VNiVERSiDAD
D SALAMANCA

DEPARTAMENTO DE ESTADíSTICA

TESIS DOCTORAL

Desarrollo de un algoritmo de *Mínimos Cuadrados Parciales* para análisis de datos de chips de ADN usando el estadístico VIP para selección de genes y clasificación binaria

Francisco Javier Burguillo Muñoz

- 2015 -

Desarrollo de un algoritmo de *Mínimos Cuadrados Parciales* para el análisis de datos de chips de ADN usando el estadístico VIP para selección de genes y clasificación binaria

---

Memoria que para optar al grado de Doctor, por el Departamento de Estadística de la Universidad de Salamanca, presenta:

Francisco Javier Burguillo Muñoz

Salamanca – 2015

VNiVERSiDAD
D SALAMANCA

DEPARTAMENTO DE ESTADíSTICA

INMACULADA BARRERA MELLADO

Profesora Titular del Departamento de Estadística

Universidad de Salamanca

JAVIER MARTÍN VALLEJO

Profesor Titular del Departamento de Estadística

Universidad de Salamanca

CERTIFICAN:

Que Francisco Javier Burguillo Muñoz, Licenciado en Ciencias Químicas, ha realizado en el Departamento de Estadística de la Universidad de Salamanca, bajo su dirección, el trabajo que para optar al Grado de Doctor presenta con el título: **Desarrollo de un algoritmo de *Mínimos Cuadrados Parciales* para el análisis de datos de chips de ADN usando el estadístico VIP para selección de genes y clasificación binaria.**

Y para que conste, firman el presente certificado en Salamanca, en Marzo de 2015.

A mis padres y esposa, que me quisieron tanto, algo que nunca podré olvidar… estarán siempre conmigo.

A mis hijos Javier e Isabel, que han estado constantemente a mi lado, me han dado tantas alegrías, tanta ayuda y muchos ánimos para seguir siempre adelante.

En memoria de Luis Pérez del Villar Moro, por haber podido disfrutar de su amistad durante varios años, y por darme ejemplo de fortaleza y optimismo en la salud y en la enfermedad.

Agradecimientos

Mi primer agradecimiento es para la profesora Purificación Galindo, que no tuvo ningún inconveniente en aceptar a un químico en el Máster de Estadística Multivariante Aplicada, del que a la sazón era directora allá por el curso 2006-2007. Su amable acogida me alentó a aprender nuevas cosas y a terminar haciendo esta Tesis doctoral.

Muchas gracias también a Inmaculada Barrera y Javier Martín, que tuvieron la amabilidad de acceder a dirigir una tesis sobre análisis de datos de chips de ADN. Ellos me han ayudado muchas veces a adaptar mis costumbres de investigación experimental al estilo y formalismo de las matemáticas.

Vaya también mi recuerdo agradecido a todos los profesores del Máster de Estadística Multivariante que, durante el curso 2006-2007, nos enseñaron tantas técnicas estadísticas. Y, cómo no dar las gracias a todos los compañeros de aquel Máster, con aquellos cafés tan animados que tomábamos en los descansos.

Y gracias también a todos aquellos que me han ayudado en este trabajo: Luis Corchete, por su infatigable apoyo con la informática y la biología; Antonio de Juan, Gonzalo Abruña, Javier San Pablo, Pablo Tirado y Sergio de la Cruz, por su ayuda con el Fortran *y el código fuente. Gracias también a los doctores Norma Gutiérrez y Jesús San Miguel por facilitarme sus datos de* microarrays *y brindarme siempre sus sugerencias.*

Más que agradecimiento, digamos afecto y admiración, debo al profesor William G. Bardsley de la Universidad de Manchester (U.K.), con el que tuve la suerte de trabajar en el año ochenta y dos y cuya amistad y ayuda ha perdurado a lo largo de tantos años. Fue él quien me contagió su entusiasmo por el ajuste de curvas y la estadística aplicada. Gracias también por su generosidad en dejar libre todo el código del Paquete Estadístico SIMFIT, sin cuyas rutinas no hubiera sido posible este trabajo.

ÍNDICE GENERAL

1. INTRODUCCIÓN..... 1

2. OBJETIVOS..... 7

3. METODOLOGÍA..... 11

3.1. Tipos de datos..... 13

3.1.1. Datos de expresión génica en chips de ADN (*microarrays*)..... 13

3.1.2. Datos clínicos categóricos y continuos..... 13

3.2. Obtención de datos de expresión génica en chips de ADN..... 17

3.3. Tratamiento estadístico de datos de chips de ADN..... 27

3.3.1. Técnicas para la selección de genes diferencialmente expresados entre clases 27

3.3.2. Técnicas para el análisis exploratorio de los grupos experimentales..... 31

3.3.3. Métodos predictores con datos de *microarrays*..... 41

3.4 Método de Mínimos Cuadrados Parciales (*Partial Least Squares, PLS*) para análisis de datos clínicos y génicos..... 55

3.4.1. Qué es y para qué sirve PLS..... 55

3.4.2. Algoritmos más usuales en PLS..... 61

3.4.3. Interpretación de los modelos PLS..... 67

3.4.4 Determinación del número óptimo de factores latentes..... 79

3.4.5. PLS con variables categóricas..... 87

3.4.6. Aplicaciones de PLS en Genómica..... 89

3.4.7. Análisis conjunto de datos clínicos y génicos en Genómica..... 97

3.4.8. Programa *PLS-VIP* desarrollado en este trabajo.....103

3.5. Simulaciones con el programa *SIMDATA* desarrollado en este trabajo..... 113
3.5.1. Simulación de variables clínicas categóricas y continuas..... 113
3.5.2. Simulación de variables de expresión génica en *microarrays*..... 119

3.6. Programas estadísticos y de clasificación utilizados..... 125

4. RESULTADOS Y DISCUSIÓN..... 127

4.1. Resultados con datos simulados de *microarrays*..... 129
4.1.1. Escenarios simulados para probar el programa *PLS-VIP*..... 129
4.1.2. Funcionamiento del algoritmo *PLS-VIP*..... 131
4.1.3. Análisis de *PLS-VIP* bajo diferentes escenarios..... 141
4.1.4. Comparación de *PLS-VIP* con otros métodos de predicción..... 145

4.2. Resultados de *PLS-VIP* con datos reales de microarrays..... 149
4.2.1. *"Macroglobulemia de Woldenstrom frente a Leucemia Linfocítica Crónica"* y *"Mieloma Múltiple con ganancia 1q frente a Mieloma Múltiple sin ganancia 1q"*..... 149

4.3. Construcción de modelos predictores que combinan variables clínicas y génicas usando el algoritmo *PLS-VIP-Consecutivo*..... 153

4.3.1 Predictores clínicos:Escenarios simulados con variables clínicas... 153
4.3.2 Predictores genómicos: Escenarios simulados con variables génicas de *microarrays*..... 163
4.3.3 Predictores clínico-genómicos: Combinación de los modelos óptimos obtenidos en los apartados 4.3.1 y 4.3.2. 171
4.3.4 Estudio comparativo de *PLS-VIP-Consecutivo* frente a otros métodos predictivos usando variables clínicas más génicas..... 185

4.4. Comportamiento de *PLS-VIP-Consecutivo* con datos reales clínicos y génicos combinados en un predictor..... 189

4.4.1. *Pacientes con Mieloma Múltiple tratados con 6 ciclos de quimioterapia categorizados como respuesta incompleta (RI) frente a respuesta completa (RC)*..... 189

5. CONCLUSIONES..... 201

6. BIBLIOGRAFÍA..... 209

ÍNDICE DE TABLAS

Tabla 1. Escenarios simulados con sólo variables clínicas....118

Tabla 2. Escenarios simulados con datos de *microarrays*.... 123

Tabla 3. Escenarios simulados con sólo variables génicas.... 124

Tabla 4. Funcionamiento del propuesto algoritmo *PLS-VIP*....132

Tabla 5. Prueba del algoritmo *PLS-VIP* bajo diferentes escenarios simulados....144

Tabla 6. Comparación de métodos habituales de clasificación frente a *PLS-VIP* usando datos simulados....148

Tabla 7. Comparación de diferentes métodos de clasificación frente a *PLS-VIP* usando datos reales.... 152

Tabla 8. Aplicación de *PLS-VIP* a escenarios simulados con sólo variables clínicas....158

Tabla 9. *PLS-VIP* con distintos escenarios simulados con sólo variables génicas....166

Tabla 10. *PLS-VIP* sobre escenarios simulados con variables clínicas y génicas unidas en sus óptimos....175

Tabla 11. Proporciones de error de clasificación en diferentes escenarios y comparación con la prueba U de Mann-Whitney....180

Tabla 12. Comparación de métodos de clasificación frente a *PLS-VIP-Consecutivo* promediando 50 repeticiones de datos simulados de un escenario....187

Tabla 13. Comparación de métodos de clasificación frente a *PLS-VIP* usando datos de Mieloma con sólo variables clínicas....193

Tabla 14. Comparación de métodos de clasificación frente a *PLS-VIP* usando datos de Mieloma con sólo variables génicas....195

Tabla 15. Comparación de métodos de clasificación frente a *PLS-VIP* usando datos de mieloma con variables clínicas + génicas génicas....197

Tabla 16. Recopilación de los errores de clasificación observados con las tres estrategias de variables usadas....199

ÍNDICE DE FIGURAS

Figura 1. Esquema de un *spotted microarray*..... 18
Figura 2. Fabricación de un *microarray* de oligonucleotidos 19
Figura 3. Esquema de los *microarrays* de dos canales y de un canal..... 20
Figura 4. Esquema del proceso completo de varios *microarrays* de un canal (Affymetrix)..... 21
Figura 5. Diagrama de flujo para el análisis de datos en *microarrays*..... 21
Figura 6. Ejemplo de una matriz final con datos de varios chips de ADN..... 25
Figura 7. Ejemplo de dendrogramas con datos de *microarrays* de mieloma..... 32
Figura 8. Ejemplo de gráficos 2D obtenidos mediante MDS con datos de mieloma..... 34
Figura 9. Ejemplo de biplot con datos reales de mieloma..... 37
Figura 10. Esquema general de los resultados que proporciona PLS..... 67
Fig. 11. Varianza explicada en promedio frente al número de factores..... 69
Fig. 12. Puntuaciones para las variables **X** bajo los 2 factores PLS..... 70
Figura 13. Cargas de las variables **X** para los dos factores PLS..... 71
Figura 14. Puntuaciones para las **Y** en los dos factores PLS..... 71
Figura 15. Cargas para las variables Y bajo los 2 factores PLS..... 72
Figura 16. Correlaciones de las puntuaciones t y u para los 2 factores PLS..... 73
Figura 17. Número de factores frente a la "varianza Y residual"..... 85
Figura 18. Número de factores frente a la "varianza Y explicada"..... 86
Figura 19. Diagrama de flujo de *PLS-VIP* en su opción de datos de *microarrays*..... 107
Figura 20. Superior: opciones de PLS. Inferior: Opciones de resultados..... 108
Figura 21. Diagrama de flujo del análisis en secuencia de datos clínicos y génicos y clínicos + génicos..... 111
Figura 22. Opciones de entrada de datos en programa para variables dicotómica..... 115
Figura 23. Ejemplo de simulación de variables dicotómicas..... 115
Figura 24. Opciones de entrada de datos para variables clínicas continuas..... 117
Figura 25. Valores simulados de entrenamiento y prueba con variables continuas..... 117
Figura 26. Pantallas de entrada de datos para las expresiones génicas..... 121
Figura 27. Serie de entrenamiento para las expresiones génicas..... 122

Figura 28. Funcionamiento del propuesto algoritmo *PLS-VIP*.
(A) Variación del nºde factores en dos iteraciones.
(B) Variación del nº de genes con las iteraciones..... 134
Figura 29. Puntuaciones y cargas de las variables **X** para la iteración 5 y 2 factores
PLS con la serie de entrenamiento y un escenario potencia baja..... 136
Figura 30. Puntuaciones de las variables **Y** para la iteración 5 y 2 factores
PLS con la serie de entrenamiento y un escenario potencia baja..... 137
Figura 31. Puntuaciones y cargas de las variables **X** para la iteración 6 y 2 factores
PLS con la serie de entrenamiento y el escenario de potencia media..... 138
Figura 32. Representaciones de puntuaciones u_i frente a t_i
(A) para el factor 1 y (B) para el factor 2..... 140
Figura 33. Puntuaciones y cargas de las variables clínicas **X** de la
serie 17 de CLIN4..... 160
Figura 34. Puntuaciones de las variables respuesta **Y** de la serie 17 de CLIN4..... 161
Figura 35. Representaciones $\mathbf{u_i}$ frente a $\mathbf{t_i}$ para dos factores de la serie 17 de
CLIN4. (A) Factor 1 y (B) factor 2..... 162
Figura 36. Puntuaciones y cargas para las variables **X** de la serie 17 de GEN3..... 168
Figura 37. Representaciones de $\mathbf{u_i}$ frente a $\mathbf{t_i}$ para los factores la serie 17 de GEN3..... 170
Figura 38. Puntuaciones de las variables **X** para la serie 17 de CLIGEN4-3..... 182
Figura 39. Cargas de las variables **X** para la serie 17 de CLIGEN4-3..... 183

ABREVIATURAS

PLS: *Partial Least Squares*

PLSR: *Partial Least Squares Regression*

VIP: *Variable Influence in Projection*

PLS-VIP: *Partial Least Squares-Variable Influence in Projection*

PLS-VIP-Consecutivo: *Partial Least Squares-Variable Influence in Projection-aplicado consecutivamente a diferentes tipos de variables.*

NIPALS: *Nonlinear Iterative Partial Least Squares*

PCA: *Principal Component Analysis*

DA: *Discriminant Analysis*

LDA: *Linear Discriminant Analysis*

PAM: *Prediction Analysis of Microarrays*

SVM: *Support Vector machines*

RF: *Random Forest*

FDR: *False Discovery Rate*

CA: *Cluster Analysis*

MDS: *Multi Dimensional Scaling*

dif. exp.: *Genes diferencialmente expresados*

PEC: *Proporción de Error de Clasificación*

FP: *Falso positivo*

PGFP: *Proporción de genes falsos positivos*

M (Q_1, Q_3): *Mediana (primer cuartil, tercer cuartil)*

CFS: *Correlation Feature Selection*

LOO: *Leave one out*

CV (3 fold): *Cross Validation (3 fold)*

WM: *Waldenström Macroglobulemia*

MM: *Multiple Mieloma*

CLL: *Chronic Lymphocytic Leukemia*

BL: *B Lymphocytes*

PC: *Plasma Cells*

RC: *Respuesta Completa*

RI: *Respuesta Incompleta*

RR: *Riesgo relativo*

TE: *Tamaño del efecto*

NA: *No disponible*

1. INTRODUCCIÓN

1. INTRODUCCIÓN

La presente Tesis doctoral se justifica por el interés despertado en los últimos años por el análisis de los datos obtenidos en los llamados chips de ADN o *microrarrays*. Esta nueva tecnología ha supuesto un nuevo paradigma en la investigación biomédica, especialmente en el estudio del cáncer. Estos chips miden los niveles de expresión de miles de genes simultáneamente, que se usan luego para caracterizar el perfil génico de las enfermedades, la respuesta a los tratamientos o la evaluación de pronósticos. Una de las aplicaciones más comunes consiste en comparar los niveles de expresión génica en dos condiciones diferentes, tal como células sanas frente a células tumorales. Normalmente se acometen dos tareas: la identificación de genes marcadores de la enfermedad (*gene selection*) y la construcción de modelos predictivos que permitan la asignación de nuevos pacientes a los grupos analizados (*class prediction*).

La dificultad que surge con este tipo de datos es que el número de casos disponibles (del orden de 40) es mucho menor que el de variables (del orden de 20000 o más), lo que conlleva una matriz de predictores $\mathbf{X\ (n \times m)}$, donde $\boldsymbol{n}$ es el número de casos y $\boldsymbol{m}$ el número de variables de expresión génica, siendo $\boldsymbol{n} << \boldsymbol{m}$ y donde las variables $\boldsymbol{m}$ podrían estar correlacionadas y presentar problemas de multicolinealidad. Esta característica hace que muchos métodos multivariantes no sean directamente aplicables y haya que proceder bien a una selección de variables buscando aquellos genes que estén más diferencialmente expresados (por ej. con *t-Student*), o bien a una reducción de la dimensión mediante factores latentes (por ej. con *Componentes Principales, PCA*). Otro inconveniente añadido es que los genes más diferencialmente expresados no forman

necesariamente el conjunto más discriminante a efectos de predicción de clase, ya que normalmente dichos genes suelen estar muy correlacionados, de manera que otro conjunto de genes con menor expresión diferencial podría servir mejor para construir un modelo predictivo. La reducción de la dimensionalidad podría resolver este problema, pero los factores latentes no resultan informativos para la selección de genes y tienen una interpretación biomédica limitada.

El presente trabajo se centra principalmente en la construcción de modelos predictivos con datos de microarrays y su motivación ha sido la de implementar un algoritmo que realice, a la vez, las dos tareas arriba mencionadas: selección de los genes más discriminantes entre clases y obtención de un modelo predictivo de dimensión reducida con el número óptimo de factores latentes.

Han sido varias las técnicas estadísticas que se han propuesto para construir modelos predictivos con datos de microarrays, entre ellas ha cobrado gran interés recientemente la *Regresión por Mínimos Cuadrados Parciales* (en inglés *Partial Least Squares, PLS*). Esta técnica es un método multivariante originalmente propuesto por Herman Wold a finales de los sesenta y que ha sido aplicada en Quimiometría y Monitorización de Procesos, con el fin de realizar una reducción de la dimensión de variables en unos casos y multicalibracion en otros.

PLS es una técnica útil cuando una matriz de respuestas $\mathbf{Y(n\times 1)}$ es observada con una matriz de variables predictoras $\mathbf{X(n\times m)}$. A diferencia de otras técnicas de reducción de la dimensión, como PCA, la aproximación PLS calcula cada variable latente a partir de **X** pero basándose en **Y**. El objetivo es maximizar la covarianza de **Y** con **X**, a diferencia de PCA que maximiza la varianza de las variables ***x*** solamente. La idea que subyace en

PLS es expresar las matrices **X** e **Y** en términos de una serie de ***k*** factores latentes, obtenidos a partir de las matrices **X** e **Y** por técnicas de proyección y regresión. Una vez que se han obtenido las expresiones que aproximan **X** e **Y** usando los factores latentes, estos se pueden utilizar para tratar **X** como una matriz de entrenamiento, de la que poder predecir qué nueva **Y** resultaría de una nueva **X** que esté expresada en las mismas variables que la matriz **X** de entrenamiento.

La propia metodología de PLS dispone de parámetros que pueden servir para la selección de variables, como son los pesos de las variables ***x (w)***, los coeficientes de regresión ***(β)*** o las puntuaciones del estadístico ***VIP*** (del inglés:*Variable Influence on Projection*). Algunos de estos parámetros se han probado ya con datos de *microarrays*, pero todavía era conveniente y motivador seguir investigando su potencia en distintos escenarios de tamaño de muestra, número y potencia discriminante de los genes discriminantes, número de genes ruido, etc., aspectos que se pensó analizar en el presente trabajo. Así mismo, otra motivación, de interés en clínica, ha sido investigar si los datos de microarrays suponen una información adicional a las variables clínicas clásicas a la hora de construir modelos predictivos, o si por el contrario las variables clínicas son suficientes.

Por los motivos anteriormente expuestos, en esta Tesis se ha implementado un nuevo algoritmo PLS con selección de variables mediante el estadístico VIP, que de modo iterativo optimiza la selección de variables y el número de factores PLS del modelo predictor, con el fin de obtener un valor mínimo del error de clasificación en el caso de clasificación binaria que es la habitual en Biomedicina. Este algoritmo se ha probado de forma sistemática tanto con datos simulados como con datos reales. Se ha investigado

también el funcionamiento del algoritmo con modelos predictores que combinan variables clínicas y génicas simultáneamente. Por último, este nuevo algoritmo se ha comparado con otros procedimientos de clasificación habituales en Genómica, tales como *K Nearest Neighbours (KNN)*, *Prediction Análisis of Microarrays (PAM)*, *Support Vector Machines (SVM)* o *Random Forest (RF)*.

2. OBJETIVOS

2. OBJETIVOS

Objetivo general

Implementar un programa dedicado al análisis de datos de microarrays que esté basado en la Regresión por Mínimos Cuadrados Parciales (*Partial Least Squares, PLS*). El programa ha de permitir la reducción de la dimensionalidad mediante la selección de las variables con mayor poder predictivo usando las puntuaciones del estadístico VIP, así como la determinación del número óptimo de factores latentes PLS. Estará diseñado para operar de forma iterativa hasta alcanzar el mínimo de error de predicción para el caso de una clasificación supervisada con dos categorías. La bondad del modelo se medirá a través de dos series independientes de datos, una de *entrenamiento* y otra de *validación*, sin recurrir al método clásico de validación cruzada que presenta mayores sesgos.

Objetivos específicos

- Escribir el código fuente de las rutinas que formaran el nuevo programa antes mencionado *(PLS-VIP)*. Se utilizará el lenguaje Fortran 95 para Windows y se hará uso de diferentes subrutinas existentes en las dlls del Paquete Estadístico *SIMFIT*.

- Desarrollar un programa de simulación de datos de microarrays y de datos clínicos *(SIMDATA)*. Se utilizarán técnicas de simulación al azar haciendo uso de las distribuciones binomial, uniforme y normal a través de subrutinas *SIMFIT*.

- Analizar el comportamiento del nuevo algoritmo *PLS-VIP con* datos de microarrays bajo diferentes escenarios simulados, contemplando diferentes tamaños de muestra, número de genes discriminantes y su potencia discriminante, así como diferente número de genes basales o genes "ruido".

- Comparar la bondad de la nueva propuesta de PLS, con selección por las puntuaciones VIP, frente a otros algoritmos predictivos de uso habitual en Genómica.

- Probar el nuevo algoritmo con datos reales de microarrays en pacientes con cáncer.

- Construir modelos predictores *PLS-VIP* que combinen a la vez variables clínicas clásicas y génicas de microarrays, con el fin de analizar en qué condiciones los datos génicos añadirían potencia predictora a las variables clínicas. Tanto con datos simulados como reales.

3. METODOLOGÍA

3.1. Tipos de datos

Conviene empezar describiendo el tipo de datos que se van a utilizar en el presente trabajo, ya que de su naturaleza dependerán las propiedades de las técnicas estadísticas que se utilicen con ellos.

3.1.1. Datos de expresión génica en chips de ADN (*microarrays*)

Estos datos miden la cuantía de la hibridación entre la sonda y el gen correspondiente, hibridación que es medida por fluorescencia a través de un escáner y su valor es expresado normalmente en log(2). Se trata, por tanto, de datos de tipo continuo de intervalo, que normalmente se mueven en un margen de 4.0 a 12.0. Los datos así expresados constituirán la matriz de predictores $\mathbf{X(n \times m)}$ de los métodos multivariantes, donde $\boldsymbol{n}$ es el número de casos y $\boldsymbol{m}$ el número de variables de expresión génica (genes), siendo $\boldsymbol{n} >> \boldsymbol{m}$.

