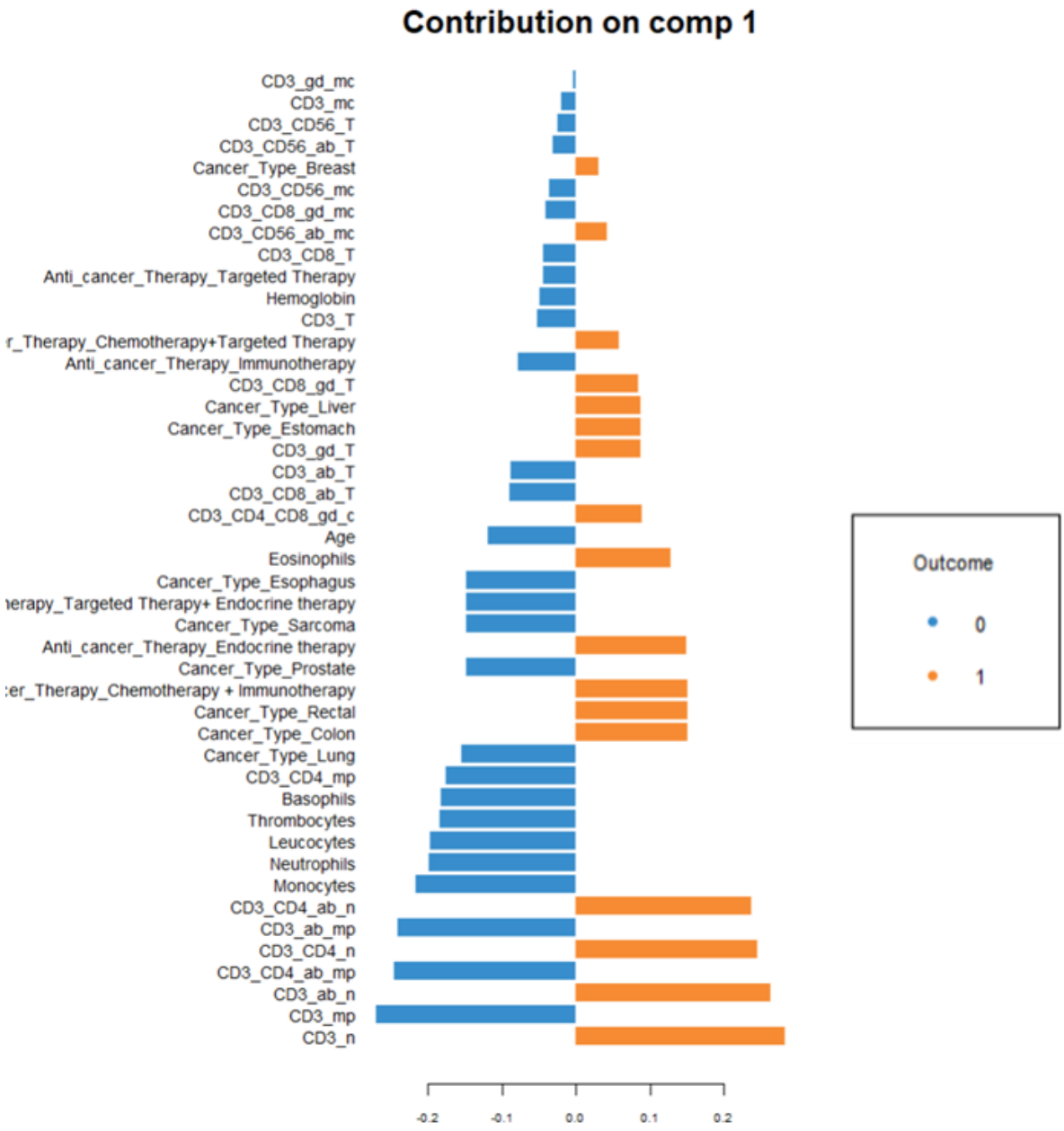

Figura 46. Carga las variables en la primera componente principal PLS-DA

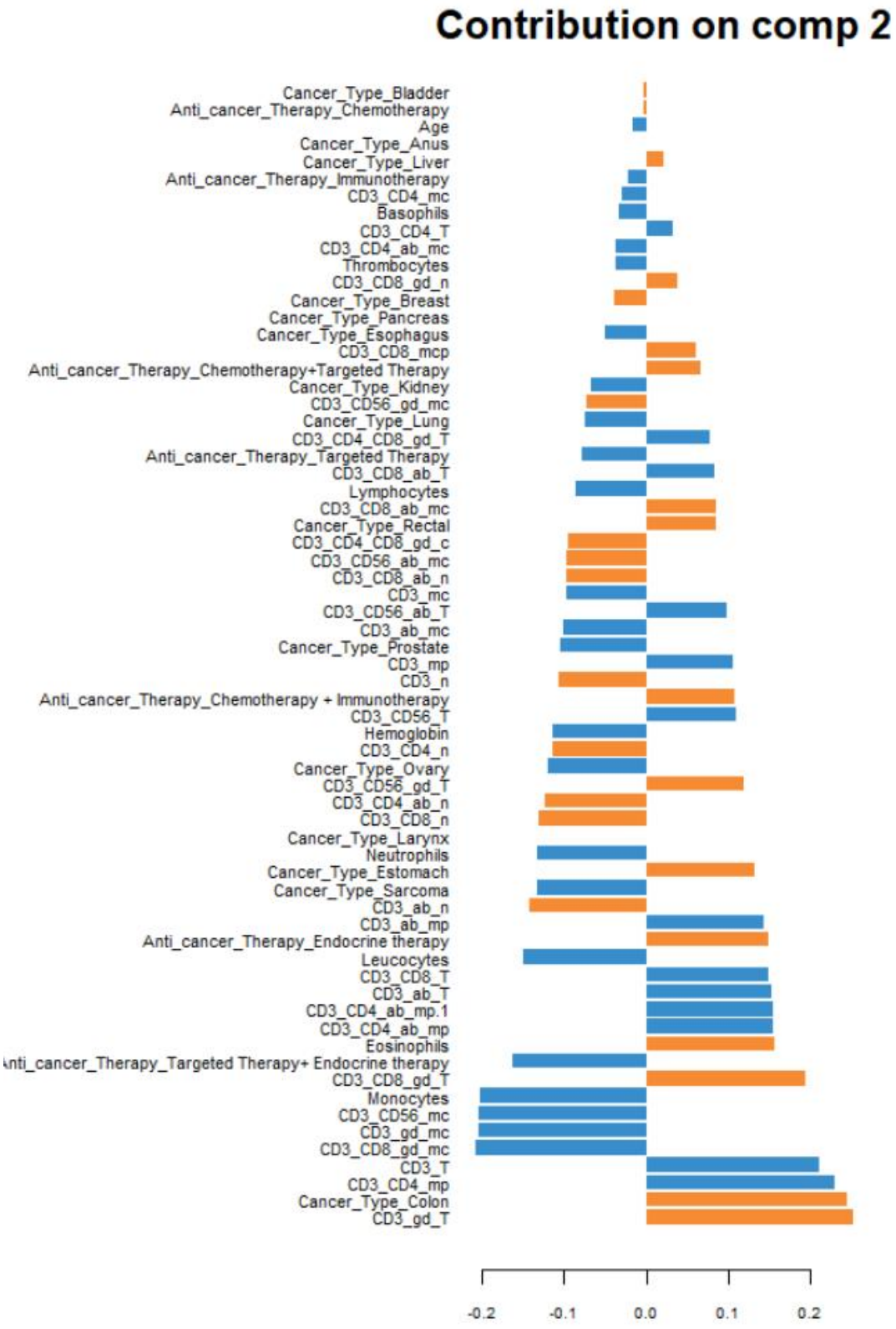

Figura 47. Carga las variables en la segunda componente principal PLS-DA

Para tratar mejorar el análisis, a pesar de lo que indican los resultados previos, y seleccionar realmente aquellas variables más influyentes se realiza un **Análisis Discriminante de Mínimos Cuadrados Parciales Disperso (sPLS-DA)**. Para ello, partiremos un PLS-DA con 3 componentes ya que es el número óptimo que indica la Figura 43.