En cuanto al vector de respuestas **Y**, estará referido en el presente trabajo a datos de clasificación binaria, que es una de las aplicaciones más frecuentes en Biomedicina. Cada sujeto pertenecerá a una de dos clases posibles (por ej. control o tumor), que estarán indicadas por un vector $\mathbf{Y(n \times 1)}$ que usará codificación ficticia, asignando **0** a las muestras de control y **1** a las muestras de tumor. Estos aspectos se describirán con mayor detalle en el apartado 3.2.

3.1.2. Datos clínicos categóricos y continuos

Desde el principio uno de los grandes paradigmas con datos de *microarrays* fue el poder usar la expresión de ciertos genes marcadores para construir modelos

predictores con ellos, normalmente para clases de tipo dicotómico (por ej. metástasis si, metástasis no). Se han intentado desarrollar algoritmos que funcionaran con bajos errores de predicción, pero en numerosas ocasiones los genes variaban entre unos estudios y otros, de tal forma que, modelos predictores que funcionaban bien en unas determinadas condiciones no han podido ser validados a veces en otras condiciones semejantes. Por otra parte, la tecnología de *microarrays* es todavía cara comparada con los factores pronóstico convencionales.

Teniendo en cuenta los anteriores condicionantes, han existido siempre dos preguntas a las que dar respuesta: a) ¿Las variables clínicas no serán suficientes y más asequibles como predictoras que los datos génicos de los *microarrays*?, b) ¿No podrían complementar los datos de *microarrays* a las variables clínicas y conseguir así mejores predictores? Para tratar de responder a estas preguntas se han estudiado también en este trabajo las variables clínicas, las cuales se describen brevemente a continuación.

Los datos clínicos que se utilizarán en el presente trabajo serán de dos tipos: continuos y categóricos. La mayoría de los datos continuos en la práctica clínica suelen ser variables de tipo razón, como edad, presión sistólica, concentración de calcio, tamaño del tumor, etc. También puede haber variables continuas de tipo intervalo como temperatura en grados centígrados o pH de la sangre u orina. Por elección de las unidades de medida, los clínicos utilizan sus valores de variables continuas en un margen aproximado de 1 a 80, y estos serán los límites que se tendrán de referencia para las simulaciones de estas variables en el presente trabajo.

En cuanto a las variables categóricas, éstas son frecuentemente de tipo dicotómico, tales como el género (hombre, mujer), sobrevive (si o no), tabaquismo (si o no), etc. Si bien puede haber variables categóricas nominales con más de dos categorías, tales como grupo sanguíneo o incluso variables ordinales con significado de gradación, como el tipo de dolor (leve, moderado, agudo). En el presente trabajo se estudiará principalmente las variables de tipo dicotómico, asignando valores ficticios de **0** para una clase y **1** para la otra.

Se analizarán diferente número de variables clínicas, combinando como máximo 4 variables categóricas dicotómicas y 6 variables continuas de tipo razón en los estudios simulados, si bien se permitirán hasta 4 variables categóricas con dos o más categorías y 10 continuas en los estudios con datos reales.

3.2. Obtención de datos de expresión génica en chips de ADN

Una descripción completa de esta tecnología de chips de ADN está fuera del alcance del presente trabajo, pero si resulta conveniente el presentar los aspectos más relevantes de dicha técnica (para más detalles ver Stekel (2003)).

Tipos de chips de ADN

Un chip de ADN consiste de una matriz ordenada de miles de pocillos microscópicos de clones ADN u oligonucleótidos, cada uno conteniendo unos picomoles de una secuencia específica de ADN. Estos segmentos de ADN se utilizan como sondas para hibridar las muestras de cADN o cRNA (llamadas dianas). Esta hibridación sonda-diana es cuantificada a partir de la fluorescencia de las dianas unidas, que han sido previamente marcadas con un compuesto fluorogénico.

Los *microarrays* se pueden fabricar usando principalmente dos tipos de tecnologías. Así, en "spotted *microarrays*" las sondas son cADN o pequeños fragmentos de productos PCR que corresponden a mARNs. Las sondas son sintetizadas y luego depositadas (*spotted*) sobre una placa de vidrio. Para ello se utilizan normalmente pequeñas agujas controladas por un brazo robótico como se aprecia en la Figura 1.

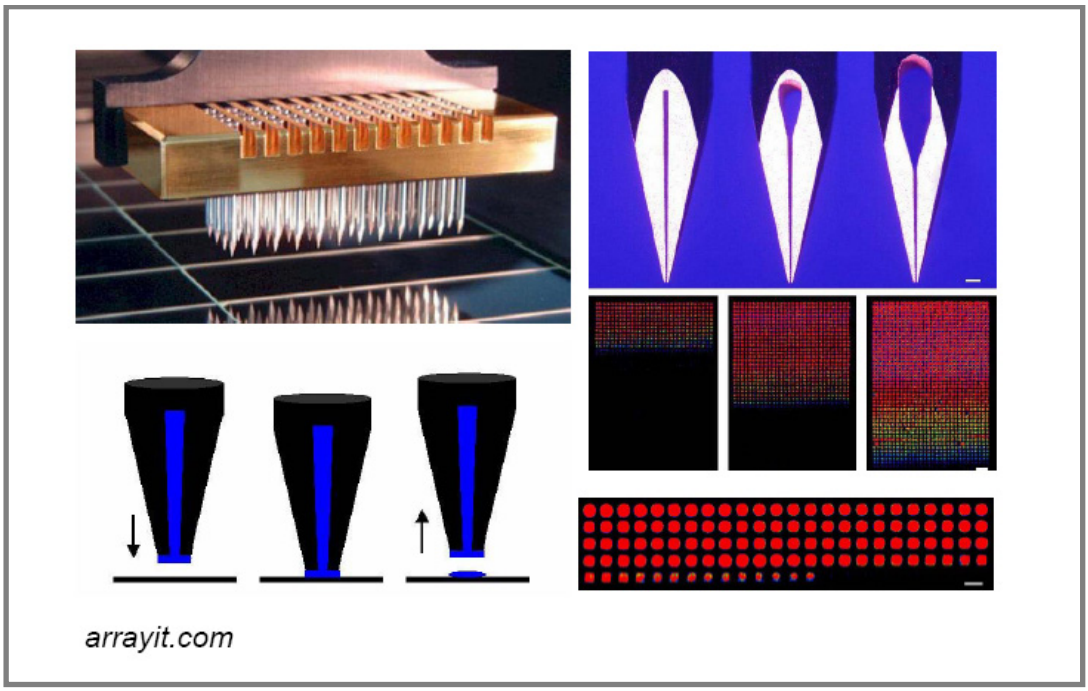

Figura 1. Esquema de un "spotted microarray" (www.arrayit.com).

Por el contrario, en los *microarrays* de oligonucleótidos las sondas son secuencias cortas diseñadas para hibridarse con partes de la secuencia conocida de un gen. Estas sondas se construyen principalmente sintetizando directamente la secuencia deseada sobre la superficie del chip en lugar de depositar las secuencias completas sobre la superficie sólida. Las secuencias son normalmente oligos de 25-mer (Affymetrix). La Figura 2 muestra un esquema de cómo se sintetizan los oligonucleótidos de Affymetrix sobre la superficie de vidrio, usando para ello una máscara y un procedimiento de fotolitografía.

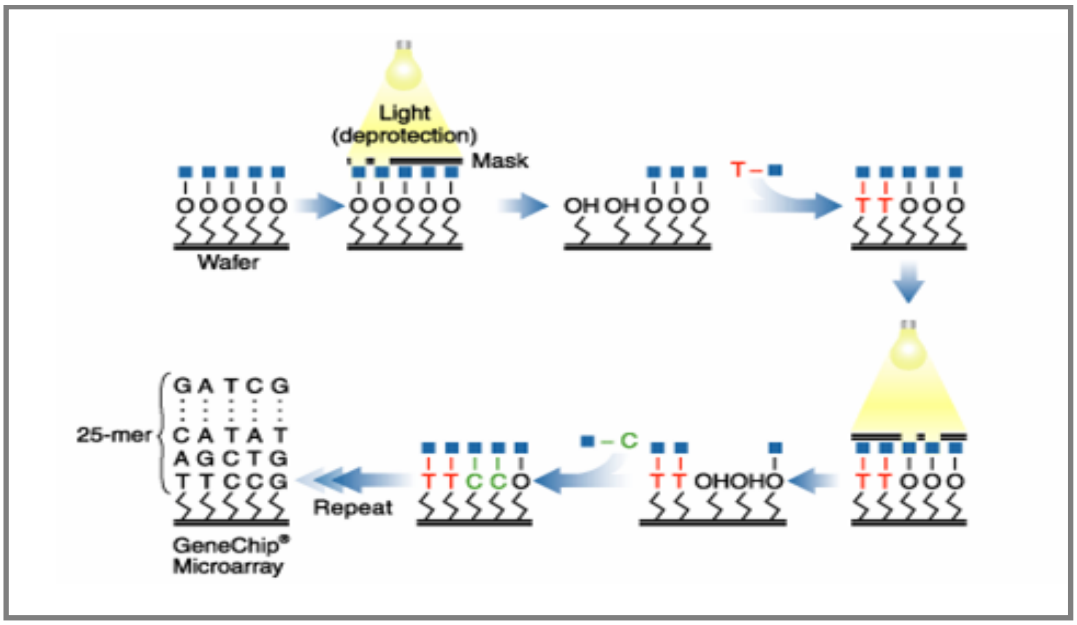

Figura 2. Fabricación de un *microarray* de oligonucleotidos (www.affymetrix com).

Con respecto a la detección de un chip tenemos la opción de "dos canales" frente a la de "un canal". Los *microarrays* de "dos canales" o "dos colores" son normalmente hibridados con cADN utilizando dos muestras a comparar (por ej. tejido tumoral frente a tejido sano). Las muestras están marcadas con dos diferentes fluorógenos (por ej. los colorantes Cy3 y Cy5). Las dos muestras marcadas son mezcladas e hibridadas en un único *microarray* de cDNA, que es luego escaneado con un detector de fluorescencia. Las intensidades relativas de cada fluoróforo se analizan luego como un cociente para identificar genes sobre-expresados o infra-expresados. Empresas que fabrican este tipo de *microarrays* incluyen Agilent, Eppendorf y Arrayit.

En los "*microarrays* de un canal" o "*microarrays* de un color", los *arrays* están diseñados para dar los valores de los niveles absolutos de la expresión génica. Por tanto, la comparación de las dos muestras requiere dos hibridaciones por separado (control y

caso). Como solamente se usa un colorante en cada ensayo (biotina marcada), los datos que se obtienen representan valores absolutos de la expresión génica. Estos valores usualmente son referenciados a sondas de control normalizantes (ARN incluido en el microarray (*spike in*)), con el fin de calibrar los datos a lo largo de cada *array* y entre varios *arrays* entre sí. En la Figura 3 se muestra un esquema de los dos procedimientos anteriores.

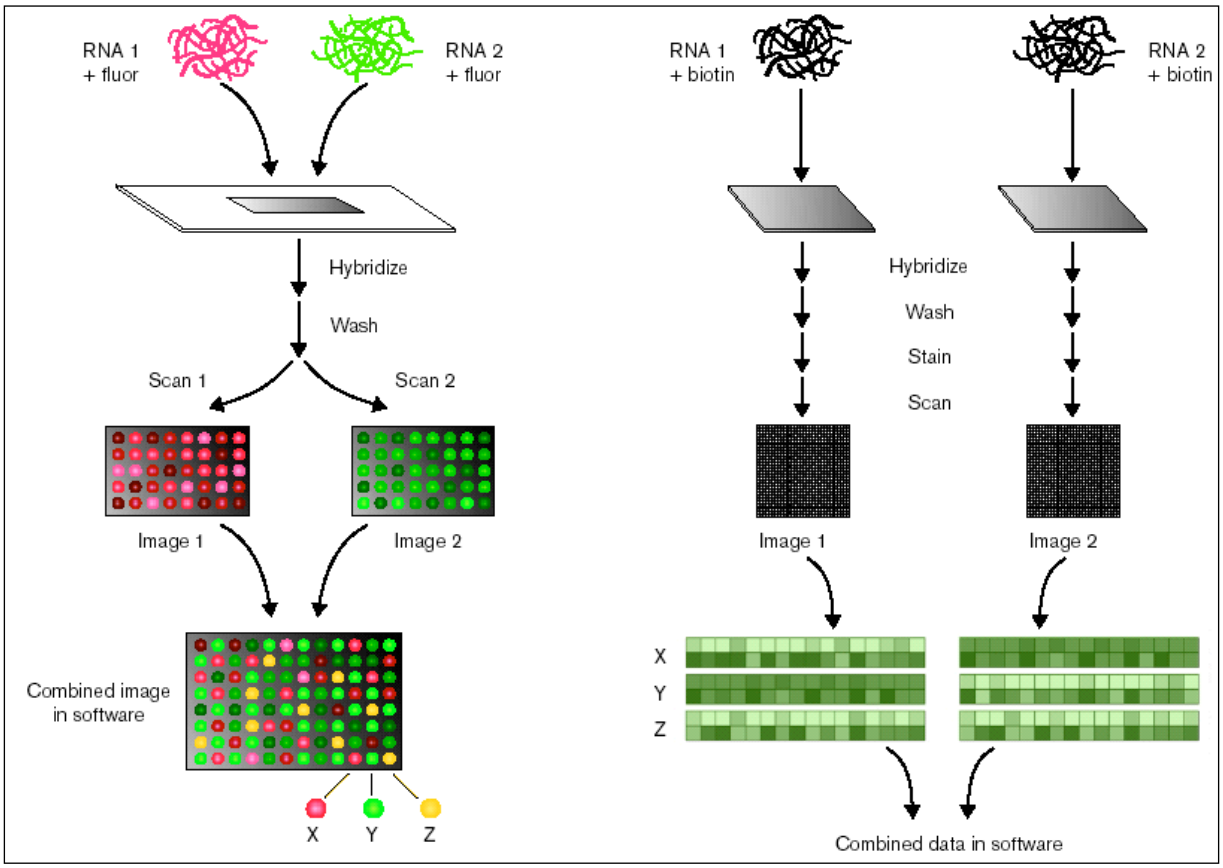

Figura 3. Esquema de los *microarrays* de dos canales y de un canal.

Los *microarrays* de un canal más populares son los de la serie "GeneChip" de Affymetrix. En el presente trabajo todas las simulaciones y análisis de datos se referirán a estos chips de Affymetrix. Un esquema del proceso completo de esta tecnología se muestra en la Figura 4 (recuérdese que se hace un chip para cada muestra y que finalmente el software del equipo construye una matriz final con los datos de todos los chips analizados en un estudio (controles y pacientes)).

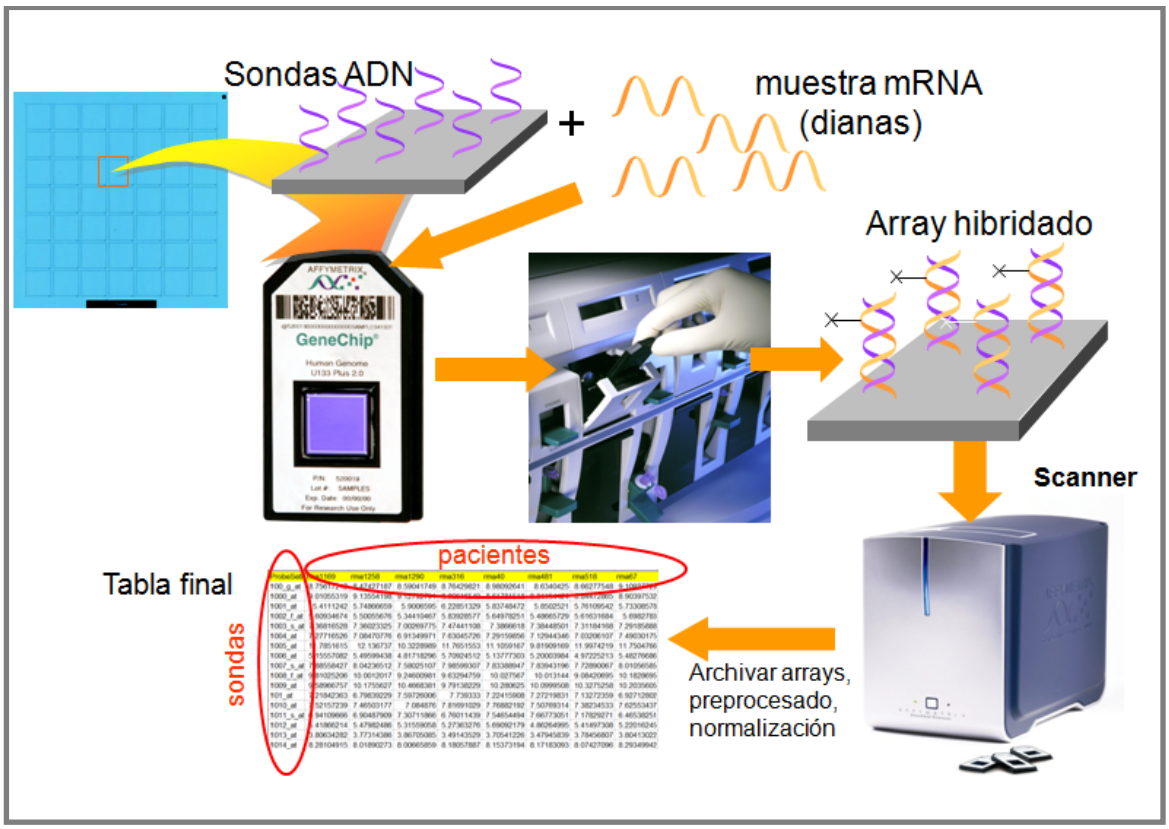

Figura 4. Esquema del proceso completo de varios *microarrays* de un canal (Affymetrix).

Etapas en la obtención de los datos de un chip de ADN

La obtención de datos incluye varios procesos que se resumen en la Figura 5.

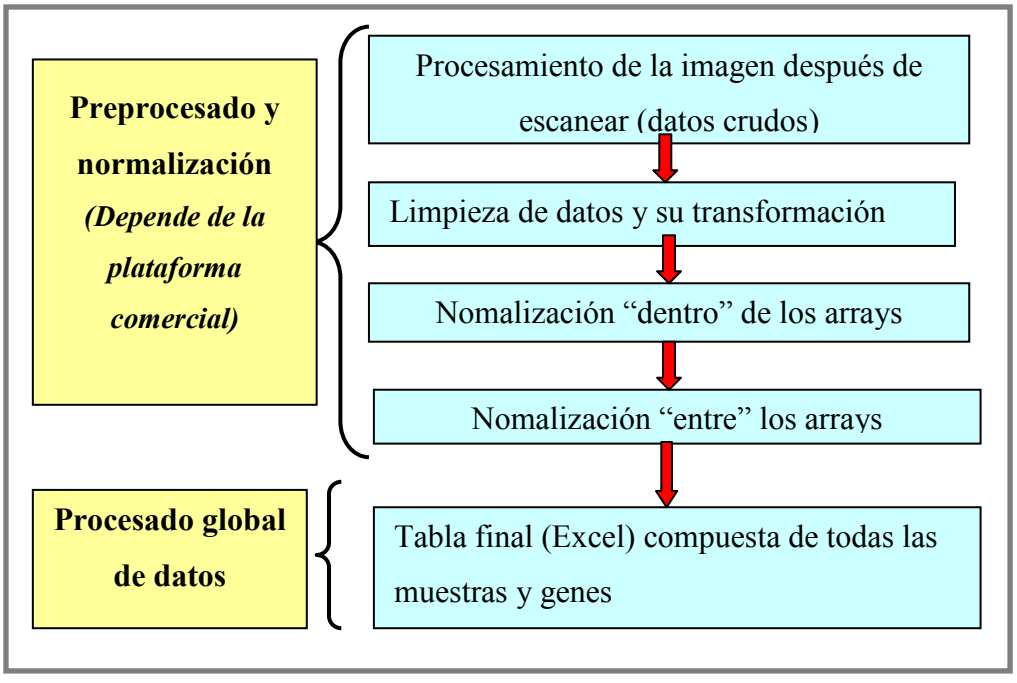

Figura 5. Diagrama de flujo para el análisis de datos en *microarrays*.

El preprocesado y normalización son dos mundos en sí mismos, ya que están íntimamente unidos al diseño del chip, escáneres y otros aspectos técnicos. Esta temática se aparta del ámbito del presente trabajo, pero algunos de los aspectos más importantes se exponen a continuación

Procesamiento de la imagen

Los algoritmos conocidos como "feature extraction software", convierten la imagen capturada por el escáner en números que cuantifican la expresión génica. Este software depende de la plataforma comercial utilizada. En los cDNA de dos colores existen cuatro pasos (Stekel (2003)):

a) Identificar la posición de las manchas en el microarray.

b) Identificar los "pixels" en la imagen que son propiamente parte de cada mancha.

c) Para cada mancha, identificar los "pixels" que se usaran para el cálculo del fondo.

d) Calcular los números que representan la intensidad de la mancha, la intensidad del fondo y la calidad de la información.

En la tecnología de Affymetrix los algoritmos del procesamiento de imagen han sido integrados en el proceso experimental de sus *microarrays*, de forma que el usuario no tiene que tomar ninguna decisión.

Limpieza de datos y su transformación

El objetivo de estos dos procedimientos, es el de limpiar y transformar los datos generados por el software de lectura de la imagen, antes de que cualquier normalización tenga lugar. En esencia abarca tres fases (Stekel (2003)):

a) En los arrays de 2 colores hay que eliminar las manchas marcadas como negativas, oscuras, etc.

b) En los *arrays* de 2 colores hay que restar el fondo (restar la señal del fondo de la intensidad de la mancha). Esta señal de fondo se considera que representa la contribución de hibridaciones no específicas. Los *arrays* de un color de Affymetrix sufren de un problema similar, en este caso la expresión génica se determina comparando la señal de la hibridación con la señal de sondas control de hibridación ("mismatched probes") que son oligos en los que una base es remplazada por una base falsa, de forma que la expresión génica se calcula en función de las diferencias entre las sondas verdaderas y sus correspondientes sondas control.

c) Tomar logaritmos. Esta es una práctica común para transformar los datos de un microarray, que están en intensidades brutas de fluorescencia, a sus correspondientes logaritmos antes de proceder a ningún otro tipo de análisis. En los arrays de 2 colores lo que se calcula es el log(2) del cociente de las intensidades de cada gen en las dos muestras (tumor, control), mientras que en los arrays de un color se calcula el log(2) de la intensidad de cada gen en cada muestra por separado. En base a este criterio todos las variables de expresión génica que se usarán en el presente trabajo corresponderán a valores expresados en log(2) relativos a *microarrays* de un color tipo Affymetrix.

Normalización dentro de los microarrays

Este proceso solo tiene relevancia para los *arrays* de dos colores tipo "spotted microarrays". Así, cuando se quiere medir la expresión diferencial entre genes para el

caso de dos muestras mediante la técnica de dos canales, se ha de eliminar cualquier sesgo y error introducido por el método experimental (Stekel (2003)).

Normalización entre arrays

Este apartado es aplicable tanto para los *arrays* de dos colores tipo "spotted microarrays" como los de un color tipo Affymetrix. Este tipo de normalización surge con el fin de comparar las muestras hibridadas a diferentes *arrays* bajo una misma referencia, es necesario por tanto corregir por la variabilidad introducida al utilizar diferentes arrays. Stekel (2003) recomienda empezar haciendo diagramas de caja de todos los genes en los diferentes arrays. Después de visualizar los datos en los diagramas de caja, el usuario debe decidir si aplica o no alguno de los dos métodos estándar. Ambos parten de la misma hipótesis: la variación en las distribuciones de los genes son el resultado de condiciones experimentales y no representan variabilidad biológica. Estos métodos estándar son:

a) *Centrado*: cuando los datos originales son transformados restando la media del *array* del valor de cada gen particular. Por tanto, las medias de todos los *arrays* se hacen igual a cero.

b) *Centrado y escalado*: A cada medida del *array* se le resta la media del *array* y se le divide por la desviación estándar, de modo que las nuevas medias de todos los *arrays* serán cero y sus desviaciones estándar serán todas iguales a uno. Este es el método que se usa normalmente para normalización de *microarrays*.

Tabla final

Una vez preprocesados y normalizados la serie completa de chips del estudio, se compone una tabla con todos los *arrays* y las intensidades de los genes expresadas

logarítmicamente, bien como el log(2) del cociente de la expresión génica en dos muestras (tumor, control) en el caso de chips de 2 colores, o como el log(2) de cada muestra por separado para los chips de un color, como se hace en el presente trabajo. Esta tabla es la matriz de partida para el análisis estadístico propiamente dicho.

A modo de ejemplo, véase la Figura 6 relativa a los datos reales de una serie de chips de Affymetrix en su forma final, en la que las filas se refieren a la expresión de los

	A	B	C	D	E	F	G	H	I	J	K	L
1	#Trying set from Leipzig in CROSSMANN 2005											
2	#Name	rma1169	rma1258	rma1290	rma316	rma40	rma481	rma518	rma67	rma695696	rma70	rma96
3	UNIGascendir	R	NR	NR	NR	R	R	R	NR	R	NR	R
4	HS.102598	9.79588461	9.79669311	9.4621595	10.169772	9.7130686	9.80144403	9.77101976	9.84276517	9.76343596	9.65126101	9.7766686
5	HS.10458	6.37606614	6.41981272	6.41318329	6.36875797	6.5590354	6.33426426	6.62449155	6.33709293	6.65773544	6.37163814	6.49727132
6	HS.104636	5.72245706	5.96944638	5.37371549	6.24287894	5.64087029	5.73205488	5.65837533	5.98181498	5.73258288	5.43213554	5.87450716
7	HS.106469	5.36910815	5.59793216	5.02237572	5.54376367	5.74205873	5.23212315	5.13196165	6.05746363	5.44337363	5.65584905	5.598189
8	HS.106674	8.46312247	8.67640297	8.40549515	8.77674956	8.59837582	8.63524835	8.4920633	8.80345121	8.53973879	8.48408198	8.30155383
9	HS.106876	10.7014263	10.4042649	10.2395934	10.0149125	10.1538023	10.051292	10.3612531	10.7100687	9.94711561	10.3124383	10.2318866
10	HS.109059	6.97547697	7.09965667	7.50917069	7.81484269	7.42524479	7.3224216	7.48691664	7.21173739	7.09298469	7.33923587	7.77672311
11	HS.109760	8.316625	8.37146622	8.45658368	7.70216145	8.1648703	8.50782731	8.38974019	6.84206112	8.8676988	8.3475094	8.52645542
12	HS.112023	6.48708179	6.53043444	6.5648483	6.43797181	6.55434504	6.65944424	6.64184312	6.56692089	6.4112731	6.9344522	6.59289079
13	HS.113094	7.26976659	7.63014216	7.51156223	7.82363126	7.70757539	7.99355871	8.19824592	6.84517636	7.66953276	7.65903761	8.09910633
14	HS.113290	4.98424662	5.09938979	5.00193554	5.01573722	5.3842097	4.99670206	4.98171814	5.25060282	5.13545043	5.21481932	4.95367865
15	HS.11392	3.3175927	3.22034926	3.31795512	3.33283612	3.40390357	3.35553576	3.45609279	3.59052575	3.46012479	3.35224203	3.36099335
16	HS.11417	10.0797264	10.4747263	10.1410411	9.79629877	10.0454949	9.67224055	9.99864788	9.70751716	9.6070036	10.0358205	9.94893158
17	HS.118110	8.48995981	8.58313112	8.66636536	7.9248339	8.33263098	7.56843862	8.39015593	7.33219079	7.37376726	8.05471768	8.30686831
18	HS.119689	6.1188121	6.2737336	5.68785856	6.09991272	6.2464185	6.14336336	6.05962812	6.3778967	6.21705389	5.94255151	6.26057467
19	HS.12	8.62752118	8.8935291	8.35656399	8.20328901	8.27459582	8.40052361	8.56133857	7.25534333	8.26284908	8.30223317	8.90917197
20	HS.12114	5.97516501	5.31396562	6.20469416	5.53936045	5.80829904	6.10403333	6.38756214	5.72521626	7.50559767	5.76486623	5.74759469
21	HS.123070	6.61674363	6.59425222	6.13390417	7.40384912	6.61096803	6.41338214	6.21908833	6.60162354	6.46448175	6.45561463	6.29734323
22	HS.125231	8.665331	8.73688627	8.39267281	8.87130334	8.78483828	8.76570427	8.60765982	8.82210315	8.810249	8.65896627	8.72592266
23	HS.127022	8.46805106	9.38159728	8.22469325	8.18252382	7.80563691	7.68551684	8.16182155	10.184925	8.05604688	7.81301761	7.64645908
24	HS.12707	7.55549412	7.32315175	7.64365493	7.22511534	7.42929647	7.40491041	7.57512748	5.91237764	7.66691827	7.75837629	7.21416116
25	HS.127799	5.79943626	5.98028426	5.37907406	4.52427578	4.44971089	4.24268292	4.47586475	7.48990546	4.91441197	4.89041982	4.32156436

Figura 6. Ejemplo de una matriz final con datos de varios chips de ADN.

genes en log(2) y las columnas a los chips o pacientes (**R** para los que responden a un fármaco (controles), y **NR** para lo que no responden (casos)). Estas matrices suelen tener del orden de 20 a 80 muestras en columnas y del orden de 20000 a 30000 sondas en filas. Obsérvese que esta disposición es la contraria a la empleada habitualmente en estadística, donde los casos se sitúan en las filas y las variables en las columnas. Esto vino motivado por las capacidades de las pantallas informáticas, que no presentaban dificultad para crecer hacia abajo pero si hacia la derecha. Los ordenadores actuales presentan mayores facilidades para solucionar este problema y en el presente trabajo las matrices se

procesarán siempre en el formato estadístico clásico de casos en filas y variables en columnas. En cuanto a la nomenclatura de sondas o genes, hay herramientas que convierten unas en los otros y en este trabajo nos referiremos siempre a genes como variables por razones de simplicidad, excepto en el caso de datos reales en los que se hablará de sondas.

3.3. Tratamiento estadístico de datos de Chips de ADN

3.3.1. Técnicas para la selección de genes diferencialmente expresados entre clases

Una vez obtenida la tabla final de partida de los *microarrays*, el paso siguiente suele ser determinar si cada gen se encuentra diferencialmente expresado en los dos tipos de muestras analizadas (por ej. cáncer y control sano) o por el contrario su expresión es la misma o el gen no se ha expresado. Para ello se suele emplear alguno de los tres métodos estadísticos que se comentan a continuación.

Test t de Student con permutaciones

Este test se aplica normalmente bajo la hipótesis de que las varianzas son distintas en los dos grupos y los cálculos que se realizan se exponen a continuación.

Sean ***x*** e ***y*** las dos condiciones experimentales (por ej. enfermedad y control) y sean ***i= 1, …., n_x*** e ***i= 1, …., n_y*** muestras independientes para los grupos de enfermedad y control, respectivamente. Así, para cada gen ***j*** de los ***g*** genes totales en el microarray, se tiene el log(2) del nivel de expresión de dicho gen en las muestras ***i*** de la enfermedad y el control. El estadístico ***t*** para cada gen se calcula con la conocida expresión:

$$t = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{n_x + n_y}{n_x n_y}}}$$

A diferencia del test clásico, en el test ***t*** con permutaciones el ***p-valor*** se calcula comparando el ***t-valor*** observado frente a una distribución de hipótesis nula obtenida de

los propios datos actuales. El algoritmo permuta al azar las etiquetas de las clases (enfermo, sano) y computa el estadístico ***t*** de los datos permutados, y esto se repite digamos 10000 veces. De esta forma se obtiene la distribución nula del estadístico ***t*** bajo la hipótesis de que las etiquetas de las clases fueron asignadas al azar independientemente de la expresión génica. Normalmente se calcula un ***p-valor*** de dos colas ya que se está interesado en que la expresión sea diferente.

Pero como hay muchos genes y se realizan muchas comparaciones hay que corregir el ***p-valor*** por algún método que tenga en cuenta estas multicomparaciones. La corrección clásica de Bonferroni no es de gran utilidad en este escenario, ya que al tratarse de miles de comparaciones este procedimiento resulta demasiado restrictivo. En su lugar se calcula para cada gen ***j*** un índice llamado "False discovery rate". Para este cálculo, se ordenan los genes en orden ascendente de ***p*** y se utiliza la siguiente expresión propuesta por Benjamini and Hochberg (1995):

$$FDR_{S_j} = p_{S_j} \frac{g}{m_j}$$

siendo ***g*** el número total de genes (comparaciones) y m_j el rango del gen en la lista ascendente S_j de los ***g*** genes totales. Finalmente, se elige un umbral para el valor de FDR_{S_j} (por ej. 0.05) y los genes con FDR por debajo del umbral son considerados como diferencialmente expresados.

Método LIMMA

Este procedimiento es una variante del test-t para comparar los datos de expresión génica entre dos grupos. Trata de encontrar un estimador de la varianza con mayor precisión que el que proporcionarían los métodos clásicos. Está indicado especialmente para el caso en el que se dispongan de menos observaciones de las que realmente se necesitarían para hacer una buena estimación de la varianza.

Este método coge información prestada de todos los otros genes para obtener mejores estimas de la varianza de un gen usando una estrategia de Bayes empírica. Esto se hace mediante un modelo específico para las varianzas utilizando un método de Bayes empírico descrito en Smyth (2004). El test LIMMA no es un método con permutaciones, sino que se usa la distribución ***t*** teórica para calcular los ***p*** valores.

Algoritmo SAM

Los métodos basados en test ***t*** convencionales proporcionan la probabilidad (p) de que ocurra una diferencia en la expresión génica por azar. Aunque una p = 0.01 podría ser significativa en un experimento diseñado para evaluar un pequeño número de genes, un experimento de microarray con 10000 genes identificaría 100 genes significativos por azar. Este problema llevó a Tusher et al. (2001) a desarrollar un método estadístico adaptado específicamente para *microarrays* que denominó "Significance Analysis of Microarrays (SAM)".

SAM identifica los genes con cambios estadísticamente significativos en la expresión asimilando una serie de t-test "específicos del gen". Así, a cada gen se le asigna una puntuación sobre la base de su cambio en la expresión génica relativa a la desviación estándar de las medidas repetidas (distintos pacientes) de dicho gen. Luego, los genes con puntuaciones más grandes que un umbral son considerados potencialmente significativos. El porcentaje de dichos genes que podrían haber sido identificados por azar se estima por el estadístico FDR. Para estimar este FDR, genes significativos por azar, se hace un análisis de permutaciones con las medidas. El umbral de FDR se puede ajustar para identificar series de genes significativos más pequeñas o más grandes, y los FDRs se calculan para cada serie elegida (ver un ejemplo en Tusher et al. (2001)).

3.3.2. Técnicas para el análisis exploratorio de los grupos experimentales

Una vez detectados los genes diferencialmente expresados entre las dos clases estudiadas (por ej. muestras de control y de cáncer), se suelen aplicar diferentes técnicas estadísticas multivariantes con el fin de obtener más información acerca de las características de los grupos en estudio, es decir con fines exploratorios. Entre ella las más utilizadas son el Análisis de Conglomerados o "Cluster Analysis (CA)", Análisis de Escalado Múltiple Dimensional o "Multi Dimensional Scaling (MDS)" y el denominado Análisis Biplot.

Análisis de conglomerados (cluster analysis (CA))

Uno de los métodos más utilizados es el análisis de *clusters*. Un estudio detallado de las diferentes variantes de esta técnica no es necesaria para el enfoque del presente trabajo, para una información más completa puede consultarse publicaciones específicas como Lebart et al. (1982), pero si conviene exponer su fundamento brevemente.

Esta técnica es valiosa en el proceso de minería de datos para revelar las estructuras naturales e identificar patrones interesantes subyacentes en los datos. El análisis de conglomerados busca dividir un conjunto en grupos de datos determinados en base a las características específicas de cada grupo, de manera que los puntos de datos dentro de un grupo son más similares entre sí que los puntos de los otros grupos. Finalmente, este procedimiento representa los objetos y los clústeres formados en un gráfico que se denomina dendrograma.

Las particiones se consiguen cortando el *dendrograma* con una línea paralela al eje de abscisas. El problema es cómo elegir esta línea de corte. Existen procedimientos estadísticos para hacer esta elección, como el cociente ***E(2)/E(1)***, el índice de Calisnki o los criterios ***RMSSTD*** y $\boldsymbol{R^2}$ (ver Everitt (1993)), pero con datos de *microarrays* generalmente se opta por una línea de corte arbitraria en base a la homogeneidad de los clústeres y la experiencia biomédica del investigador (Eisen et al. (1998), Sturn et al. (2002)).

En estudios de *microarrays*, se suele hacer un CA de los casos y un CA de los genes. Así, a modo de ejemplo se muestran en la Figura 7 los correspondientes *dendrogramas* de unos datos reales de mieloma tomados de Gutiérrez et al. (2010) y realizados con el *Paquete estadístico SIMFIT* (Bardsley (2013)). Como puede verse en la Figura 7(A), los dos grupos de pacientes se separan perfectamente en dos *clusters* bien diferenciados. Respecto al CA de los genes puede apreciarse en la Figura 7(B) la formación de distintos conglomerados, aquí la interpretación es de tipo genómico y escapa al enfoque del presente trabajo.

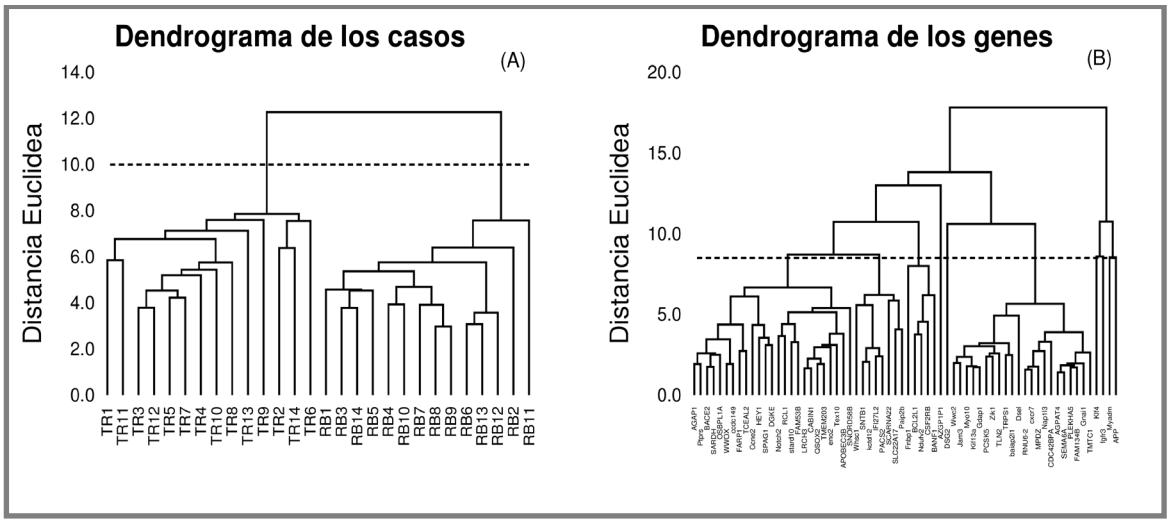

Figura 7. Ejemplo de dendrogramas con datos de *microarrays* de muestras de mieloma.

Escalado múltiple dimensional (Multi Dimensional Scaling (MDS))

El MDS es una técnica multivariante de interdependencia que trata de representar en un espacio geométrico de pocas dimensiones, generalmente 2, las proximidades existentes entre un conjunto de objetos o de estímulos. El MDS está basado en la comparación de objetos, de forma que, si dos objetos A y B son los más similares del estudio entonces las técnicas de MDS colocarán a los objetos A y B en el gráfico de forma que la distancia entre ellos sea más pequeña que la distancia entre cualquier otro par de objetos. La ventaja de esta técnica con respeto a la anterior de análisis de clústeres es que suministra información, para cada individuo, de las distancias que le separan del resto de los individuos. Esta técnica es, pues, una alternativa, o bien un complemento, a los dendrogramas vistos anteriormente.

Desde un punto de vista estadístico existen diferentes tipos de MDS, cuyos fundamentos pueden verse en diferentes publicaciones (Torgerson (1952), Takane et al. (1977), Kruskal and Wish (1978)) siendo el "MDS métrico" el más frecuente con datos de *microarrays*.

Con datos de *microarrays*, se suele hacer un MDS de los casos y un MDS de los genes. Así, a modo de ejemplo, se muestran en la Figura 8 las dos correspondientes representaciones de los mismos datos reales de mieloma utilizados en el apartado anterior. En ambos MDS, los datos se han tratado sin transformación alguna, la distancia elegida ha sido la "Euclidea", el algoritmo de fusión fue el de "promedio de grupo" y los cálculos y gráficas se han hecho con *SIMFIT*. Recordemos que se trata de dos tipos de pacientes: mielomas con "traslocación 4(14) + RB deletion" etiquetados como "TR" y mielomas con "sólo RB deletion" con etiqueta "RB". Como puede verse en la Figura 8(A), los dos tipos de mielomas se separan perfectamente en dos *clusters* bien separados.

Respecto al MDS de los genes puede verse en la Figura 8 (B) que existe un conglomerado principal y otros secundarios, aquí la interpretación vuelve a ser de tipo biomédico y va más allá de la orientación del presente trabajo.

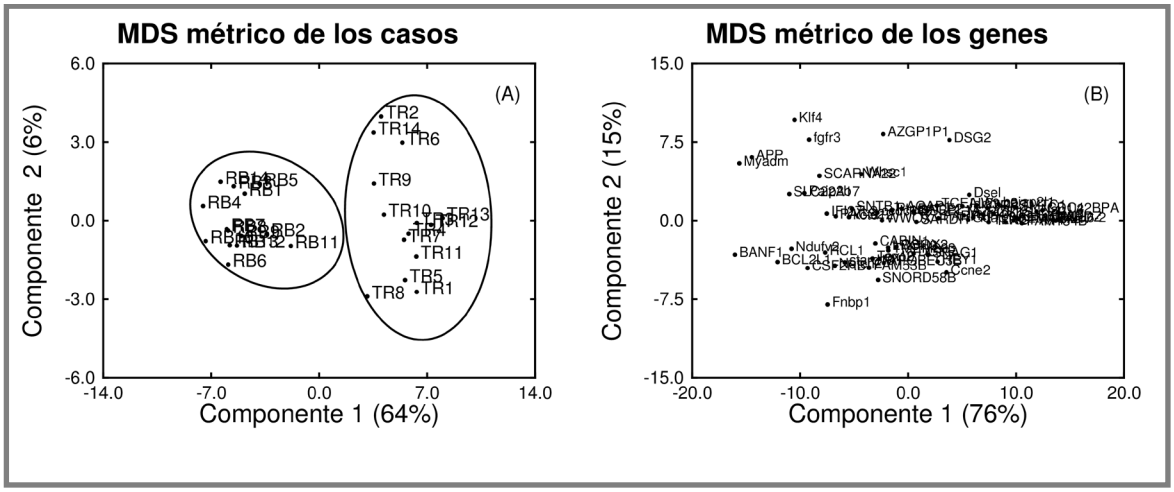

Figura 8. Ejemplo de gráficos 2D obtenidos mediante MDS usando datos de mieloma.

Análisis Biplot

De particular interés en la fase de interpretación resulta el análisis Biplot (Gabriel (1971)), ya que nos va a permitir no sólo identificar los individuos que pertenecen a cada grupo, sino saber qué genes han sido responsables de esa agrupación. Si bien el análisis de conglomerados es un método popular para asignar genes a los grupos con expresión similar (Eisen et al. (1998), Celis et al. (2000)), recientemente han sido aplicados a los datos de expresión génica métodos basados en la técnica de "Descomposición en Valores Singulares", en inglés "Singular Value descomposition o SVD*"* (Eckart and Young (1936), Alter et al. (2000)).

Esta técnica permite representar en un mismo sistema de referencia datos multivariantes, es decir puede representar simultáneamente ***n*** casos (individuos) y ***m*** variables (genes). Para ello se recurre a extraer la máxima información mediante 2 o 3 componentes o ejes ficticios (Biplot 2D ó 3D), obtenidos por descomposición de la matriz original con la técnica SVD.

A partir de la definición formal dada por Gabriel (1971), podemos definir Biplot de una matriz de datos **X** como una representación gráfica mediante marcadores $g_1, g_2, \ldots\ldots g_n$ para las filas de **X** y $h_1, h_2, \ldots\ldots h_m$ para las columnas de **X**, de forma que el producto interno $g_i^T h_j$ aproxime el elemento x_{ij} de la matriz de partida tan bien como sea posible, es decir:

$$\mathbf{X} \cong \mathbf{G}\mathbf{H}^{\mathbf{T}}$$

El interés práctico del Biplot es que el orden de las proyecciones de cada uno de los marcadores fila sobre un marcador columna reproducen el orden de los elementos de la matriz de partida y, por tanto, analizando las posiciones de las proyecciones de los marcadores fila (que identifican a los individuos) sobre cada marcador columna (que representa a una variable), nos permite ordenar los individuos según los valores que éstos toman en esa variable y eso puede hacerse para cada una de las variables.