Para seleccionar las variables a introducir en cada componte se calcula la tasa de error de clasificación en cada componte,

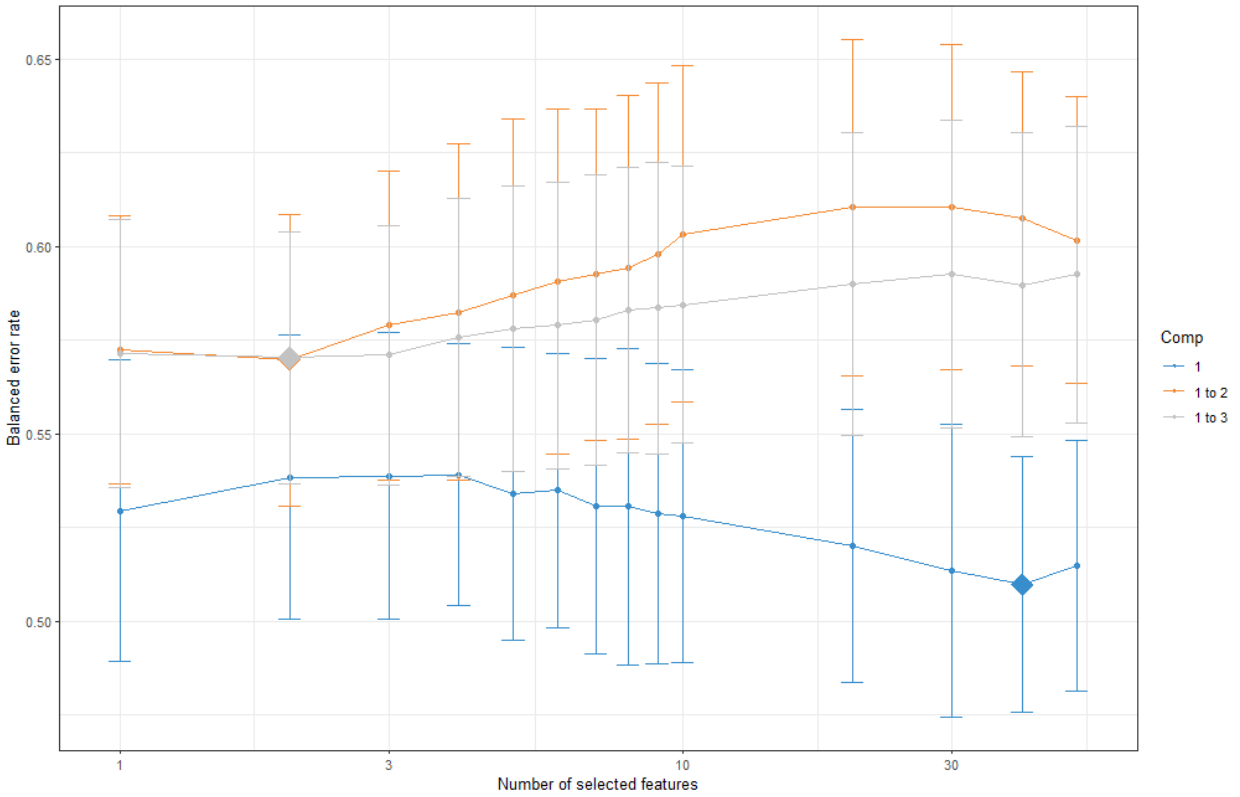

Figura 48. Selección del número de variables en cada componente sPLS -DA

El gráfico muestra en cada línea de color la tasa de error equilibrada por componente (Y) dependiendo del número de variables introducida (representadas en el eje X) después de aplicar validación cruzada con 5 pliegues. El rombo por otro lado, indica el número óptimo de componentes a introducir en cada componente. En este caso habría que seleccionar 35 variables en la primera componente y 2 en la segunda y en la tercera. A pesar de ello, se debe tener en cuenta que el gráfico está indicando que la introducción de más de una componente en el modelo hace que aumente el error de clasificación, por lo que se decide prescindir de la tercera componente.

Para poder visualizar el resultado del modelo se procede a ajustar un modelo sPLS-DA con dos componentes, donde la primera tendrá 35 variables y la segunda 2 variables.

Tal y como mostraba el modelo PLS-DA del apartado anterior no se obtiene una distinción de los pacientes que tienen anticuerpos (1) de los que no tienen (0). Sin embargo, mediante la introducción del modelo disperso podremos observar, cual ha sido la selección de las variables más significativas a la hora de distinguir estos grupos de pacientes.

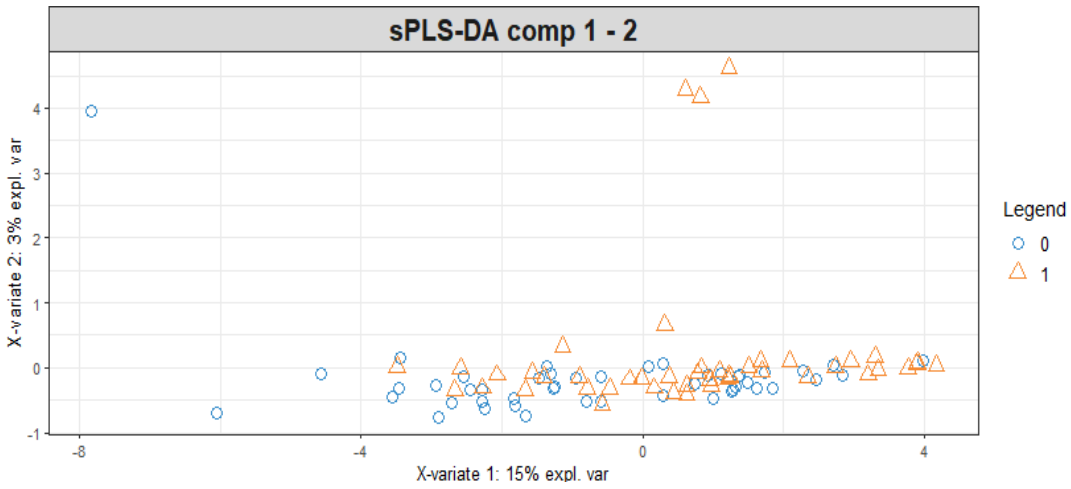

Figura 49. Score plot sPLS-DA 1º y 2º Comp

A continuación, se muestra el gráfico de las puntuaciones factoriales donde se aprecia el peso de las variables en la primera y segunda componente:

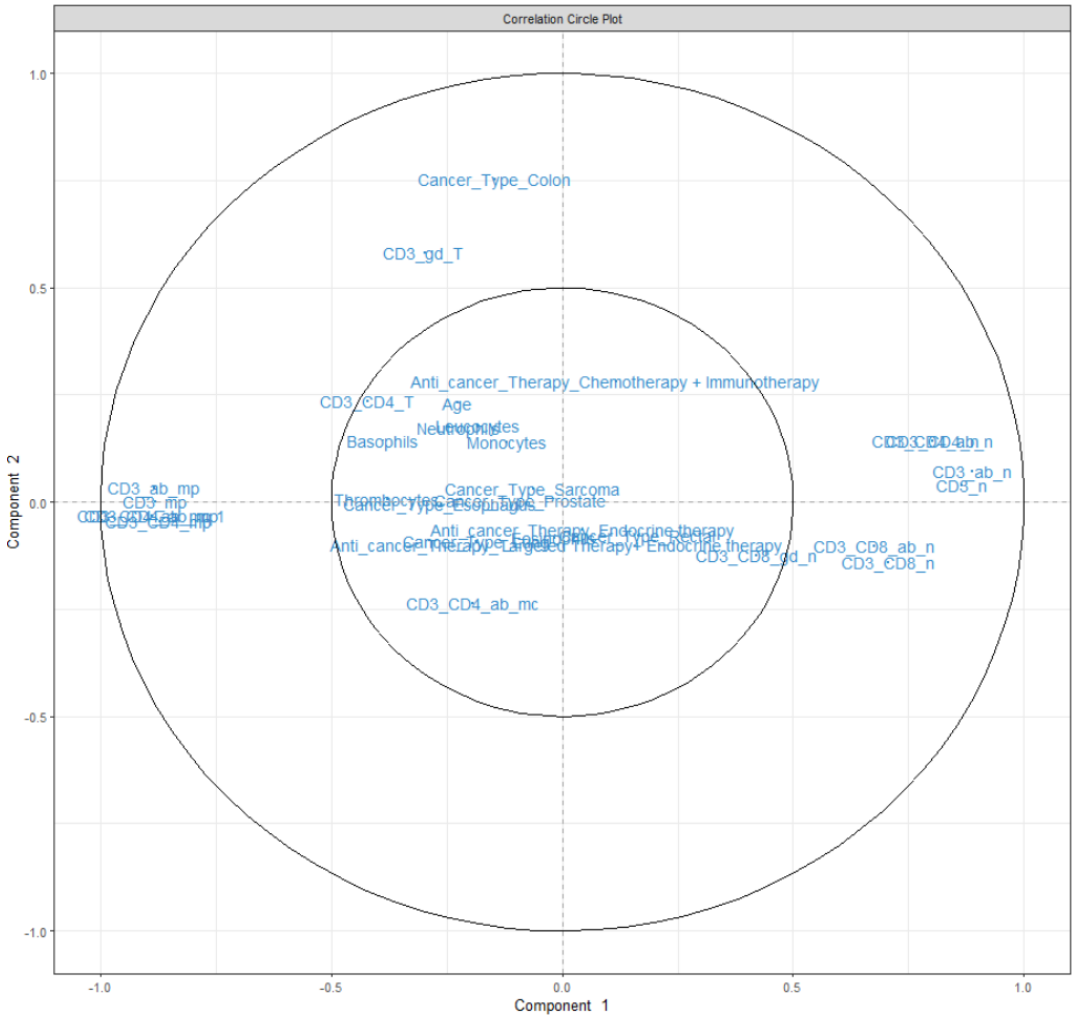

Figura 50. Score plot 1º y 2ª Componentes sPLS-DA

Asimismo, se muestra la contribución de las variables tanto en la primera como en la segunda componte principal.

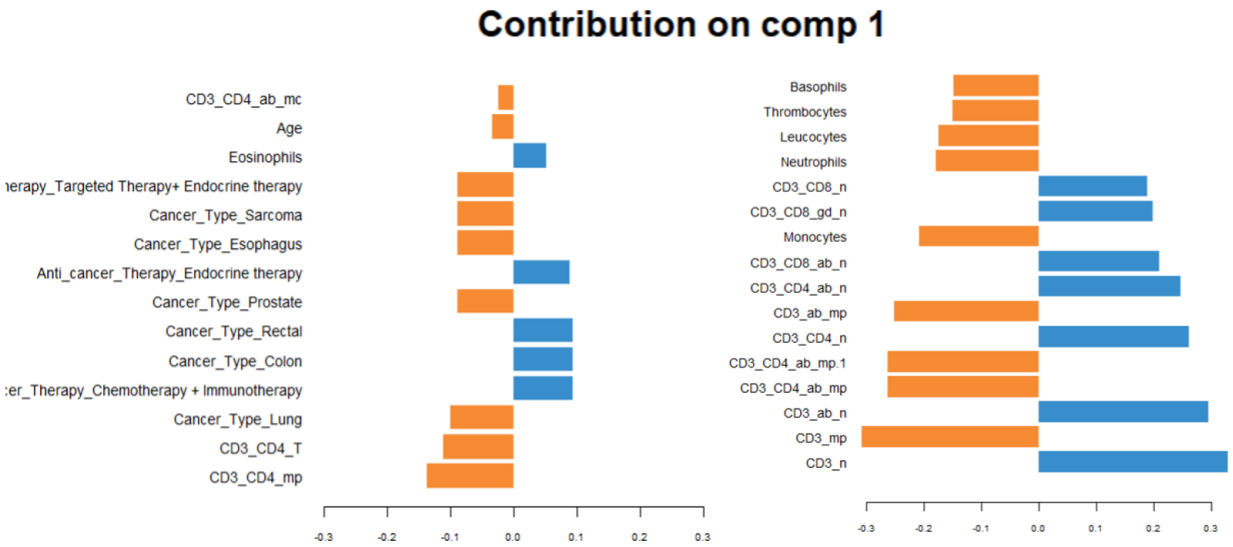

Figura 51. Peso de las variables en la primera componente sPLS-DA

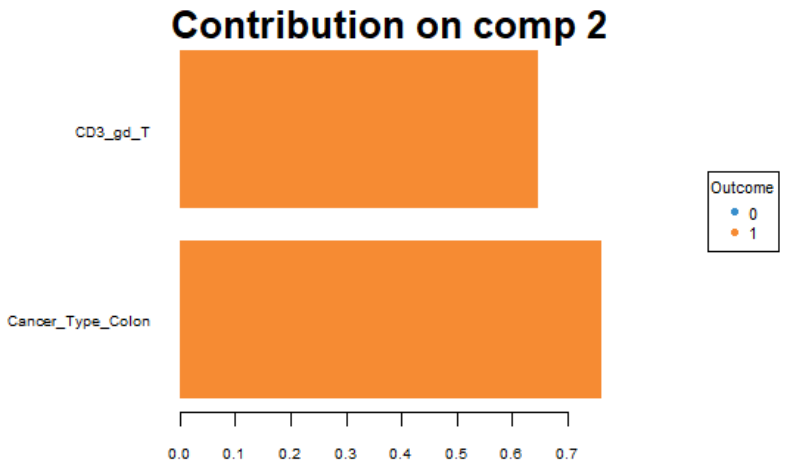

Figura 52. Peso de las variables en la segunda componente sPLS-DA

Por un lado, en la primera componente se observa como las variables más influyentes son *CD3 Naive, CD3 AB Naive* y *CD3 Memoria Periférica* al igual que mostraba la Figura 46. El aumento de los niveles de CD3 puede reflejar la activación de los linfocitos T y su participación en la respuesta inmunitaria contra los antígenos específicos de la vacuna, recordando que estos desempeñan un papel fundamental en la respuesta inmunitaria adaptativa.