En cuanto a la factorización de la matriz **X**, el problema se reduce a una descomposición en valores y vectores propios de **X** según la expresión:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{T}$$

donde **U** es una matriz cuyos vectores columna son ortonormales y vectores propios de $\mathbf{X}\mathbf{X}^{\mathbf{T}}$, **V** es una matriz ortonormal cuyos vectores columna son vectores propios de $\mathbf{X}^{\mathbf{T}}\mathbf{X}$

y **D** es la matriz diagonal de valores singulares de **X**, que son las raíces cuadradas no negativas de los valores propios de $\mathbf{X^TX}$.

Imaginemos una matriz **X** de rango ***s*** y que sea lo más próxima posible a **X** en el sentido de los mínimos cuadrados; normalmente de lo que se trata es de buscar una matriz **X(2)** de rango 2 que aproxime **X.**

La forma general de los marcadores es la siguiente:

$$\mathbf{G = UD\gamma} \qquad \text{y} \qquad \mathbf{H = VD(1-\gamma)}$$

Gabriel (1971) propone diversas elecciones de γ a las que da diversos nombres y para las cuales demuestra algunas de sus propiedades:

- **JK^T-Biplot o RMP-Biplot** (*Row Metric Preserving Biplot*): Con $\gamma = 1$, entonces, **G = UD** y **H = V** y se verifica que $\mathbf{H^TH = I}$. En esta aproximación, son las filas las que tienen una calidad de representación óptima.
- **GH^T-Biplot o CMP-Biplot** (*Column Metric Preserving Biplot*): Con $\gamma = 0$ **,** obtenemos entonces: **G = U** y **H = VD,** verificándose que $\mathbf{G^TG = I}$. En este Biplot son los marcadores columnas los que tiene una calidad de representación óptima.

Galindo (1986) propone el **HJ-Biplot**, donde tanto los marcadores fila como columna presentan calidad de representación óptima. La factorización se realizaría de la siguiente manera: **J=UD** y **H=VD**. Este Biplot presenta las buenas propiedades de los Biplots de Gabriel (1971), sin embargo, no reproduce los datos originales de la matriz **X** y, por lo tanto, no es un Biplot en el sentido estricto de la definición dada por Gabriel.

En el caso de los GH y JK Biplots con datos centralizados, la puntuación del valor original del individuo para cada variable analizada se puede obtener proyectando

perpendicularmente los distintos asteriscos (individuos) sobre las líneas de los vectores (variables) y recordando que en la dirección del vector los valores de las variables son positivos y en la dirección contraria son negativos.

En el campo de los *microarrays* la elección del tipo de Biplot se realizará en función del énfasis que se quiera dar, bien sea a los individuos (filas, JK Biplot) o a los genes (columnas, GH Biplot). Así, a modo de ejemplo se ha realizado un GH Biplot de datos centralizados (Figura 9) con los datos reales de *microarrays* ya comentados anteriormente, con dos tipos de pacientes: mielomas con "traslocación 4(14) + RB deletion" etiquetados como "TR" y mielomas con sólo "RB deletion" con etiqueta "RB". Los pacientes serían las filas y son representados por un asterisco rojo. Por su parte, los genes serían las columnas y aparecen con su símbolo internacional y se representan con una flecha azul. Los cálculos y gráfico se han realizado con el Paquete *SIMFIT*.

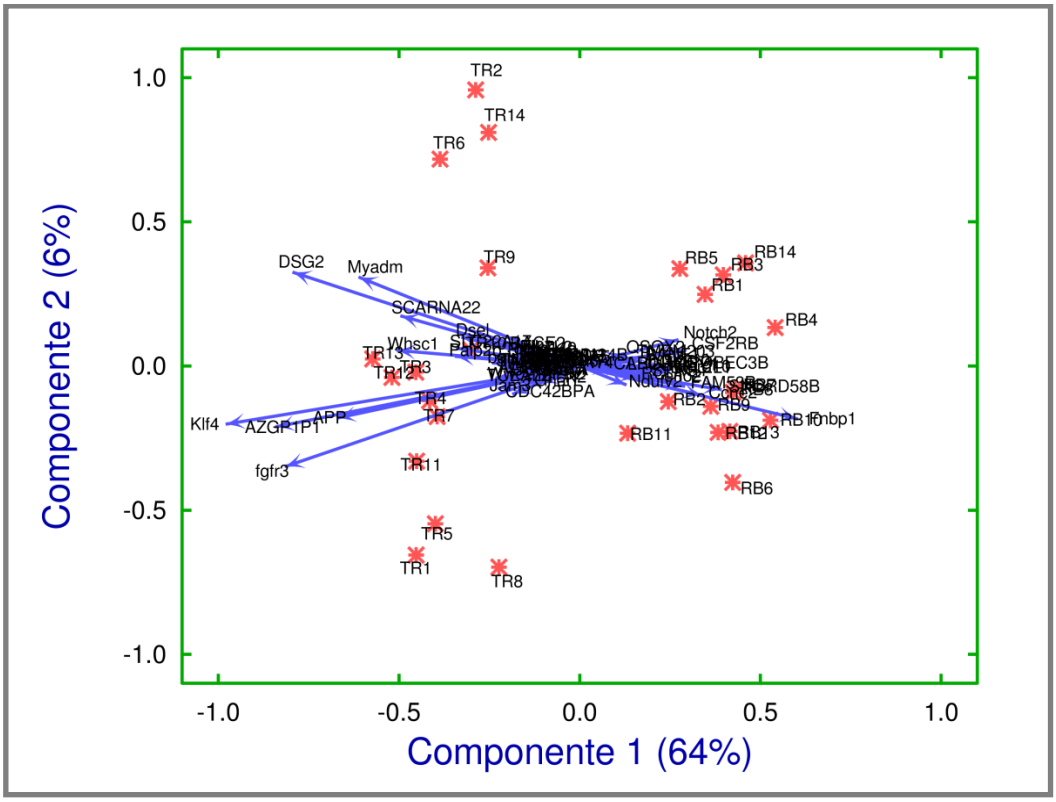

Figura 9. Ejemplo de biplot con datos reales. Los datos fueron centralizados por columna y el Biplot es del tipo "énfasis en columnas". Los asteriscos representan los casos y los vectores los genes. TR = traslocación 4(14). RB = RB deletion.

A partir de esta Figura 9 y de las características de este tipo de Biplot, se pueden interpretar diferentes propiedades de los pacientes y de los genes simultáneamente:

a) Los pacientes que quedan próximos entre sí significa que tienen valores semejantes en la expresión de sus genes. En este caso, los pacientes se agrupan en dos *clusters* claramente separados, uno a la izquierda reuniendo los pacientes con "translocación 4:14" (TR) y otro a la derecha agrupando a los pacientes con "sólo RB deletion" (RB).

b) Los vectores que "apuntan" a la izquierda están próximos a los pacientes TR, mientras que su proyección desde el (0,0) hacia la derecha estaría cercana a los pacientes RB. Esto significa que genes como *DSG2, Miadm, SCARNA22, etc.* están sobre-expresados en los pacientes TR e infra-expresados en los pacientes RB. Lo contrario sería para los vectores (genes) que "apuntan" hacia la derecha, de manera que genes como *Notch2*, *CSF2RB*, *fnbp1*, etc. están sobre-expresados en RB e infra-expresados en TR.

c) Si los genes tienen vectores en la misma dirección y forman entre sí un ángulo pequeño, significa que dichos genes están correlacionados positivamente, por ejemplo *DSG2* y *SCARNA22*. Si los genes tienen vectores perpendiculares entre si significa que existe una correlación nula entre ellos. Y cuando los genes presentan vectores en sentido contrario tienen máxima correlación negativa, por ejemplo *DSG2* y *fnbp1*. En resumen, se puede afrimar que existe igualdad entre los cosenos de los ángulos formados por los vectores y los coeficientes de correlación entre los genes que representan.

d) La longitud del vector puede entenderse como la desviación estándar de ese gen en todas las muestras (TR y RB). Genes con vectores de gran longitud, como *DSG2*, significa que se encuentran muy diferencialmente expresados en los dos tipos de pacientes.

3.3.3. Métodos predictores con datos de *microarrays*

Una vez que se han seleccionado los genes que difieren en los grupos experimentales y se ha explorado su comportamiento en la clasificación de los individuos, sería interesante poder utilizar esta información para predecir, a partir de los genes sobre-expresados o infra-expresados, a qué clase pertenecería un nuevo paciente al que se diagnostica la patología en custión.

La llamada "predicción de clase" mediante datos de *microarrays* trata de predecir la pertenencia a un subtipo de enfermedad o de prever resultados futuros como las recaídas en cáncer, la respuesta al tratamiento, etc. La estrategia es sencilla, se utilizan datos de expresión génica de una serie de casos conocidos ("training set") para calibrar algún modelo predictivo, posteriormente se utiliza dicho modelo para predecir la clase o el pronóstico de nuevos casos a partir de sus correspondientes datos de expresión génica en *microarrays*. Este objetivo es distinto del llamado "comparación de clases", cuyo enfoque es encontrar una lista de genes diferencialmente expresados entre las clases, digamos una "firma génica", por ejemplo entre pacientes de un tipo de leucemia en comparación con controles de personas sanas.

Estos dos objetivos son diferentes, aunque en la práctica a veces se encuentra relacionados, en el sentido de que se suelen buscar primero los genes con mayor diferencia de expresión entre las clases y este subgrupo de genes es el que se utiliza para "entrenar" el método predictivo elegido. No obstante conviene hacer hincapié en que, no necesariamente, los genes más diferencialmente expresados son los más discriminantes a efectos predictivos. La razón es que el subgrupo de genes con mayor diferencia de expresión pueden estar muy correlacionados y proporcionar una información redundante,

mientras que genes con una expresión diferencial menor, pero menos redundantes, podrían ser más útiles a efectos de discriminación entre clases (pag.3 en Diaz-Uriarte (2005), pág. 79 en Boulesteix et al. (2008a))

En base a Boulesteix et al. (2008a), se podrían considerar tres aproximaciones en cuanto a los métodos de clasificación con datos de *microarrays*:

1) **Métodos con "selección extrínseca de variables"**. Incluyen dos etapas: primero se selecciona un número de genes diferencialmente expresados mediante alguno de los métodos univariantes ya comentados en el apartado 3.3.1, tales como t-test permutaciones, LIMMA, SAM, etc., y segundo, los genes seleccionados más significativos se pasan a cualquier método tradicional, como son regresión logística binaria, análisis discriminante o el método de los "k-vecinos más cercanos" (Diaz-Uriarte (2005)).

2) **Procedimientos que abordan una "selección intrínseca de variables"**. Se pueden considerar de dos tipos:

   a) Métodos de regularización, tales como "Regresión Logística Penalizada" (Zhu and Hastie (2004)), o los basados en centroides contraídos como "Prediction Analysis of Microarrays (PAM)" (Tibshirani et al. (2002)), o en funciones kernel como "Support Vector Machines (SVM)" (Vapnik (1995)).

   b) Métodos ensambladores, como procedimientos de empaquetamiento o *bagging* (Breiman (1996), Dudoit et al. (2002)) o de aumento (*boosting*) mediante árboles de decisión, tales como "Random Forest (RF)" ((Breiman (2001), Diaz-Uriarte and Alvarez de Andres (2006)).

3) Métodos basados en una "reducción de la dimensión". Primero construyen algunas variables latentes, por ejemplo "Análisis de Componentes Principales" o "Mínimos Cuadrados Parciales (Partial Least Squares (PLS))", y luego pasan estas nuevas variables a cualquier procedimiento de clasificación multivariante (Boulesteix and Strimmer (2007)).

Los algoritmos más importantes de las categorías arriba comentadas se exponen a continuación, ya que varios de ellos se utilizarán en el presente trabajo.

1) Predicción de clase con selección extrínseca de genes

K vecinos más cercanos (K nearest neighbours (KNN))

KNN es el método más sencillo de predicción para decidir a qué clase pertenece una muestra nueva (Stekel (2003), Man et al. (2004)). Se basa en una medida de similaridad, que con frecuencia suele ser una medida de distancia de la nueva muestra a las observaciones de una serie de muestras de entrenamiento cuya clase es conocida. La idea detrás de KNN es que las muestras que "caen" juntas en el espacio de todas las variables pertenecen a la misma clase. El procedimiento consta de las siguientes etapas:

1) Se elige una métrica para la distancia que, para variables continuas como es la expresión génica, suele ser la Euclidea.

2) Propiamente no es necesario una etapa de entrenamiento, sino que el algoritmo procede directamente a encontrar las K muestras de la serie de entrenamiento que quedan más cercanas (menor distancia) a la nueva muestra. Finalmente se asigna a la nueva muestra la clase que es más frecuente entre las K muestras más cercanas.

3) La mejor elección del parámetro K se determina por técnicas de validación cruzada, aunque a veces se fija *a priori* en 3 ó 5.

El método KNN presenta la ventaja de que es intuitivo y fácil de aplicar pero presenta también algunos inconvenientes:

a) Si la distribución de clases está sesgada, es decir predomina una clase sobre otras, ocurre que la clase más frecuente tiende a dominar la predicción de la nueva muestra.
b) Es un método muy sensible a las observaciones atípicas.

Análisis discriminante (Discriminant Analysis (DA))

El método discriminante clásico, llamado Análisis Discriminante Lineal ("Linear Discriminant Analysis (LDA)"), fue desarrollado por Fisher en 1936 para el caso de clasificación en dos grupos y asignación de nuevos casos a dichos grupos. Se basa en identificar las combinaciones lineales entre los genes que producen la mayor discriminación entre los vectores de media para los grupos experimentales. Este método parte de los supuestos de que las variables siguen distribuciones normales y que los grupos presentan igualdad de matriz de covarianzas.

Una variante del método anterior es el "Análisis Discriminante Cuadrático" (QDA por sus iniciales en inglés), que asume también dos clases, pero ahora las variables tienen distintas distribuciones multivariantes normales y distintas matrices de covarianza en las dos clases.

Hay otras dos alternativas del DA que merecen ser mencionadas, ambas asumen que las clases tienen matrices de covarianza diagonales (es decir cada par de variables tienen una correlación igual a cero), lo que conduce a análisis más simplificados que

suelen resultar en una mayor exactitud de predicción que los métodos clásicos anteriores (Simon et al. (2003)). Una versión es la llamada "Análisis Discriminante Lineal Diagonal" (DLDA en inglés), en la cual las matrices de covarianza diagonales son asumidas iguales en los grupos experimentales. La otra versión es denominada "Análisis Discriminante Cuadrático Diagonal (DQDA)", que es una modificación de la anterior donde no se supone que la diagonal de la matriz de covarianzas es la misma para los grupos.

En la versión DLDA, para predecir la clase a la que pertenece un nuevo caso, representado por ejemplo por un vector $\boldsymbol{x}$ de variables con los valores de expresión génica de $\boldsymbol{g}$ genes, se hace lo siguiente: Se le asigna la clase **1** si

$$\sum_{i=1}^{g}\left[\frac{\left(x_i-\bar{x}_i^{(1)}\right)^2}{s_i^2}\right]\leq\sum_{i=1}^{g}\left[\frac{\left(x_i-\bar{x}_i^{(2)}\right)^2}{s_i^2}\right]$$

y en caso contrario se le asigna la clase **2**. En dicha expresión, s_i^2 denota la estima mezclada de la varianza intra-clase para el gen $\boldsymbol{i}$, $\bar{x}_i^{(1)}$ y $\bar{x}_i^{(2)}$ son la expresión media del gen $\boldsymbol{i}$ en las clases 1 y 2, respectivamente, y x_i es la expresión del gen $\boldsymbol{i}$ en el nuevo caso a clasificar.

Por último, cabe mencionar que se han desarrollado otras variantes de análisis discriminante. Así, algún autor prefiere utilizar técnicas Bayesianas para asignar los nuevos casos a las clases cuando se puede asumir una distribución multivariante normal (manual de *SIMFIT* versión 7.0.6., páginas 217-219).

Regresión logística binaria (Binary Logistic Regression (BLR))

Una situación frecuente con datos de *microarrays* es el de la clasificación binaria, en la que sólo existen dos clases, digamos una de fracaso y otra de éxito (por ej. fallece o sobrevive). La regresión logística binaria es la técnica que aborda esta situación. Su objetivo es encontrar el mejor ajuste para describir la relación entre una variable dependiente dicotómica y un conjunto de variables independientes.

Asumiendo un valor de "1" para la presencia del evento favorable (éxito) y de "0" para su ausencia (fracaso), el modelo BRL se puede expresar como:

$$\ln\left(\frac{p(1)}{1-p(1)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_j x_j$$

Los supuestos en los que se basa la regresión BLR son los de distribución binomial e independencia entre las muestras. En cuanto al método de ajuste se utiliza el procedimiento de "Máxima Verosimilitud", en lugar del criterio de "Mínimos Cuadrados" que se utiliza en la regresión convencional.

El proceso de predicción de una muestra nueva se lleva a cabo obteniendo para dicha muestra los valores de las variables x_i y estimando la probabilidad binomial de "éxito" mediante la expresión:

$$p(1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_j x_j)}}$$

2) Métodos con selección intrínseca de variables (genes)

Se trata de algoritmos que incorporan una selección de variables (genes) junto con la construcción del predictor, lo que resulta de gran interés con datos de *microarrays*, ya que normalmente los genes de partida son del orden de los miles. A continuación se exponen los métodos más habituales en la actualidad.

Centroides contraídos (método PAM)

Este procedimiento fue propuesto por Tibshirani et al. (2002) y supone una modificación del método del "centroide más cercano". La modificación consiste en una compresión de todos los centroides de las clases hacia el centroide global de todas las clases. La cantidad de contracción (denominado "threshold") es determinada por validación cruzada.

En esencia el algoritmo sigue las siguientes etapas:

1) Sea x_{ij} la expresión de los genes $i = 1, 2, ..., m$ en los casos $j = 1, 2, \ldots, n$. Se tienen las clases *1, 2,...K*, y sea C_k el índice de los n_k casos en la clase *k*. El ***i-ésimo*** componente del centroide para la clase *k* es:

$$\bar{x}_{ik} = \sum_{j \in C_k} x_{ij} \Big/ n$$

y el ***i-ésimo*** componente del centroide global es:

$$\bar{x}_{i} = \sum_{j=1}^{n} x_{ij} \Big/ n$$

2) El método consiste en contraer los centroides de clase hacia el centroide global, después de hacer una estandarización por la desviación estándar "dentro-de las clases" para cada gen. Se empieza calculando:

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k (s_i + s_0)}$$

donde s_i es la desviación estándar combinada "dentro de la clases" para el gen i y $m_k = \sqrt{1/n_k + 1/n}$ hace $m_k s_i$ igual al error estándar del numerador en d_{ik}. El valor s_0 es una constante positiva (igual para todos los genes) que se fija igual a la mediana de los s_i correspondientes a todos los genes.

3) Observando la ecuación anterior, se concluye que d_{ik} viene a ser el estadístico t para el gen i, comparando la clase k con el centroide global. Reescribiendo dicha ecuación se tiene que:

$$\bar{x}_{ik} = \bar{x}_i + m_k (s_i + s_0) d_{ik}$$

El método contrae ahora cada d_{ik} hacia el cero, dando un d'_{ik} que produce unos centroides contraídos:

$$\bar{x}_{ik} = \bar{x}_i + m_k (s_i + s_0) d'_{ik}$$

Cada d_{ik} es reducido a d'_{ik} por una cantidad Δ en valor absoluto y es fijado a cero si este valor absoluto es menor que cero.

4) La cantidad de contracción (Δ) se elige por validación cruzada usando un procedimiento de 90% "training" y 10% "test" y repitiendo el proceso 10 veces para medir un porcentaje de error de clasificación promedio. El valor óptimo de Δ

(menor error de clasificación) es el que se utiliza para fijar el número de genes activos con al menos un valor de d'_{ik} distinto de cero. El resto de los genes son eliminados como no significativos. A la vez se identifican las subseries de genes que mejor identifican cada clase.

5) Una muestra nueva se clasificará por la habitual regla del centroide más cercano, pero usando los centroides contraídos de clase. La clase cuyo centroide es el más cercano, como distancia al cuadrado, es la clase predicha para la nueva muestra.

Máquinas de Vectores Soporte (Support Vector Machines (SVMs))

SVMs son unos algoritmos de clasificación que fueron propuestos por el grupo de Vladimir Vapnik en los laboratorios *AT&T Bell* (Cortes and Vapnik (1995)) y que han sido aplicados con éxito al análisis de datos de microarrays (Furey et al. (2000)).

La idea que subyace en SVMs es la de encontrar el mejor hiperplano en el espacio de los genes que separe las dos clases de la serie de entrenamiento mediante un método de maximización. Normalmente, se trata de maximizar el llamado "margen", es decir la suma de las distancias desde el hiperplano a las observaciones más cercanas de las dos clases correctamente clasificadas, mientras que se penaliza el número de las observaciones mal clasificadas. La búsqueda del hiperplano se puede hacer en el espacio original, dando lugar a los llamados SVMs lineales, o en un espacio de dimensión más alta, los llamados SVMs no lineales (ver detalles en Burges (1998), Hastie et al. (2001))

SVMs lineales

Caso linealmente separable

Es la situación más sencilla, aquí las dos clases en la serie de entrenamiento se pueden separar completamente por un hiperplano. La

búsqueda de este hiperplano supone un problema de programación lineal. En la práctica el problema es reformulado para encontrar un par de hiperplanos H_1 y H_2 que son paralelos por construcción y entre los que no cae ninguna de las observaciones de entrenamiento, es decir separan perfectamente las muestras de las dos clases. Los puntos que "caen" sobre alguno de los dos hiperplanos se les llama "vectores de soporte" (ver Dudoit and Fridlyan (2003)).

Caso no linealmente separable

Cuando el algoritmo anterior se aplica a datos no separables, la función objetivo a maximizar se hace muy grande y la solución no es factible. Se necesita en este caso relajar las restricciones sobre los dos hiperplanos introduciendo algún tipo de penalización. Cortes and Vapnik (1995) introdujeron una variable flexible $\xi_i \geq 0$ entre las restricciones, esta variable es adaptada por un escalar C denominado "coste o penalti", de forma que cuanto mayor es C mayor es la penalización a los errores (Dudoit and Fridlyan (2003)). Este valor C se calcula empíricamente por validación cruzada.

SVMs no lineales

En muchas ocasiones, es útil considerar la transformación de los datos en un espacio de dimensión superior, de forma que reduciendo luego el problema al caso lineal se alcanza una solución sencilla. La representación por medio de funciones Kernel ofrece una solución a esta estrategia, consiste en transformar el espacio de las variables de entrada x en un espacio de mayor dimensionalidad:

$$x = \{x_1, x_2, ..., x_n\} \rightarrow \phi(x) = \{\phi(x)_1, \phi(x)_2, ..., \phi(x)_n\}$$

Existen diferentes funciones kernel, las más usadas son: "Polinomio de grado p", "Función de Base Radial Gaussiana (RBF)" o "Red Neuronal Sigmoidea de Dos Capas" (Dudoit and Fridlyan (2003)).

En estos SVMs el usuario tiene que controlar los siguientes aspectos: especificar la función Kernel a emplear y elegir el parámetro de coste ***C*** de penalización por validación cruzada.