Por otro lado, es importante destacar el gran impacto que tiene el Cáncer de Colon en la segunda componente, lo cual encaja con el resultado obtenido en la Figura 47. Esto indica que los pacientes que sufren este tipo de cáncer tienen mayor tendencia a generar niveles más altos de respuesta inmune.

Por último y para ver estos resultados, si observamos la variable CD3_n en función de si ha creado o no anticuerpos el paciente se podrías decir que, a pesar de ser similar, cuando el paciente recibe la vacuna del Covid los niveles de CD3 aumentan en el paciente. Asimismo, se observa como también aumenta el valor medio de la variable CD3_ab_n.

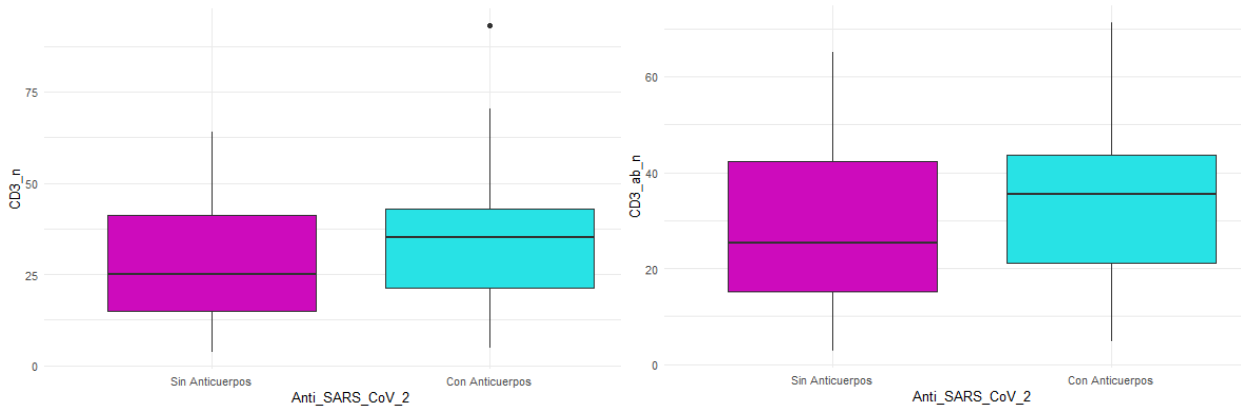

Figura 53. Distribución de CD3_n y de CD3_ab_n en función de los anticuerpos

5. CONCLUSIONES

En la presente memoria, se ha logrado avanzar en la comprensión y funcionamiento del sistema inmunitario en pacientes oncológicos en tratamiento activo después de recibir la vacuna de ARNm contra el SARS-CoV-2. Además, se ha obtenido una mayor claridad acerca de los efectos del tratamiento recibido y del tipo de cáncer en la respuesta inmunitaria.

Los datos recopilados han proporcionado evidencia clara de la generación de anticuerpos específicos contra el COVID19 después de la vacunación, lo que ha permitido identificar patrones y características dentro de los grupos de respuesta inmunitaria.

En cuanto al estudio sobre la influencia del tratamiento se ha observado que los pacientes que reciben tratamiento de inmunoterapia presentan niveles más altos de ciertos indicadores hematológicos, como la Hemoglobina, Linfocitos y Eosinófilos. Por otro lado, los pacientes sometidos a quimioterapia muestran niveles más bajos en la respuesta inmunitaria, lo cual puede deberse a la naturaleza agresiva de este tipo de tratamiento.

El estudio sobre la influencia del tipo de cáncer sugiere que los pacientes que padecen cáncer de Colon tienden a crear mayor nivel de anticuerpos que otros pacientes oncológicos. Sin embargo, resulta desafiante confirmar de manera absoluta esta conclusión debido a la limitada cantidad de casos presentes en cada categoría de cáncer estudiada.

Al analizar la respuesta celular en los distintos niveles de anticuerpos, la aplicación de técnicas dispersas como el sPLS ha permitido reducir los datos de 72 variables a dos componentes de 45 variables cada una. A pesar de que estas no consigan explicar mucho, se observa como principalmente destaca el CD3 o combinaciones de CD3. Asimismo, se aprecia que los niveles de las células de memoria central y TEMRA resultan más significativas en cuanto a la diferencia en los niveles de anticuerpos contra el virus.

Finalmente, se han categorizado los niveles de anticuerpos IgG específicos contra el virus SARS-CoV-2 para estudiar la generación de anticuerpos en función de las características individuales y de la respuesta inmunitaria. La aplicación del sPLS-DA en este apartado ha permitido simplificar la comprensión de los resultados pasando de 67 variables a 3 componentes 35 variables en la primera componente y 3 en las dos siguientes. Nuevamente, ha revelado la presencia de una memoria inmunológica en otros componentes del sistema inmune, especialmente un aumento en los niveles de CD3. Asimismo, destacan los pacientes que padecen cáncer de Colon con niveles más altos de respuesta inmune.

En conclusión, este estudio destaca la utilidad del empleo de los métodos multivariantes dispersos para la evaluación de la respuesta inmunitaria en este grupo de población, considerando sus particularidades y los posibles beneficios adicionales que puede proporcionar.

Como perspectiva de investigación futura, sería altamente interesante explorar la incorporación de la N-integración del paquete MixOmics en el análisis. Esto permitiría considerar la estructuración de los pacientes en bloques según variables como la edad, tipo de cáncer, tipo de tratamiento, entre otros. La aplicación de esta metodología enriquecería aún más el conocimiento al considerar la influencia conjunta de múltiples factores en la respuesta inmunológica de los pacientes oncológicos en tratamiento activo.

REFERENCIAS

Abbas, A. K., Lichtman, A. H., & Pillai, S. (2020). Inmunidad celular y molecular. En Inmunología básica: funciones y trastornos del sistema inmunitario. Barcelona: Elsevier.

Abbas, A., Lichtman, A., & Pillai, S. (2007). Cellular and molecular immunology. Saunders Elsevier.