Bosque al azar (Random forest (RF))

RF es un algoritmo de clasificación desarrollado por Breiman (2001) basado en el agrupamiento de muchos árboles de clasificación. El algoritmo de Breiman (2001) estaría formado por las siguientes etapas:

1) Se va a construir una colección muy grande de árboles de clasificación (cientos).
2) Sea ***N*** el número de casos (muestras) en la serie de entrenamiento y ***M*** el número de variables (expresiones de genes en nuestro caso).
3) Cada árbol se crece con una muestra "bootstrap" (con reemplazamiento), seleccionando ***n*** casos de los ***N*** posibles en la serie de entrenamiento. Los restantes ***N-n*** casos se usarán como serie de prueba para predecir sus clases.
4) Por cada nodo de un árbol se usará un número de variables ***m***, seleccionadas al azar de las ***M*** disponibles (normalmente $m << M$). Estas variables ***m*** servirán para tomar la decisión en ese nodo.
5) Cada árbol se crece completamente, no estando permitido el truncar o "podar".
6) Para predecir una nueva muestra, se introduce dicha muestra en un árbol y se le asigna la etiqueta en el nodo terminal de dicho árbol, la que se obtuvo con la serie de entrenamiento. El proceso se repite con el resto de los árboles del agrupamiento

(bosque) y a la muestra se le acaba asignando aquella clase que haya obtenido el mayor número de votos a partir de todos los árboles.

RF se ha aplicado con éxito a datos de *microarrays*, observándose que tiene un funcionamiento comparable al de otros métodos de clasificación como KNN, DA o SVM (Diaz-Uriarte and Alvarez de Andres (2006)).

3) Métodos basados en reducción de la dimensión

Persiguen extraer la información de cientos de variables en unas pocas variables ficticias, llamadas componentes o factores. Estas nuevas variables son las que se utilizan luego con algún procedimiento predictor clásico. Existen dos grandes métodos de extracción, el "Análisis en Componentes Principales" que se expone a continuación y la "Regresión por Mínimos cuadrados Parciales (*Partial Least Squares (PLS)*) que se verá en un capítulo específico más adelante.

Análisis en componentes principales (Principal Component Analysis (PCA))

Se dispone de una matriz de n casos por m variables $x_1, x_2, x_3, \ldots, x_m$, inicialmente correlacionadas, para posteriormente obtener a partir de ellas un número k ($k < m$) de variables incorreladas $C_1, C_2, \ldots, C_k$ denominadas componentes principales. Estas componentes serán combinación lineal de las variables originales y han de resumir lo mejor posible dichas variables originales con la mínima perdida de información.

Inicialmente se tienen tantas componentes como variables originales, pero sólo se retienen las k componentes que expliquen un porcentaje alto de las variables de partida (Perez-Lopez (2005)). Usualmente las variables originales se encuentran estandarizadas (media cero y varianza unidad), con el fin de evitar problemas de escala entre las variables.

Para obtener las componentes principales se analiza normalmente la matriz de varianza-covarianza y como medida de la cantidad de información incorporada en cada componente se utiliza su varianza. Es decir, cuanto mayor sea su varianza mayor es la información que lleva incorporada dicha componente. Por este motivo, se selecciona como primera componente aquella que tenga mayor varianza, mientras que la última componente elegida será la de menor varianza. La ventaja de PCA es que cuando las variables originales están muy correlacionadas entre sí, ocurre que la mayor parte de su variabilidad se puede explicar con muy pocas componentes principales.

Una vez seleccionado el número de componentes, se utilizan las puntuaciones en dichas componentes como unas nuevas variables, con el fin de aplicarlas a algún otro procedimiento de predicción clásico que requiera pocas variables, como son la regresión logística binaria (PCA-LR) o el análisis discriminante (PCA-DA). Esta estrategia puede funcionar bien a efectos predictivos en algunos casos, pero tiene dos inconvenientes:

a) Las componentes principales son una combinación lineal de todas las variables (genes), lo que a veces no resulta muy ilustrativo biológicamente, a pesar de que los pesos o cargas de las variables podrían aportar alguna información.

b) PCA no usa para construir sus componentes la información de las clases a las que pertenecen los individuos. Es lo que se llamaría un método no supervisado, que a efectos de predicción podría funcionar bien con buenas series de datos pero no ser adecuado para datos complejos. Por tanto, PCA debiera ser considerado inapropiado para clasificación frente a métodos supervisados como PLS (Boulesteix (2004)).

Como se verá en el capítulo 3.4, PLS utiliza en su desarrollo una combinación lineal de las variables predictoras pero, además, considera la información de las

clases a las que pertenecen los individuos. Según Boulesteix (2004), esas características sitúan a PLS como el único método supervisado de reducción de la dimensión que puede manejar un elevado número de genes como variables, lo que le convierte en una herramienta idónea para el análisis de datos de *microarrays*.

3.4. Método de Mínimos Cuadrados Parciales ("Partial Least Squares (PLS)") para análisis de datos clínicos y génicos

3.4.1. Qué es y para qué sirve PLS

PLS es un método multivariante que fue originalmente propuesto por Herman Wold (Wold (1966)) y que fue luego desarrollado en disciplinas tales como Econometría, Quimometría y Monitorización de Procesos, con el fin de realizar una reducción de la dimensión de variables en unos casos y para multicalibración en otros. Los fundamentos y aplicaciones detalladas de PLS pueden verse en algunos manuales (Esbensen (2010), Eriksson et al. (2006)) o en diferentes publicaciones (Abdi and Williams (2013), Boulesteix and Strimmer (2007)). No obstante, parece necesario exponer a continuación la metodología en la que se basa la creación de modelos PLS, su interpretación, las diferencias entre las fases de calibración y predicción, así como la elección del número óptimo de factores latentes.

Los métodos PLS se propusieron en un principio como técnicas para resolver las limitaciones que aparecieron en los llamados "path analysis" o modelos de ecuaciones estructurales. Posteriormente se utilizaron en los modelos espectrométricos dentro del campo de la Quimiometría. Los PLS fueron propuestos como técnicas con menos limitaciones que las técnicas clásicas en la calibración multivariante, y se aplicaron principalmente desde la perspectiva de los modelos de regresión, más que de la obtención de factores latentes o modelos de reducción de la dimensionalidad. Actualmente se aplican a un gran número de campos de investigación, como son la Farmacología, Psicología, Medicina, Bioinformática, Ciencias Sociales, etc. (Esbensen (2010), Eriksson et al. (2006)). La semejanza entre las características del análisis de la expresión génica

con los problemas planteados dentro de la Quimiometría hace que esta técnica se esté aplicando cada vez más en el área de la Genómica y Proteómica.

PLS es una técnica útil cuando una matriz de respuestas **Y(n×r)** es observada con una matriz de variables predictoras **X(n×m)**, en el mismo número de casos ***n*** que es más pequeño que el número de variables predictoras ***m*** y, además, las variables ***x*** podrían estar correlacionadas. Pero existen muchas otras técnicas para relacionar dos conjuntos de variables que hacen referencia a los mismos individuos o unidades de muestreo. Las diferencias entre las distintas técnicas que estudian las relaciones entre dos conjuntos de variables, se deben al objetivo que pretenden cubrir, así como a la matriz elegida para la descomposición espectral.

Una primera división sobre las técnicas que analizan las relaciones de dos conjuntos de datos se puede establecer dependiendo de la dirección de las asociaciones que se pretende analizar. Así, aquellas técnicas que pretendan explorar las relaciones entre las variables de los dos conjuntos de datos sin darle diferente *status* a dichas matrices, se denominan análisis simétricos:

$$\mathbf{X_1} \leftrightarrow \mathbf{X_2}$$

Por otra parte, aquellas técnicas que pretender explicar el comportamiento de un conjunto de variables (variables respuesta) a partir de otro conjunto de variables (variables explicativas o regresoras) se denominan análisis asimétricos:

$$\mathbf{Y} \leftarrow \mathbf{X}$$

En ambas clases de análisis se parte de la matriz de covarianzas. Esta matriz la podemos expresar de la siguiente manera:

$$\text{cov}(Y, X) = \mathbf{M}_{YX} = \begin{pmatrix} \mathbf{M}_{YY} & \mathbf{M}_{YX} \\ \mathbf{M}_{XY} & \mathbf{M}_{\text{XX}} \end{pmatrix}$$

donde:

$\mathbf{M}_{\text{YY}}$ es r×r matriz de covarianzas de **Y.**

$\mathbf{M}_{\text{XX}}$ es m×m matriz de covarianzas de **X.**

$\mathbf{M}_{\text{YX}}$ es r×m matriz de covarianzas de entre **X** e **Y.**

$\mathbf{M}_{\text{YX}}=\mathbf{M}_{XY}^{T}$

En función de lo que se pretende conseguir y como se descompone esta matriz de covarianzas, se puede realizar una clasificación de las técnicas que relacionan dos conjuntos de datos de la siguiente manera:

- ***Regresión multivariante***. Es un análisis asimétrico donde se quiere conocer el comportamiento de un conjunto de variables respuesta en función de un conjunto de variables explicativas. Presenta las limitaciones de la colinealidad y la necesidad de mayor número de individuos que de variables.

$$\mathbf{M}_{YX}\mathbf{M}_{XX}^{-1} = \mathbf{B}$$

- ***Análisis de Correlación Canónica (en inglés "CA") (Hotelling (1936))***. Es un análisis simétrico, donde se quiere conocer las relaciones entre dos conjuntos de variables. Esta técnica maximiza la matriz de correlaciones entre las dos matrices y pretende encontrar los primeros factores dentro de cada matriz que presentan máxima correlación. La matriz que se descompone en valores singulares sería:

$$\mathbf{M}_{YY}^{-1/2}\mathbf{M}_{YX}\mathbf{M}_{XX}^{-1/2} = \mathbf{U\Sigma V}^{T}$$

Pero presenta el problema de colinealidad.

- ***Análisis de la redundancia ("RA" en Inglés) (Van Den Wollenberg (1977)) o regresión de bajo rango (Izenman (1975))***. Busca los factores dentro de la matriz de variables dependientes que sea explicada por las variables independientes iniciales. Es un análisis asimétrico. Su formulación sería:

$$\mathbf{M}_{YX}\mathbf{M}_{XX}^{-1}\mathbf{M}_{XY} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{T}$$

Presenta también el problema de colinealidad y las predicciones no suelen ser muy exactas (Tobias (1995)). Por otra parte, si en lugar de considerar una respuesta lineal se definiera una respuesta gaussiana estaríamos ante el método del "Análisis Canónico de Correspondencias (CCA)" (Ter Braak (1986)), y en lugar de utilizar las componentes principales se emplearía el análisis de correspondencias.

- ***Regresión de componentes principales.*** Es un análisis asimétrico donde se quiere explicar un conjunto de variables respuesta en función de un conjunto de variables explicativas. Se realizan unas componentes principales de la matriz de variables independientes y en un segundo paso los factores obtenidos se utilizan para explicar las variables respuesta. Su formulación sería:

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{T}$$

$$\mathbf{X}^{T} = \mathbf{U}\boldsymbol{\Sigma}$$

$$\mathbf{M}_{\mathrm{YX}^{T}}\mathbf{M}^{-1}{}_{\mathrm{X}^{T}\mathrm{X}^{T}} = \mathbf{B}$$

No existe el problema de la colinealidad ni la limitación del número de observaciones, pero no se garantiza que los factores obtenidos del análisis de

componentes principales sean los que presentan mayor relación con las variables respuesta.

- ***Partial Least Squares (PLS).*** Esta técnica puede utilizarse como un análisis tanto simétrico como asimétrico. En el primer caso, se utiliza simplemente como reducción de la dimensionalidad, bajo el criterio de que los factores obtenidos a partir de cada matriz presenten máxima covarianza. El objetivo es encontrar la información que comparten ambas matrices **X** e **Y**:

$$\mathbf{M}_{YX} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

Este tipo de PLS suele tener varias denominaciones en función de los autores y del campo que lo utiliza. Así, Abdi and Williams (2013) le denominan "Partial Least Square Correlation (PLSC)". En el campo de la Ecología se denomina "Análisis de la Coinercia (CoA)**"** (Doledec and Chessel (1994)) y en el campo de Psicología se denomina "Inter-Battery Analysis" (Tucker (1958)). Cuando los PLS se aplican en el campo de los modelos de ecuaciones estructurales o "path modeling" se denominan PLS-PM.

Cuando PLS se plantea como un análisis asimétrico se le denomina *Partial Least Square Regressión (PLSR)*. En este segundo caso, se pretende explicar (o predecir) las variables respuesta a partir de un conjunto de variables explicativas. Esta técnica resuelve las limitaciones de la regresión multivariante en relación a la colinealidad y a la necesidad de que el número de observaciones sea mayor que el número de variables. Además, presenta estimaciones más exactas que otras técnicas como el *análisis de la redundancia* y la *regresión de componentes*

principales (Yeniay and Goktas (2002)). En este caso, los algoritmos de cálculo de la solución ***PLS*** se modifican en base a ese carácter asimétrico del análisis. Así, el "Análisis de la Coinercia" se puede considerar como el primer paso de un PLSR (Lindgren et al. (1993)).

A diferencia de otras técnicas de reducción de la dimensión, como PCA, la aproximación PLSR calcula cada variable latente a partir de **X** pero basándose en **Y**. El objetivo es maximizar la covarianza de **Y** con **X**, a diferencia de PCA, que maximiza la varianza de las variables ***x*** solamente. La idea que subyace en PLSR es expresar las matrices **X** e **Y** en términos de una serie de ***k*** factores latentes, obtenidos a partir de las matrices **X** e **Y** por técnicas de proyección y regresión. Una vez que se han obtenido las expresiones que aproximan **X** e **Y** usando los factores latentes, éstos se pueden utilizar para tratar **X** como una matriz de entrenamiento, de la que poder predecir qué nueva **Y** resultaría de una nueva **X** que esté expresada en las mismas variables que la matriz de entrenamiento **X**.

En el presente trabajo se utiliza siempre la metodología de PLSR pero, en aras de la sencillez, a partir de ahora, la denominaremos simplemente PLS.

3.4.2. Algoritmos más usuales en PLS

En términos generales se asume que **X** es una matriz **n×m** e **Y** una matriz **n×r**. La técnica se basa en extraer sucesivamente factores de ambas matrices de manera que la covarianza entre dichos factores es máxima. El PLS puede trabajar tanto si en la matriz **Y** ***r = 1*** o ***r > 1***.

Formalmente esta técnica intenta encontrar una descomposición lineal de **X** e **Y**, tal que:

$$\mathbf{X}=\mathbf{T}\mathbf{P}^{\mathbf{T}}+\mathbf{E}$$

$$\mathbf{Y}=\mathbf{U}\mathbf{Q}^{\mathbf{T}}+\mathbf{F}$$

donde:

$\mathbf{T}_{n\times k}$ son las puntuaciones de **X**

$\mathbf{U}_{n\times k}$ son las puntuaciones de **Y**

$\mathbf{P}_{m\times k}$=son las cargas de **X**

$\mathbf{Q}_{r\times k}$=son las cargas de **Y** (si sólo tenemos una variables dependiente r=1)

$\mathbf{E}_{n\times m}$= Los residuos de **X**

$\mathbf{F}_{n\times r}$= Los residuos de **Y**.

haciéndose la descomposición de manera que la covarianza sea máxima entre **T** y **U**.

Existen un gran número de algoritmos PLS. Herman Wold (Wold (1975)) desarrolló un primer algoritmo, sencillo pero eficiente, denominado ***NIPALS*** (*Nonlinear Iterative Partial Least Squares*). Este procedimiento se describe a continuación.

Algoritmo NIPALS

En el método NIPALS se van obteniendo los factores de manera secuencial, de forma que se calcula la primera componente y a partir de ella se reconstruyen las matrices

estimadas de $\widehat{\mathbf{X}}$ e $\widehat{\mathbf{Y}}$; se calculan luego las matrices de residuos como $\mathbf{X} - \widehat{\mathbf{X}}$ e $\mathbf{Y} - \widehat{\mathbf{Y}}$, y a partir de estas nuevas matrices se construye la siguiente componente, y así sucesivamente.

Sin entrar en aspectos numéricos particulares, el algoritmo propuesto en Esbensen (2010), página 140, se muestra en el siguiente recuadro:

0) Centralizar y escalar ambas matrices **X** e **Y**.

Indice de inicialización, **f :** $\mathbf{f = 1} \quad \mathbf{X_f = X} \quad ; \quad \mathbf{Y_f = Y}$

1) Para el vector cebador $\mathbf{u_f}$ se elige cualquier columna de **Y**

2) $\mathbf{w_f} = \mathbf{X^T u_f} / |\mathbf{X^T u_f}|$ (**w** es normalizado)

3) $\mathbf{t_f = X_f w_f}$

4) $\mathbf{q_f} = (\mathbf{Y_f})^\mathbf{T} \mathbf{t_f} / |(\mathbf{Y_f})^\mathbf{T} \mathbf{t_f}|$ (**q** es normalizado)

5) $\mathbf{u_f = Y_f q_f}$

6) Convergencia: Si $|\mathbf{t_{f,nuevo} - t_{f,viejo}}| <$ límite convergencia, parar, sino ir a paso 2).

7) $\mathbf{p_f} = (\mathbf{X_f})^\mathbf{T} \mathbf{t_f} / (\mathbf{t_f})^\mathbf{T} \mathbf{t_f}$

8) $\beta_\mathbf{f} = (\mathbf{u_f})^\mathbf{T} \mathbf{t_f} / (\mathbf{t_f})^\mathbf{T} \mathbf{t_f}$ (Relación interna en PLS)

9) $\mathbf{X_{f+1} = X_f - t_f} (\mathbf{p_f})^\mathbf{T} \qquad \mathbf{Y_{f+1} = Y_f - \beta_f t_f} (\mathbf{q_f})^\mathbf{T}$

10) $\mathbf{f = f + 1}$

Repetir de 1) a 10) hasta que $\mathbf{f = k_{opt}}$ (óptimo número de factores PLS por validación)

Los pasos 2 a 5 corresponden al formalismo de una regresión convencional, por eso al algoritmo NIPALS se le ha descrito a veces como regresiones "criss-cross" en los espacios **X** e **Y**. Como puede apreciarse, se trata de un algoritmo iterativo. Suele

converger normalmente en una solución estable en menos iteraciones que la situación equivalente de PCA, ya que la estructura correlacionada de los datos en ambos espacios (**X** e **Y**) se apoya mutuamente. La llamada "relación interna" del modelo PLS se refiere a la representación gráfica de **T** frente a **U,** que constituye la representación central de PLS.

A partir de este algoritmo se han presentado varias soluciones basadas principalmente en los procesos para el cálculo de las puntuaciones (***t*** , ***u***) y los pesos (***w***). Una de las soluciones está basada en el fundamento del algoritmo NIPALS y consiste en la deflación de las matrices, o lo que es lo mismo de la reducción del rango de las matrices paso a paso. La otra solución se basa en la obtención de los factores en un sólo paso mediante la descomposición en valores singulares de la matriz de covarianzas.

Rosipal and Kraemer (2006) hicieron la siguiente clasificación de las formas de PLS en función de cómo se deflactan las matrices:

- **PLS Modo A**. el algoritmo está basado en la reducción secuencial del rango de las matrices usando las correspondientes puntuaciones y cargas de los vectores.

$$\mathbf{X}_{i+1} = \mathbf{X}_i - \mathbf{t}_i\mathbf{p}_i^T \quad y \quad \mathbf{Y}_{i+1} = \mathbf{Y}_i - \mathbf{u}_i\mathbf{q}_i^T$$

  Corresponde al algoritmo propuesto por Wold (Wold (1975)). En este caso la relación entre ambos conjuntos de datos es simétrica.

- **PLS1 y PLS2**. El algoritmo se denomina PLS1 cuando sólo hay una variable dependiente y PLS2 cuando hay más de una variable dependiente. Son los métodos generalmente utilizados en PLSR. La relación entre **X** e **Y** es asimétrica. Considera que hay una relación interna entre los vectores **t** y **u** :

$$\mathbf{U} = \mathbf{TD} + \mathbf{H}$$

donde **D** es una matriz diagonal **r**×**r** y **H** es la matriz residual. Por lo tanto:

$$\mathbf{X}_{i+1} = \mathbf{X}_i - \mathbf{t}_i\mathbf{p}_i^T \quad y \quad \mathbf{Y}_{i+1} = \mathbf{Y}_i - \mathbf{t}_i\mathbf{t}_i^T\mathbf{Y}_i / \mathbf{t}_i^T\mathbf{t}_i = \mathbf{Y}_i - \mathbf{t}_i\mathbf{q}_i^T$$

donde **q** es el vector de cargas de ***Y*** no normalizado del paso 4 del NIPALS.

Este esquema de deflación de las matrices garantiza la ortogonalidad de los vectores de puntuaciones **t**, restricción que no se produce en el PLS Modo A y que permite la comparación de las componentes principales.

- **PLS-SB**. Se puede demostrar que la descomposición en valores singulares (*Single value descomposition, SVD*) de la matriz $\mathbf{X^TY}$ resuelve también el problema (Le Cao et al. (2008)). Además la extracción de las componentes se realiza a la vez. Los PLS-SB utilizan la descomposición de la matriz $\mathbf{X^TY}$ en vectores singulares. En contraste con los algoritmos PLS1 y PLS2, los vectores ***t*** no son, por lo general, mutuamente ortogonales.

- **SIMPLS**. Este método se propone para evitar la reducción secuencial del rango de la matriz que presentan los algoritmos PLS1 y PLS2. Este método busca un vector de ponderaciones que se aplica a la matriz original X sin reducir su rango. Se puede demostrar que la solución de SIMPLS es la misma cuando aplicamos PLS1 pero difiere para PLS2 (Boulesteix and Strimmer (2007)).

A continuación se describe el algoritmo utilizado por la librería NAG (NAG (2012)) que emplea la técnica SVD, que es la que utiliza el paquete *SIMFIT* y que servirá también de base al algoritmo *PLS-VIP* propuesto en este trabajo.

Algoritmo basado en SVD

Consta de los siguientes pasos que se muestran en el siguiente recuadro.

1) Sean $\mathbf{X_1}$ y $\mathbf{Y_1}$ las matrices centradas por substracción de las medias de las columnas **X** e **Y** de partida a los valores originales (las columnas también pueden ser escaladas a varianza unidad si se considera adecuado).

2) Desde ***i = 1*** a ***k***, siendo ***k*** el número deseado de factores, se deben llevar a cabo los siguientes procedimientos:

a) Computar la dirección de máxima covarianza llevando a cabo una descomposición SVD de $\mathbf{(X_i)^T Y_i}$ y cogiendo el primer vector singular de la izquierda para definir los pesos de **x** ($\mathbf{w_i}$) para las variables predictoras.

b) Calcular el vector de puntuaciones de **x:** $\mathbf{t_i = X_i w_i}$

c) Calcular las cargas de ***x*** por mínimos cuadrados: $\mathbf{(p_i)^T = (t_i)^T X_i}$

d) Calcular las cargas de ***y*** por mínimos cuadrados: $\mathbf{(q_i)^T = (t_i)^T Y_i}$

e) Calcular el vector de puntuaciones de ***y***: $\mathbf{u_i = Y_i q_i}$

f) Calcular las estimaciones de $\mathbf{X_i}$ and $\mathbf{Y_i}$:

$\mathbf{\hat{X}_i = t_i (p_i)^T}$; $\mathbf{\hat{Y} = t_i (q_i)^T}$

g) Deflactar las matrices previas:

$\mathbf{X_{i+1} = X_i - \hat{X}_i}$; $\mathbf{Y_{i+1} = Y_i - \hat{Y}}$

h) Se hace ***i = i+1*** y se vuelve al paso a)

3) finalmente, las puntuaciones **t** y **u** son usadas para calcular los parámetros de regresión usando los ***k*** factores latentes. Estos parámetros vienen dados por:

$$\beta = \mathbf{W(P^T W)^{-1} Q^T}$$

donde **W** es la matriz **m×k** de los ***pesos de x***, **P** es la matriz **m×k** de ***cargas de x***, y **Q** es la matriz **r×k** de ***cargas*** de ***y***. Nótese que los valores $\boldsymbol{\beta}$ calculados de esta forma corresponden a las matrices centradas de las **X** and **Y** originales, pero también se pueden calcular referidos a los datos originales deshaciendo el paso de centrado (y escalado si lo hubiera).

3.4.3 Interpretación de los modelos PLS

La técnica PLS proporciona varios procedimientos de diagnóstico que facilitan la interpretación del modelo que se haya construido, así como de la evaluación de su comportamiento y de la importancia de las distintas variables. En la Figura 10 se presenta un esquema de los resultados que se obtienen en un análisis de PLS.

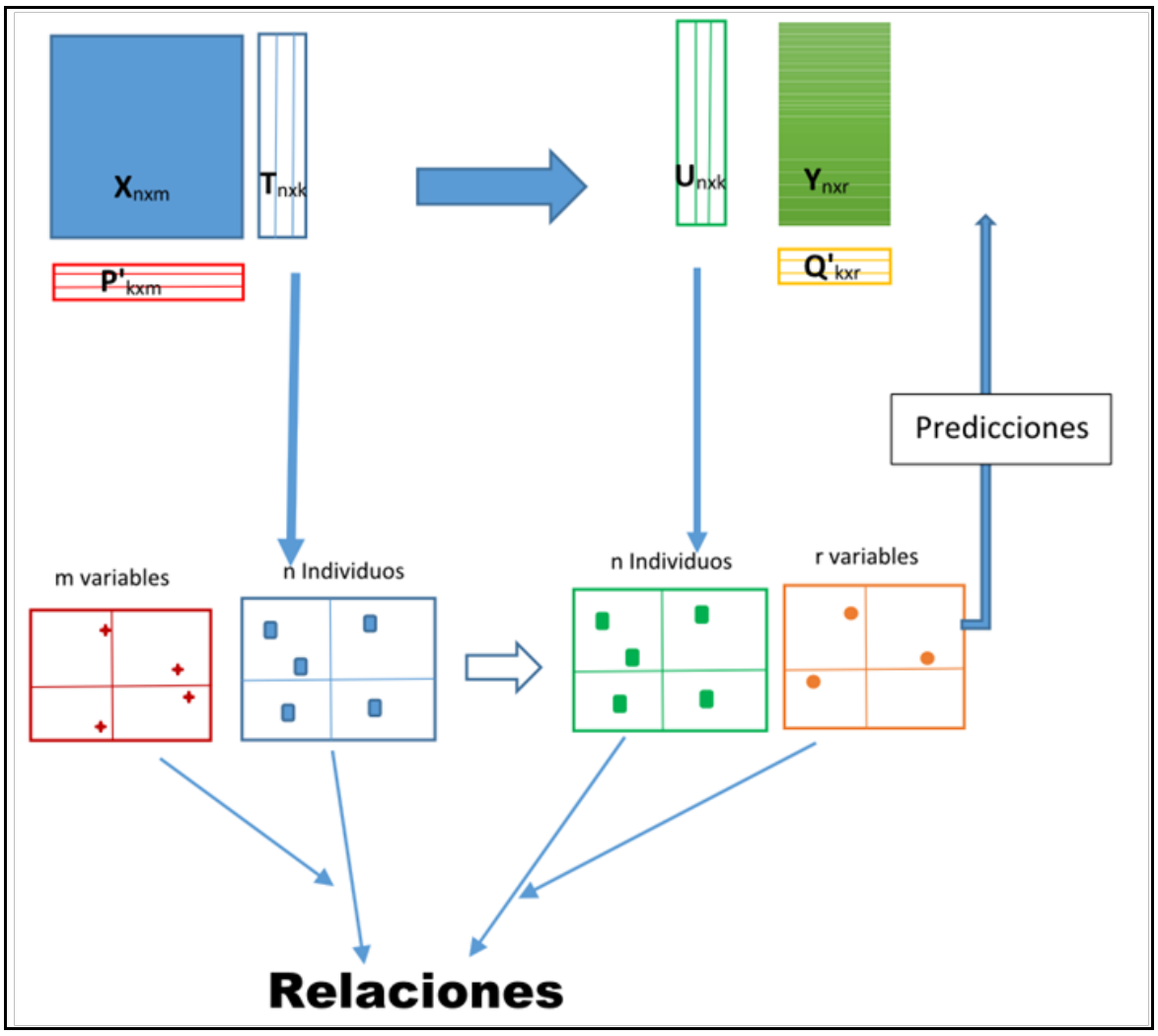

Figura 10. Esquema general de los resultados que proporciona PLS. La operación de transponer es aquí indicada por una prima en lugar de T, para evitar confusiones con la matriz de puntuaciones T.

A continuación se van a exponer estos resultados adaptando un ejemplo sencillo sobre las características de ciertos vinos en relación con determinadas variables dependientes (Abdi (2010). El algoritmo utilizado será el basado en SVD comentado anteriormente, siendo los resultados los que presenta el Paquete Estadístico *SIMFIT* que implementa dicho algoritmo. Los datos, además de centrados, han sido escalados a varianza unidad de acuerdo con Abdi (2010).

Considérese la matriz de variables predictoras **X** compuesta por 5 marcas de vino y 4 variables que incluyen su precio, su contenido en azúcar y alcohol y su acidez:

Vino\variable	Precio	Azúcar	Alcohol	Acidez
Vino 1	**7**	**7**	**13**	**7**
Vino 2	**4**	**3**	**14**	**7**
Vino 3	**10**	**5**	**12**	**5**
Vino 4	**16**	**7**	**11**	**3**
Vino 5	**13**	**3**	**10**	**3**

La correspondiente matriz de respuestas **Y** consiste en las puntuaciones que los expertos han dado a dichos vinos respecto a su calidad y su uso en gastronomía:

Vino\variable	Calidad	Para carnes	Para postres
Vino 1	**14**	**7**	**8**
Vino 2	**10**	**7**	**6**
Vino 3	**8**	**5**	**5**
Vino 4	**2**	**4**	**7**
Vino 5	**6**	**2**	**4**

Más adelante se analizará en detalle cómo determinar el número óptimo de factores PLS en un modelo. A los efectos ilustrativos del presente ejemplo, asumiremos que el número óptimo de factores es 2 y nos centraremos solamente en los aspectos de calibración o modelización, dejando para más adelante los aspectos de predicción.

Representaciones gráficas

PLS es una técnica que posee diferentes tipos de gráficos que ayudan a interpretar el comportamiento de las muestras y variables dentro del modelo. Se analizan a continuación las características de estas representaciones utilizando el Paquete *SIMFIT*.

En la Figura 11 se muestra la varianza explicada, de tipo acumulativa, de los factores latentes, tanto para las variables **X** como las **Y.** Como puede apreciarse, ya con 2 factores se ha explicado un 98 % de la variabilidad en **X** y un 85 % en **Y**. En base a que esta captura de variabilidad es aceptable podemos admitir a efectos de calibración que 2 factores PLS son suficientes para el modelo y este es el número elegido para los cálculos que seguirán después.

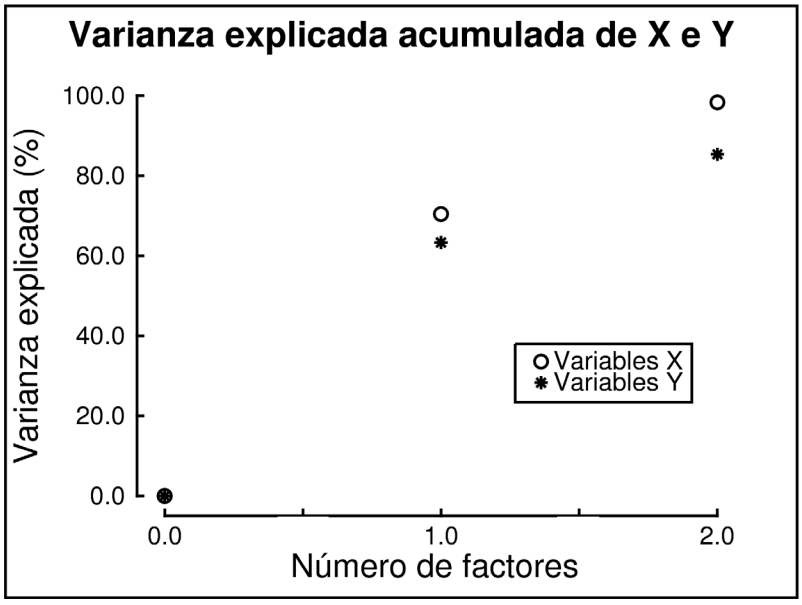

Fig. 11. Varianza explicada en promedio frente al número de factores.

En la Figura 12 se representan las puntuaciones ("scores **t"**) que los vinos presentan en su proyección a los dos factores latentes. Los distintos vinos están bastante separados y no se detectan agrupamientos (*clusters*). El vino 3 queda en el origen de coordenadas y sus variables **x** no estarían capturadas por los 2 factores seleccionados. El primer factor separa los vinos 1 y 2 de los 4 y 5, mientras que el segundo factor separa los vinos 2 y 5 de los 1 y 4.

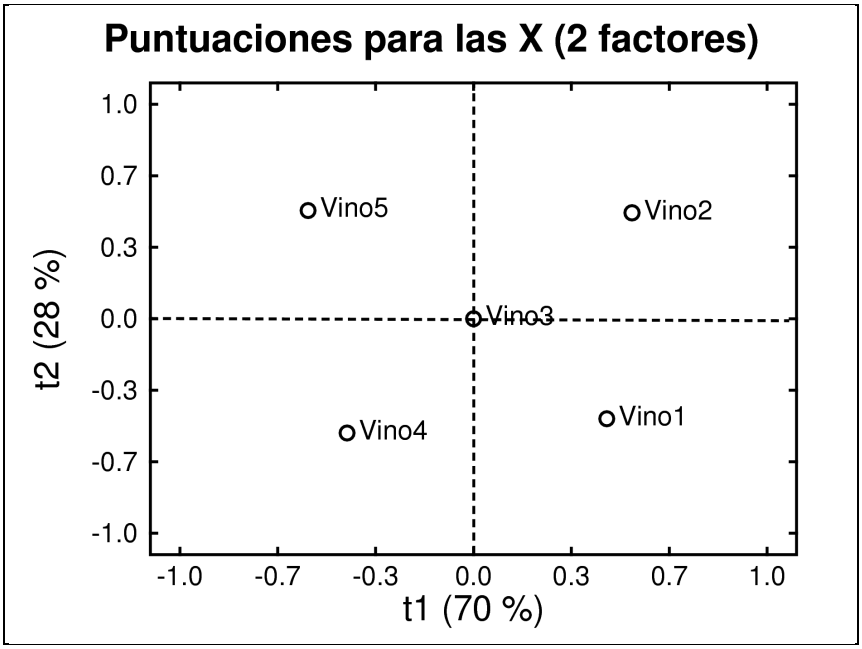

Fig. 12. Puntuaciones para las variables X bajo los 2 factores PLS.

En la Figura 13 se muestran las cargas (*loadings*) de las variables ***x*** para los dos factores. La acidez, el alcohol y precio contribuyen al factor 1, mientras que el contenido en azúcar es la principal contribución del factor 2. Asimismo, la acidez y el alcohol están fuertemente correlacionadas entre sí e inversamente correlacionadas con el precio (paradójicamente en este ejemplo).

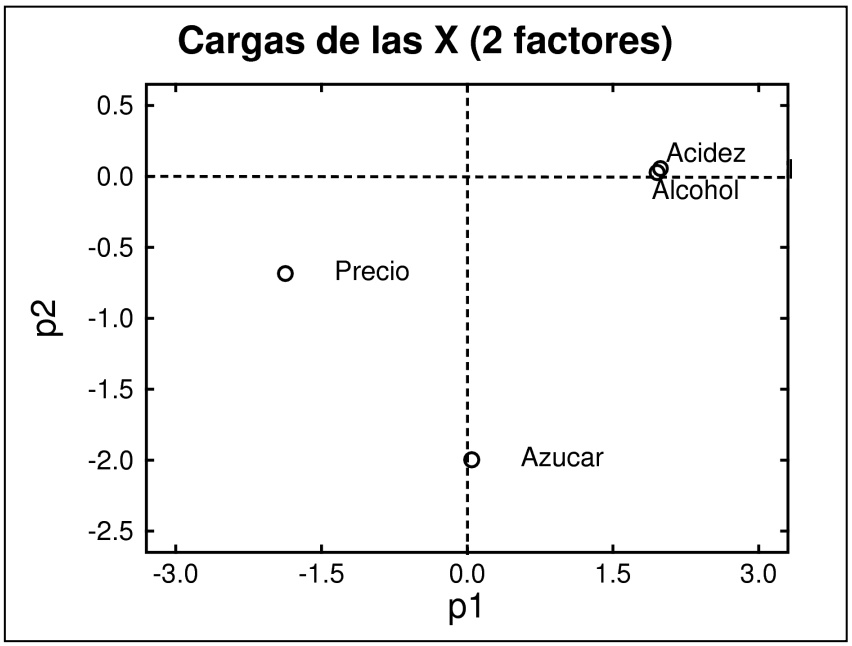

Figura 13. Cargas de las variables X para los dos factores PLS.

La Figura 14 recoge las puntuaciones de las **Y** para los diferentes vinos. La distribución de los vinos no presenta aquí tampoco agrupaciones destacables, sino que se encuentran bastante separados.

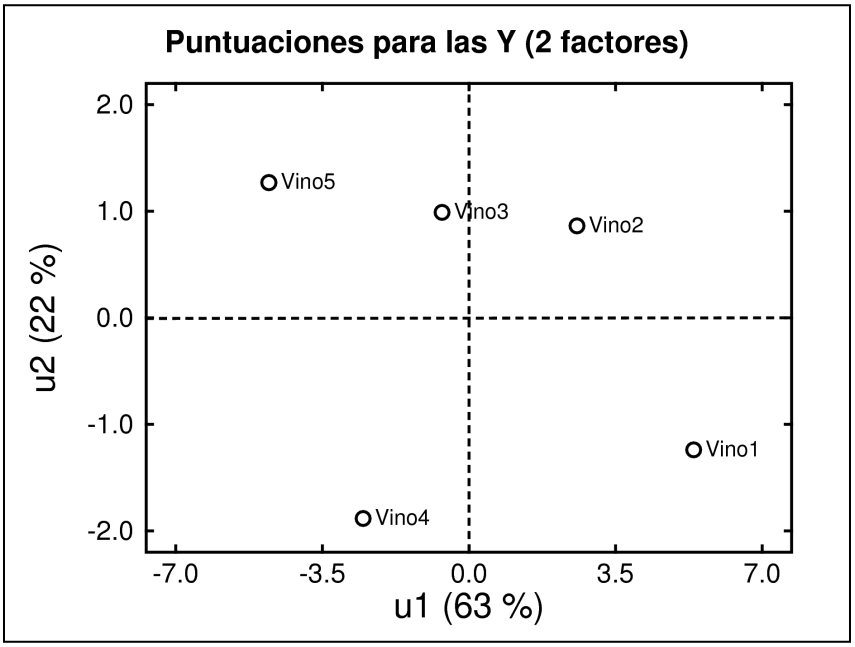

Figura 14. Puntuaciones para las Y en los dos factores PLS.

En la Figura 15 se han representado las cargas de los factores para las **Y**. Se puede apreciar como las 3 variables están bien representadas por el factor 1. El factor 2 captura principalmente el uso para postres, apreciándose que la calidad y el uso para carnes están ligeramente correlacionadas entre si y la calidad está inversamente correlacionada con el uso para postres.

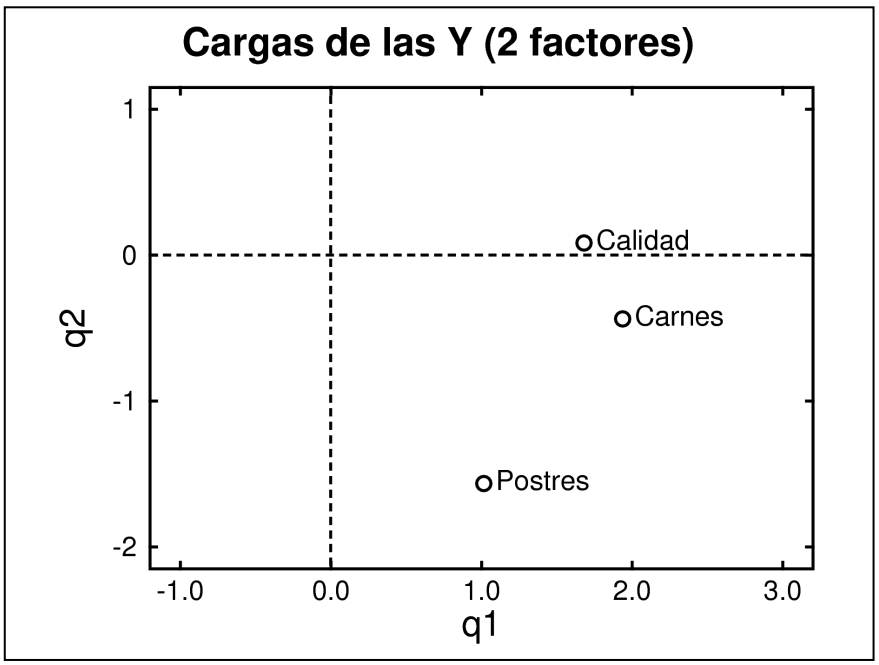

Figura 15. Cargas para las variables Y bajo los 2 factores PLS.

En la Figura 16 se muestran las correlaciones entre las puntuaciones de **X** (es decir ***t***) y las de **Y** (es decir ***u***) para los dos factores PLS. Las dos líneas de ajuste corresponden a las rectas de regresión de ***u*** sobre ***t*** (trazo continuo) y de ***t*** sobre ***u*** (trazo discontinuo), observándose un buen solapamiento entre ambas lo que significa un buen ajuste del modelo para los 5 vinos, lo que concuerda con valores de $p < 0.05$ para los coeficientes de correlación.

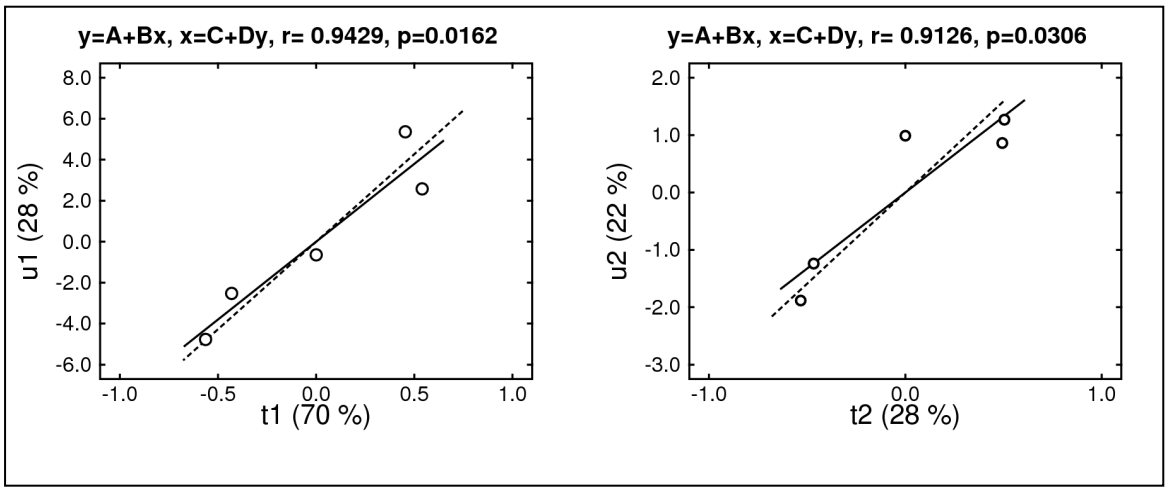

Figura 16. Correlaciones de las puntuaciones *t* y *u* para los 2 factores PLS. Línea continua = regresión de y sobre x. Línea discontinua = regresión de x sobre y.

Parámetros de regresión

Como se ha expuesto más arriba, al hablar de los algoritmos, un modelo PLS expresado en sus factores latentes (puntuaciones, cargas, etc.), se puede re-expresar en forma de un modelo de regresión multivariante relativo a las variables centradas y escaladas. En este caso, el tamaño y el signo de los parámetros indican la influencia de cada variable en el modelo, facilitando su interpretación. El formalismo matemático es el de una regresión lineal múltiple, lo mismo que su interpretación. Cada coeficiente se interpretará como el incremento que se espera que se produzca en la ***y*** por incremento unitario de ***x*,** suponiendo el resto de la regresoras constantes. Interpretación que puede hacerse, ya que ahora están libres del fenómeno de colinealidad. Este formalismo se expresa como:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

Los parámetros para el presente ejemplo con 2 factores y expresados en unidades internas (centradas y escaladas) son las siguientes:

Parámetros (β) / y	Calidad	Carnes	Postres
Precio	**-0.27**	**-0.25**	**0.019**
Azúcar	**0.063**	**0.32**	**0.79**
Alcohol	**0.29**	**0.36**	**0.27**
Acidez	**0.31**	**0.37**	**0.24**

Así la puntuación de un vino para postres en unidades estandarizadas vendría dada por:

$$postres = -0.019\,precio + 0.79\,azucar + 0.27\,alcohol + 0.24\,acidez$$

En este caso, se aprecia cómo el contenido en azúcar es el principal responsable que confiere al vino su idoneidad para postres, ya que presenta el parámetro positivo de mayor tamaño (0.79).