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). Wiley Computational Statistics. doi:10.1002/wics.51

Baden, L. R., Sahly, H. M., Essink, B., Kotloff, K., Frey, S. N., R., D. D., . . . Corey, L. (2021). Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *The New England Journal of Medicine, 384*, 403-416.

Castellanos-Bueno, R. (2020). Immune response. Revista Colombiana De Endocrinología, Diabetes &Amp; Metabolismo.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Germany: Geoscientific Model Development. doi:10.5194/gmd-7-1247-2014

Chin, W. (1998). The Partial Least Squares Approach to Structural Equation Modeling. Lawrence Erlbaum Associates Publishers.

Chung, D., & Keles, S. (2010). *Sparse partial least squares classification for high dimensional data* (Vol. 9(1)). Stat Appl Genet Mol Biol. doi:10.2202/1544-6115.1492

Conchado, A., Fernandez-Murga, L., Garde, J., Serrano, L., Portero, M. L.-C., & Martin, N. (2023). Sparse multivariate methods to assess immune response in actively treated oncology patients after COVID19 vaccination. *Mathematical Modelling in Engineering & Human Behaviour 2023 (MME&HB2023).* Valencia.

Dastan, H. M., & Adnan, M. A. (2020). A Review on Linear Regression Comprehensive in Machine Learning. Applied Science and Technology Trends.

Desai, C., Pathak, H., & Madamwar, D. (2010). Advances in molecular and ‘‘-omics” technologies to gauge microbial communities and bioremediation at xenobiotic/anthropogen contaminated sites. Elsevier. doi:10.1016/j.biortech.2009.10.080

Dheri, G., Soumen, P., Varinderpal, S., Sudeep, M., & Choudhary, O. (2019). Hands-on Training on "Statistical Tools and Database Management in Agriculture. LUDHIANA.

Dougherty, E. R., Sima, C., Hua, J., Hanczar, B., & Braga-Neto, U. M. (2010). Performance of Error Estimators for Classification. Current Bioinformatics. doi:10.2174/157489310790596385

Ferrer, A., Aguado, D., Vidal-Puig, S., Prats, J. M., & Zarzo, M. (2008). PLS: A versatile tool for industrial process improvement and optimization. (A. S. Industry, Ed.) *24(6)*, 551-567.

Fordellone, M., Bellincontro, A., & Mencarelli, F. (2020). Partial least squares discriminant analysis s: A dimensionality reduction method to classify hyperspectral data. talian Journal of Applied Statistics. doi:10.26398/IJAS.0031-010

Geladi, P., & Kowalski, B. (1986). PARTIAL LEAST-SQUARES REGRESSION: A TUTORIAL. Amsterdam: Elsevier. doi:10.1016/0003-2670(86)80028-9

Greenacre, M., Groenen, P., Hastie, T., Iodice D'Enza, A., Markos, A., & Tuzhilina, E. (2022). Principal Component Analysis. 100. Nature Reviews Methods Primers.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical Learning with Sparsity, Lasso and Generalization. doi:10.1201/b18401

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. En *Data Mining, Inference, and.* Springer.

Hoebe, K., Janssen, E., & Beutler, B. (2004). The interface between innate and adaptive immunity. Nature Publishing Group. doi:10.1111/j.0022-202X.2005.23856.x

Hoerl, A. E., & Kennard, R. W. (2000). Ridge Regression: Biased Estimation for Nonothogonal Problems. Technometrics. doi:10.2307/1271436

Huang, H. S. (2008). Sparse principal component analysis via regularized low rank matrix approximation. Journal of Multivariate Analysis. doi:10.1016/j.jmva.2007.06.007

Indhupriya, S., Srikant, V., & Shiva, K. (2020). Multi-omics Data Integration, Interpretation, Abhay Jere and Krishanpal Anamika. New Delhi: Bioinformatics and Biology Insights. doi:10.1177/1177932219899051

Instituto Europeo & Instituto Europeo de Salud y Bienestar Social. (2020). Obtenido de Vacuna: la respuesta inmunitaria.: https://institutoeuropeo.es/articulos/insights/vacuna-respuesta-inmunitaria/

Ji-Hoon, C., Jong-Min, L., Sang Wook, C., Dongkwon, L., & In-Beum, L. (2005). Fault identification for process monitoring using kernel principal. ELSEVIER. doi:10.1016/j.ces.2004.08.007

Kennard, A. E. (2000). Ridge Regression: Biased Estimation for Nonorthogonal Problems. American Statistical Association and American Society for Quality.

Lê Cao, K., Rossouw, D., Robert-Granié, C., & Besse, P. (2008). A sparse PLS for variable selection when integrating omics data. Stat Appl Genet Mol Biol.

Lê Cao, K.-A., & Welham, Z. (2021). Multivariate Data Integration Using R. Methods and Applications with the mixOmics Package. New York: CRC Chapman & Hall. doi:10.1201/9781003026860

Lê Cao, K.-A., Boitar, S., & Besse, P. (2011). Sparse PLS discriminant analysis: biologically elevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics. doi:https://doi.org/10.1186/1471-2105-12-253

Lee, W., Lee, D., Lee, Y., & Pawitan, Y. (s.f.). *Sparse canonical covariance analysis for high-throughput data. Statistical Applications in Genetics and Molecular Biology* (Vol. 10(1)). 2011. doi:https://doi.org/10.2202/1544-6115.1638

Mackay, M., S., F., & Rosen, M. (2000). Advances in Immunology. INNATE IMMUNITY. The New England Journal of Medicine. doi:10.1056/NEJM200008033430506

Montoya, A. F. (2021). MEMORIA INMUNOLÓGICA, STRESS Y EMOCIONES. Le corps et l'analyse. Revue des societés francophones d'analyse bioénergétique.

Noell G, F. R. (2018). From systems biology to P4 medicine: applications in respiratory medicine. European Respiratory Review. doi:10.1183/16000617.0110-2017

Palsson, A. R. (2006). The model organism as a system: integrating 'omics' data sets. Nature Publishing Group. doi:10.1038/nrm1857

Paul, G., & Johan , L. (2020). Principal component analysis. En *Comprehensive chemometrics : chemical and biochemical data analysis* (Vol. 2, págs. 17–37).

Petrellis, D., & Skoupil, D. (2023). Ridge regression for minimizing hyperon resonances' couplings in the K+Λ photoproduction. Czech Republic: Nuclear Physics Institute. doi:10.48550/arXiv.2212.14305

Polack, F. P., Thomas, S. H., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., . . . Hammitt, L. L. (2020). Safety and Efficacy of the BNT162b2 mRNA COVID19 Vaccin. *The New England Journal of Medicine, 383*, 2603-2615. doi:10.1056/nejmoa2034577

Poland, G., Ovsyannikova, I., Jacobson, R., & Smith, D. (2007). Heterogeneity in Vaccine Immune Response: The Role of Immunogenetics and the Emerging Field of Vaccinomics. Nature publishing group. doi:10.1038/sj.clpt.6100415

Ringnér, M. (2008). What is principal component analysis? Nature Publishing Group. doi:10.1038/nbt0308-303

Rocke, D. V., & David, M. (2002). Tumor classification by partial least squares using microarray gene expression data. Bioinformatics.