En la práctica, para evitar confusiones, los cálculos se suelen hacer en unidades originales (deshaciendo el centrado y el escalado), y ahora la regresión lineal múltiple presenta un término independiente:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

Los parámetros para el presente ejemplo con 2 factores y unidades originales son:

Parámetros (β) / y	Calidad	Carnes	Postres
Independiente	**-3.3**	**-3.4**	**-1.4**
Precio	**-0.26**	**-0.11**	**0.0063**
Azucar	**0.14**	**0.34**	**0.62**
Alcohol	**0.81**	**0.49**	**0.27**
Acidez	**0.69**	**0.40**	**0.19**

Así, la puntuación de un vino para postres en unidades originales vendría dada por:

$$postres = -1.4 - 0.0063 precio + 0.62 azucar + 0.27 alcohol + 0.19 acidez$$

Estos parámetros en unidades originales son los que se utilizarán para hacer predicciones de las variables respuesta que tendrán nuevos vinos. Pero antes habría que haber realizado un proceso de validación del modelo (2 factores), bien con una serie de prueba o por validación cruzada, pero de esto nos ocuparemos más adelante.

Estadístico VIP

La interpretación de un modelo PLS que tenga varias variables ***x*** y varias variables ***y*** puede resultar compleja. Un estadístico que viene a resumir la importancia de cada variable ***x*** fue propuesto por Wold (1994) con el nombre de VIP (de las siglas en inglés para "Variable Influence on Projection"). La puntuación de este estadístico para cada variable ***x*** se puede calcular de forma que refleje la influencia de dicha variable predictora sobre cada una de las variables individuales ***y*** por separado, o bien sobre todas las variables ***y*** en promedio. Para una variable ***x*** cualquiera "ℓ", su valor VIP, referido en promedio a todas las variables ***y***, que es lo habitual, viene dado por la expresión:

$$VIP_{\ell} = \sqrt{m \cdot \frac{\sum_{j=1}^{r} \sum_{i=1}^{k} \frac{w_{\ell,i}^2 \left(YCV_{i,j} - YCV_{i-1,j} \right)}{YCV_{k,j}}}{r}}$$

Donde el índice ***i*** se usa para indicar el factor latente, siendo ***k*** el número total de factores, refiriéndose el índice ***j*** a las variables ***y***, siendo ***m*** el número total de variables ***x,*** y siendo ***r*** el número total de variables ***y***. Por su parte $YCV_{i,j}$ se refiere al porcentaje acumulado de varianza capturada hasta el factor ***i*** para la variable ***j*** de las ***y***. Este valor es calculado en la

forma habitual, sumando el porcentaje de varianza capturada por dicho factor a la ya capturada por los factores anteriores.

El porcentaje de varianza capturada por un determinado factor, no es otra cosa que el cociente entre la varianza de las ***y*** estimadas por PLS para ese factor ($\hat{\mathbf{y}} = \mathbf{tq}^T$) y la varianza real de las ***y***, es decir el cociente entre $\mathbf{s^2(\hat{y})}$ y $\mathbf{s^2(y)}$ multiplicado por cien. Asimismo, $\mathbf{YCV_{i-1,j}}$ tiene un significado análogo al anterior, de manera que la diferencia $\mathbf{YCV_{i,j} - YCV_{i-1,j}}$ representa la varianza capturada al pasar del factor ***i-1*** al factor ***i*** . Este incremento de varianza capturada es luego multiplicado por $w_{\ell,i}^2$, que es el peso PLS asignado a la variable predictora ℓ en el factor ***i***.

Este producto entre el incremento de varianza explicada y el cuadrado del peso (**w**) con el que interviene la variable ***x*** en la proyección, viene a reflejar la importancia de dicha variable ***x*** en el modelo PLS con ***k*** factores. Los sumatorios y el índice ***r*** extienden los cálculos para obtener un valor VIP de la variable ***x*** que sea un promedio de su influencia en las diferentes variables ***y***.

El atractivo de la definición de VIP radica en su intrínseca parsimonia, ya que para un modelo dado existe solamente un vector que viene a resumir la importancia de las variables ***x*** en la predicción de las variables ***y*** ((Eriksson et al., 2006), pag 79). Así, en nuestro ejemplo el vector de valores VIP es el siguiente:

Variable / VIP	Estadístico VIP
Precio	**0.949**
Azúcar	**1.02**
Alcohol	**0.983**
Acidez	**1.05**

Un valor de VIP grande indica que la variable ***x*** tiene una importante influencia en la proyección. Además, ya que la media de los valores VIP al cuadrado es igual a 1, se suele tomar este valor como el punto de corte para interpretar si una variable es relevante o no. Así, aquellas variables que tienen valores de VIP más grandes que 1 son consideradas como las variables predictoras más importantes en la matriz **X**. En nuestro caso, la variable acidez seguida de la variable azúcar serían las más relevantes en el modelo, si bien las diferencias entre los valores VIP de todas las variables no son grandes.

A modo de ejemplo, podríamos mostrar el cálculo detallado del estadístico VIP para la variable "precio" en nuestro modelo PLS de 2 factores. Para ello necesitamos los valores calculados para las varianzas acumuladas de las variables ***y*** por separado, así como los pesos de las variables predictoras ***x*** en cada factor.

Las varianzas acumuladas para las ***y*** presentan los siguientes valores:

Factor/y	Calidad	Carnes	Postres	Media
Factor 1	70.53	93.74	25.72	63.33
Factor 2	70.71	98.51	86.97	85.40

Los pesos (**w**) obtenidos para las variables ***x*** son los siguientes:

Factor/w	Precio	Azúcar	Alcohol	Acidez
Factor 1	-0.5137	0.2010	0.5705	0.6085
Factor 2	-0.3379	-0.9400	-0.01878	0.04286

Aplicando la expresión dada más arriba, el cálculo del estadístico VIP de la variable "precio" sería:

$$VIP = \sqrt{4\frac{\left(\begin{array}{l}\frac{70.53(-0.5137)^2 + (70.71-70.53)(-0.3379)^2}{70.71} + \\ + \frac{93.74(-0.5137)^2 + (98.51-93.74)(-0.3379)^2 + }{98.51} \\ + \frac{25.72(-0.5137)^2 + (86.97-25.72)(-0.3379)^2}{86.97}\end{array}\right)}{3}}^2 = 0.951$$

Como era de esperar, el valor obtenido (0.951) cierra perfectamente con el que aparece en la correspondiente tabla (0.949) mostrada más arriba.

3.4.4. Determinación del número óptimo de factores latentes

Es fácil de apreciar que cuanto mayor es el número de de factores latentes que se incluyan en el modelo, mayor será la bondad del ajuste para la serie de datos llamada de entrenamiento, pero se puede llegar a una situación de hiperajuste que no sería deseable, sobre todo si se desea utilizar el modelo PLS para la predicción de las respuestas de una nueva matriz **X**. Las estrategias para encontrar el óptimo número de factores es un tema controvertido que ha dado lugar a numerosas investigaciones, por lo que merece una revisión detallada que se aborda a continuación.

Introducción al concepto de calibración y validación

El punto de partida de un modelo PLS son 2 matrices, una matriz **X** de variables predictoras y una matriz **Y** de variables respuesta, ambas conteniendo el mismo número de casos. Estas 2 matrices constituyen la llamada serie de entrenamiento o calibración ("training set") y debe ser representativa de la población de la que posteriormente se sacaran nuevas matrices **X** para predecir las respuestas. Otra condición es que la serie de calibración debe abarcar el rango de las variables ***x*** e ***y*** tanto como sea posible, ya que estos márgenes definirán la región de aplicación del modelo en las predicciones posteriores (Esbensen (2010)).

En un modelo de predicción, la "validación" del mismo significa analizar su capacidad predictiva con una nueva serie de datos. Esta nueva serie se la denomina serie de prueba ("test set"), que debe haber sido muestreada de la misma población diana de la que se extrajo la serie de entrenamiento.

A partir de una validación se obtienen resultados cuantitativos importantes, en especial la determinación del número óptimo de factores latentes a usar en el modelo de calibración, así como la estima estadística de los errores de predicción.

Validación con serie de prueba o "test set"

Se entiende por serie de prueba una serie completamente nueva de datos compuesta por una matriz **X** de predictores y su correspondiente matriz **Y** de respuestas. La validación consiste en permitir que el modelo calibrado prediga las estimas de las respuestas ($\mathbf{Y}_{\mathbf{predicha}}$) y las compare con los valores reales ($\mathbf{Y}_{\mathbf{real}}$). Los resultados de esta comparación se pueden expresar como errores de predicción o varianza residual, que cuantifican tanto la exactitud como la precisión de los valores ***y*** predichos, es decir los niveles de error que cabe esperar en predicciones futuras.

Validación cruzada (Cross Valiation (CV))

Según Esbensen (2010) no existe mejor validación que la validación con serie de prueba anteriormente expuesta. No obstante el precio que hay que pagar es que se necesita el doble de muestras de las que serían necesarias si sólo se utilizase para la validación la propia serie de entrenamiento. A veces hay situaciones en la que obtener el doble de muestras no es posible por razones éticas o de coste. En estas situaciones una alternativa de validación es la validación cruzada. Ésta consiste en apartar una porción de casos de la serie de entrenamiento, el resto de los datos se utiliza para la construcción del modelo PLS y posteriormente se utilizará este modelo para predecir las respuestas de los casos apartados.

Hay tres grandes variantes de validación cruzada (Esbensen (2010), la denominada *dejar uno fuera* (l*eave one out*), la *validación cruzada segmentada* (*k-fold cross validation*) y la *división de la serie de entrenamiento en dos partes* (s*plitting the data*):

- La opción *dejar uno fuera* consiste en apartar un caso cada vez. Con los casos restantes se forma una serie de entrenamiento con la que se seleccionan las variables y se construye el modelo predictor deseado. Luego con dicho modelo se predicen los valores de ***y*** del caso que se dejó fuera. Este mismo procedimiento se repite para cada caso y al final se hace un promedio de los residuales entre los valores predichos y los reales elevados al cuadrado.

- La *validación cruzada segmentada* se basa en repartir la serie de entrenamiento en k grupos, donde k suele ser 3, 5 o 10. Pensemos, por ejemplo, que la serie de prueba la dividimos en 10 grupos (10% de casos en cada grupo), hecho esto se aparta un grupo y con los otros 9 se seleccionan las variables y se construye el modelo predictor. A continuación se predicen los valores de ***y*** de los casos que se dejaron a un lado y se hace un promedio de los residuales al cuadrado. El procedimiento se repite apartando el siguiente grupo e incorporando el anterior. Finalmente se calcula el valor promedio global de residuales al cuadrado con todos los grupos.

- Según Esbensen (2010) la alternativa ideal a la validación con serie de prueba es la llamada "división de la serie de entrenamiento en dos partes". Este procedimiento consiste simplemente en dividir los datos iniciales en dos grupos A y B. Luego se hace una calibración del modelo usando A y éste se prueba con B.

A continuación se hace un cambio, se construye el modelo con B y ahora se prueba con A. Finalmente se calcula el error de predicción como la media de los residuales al cuadrado de ambos grupos.

Se puede concluir que la economía de datos a veces sólo permite usar la serie de calibración para buscar la dimensionalidad óptima del modelo PLS y el error de predicción mediante algún método de validación cruzada. Pero en el caso ideal, incluso la determinación de la dimensión óptima del modelo (número óptimo de factores ($\boldsymbol{k_{opt}}$)), además del cálculo del error de predicción, debiera llevarse a cabo usando una serie de prueba.

Este es un punto controvertido que merece algún comentario adicional. Hay autores que sugieren optimizar el modelo (encontrar $\boldsymbol{k_{opt}}$) por validación cruzada (Osten (1988), Martens and Naes (1989)) y sólo cuando el modelo está totalmente construido validarlo con una serie de prueba que no haya participado en la optimización del modelo. Por el contrario, otros autores como Esbensen and Geladi (2010), no están de acuerdo con dicha aproximación y defienden que una segunda serie de datos (serie de prueba) resulta absolutamente necesaria incluso para optimizar el modelo, con el fin de incluir los errores de muestreo de las situaciones futuras en las que el modelo habrá de operar. Para terminar, convendría citar algunas de sus propuestas concretas:

- *Nada adverso resulta de aplicar siempre una validación por serie de prueba, lo cual proporciona una información completa en una sola operación, ya que una validación con serie de prueba de hecho suministra estimas tanto del número óptimo de factores ($\boldsymbol{k_{opt}}$) como de la* ***"varianza Y residual"****, mientras que todo supone un cierto riesgo si se usa una validación cruzada.*

- *No resultaría correcto el usar validación cruzada para la determinación de* $\boldsymbol{k_{opt}}$*. Usando una validación por serie de prueba, se obtiene el número óptimo de factores PLS y la estima más precisa de la* ***"varianza Y residual"****, ya que esta validación incluye todas las incertidumbres de muestreo y de medida originadas por cambios circunstanciales en las condiciones de trabajo.*

Diferencia entre "error de calibración" y "error de predicción"

Las técnicas de validación expuestas anteriormente están todas diseñadas para evaluar la capacidad predictora del modelo, es decir la exactitud asociada a $\mathbf{Y_{predicha}}$ en su comparación con $\mathbf{Y_{real}}$. Parece que, cuanto mayor sea el número de factores que usemos en el modelo, menor será la diferencia entre $\mathbf{Y_{predicha}}$ e $\mathbf{Y_{real}}$, pero solamente hasta un punto que es el llamado número óptimo de factores. De este aspecto nos ocuparemos a continuación.

Hasta ahora la matriz **Y** de respuestas podía estar formadas por varias variables. Normalmente es frecuente utilizar todas las variables de forma exploratoria pero en la práctica luego se hacen modelos PLS por variables aisladas. En lo que sigue, nos referiremos a un vector $\mathbf{Y(n\times 1)}$ con sólo una variable respuesta. Empecemos distinguiendo entre lo que es el error de calibración y el error de predicción.

Error de calibración (o de modelización)

Imaginemos que se ha construido un modelo PLS basado en la matriz de calibración $\mathbf{X_{cal}(n\times m)}$ y en el vector de calibración $\mathbf{Y_{cal}(n\times 1)}$ y que tentativamente

asumimos ***k*** factores latentes como correctos. Asumimos también que nosotros utilizamos los propios valores $\mathbf{X_{cal}}$ para introducirlos en el modelo y predecir unos valores $\hat{\mathbf{Y}}$. Haciendo esto obtendremos una idea del "error de modelización", debido al hecho de que solamente hemos usado ***k*** factores en el modelo y no todas las variables ***m***. Este error se puede calcular con la llamada **"varianza residual de calibración"** (Esbensen (2010) , págs. 157-160 y 201-202) dada por la expresión:

$$Varianza\ \ residual_{cal} = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}$$

Una expresión análoga en términos de varianza explicada sería:

$$Varianza\ \ \exp licada_{cal} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Error de validación

El error de predicción se suele expresar como la varianza residual promedio de **Y**, obtenida mediante validación con una serie de prueba o por validación cruzada, y análogamente a la expresión ya vista se escribe como:

$$Varianza\ \ residual_{val} = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}$$

La expresión correspondiente en términos de varianza explicada sería:

$$Varianza\ \ \exp licada_{val} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Representación de las varianzas de calibración y predicción

Las varianzas vistas anteriormente, de calibración y validación, se pueden representar tanto como varianzas residuales como varianzas explicadas, ya que ambas no son sino expresiones alternativas basadas en los mismos datos. Así, en la Figura17 se ha representado en ordenadas la varianza residual de calibración y validación frente al número de factores en abscisas para unos datos de ejemplo. Como puede apreciarse, la "varianza Y residual" de calibración disminuye progresivamente a medida que aumenta el número de factores, llegando a valer cero cuando el número de factores se hace igual al número de variables ***x***. Por el contrario la "varianza Y residual" de validación disminuye hasta un mínimo y luego crece a medida que el número de factores aumenta. Normalmente de lo que se trata es de encontrar este mínimo y considerar el número de factores en dicho mínimo como el número óptimo de factores para el modelo.

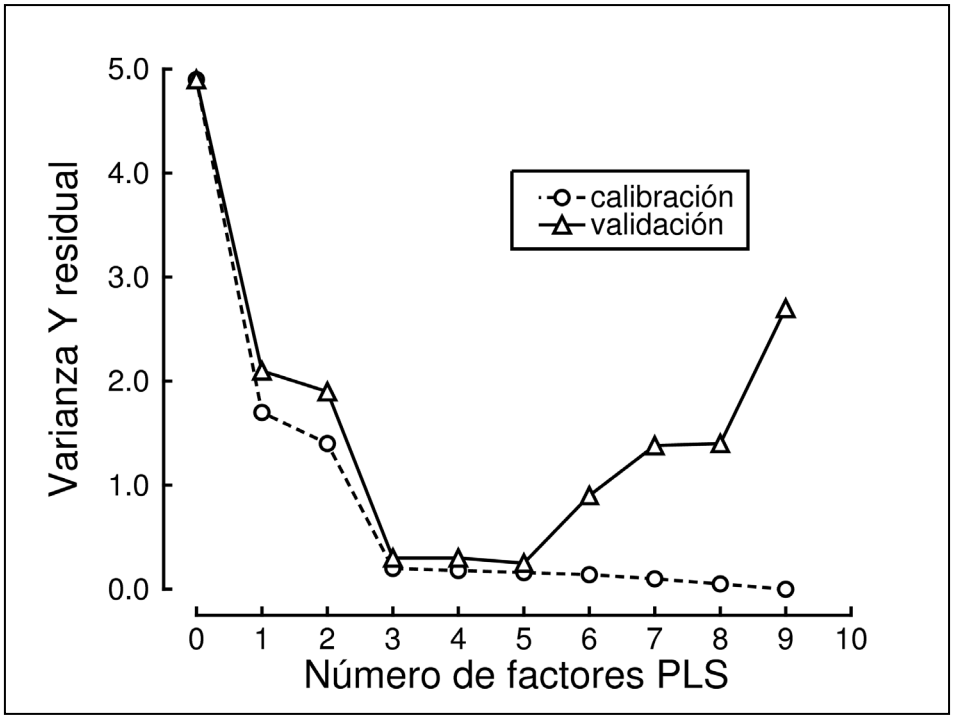

Figura 17. Número de factores frente a la "varianza Y residual".

En este ejemplo, se aprecia un mínimo claro en 3 factores, que indica que este número de factores es el óptimo, es decir el número donde la varianza Y residual de validación (predicción) ha sido minimizada. La inclusión de un número mayor de componentes podría mejorar el ajuste específico del modelo de calibración, pero claramente reduciría su capacidad de predicción, porque la varianza residual de validación aumenta a partir de este número. Por tanto, desde el punto de vista de optimizar la predicción, este mínimo representa la óptima complejidad del modelo PLS, es decir el "correcto número de factores a usar en el modelo de predicción" (Esbensen (2010), pág. 123).

Las curvas complementarias de varianzas explicadas para calibración y validación se han representado en la Figura 18. Como puede apreciarse la proporción de varianza total capturada aumenta a medida que aumenta el número de factores en el caso de la calibración y alcanza un máximo para la situación de validación.

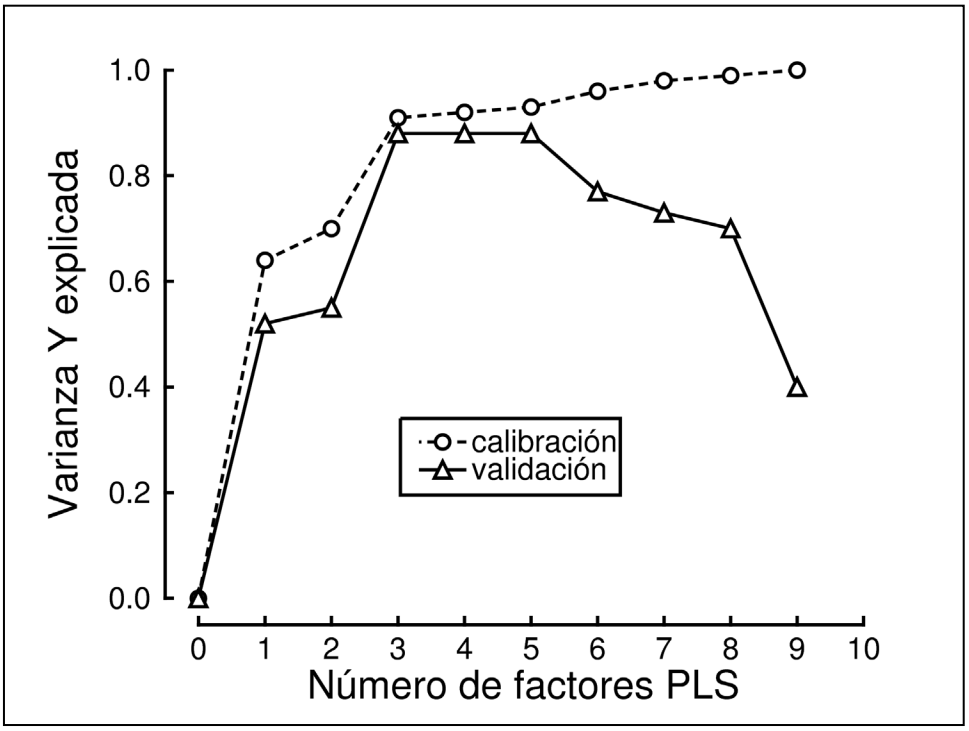

Figura 18. Número de factores frente a la "varianza Y explicada".

3.4.5. PLS con variables categóricas

Codificación en variables ficticias

En rigor PLS se basa en el uso de variables continuas, pero también se puede utilizar con variables categóricas haciendo uso de variables ficticias ("dummy"). Este tipo de codificaciones se ha visto, experimentalmente, que da buenos resultados, siempre que no haya demasiadas variables categóricas en comparación con las variables continuas.

Esta variante de PLS ha encontrado numerosas aplicaciones prácticas principalmente cuando la variable respuesta ***y*** es la categórica, ya que permite la predicción de la categoría de muestras nuevas a partir de las respectivas variables ***x*** predictoras, eligiendo un punto de corte adecuado ("cut-off") para la asignación de la categoría. Algunos autores han denominado esta modalidad como "Modelización PLS discriminante" (Esbensen (2010)).

Tasa de error de clasificación cuando la respuesta es dicotómica

Como se acaba de comentar, la técnica de PLS-Discriminante tiene aplicación práctica a efectos de clasificación. Así, si la respuesta es dicotómica la matriz **Y** de respuestas será un vector de ceros y unos y si la respuesta es politómica la matriz **Y** estaría formada por tantas columnas con ceros y unos como categorías haya, utilizando la recodificación arriba indicada. La matriz **X(n×m)** de predictores tiene por su parte la forma habitual en variables continuas, aunque también podría incluir alguna variable categorizada codificada en variables postizas.

El caso más utilizado en el presente trabajo es de una variable ***y*** respuesta dicotómica, en cuyo caso se tiene el siguiente vector: $\mathbf{Y(n\times 1)} = (0, 0, 0, 0...1, 1, 1, 1)^T$

siendo ***n*** el número de muestras. Se trabaja en la forma usual elaborando secuencialmente modelos con ***1,2….k*** factores latentes. Luego se procede a hacer la predicción de las muestras nuevas de la serie de prueba como ya se ha comentado. Sólo que ahora hay un problema, éste consiste en que el vector **Y-predicho** viene expresado en valores continuos, por lo que hay que elegir un punto de corte para clasificar una muestra como **0** o como **1**. En este caso dicotómico, el corte elegido suele ser 0.5, de forma que si el ***y-predicho*** es < 0.5 se le asigna la clase **0** y si es >= 0.5 se le asigna la clase **1**.