Rodolphe, J., Obozinski, G., & Francis, B. (2009). Structured Sparse Principal Component Analysis. Willow Project.

Rosipal, R., & Trejo, L. (2001). Kernel Partial Least Squares Regression in Reproducing.

Štruc, V., & Pavešic´, N. (2009). A comparison of feature normalization techniques for PC-Based Palmprint Recognition. Faculty of Electrical Engineering, University of Ljubljana, Slovenia.

Su, E., Fischer, S., Demmer-Steingruber, R., Nigg, S., Güsewell, S., Albrich, W. C., . . . Kahlert, C. R. (2022). Humoral and cellular responses to mRNA-based COVID19 booster Humoral and cellular responses to mRNA-based COVID19 booster. *ESMO, 7*(5). doi:10.1016/j.esmoop.2022.100587

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lassi. Journal of the Royal Statistical Society.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Wiley for the Royal Statistical Society.

Waldhorn, I., Holland, R. I., Goshen-Lago, T., Shirman, Y., Szwarcwort-Cohen, M., Reiner-Benaim, A., . . . & Ben-Aharon, I. (2021). Six-Month Efficacy and Toxicity Profile of BNT162b2 Vaccine in Cancer Patients with Solid Tumors. *Cancer Discovery,, 11*, 2430-2435. doi:10.1158/2159-8290.cd-21-1072

Weber, C., Rubio, T., Wang, L., Zhang, W., Robert, P., Akbar, R., . . . Greiff, V. (2022). eference-based comparison of adaptive immune receptor repertoires. *2*. doi:10.1016/j.crmeth.2022.100269

WHO, W. H. (s.f.). *WHO Coronavirus (COVID19) Dashboard With Vaccination Data*. Obtenido de https://covid19.who.int/?mapFilter=deaths

Wold, S., Kettaneh, N., & Tjessem, K. (1996). *Hierarchical Multiblock PLS and PC Models for Easier Model Interpretation and as an Alternative to Variable Selection* (Vol. 10). Journal of Chemometrics. doi:https://doi.org/10.1002/(sici)1099-128x(199609)10:5/6

ANEXOS

ANEXO A

Descripción de las variables de trabajo

A.1. Parámetros Hematológicos:

- **Leucocitos**: Son células sanguíneas que se produce en la médula ósea y se encuentra en la sangre y el tejido linfático. Forman parte del sistema inmunitario y ayudan a proteger el cuerpo contra infecciones y enfermedades. Los tipos de leucocitos son los granulocitos (neutrófilos, eosinófilos y basófilos), los monocitos y los linfocitos (células T y células B).

- **Neutrófilos**: Se encargan de combatir infecciones bacterianas. Son los más abundantes y suelen ser los primeros en llegar a una zona infectada.
- **Linfocitos**: desempeñan un papel fundamental en la respuesta inmunitaria. Hay diferentes tipos de linfocitos, como los linfocitos T y los linfocitos B, que desempeñan diferentes funciones en la defensa del organismo. Las células B elaboran los anticuerpos para luchar contra bacterias, virus y toxinas invasoras. Las células T destruyen las propias células del cuerpo que han sido infectadas por virus o que se han vuelto cancerosas.
- **Monocitos**: Se encargan de eliminar materiales extraños y células muertas. También pueden diferenciarse en macrófagos, que son células que fagocitan y eliminan patógenos y desechos celulares.

- **Eosinófilos**: Implicados en las respuestas alérgicas y en la defensa contra parásitos. Se caracterizan por tener gránulos que contienen sustancias tóxicas para los parásitos.

- **Basófilos**: Liberan sustancias químicas inflamatorias, como la histamina, que participan en las reacciones alérgicas.

- **Trombocitos**: Son fragmentos celulares involucrados en la coagulación sanguínea. Juegan un papel crucial en la formación de coágulos para detener el sangrado en caso de lesiones.

- **Hemoglobina**: Es una proteína presente en los glóbulos rojos que transporta el oxígeno desde los pulmones hacia los tejidos del cuerpo y ayuda a eliminar el dióxido de carbono.

A.2. Respuesta Celular:

- **CD3**: Es una proteína presente en la superficie de los linfocitos T. El CD3 se utiliza como marcador para identificar y contar los linfocitos T en muestras biológicas.

- **CD4**: Es una molécula de superficie que se encuentra principalmente en los linfocitos T auxiliares o células T colaboradoras. Estas células juegan un papel clave en la coordinación de la respuesta inmunitaria, ayudando a activar y regular otras células del sistema inmunitariO.

- **CD8**: Es una molécula de superficie que se encuentra en los linfocitos T citotóxicos o células T asesinas. Los linfocitos T CD8+ están involucrados en la respuesta inmunitaria contra células infectadas por virus y células tumorales.

- **CD56**: Es una molécula de superficie expresada en las células asesinas naturales (NK) y en una subpoblación de linfocitos T. Las células NK son células del sistema inmunitario innato que desempeñan un papel importante en la defensa del cuerpo contra infecciones y en la eliminación de células infectadas o tumorales.

Por otro lado;

- **Memoria central:** Células de memoria que se encuentran en los ganglios linfáticos y la sangre. Estas células tienen una mayor capacidad para responder rápido y eficientemente a la reexposición a antígenos encontrados previamente. Son responsables de respuestas inmunitarias a largo plazo.
- **Memoria periférica:** Se refiere a las células de memoria que se encuentran en los tejidos periféricos del cuerpo. Estas células están preparadas para reaccionar rápidamente cuando se exponen de nuevo al antígeno en sitios infectados o inflamatorios.
- **TEMRA:** Describe una subpoblación de células de memoria de linfocitos T. Estas células se caracterizan por tener una alta capacidad citotóxica y participan en respuestas inmunitarias rápidas y eficaces.
- **Naive:** Se refiere a las células T que aún no han sido expuestas a un antígeno específico. Estas células no han desarrollado una respuesta inmunitaria adaptativa. En respuesta a la exposición a un antígeno, tienen la capacidad de diferenciarse en células de memoria o células efectoras.

A.3. Codificación de las variables de Respuesta Celular:

Tabla A 1. Nombres de las variables de Respuesta Celular codificadas

Respuesta Celular	Respuesta Celular Codificada
CD3	CD3
CD4	CD4
CD8	CD8
CD3_CD56_NKT	CD3_CD56_NKT
CD3_CD56_NK	CD3_CD56_NK
CD3_AB	CD3_AB
CD4_AB	CD4_AB
CD8_AB	CD8_AB
CD56_AB	CD56_AB
CD3_GD	CD3_GD
CD4_GD	CD4_GD
CD8_GD	CD8_GD
CD56_GD	CD56_GD
CD4_CD8_GD	CD4_CD8_GD
CD19	CD19
CD3_naive	CD3_n
CD3_memoria_central	CD3_mc
CD3_memoria_periferica	CD3_mp
CD3_TEMRA	CD3_T
CD3_CD4_naive	CD3_CD4_n
CD3_CD4_memoria_central	CD3_CD4_mc
CD3_CD4_memoria_periferica	CD3_CD4_mp
CD3_CD4_TEMRA	CD3_CD4_T
CD3_CD8_naive	CD3_CD8_n
CD3_CD8_memoria_centraPerc	CD3_CD8_mcp
CD3_CD8_memoria_periferica	CD3_CD8_mp
CD3_CD8_TEMRA	CD3_CD8_T
CD3_CD56_naive	CD3_CD56_n
CD3_CD56_memoria_central	CD3_CD56_mc
CD3_CD56_memoria_periferica	CD3_CD56_mp
CD3_CD56_TEMRA	CD3_CD56_T
CD3_ab_naive	CD3_ab_n
CD3_ab_memoria_central	CD3_ab_mc
CD3_ab_memoria_periferica	CD3_ab_mp
CD3_ab_TEMRA	CD3_ab_T

CD3_CD4_ab_naive	CD3_CD4_ab_n
CD3_CD4_ab_memoria_central	CD3_CD4_ab_mc
CD3_CD4_ab_memoria_periferica	CD3_CD4_ab_mp
CD3_CD4_ab_TEMRA	CD3_CD4_ab_T
CD3_CD8_ab_naive	CD3_CD8_ab_n
CD3_CD8_ab_memoria_central	CD3_CD8_ab_mc
CD3_CD8_ab_memoria_periferica	CD3_CD8_ab_mp
CD3_CD8_ab_TEMRA	CD3_CD8_ab_T
CD3_CD56_ab_naive	CD3_CD56_ab_n
CD3_CD56_ab_memoria_central	CD3_CD56_ab_mc
CD3_CD56_ab_memoria_periferica	CD3_CD56_ab_mp
CD3_CD56_ab_TEMRA	CD3_CD56_ab_T
CD3_gd_naive	CD3_gd_n
CD3_gd_memoria_central	CD3_gd_mc
CD3_gd_memoria_periferica	CD3_gd_mp
CD3_gd_TEMRA	CD3_gd_T
CD3_CD4_gd_naive	CD3_CD4_gd_n
CD3_CD4_gd_memoria_central	CD3_CD4_gd_mc
CD3_CD4_gd_memoria_periferica	CD3_CD4_gd_mp
CD3_CD4_gd_TEMRA	CD3_CD4_gd_T
CD3_CD8_gd_naive	CD3_CD8_gd_n
CD3_CD8_gd_memoria_central	CD3_CD8_gd_mc
CD3_CD8_gd_memoria_periferica	CD3_CD8_gd_mp
CD3_CD8_gd_TEMRA	CD3_CD8_gd_T
CD3_CD4_CD8_gd_naive	CD3_CD4_CD8_gd_n
CD3_CD4_CD8_gd_central	CD3_CD4_CD8_gd_c
CD3_CD4_CD8_gd_periferica	CD3_CD4_CD8_gd_p
CD3_CD4_CD8_gd_TEMRA	CD3_CD4_CD8_gd_T
CD3_CD56_gd_naive	CD3_CD56_gd_n
CD3_CD56_gd_memoria_central	CD3_CD56_gd_mc
CD3_CD56_gd_memoria_periferica	CD3_CD56_gd_mp
CD3_CD56_gd_TEMRA	CD3_CD56_gd_T

ANEXO B

Gráficos

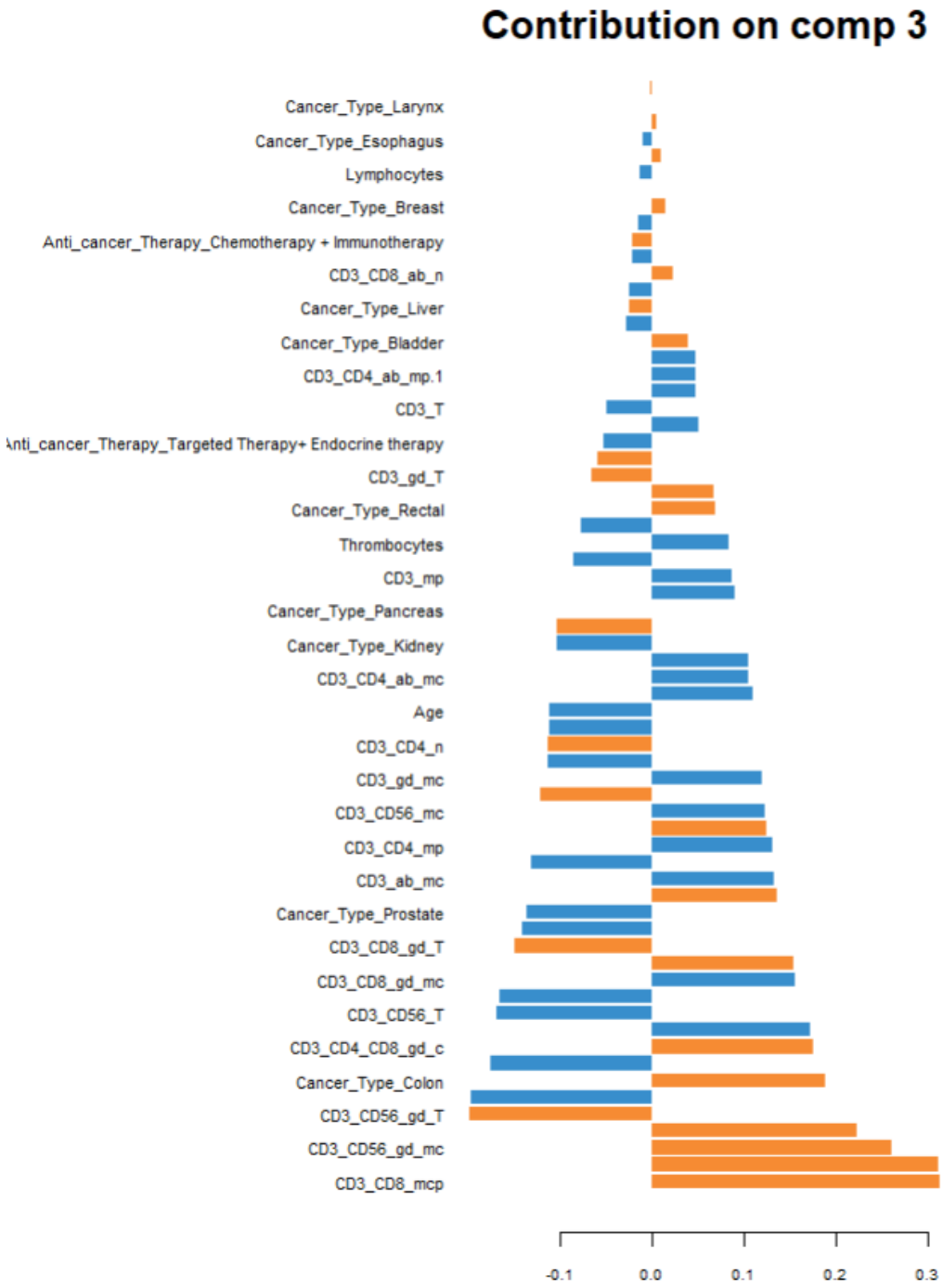

Figura B 1. Carga las variables en la tercera componente principal PLS-DA

ANEXO C

Relación del Trabajo con los Objetivos de Desarrollo Sostenible de la Agenda 2030 (ODS)

Tabla C 1. Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS)

Objetivos de desarrollo sostenible	Alto	Medio	Bajo	No Procede
ODS 1. Fin de la pobreza				X
ODS 2. Hambre cero				X
ODS 3. Salud y bienestar	X			
ODS 4. Educación de calidad				X
ODS 5. Igualdad de género				X
ODS 6. Agua limpia y saneamiento				X
ODS 7. Energía asequible y no contaminante.				X
ODS 8. Trabajo decente y crecimiento económico				X
ODS 9. Industria, innovación e infraestructuras		X		
ODS 10. Reducción de las desigualdades				
ODS 11. Ciudades y comunidades sostenibles				X
ODS 12. Producción y consumo responsables				X
ODS 13. Acción por el clima.				X
ODS 14. Vida submarina				X
ODS 15. Vida de ecosistemas terrestres				X
ODS 16. Paz, justicia e instituciones sólidas				X
ODS 17. Alianzas para lograr objetivos				X

El presente trabajo está estrechamente relacionado con el Objetivo 3 de Salud y Bienestar, ya que aborda de manera efectiva un problema real de salud al buscar una mejor comprensión del funcionamiento del sistema inmunológico en pacientes vulnerables a través de la aplicación de diversos métodos y técnicas estadísticas.

Además, este estudio también aborda el Objetivo 9 de Industria, Innovación e Infraestructuras. Se logra esto mediante la innovación en el uso de métodos dispersos, que contribuyen significativamente a la comprensión de la respuesta inmune en el cuerpo humano. Estos métodos representan un enfoque innovador en el análisis de datos y ofrecen nuevas perspectivas para mejorar la investigación y el desarrollo en el campo de la salud.

ANEXO D

Código de Métodos Multivariantes Dispersos

```
library(mixOmics)

# sPCA
tune.pca.cel <- tune.pca(Celular[,-c(1,2,3,4,5)], ncomp = 20, scale = TRUE)
tune.pca.cel$prop_expl_var$X

grid.keepX <- c(seq(5, 30, 5))
tune.spca.cel <- tune.spca(Celular[,-c(1,2,3,4,5)], ncomp = 4,  folds = 10,  test.keepX = grid.keepX, nrepeat =
10,scale = TRUE)
tune.spca.cel$choice.keepX

plot(tune.spca.cel) +
 xlab("Número de componentes seleccionadas") +
 ylab("Correlación de las componentes")

#loadings
plotVar(final.spca.cel,
      comp = c(1, 2),
      var.names = TRUE,
      cex = 5, # To change the font size
      title = 'Loading Plot')

#scores
plotIndiv(final.spca.cel,
       comp = c(1, 2),   # Specify components to plot
       ind.names = FALSE, # Show row names of samples
       group = Celular$Measure,
       title = ' Score Plot',
       legend = TRUE, legend.title = 'Measure')
```

```
#Loadings contribution

plotLoadings(final.spca.cel, comp = 1,size.title = 1,size.name = 0.8)
```

 # RIDGE & LASSO

```
fit_celular <- glmnet(x=matrix_Celular[,-c(1,12)], y=matrix_Celular[,1], standardize = TRUE)
```

 # RIDGE

```
Cv0_celular= cv.glmnet(matrix_Celular[,-c(1,12)], y=matrix_Celular[,1],alpha=0, standardize = TRUE, nfolds = 20)

Cv0_celular$lambda.min

Cv0_celular$lambda.1se

fit_RIDGE <-glmnet(x=matrix_Celular[,-c(1,12)] , y=matrix_Celular[,1],lambda = Cv0_celular$lambda.1se, standardize = TRUE)

fit_RIDGE$beta

coef(fit_RIDGE,s="lambda.1se")
```

 # LASSO

```
Cv1_celular <- cv.glmnet(x=matrix_Celular[,-c(1,12)] , y=matrix_Celular[,1],alpha=1, standardize = TRUE, nfolds = 20)

Cv1_celular$lambda.min

Cv1_celular$lambda.1se

fit_LASSO<- glmnet(x=matrix_Celular[,-c(1,12)] , y=matrix_Celular[,1],lambda = Cv1_celular$lambda.1se, standardize = TRUE)

fit_LASSO$beta

coef(fit_LASSO,s="lambda.1se")
```

 # SPLS

```
tune.pls1_celular <- pls(X=Celular_Basal[,-c(1,2,3,4,5)], Y = Celular_Basal[,5], ncomp = 10, mode = 'regression',scale = TRUE)

Q2.pls1_celular <- perf(tune.pls1_celular, validation = 'Mfold', folds = 10, nrepeat = 10)

plot(Q2.pls1_celular, criterion = 'Q2',xlab="Número de Componentes")


list.keepX <- c(5:10, seq(15, 50, 15))

tune.spls1.MAE_cel <- tune.spls(X= Celular_Basal[,-c(1,2,3,4,5)], Y =Celular_Basal[,5] , ncomp= 2, test.keepX = list.keepX,  validation = 'Mfold',
                folds = 10,nrepeat = 5,progressBar = FALSE, scale = TRUE,
```

```
                measure = 'MAE') #MSE, MAE,R2,Bias
plot(tune.spls1.MAE_cel)

choice.ncomp <- tune.spls1.MAE_cel$choice.ncomp$ncomp
choice.keepX <- tune.spls1.MAE_cel$choice.keepX[1:choice.ncomp]
choice.keepX2 <- c(rep(choice.keepX, 2))
names(choice.keepX2) <- c("Comp1", "Comp2")

spls2.cel <- spls(X= Celular_Basal[,-c(1,2,3,4,5)], Y =Celular_Basal[,5] , ncomp = 2, keepX = choice.keepX2, mode = "regression")

plotIndiv(spls2.cel,ind.names = FALSE, group = Celular_Basal$Cancer_Type,legend = TRUE)
plotVar(pls2.cel, cex = c(4,5), var.names = c(TRUE, TRUE),title = 'Loading Plot PLS')

# sPLS-DA
List.keepX <- c(1:10,  seq(20, 100, 10))
tune.splsda_anti<- tune.splsda(X=Anticuerpos_dummy[,-2],Y=Anticuerpos_dummy[,2], ncomp = 3, validation = 'Mfold', folds = 5, dist = 'max.dist', test.keepX = list.keepX, nrepeat = 50)
plot(tune.splsda_anti, sd = TRUE)
choice.ncomp_da <- tune.splsda_anti$choice.ncomp$ncomp
choice.keepX_da <- tune.splsda_anti$choice.keepX
select.keepX_SPCA <- tune.splsda_anti$choice.keepX[1: choice.ncomp_da]

splsda_anti <- splsda(X=Anticuerpos_dummy[,-2],Y=Anticuerpos_dummy[,2], ncomp = ncomp, keepX = select.keepX_SPCA)

plotIndiv(splsda_anti, comp = c(1,2),
      ind.names = FALSE,ellipse = FALSE, legend = TRUE,star = FALSE,
      title = 'sPLS-DA comp 1 - 2')
plotVar(splsda_anti, comp = c(1,2), cex = 3)
```