Como ya se ha comentado en el apartado 3.4.4, hay que determinar el número óptimo de factores PLS mediante validación con la serie de prueba. Pero ahora no se suele utilizar el concepto de varianza residual de validación, sino el de tasa de error de clasificación (*error rate*). El procedimiento es sencillo, para cada componente PLS que se va introduciendo en el modelo se calcula, en base al punto de corte, el número de muestras mal clasificadas (tasa de error) y el número de componentes PLS óptimo será aquel que presente un valor mínimo para la tasa de error.

Una vez establecido el modelo óptimo, éste se puede utilizar para clasificar muestras nuevas, lo cual tiene interés en muchas ciencias, particularmente en medicina.

Si en lugar de 2 categorías la respuesta ***y*** fuese multicategórica, se procede de una forma totalmente análoga a la expuesta más arriba, teniendo la precaución de que ahora el criterio de asignación a una clase a partir de las Y-predichas vendrá dado por el valor más alto observado en una clase, y cuanto más se aproxime a 1 mejor será su asignación.

3.4.6. Aplicaciones de PLS en Genómica

La tecnología de *microarrays* permite la medida de niveles de expresión de miles de genes simultáneamente, que se usan luego para caracterizar el perfil génico de las enfermedades, la respuesta a los tratamientos o la evaluación de pronósticos. Una de las aplicaciones más comunes es la de comparar los niveles de expresión génica en dos condiciones diferentes, tal como células tumorales frente a células sanas. Normalmente se acometen dos tareas: la identificación de genes biomarcadores y la construcción de modelos predictivos, basados en los genes identificados, que permitan la asignación de nuevos pacientes a los grupos analizados.

Usos de PLS como método de reducción de la dimensión

Una situación importante con datos de *microarrays* es la clasificación binaria, en la que se adapta PLS usando un vector **Y** codificado con valores "0" para el caso de ausencia del evento y "1" para su presencia. Esta aproximación se basa en buenos resultados empíricos pero no está justificada teóricamente, ya que ahora la respuesta es dicotómica y no es una variable continua, como requiere en rigor el método PLS. A pesar de todo, ha habido algún esfuerzo para encontrar alguna explicación teórica a los buenos resultados empíricos (Barker and Rayens (2003)). Por otra parte, la extensión de PLS a la situaciones de multiclasificación es fácil, consiste simplemente en construir una matriz con codificación postiza para las diferentes clases (Boulesteix (2004).

Las primeras aproximaciones de clasificación en Genómica estaban basadas en usar PLS como un procedimiento de reducción de la dimensión. Así, Nguyen and Rocke (2002a) analizaron datos de *microarrays* de muestras procedentes de tumores humanos para realizar clasificación binaria y predicción de clase de muestras nuevas. Después de

una selección preliminar de genes, siguieron un procedimiento en dos pasos que consistía en una reducción de la dimensión por PLS seguida de una clasificación por regresión logística binaria o análisis discriminante cuadrático. Estos investigadores también extendieron su metodología a la multiclasificación de otros tipos de cáncer (Nguyen and Rocke (2002b)). Por su parte, Boulesteix (2004), trabajando con 9 series de datos de *microarrays* con muestras tumorales, estudió un procedimiento de clasificación que consistía en una reducción de la dimensión por PLS seguida de un análisis discriminante lineal sobre los factores latentes obtenidos con PLS; a su vez hizo una comparación entre sus resultados y los obtenidos con otros métodos de clasificación tales como LDA, KNN, PAM and SVM. La autora propuso que PLS es una herramienta competitiva para problemas de clasificación que, además, puede manejar todos los genes simultáneamente. Otros autores como Sampson et al. (2011) probaron la utilidad de los métodos de clasificación basados en PLS con series de datos de proteómica clínica. Usaron PLS para una reducción de la dimensión en combinación con los métodos usuales de clasificación. Así, compararon PLS+LDA, PLS+RF, SVM y PCA+LDA. Encontraron que SVM conseguía la clasificación más eficiente en la mayoría de las series de datos probadas, pero que los clasificadores basados en PLS proporcionaban información adicional tal como las cargas de las proteínas en los factores y diferentes tipos de gráficas.

La anterior aproximación basada en la reducción de la dimensión por PLS puede resolver el problema de la predicción de clase, pero desafortunadamente los factores latentes no resultan informativos para la selección de genes y tienen una interpretación bioquímica pobre. Para mejorar esta aplicación biomédica de PLS con datos de *microarrays*, se deben abordar dos aspectos: la detección de los mejores genes discriminatorios entre los miles de genes ensayados en el microarray, y la selección del

correcto número de factores latentes que han de formar parte del modelo predictivo. Desafortunadamente, ambas tareas dependen de la estructura particular de los datos y del enorme ruido que proviene del elevado número de genes medidos en el *microarray*. Sin embargo, algunos métodos han sido propuestos para alcanzar ambos objetivos y ciertos autores han usado PLS para realizar al mismo tiempo una selección de los genes y la construcción de un modelo de predicción con los factores latentes adecuados. De la revisión de estos aspectos nos ocuparemos en el siguiente apartado.

Usos de PLS para hacer a la vez selección de genes y predicción

PLS fue designado específicamente para predicción, con el objetivo de buscar un número mínimo de factores latentes que permitiese la estimación de una nueva matriz **Y** a partir de una nueva matriz **X** , por lo que no estaba especialmente enfocado a la selección de variables, no obstante también sirve para este fin.

La reciente revisión de Mehmood et al. (2012) categoriza los métodos de selección de variables con PLS en tres categorías principales: métodos de filtrado, de envoltura y embebidos. A continuación se revisarán las aplicaciones de estos métodos publicadas en la bibliografía referida al campo de la Genómica y áreas afines, haciendo especial énfasis en los métodos de envoltura, ya que éstos son los directamente relacionados con el presente trabajo.

Métodos de filtrado

Estos usan los resultados de PLS simplemente para identificar una subclase de variables significativas. Trabajan en dos pasos: ajustan primero un modelo PLS y luego realizan una selección de variables usando un umbral de corte con algún parámetro de relevancia. Para medir la importancia de las variables, normalmente se utilizan tres

parámetros de filtrado: los pesos de las ***variables x*** (***w***), los coeficientes de regresión (***β***) o las puntuaciones VIP. Un ejemplo podría ser el trabajo de Viala et al. (2007) que estudiaron variables clínicas y efectos de calidad de vida comunicados por los pacientes para predecir supervivencia durante el tratamiento con bortezomib. Usaron PLS y filtraron las variables originales en base a la significancia de los coeficientes de regresión de dichas variables en el modelo PLS.

Métodos de envolvente

Estos algoritmos son de tipo iterativo. Usan algún método de filtrado para extraer una serie de variables relevantes y a continuación reajustan el modelo, y así sucesivamente hasta que se alcanza algún criterio de bondad. De acuerdo con Mehmood et al. (2012) hay diferentes clases de algoritmos de "envolver", tales como el algoritmo genético combinado con PLS (GA-PLS), eliminación de variables hacia atrás combinado con PLS ("Backward variable elimination PLS (BVE-PLS)"), PLS con pesos iterativos de las variables predictoras (IPW-PLS), etc. El método más relacionado con el presente trabajo es el método BVE-PLS. El fundamento de este método BVE-PLS es simple: primero se clasifican las variables de acuerdo con algún criterio de filtrado, tal como los pesos de las variables ***x***, los coeficientes de regresión o las puntuaciones VIP, y luego se usa un umbral de corte para eliminar las variables menos importantes. Finalmente se reajusta de nuevo un modelo PLS utilizando las variables restantes. Este procedimiento se repite iterativamente hasta que se alcanza un funcionamiento óptimo del modelo. Aquí nos centraremos en reunir las publicaciones aparecidas hasta la fecha de este método BVE-PLS, haciendo especial énfasis en los trabajos que se refieren a Genómica.

Cho et al. (2002) son tal vez los primeros autores que usaron el método BVE-PLS con datos de *microarrays*. Hicieron una clasificación de 2 subtipos de leucemias usando una serie de datos publicados de chips de oligonucleótidos que tenían inicialmente 7129 genes. Primero hicieron una preselección de genes con el test "t", reduciendo los genes a 1178 diferencialmente expresados. Luego esos genes los introdujeron en un algoritmo PLS iterativo que realizaba una selección de variables basada en elegir aquellas que tenían VIP > 1 y evaluaron la bondad de clasificación por validación cruzada del tipo "dejar uno fuera". Encontraron que el método BVE-PLS con selección por valores VIP mostraba una buena capacidad de clasificación y de predicción de clase.

Perez-Enciso and Tenenhaus (2003) estudiaron la técnica PLS para clasificación usando datos publicados de *microarrays* de pacientes con cáncer de mama. Analizaron 62 muestras y 1753 clones cDNA para evaluar la habilidad discriminante de PLS en base a tres criterios de cáncer, observando que el mejor comportamiento se encontraba con las muestras de "antes" y "después" del tratamiento con quimioterapia. Para esta comparación, llevaron a cabo un primer modelo PLS con el que hicieron una selección de clones jerarquizándolos por su valor VIP, con este criterio se quedaron con aquellos 18 clones cDNA de los1753 que presentaban un VIP > 2. Luego hicieron un segundo PLS usando los clones cDNA seleccionados y construyeron un modelo que presentó buenas propiedades predictivas.

Recientemente, Mehmood et al. (2011) han estudiado un nuevo algoritmo de clasificación basado en un método PLS de rango-reducido con eliminación parsimoniosa de variables bajo tres criterios diferentes: pesos, coeficientes de regresión y valores VIP. Todos dieron buenos resultados cuando se aplicaron a secuencias genómicas para clasificación del *phylum* bacteriano. En 2014, han aparecido dos publicaciones que

también utilizan PLS con selección de variables por VIP, una analizando epilepsia (Wang et al., 2014) y otra dedicada al fallo renal (Ding et al., 2014).

Los métodos para selección de variables han sido el objeto de numerosos estudios en otros campos distintos de la Genómica y algunos de ellos merecen ser mencionados. Así, Gauchi and Chagnon (2001) realizaron una amplia comparación de métodos de selección de variables en PLS dentro del marco de los procesos de manufacturación, concluyendo que una selección paso a paso hacia atrás basada en el criterio de máximo Q^2_{cum} es el más recomendable en situaciones prácticas. Este grupo extendió más tarde esta investigación a otro métodos (Lazraq et al. (2003)), incluyendo el ordenar los valores VIP en orden decreciente para retener los valores VIP más altos en el modelo, pero sin realizar más iteraciones PLS.

En el campo de la ingeniería, Chong and Jun (2005) compararon el funcionamiento de la eliminación de variables por valores VIP (método PLS-VIP) con otros métodos como LASSO y regresiones paso a paso. Usaron series de datos simuladas que incluían 500 muestras y hasta un máximo de 100 variables, encontrando que el método PLS-VIP se comportaba muy bien para identificar predictores relevantes y mejoraba incluso otros métodos conocidos.

Métodos de embeber

Aquí la selección de variables es parte del algoritmo PLS. Estos métodos encadenan la selección de variables dentro del propio algoritmo de PLS. Entre ellos existen varias aproximaciones en el marco de la regresión penalizada, que introduce penalizaciones en el modelo. Así, Huang and Pan (2003) publicaron un buen funcionamiento de la regresión PLS penalizada con datos de *microarrays* de muestras de

cáncer y clasificación en dos clases. Este grupo también probó su método con datos de *microarrays* con muestras de corazón humano (Huang et al. (2004)). Por su parte, Le Cao et al. (2008) propusieron un método de selección de variables conocido como PLS escaso o poco denso ("Sparse PLS"), donde la "escasez" se alcanza con una penalización LASSO sobre los vectores de las cargas PLS en el momento de hacer el cálculo SVD. Más tarde, este grupo extendió su método a problemas de multiclasificación (Le Cao et al (2011)). Al mismo tiempo, Chung and Keles (2010) implementaron un algoritmo de "PLS escaso" que probaron con datos simulados y reales, observando buenos resultados para la selección de variables y predicción. Recientemente se ha publicado una variante de "Sparse PLS" que introduce ciertos parámetros de ajuste adicionales (Olson Hunt et al. (2014)).

Como se ha comentado más arriba, la regresión PLS con selección de variables por los valores del estadístico VIP (PLS-VIP) ha sido aplicada en otras ocasiones, especialmente en Quimiometría, pero su uso ha sido más limitado en análisis de datos de *microarrays*, donde existe menos información acerca de su funcionamiento. El objetivo del presente trabajo ha sido el de llevar a cabo un estudio sistemático de PLS sobre datos simulados por ordenador, con el fin de explorar como opera la estrategia PLS-VIP bajo diferentes condiciones de tamaño de muestra, número de genes discriminatorios y genes ruido. Asimismo se ha estudiado las características de este procedimiento PLS-VIP con variables clínicas y genómicas, tanto separadas como combinadas. Para investigar todo esto, se ha implementado un algoritmo PLS que realiza una selección de variables mediante un PLS iterativo que usa eliminación hacia atrás en base a los valores VIP. El algoritmo encuentra el número óptimo de variables discriminatorias y de factores PLS, permite la identificación de dichas variables y presenta un buen funcionamiento para la predicción de nuevas muestras. También se han realizado comparaciones con otros

métodos de clasificación y se han analizado datos reales para probar la bondad del algoritmo propuesto.

3.4.7. Análisis conjunto de datos clínicos y génicos en Genómica

El empleo de *microarrays* se ha utilizado en la investigación de diferentes patologías, especialmente en el estudio del cáncer. Desde el principio, uno de los grandes paradigmas fue el poder usar la expresión de ciertos genes para que actuaran como marcadores de la enfermedad. Para ello, se construían modelos predictores con esos genes, normalmente para clases de tipo dicotómico (no responde, responde), y se trataba de predecir a que clase pertenecería un nuevo paciente al que se medía la expresión de dichos genes marcadores usando un *microarray*.

Desafortunadamente el construir estos modelos de predicción de clase con datos de *microarrays* es una tarea difícil. Se ha llevado a cabo un gran esfuerzo para desarrollar algoritmos que funcionaran con bajos errores de predicción, pero en numerosas ocasiones los genes y su potencia predictiva variaba entre unos estudios y otros, debido al ruido procedente de los miles de genes de un microarray, los pequeños tamaño de muestra empleados, al elevado coste de los *microarrays* y a la variabilidad entre los pacientes. De tal forma que, modelos predictores que funcionaban bien en unas determinadas condiciones, no han podido ser validados a veces en otras condiciones semejantes. Todo esto ha hecho que el uso de *microarrays* como herramienta de pronóstico no se haya convertido, de momento, en un procedimiento rutinario en la práctica clínica. Además, la tecnología de *microarrays* es todavía cara comparada con los factores pronóstico convencionales.

Teniendo en cuenta los anteriores condicionantes, han existido siempre varias preguntas a las que dar respuesta: a) ¿Las variables clínicas no serán suficientes y más asequibles como predictoras que los datos génicos de los *microarrays*?, b) ¿No debieran

tener más potencia los modelos predictores basados en datos de *microarrays* que los basados en marcadores clínicos?, c) ¿No podrían complementar los datos de *microarrays* a las variables clínicas y así conseguir mejores modelos predictores, o por el contrario debieran ser reemplazados los predictores clínicos por los génicos? En la bibliografía, algunas publicaciones han abordado estas cuestiones, si bien las respuestas no están claras y algunas veces han sido contradictorias. A continuación se expone una revisión de los trabajos más importantes publicados sobre este tema.

Eden et al. (2004) publicaron que los marcadores clínicos tenían una potencia similar a los datos de expresión génica de *microarrays* para predecir pronósticos en cáncer de mama. Usaron los datos de 97 pacientes publicados por van 't Veer et al. (2002) y analizaron la predicción de metástasis comparando marcadores clínicos habituales y algunos índices clínicos como NPI y NIH con los marcadores génicos. Concluyeron que, a pesar de la emergencia de los datos de *microarrays*, todavía los marcadores clínicos funcionaban igual o mejor como predictores de pronóstico que los datos génicos.

Por su parte, Gevaert et al. (2006) propusieron una estrategia basada en redes Bayesianas para integrar datos clínicos y de *microarrays*. Después de analizar los datos de cáncer de mama de van 't Veer et al. (2002) para predecir mal o buen pronóstico, propusieron que su método, en la modalidad de "integración parcial", daba mejores resultados que los índices clínicos convencionales por separado.

Un estudio de Sun et al. (2007) abordó asimismo la predicción del pronóstico del cáncer de mama a partir de la combinación de marcadores clínicos y génicos, utilizando también los datos de van 't Veer et al. (2002), pero esta vez utilizando el algoritmo "I-RELIEF" desarrollado por ellos mismos. Publicaron que este algoritmo identifica una serie híbrida que combina las variables clínicas y génicas que tiene una capacidad

predictora para "recurrencia" y "metástasis" apreciablemente mejor que los marcadores clínicos o génicos por separado.

A.L. Boulesteix y colaboradores han publicado tres artículos importantes sobre el valor predictivo adicional que podrían tener los datos de *microarrays* cuando se combinan con datos clínicos.

En un primer trabajo Boulesteix et al. (2008b) proponen un nuevo procedimiento en dos etapas, una primera de reducción de la dimensión por PLS con prevalidación seguida del método de Random Forest.

En el siguiente trabajo Boulesteix and Hothorn (2010) publicaron un nuevo método que combina dos procedimientos estadísticos: regresión logística y regresión "boosting" (con pesos estadísticos). El método funciona en dos pasos. En el primero se ajusta un modelo de regresión logística a las variables clínicas y los coeficientes de esta regresión se pasan al segundo paso como una línea base fija ("offset"). En el segundo paso se acepta la anterior función "offset" de las variables clínicas y se ejecuta el algoritmo de regresión "boosting" con las variables génicas y se obtienen los coeficientes de las variables génicas. Finalmente se combinan ambos modelos y se calculan las predicciones. Con este método analizaron datos simulados y reales y observaron que tenía una buena potencia predictiva en diferentes situaciones.

En el tercer trabajo Boulesteix and Sauerbrei (2011), los autores realizan una revisión crítica de los distintos métodos que se podrían utilizar para evaluar y validar el posible poder predictivo adicional de datos clínicos cuando se combinan con datos de *microarrays*. Analizan dos grandes aspectos: estrategias para obtener modelos predictores combinados y procedimientos de validación del valor predictivo añadido. El primero es el de mayor relación con el presente trabajo y en él se discuten cinco estrategias:

1) "Inocente", que consistiría en construir un sólo modelo dando igual tratamiento a las variables clínicas y génicas, lo que podría ocasionar una infravaloración de las pocas variables clínicas respecto a las más numerosas génicas.

2) "Residual", que sería el otro extremo, donde se obtendría un modelo predictor fijo con las variables clínicas, por ejemplo con regresión logística, y este predictor se le consideraría luego como una línea base ("offset") que se actualizaría con las variables génicas, por ejemplo con regresión "lasso" o "boosting".

3) "Favorecer", una estrategia entre las dos anteriores que ajustaría un modelo predictor con las variables clínicas y génicas a la vez, pero "favoreciendo" los predictores clínicos, por ejemplo en términos de las probabilidades "a priori" en aproximaciones Baxesianas o a través de diferentes "penaltis" en regresión penalizada.

4) "Reducción de la dimensión", que incluye métodos como PCA o PLS. En éstos, las variables génicas son primero resumidas en forma de nuevos componentes en un paso de reducción de la dimensión. Seguidamente se construye un modelo predictor, incluyendo estos nuevos componentes y las variables clínicas como covariables, y utilizando algún método predictor estándar como "Análisis Discriminante" o "Random Forest".

5) "Reemplazamiento", que consiste en reemplazar alguna de las variables clínicas con bajo poder predictivo por alguna variable génica de mayor poder predictor.

El algoritmo que se propone en el presente trabajo se podría considerar una combinación de las estrategias 4 y 5, como se verá más adelante.

En cuanto al aspecto de validación, los autores Boulesteix and Sauerbrei (2011) analizaron muchos casos posibles. El de mayor interés para este trabajo es su aproximación "A", en la que primero se construye el modelo predictor en variables clínicas y luego se ajusta un modelo combinado de variables clínicas y génicas, utilizando en ambos casos una serie de datos de entrenamiento (*training set*) y usando algunas de las estrategias 2), 3) ó 4) arriba mencionadas. Los dos modelos así obtenidos se comparan utilizando una serie externa de validación (*test set*). La comparación dependerá del tipo de resultado que se haya considerado. Para el caso de asignación a una de dos clases posibles, se puede utilizar el llamado error de clasificación ("error rate"), así como otros indicadores tales como especificidad, sensibilidad o diferentes test estadísticos estándar.

Muy recientemente, Karlsson et al. (2012) han propuesto combinar datos de *microarrays* con datos clínicos usando una variante de PLS denominada "Canonical PLS" (Indahl et al. (2009). Esta técnica permite el uso de datos secundarios (los clínicos) como una repuesta adicional a la respuesta primaria (enfermo, sano) a la hora de encontrar los factores latentes. Se asume que los datos secundarios están disponibles para la construcción del modelo con el fin de estabilizarlo para predecir el estado dicotómico enfermo-sano a partir de los datos de *microarrays*, sin que los datos secundarios sean necesarios para predecir nuevas muestras. Los autores utilizaron este método con datos simulados y datos reales de enfermedad de Parkinson y encontraron que los modelos así obtenidos resultaban más simples y más estables que los obtenidos por el método PLS estándar.

Hace unos meses, el grupo de Boulesteix ha publicado la capacidad de predicción de modelos de supervivencia con dos casos prácticos basados en datos de cáncer (De Bin et al. (2014a)), así como los aspectos relacionados con la validación (De Bin et al. (2014b)).

3.4.8. Programa *PLS-VIP* desarrollado en el presente trabajo

El programa *PLS-VIP* ha sido diseñado para adaptar las estrategias de la técnica PLS estándar a las peculiaridades de los datos de *microarrays*, principalmente abordando el problema de la optimización de un modelo PLS de dos formas: a) la selección iterativa de los genes más predictivos mediante el estadístico VIP (de entre los miles del microarray) y b) la búsqueda del valor óptimo para el número de iteraciones y de factores PLS.

Se han utilizado como base las rutinas del Paquete Estadístico *SIMFIT* (http://www.simfit.org) que se encuentran agrupadas en las librerías "dlls" del Paquete. Estas rutinas permiten ser llamadas desde el código fuente de un nuevo programa para hacer todo tipo de procedimientos, tales como ventanas, gráficos, cálculos numéricos complejos, etc. Especial mención para este trabajo merecen las rutinas numéricas relativas a los cálculos de PLS, que han sido llamadas en diferentes puntos por nuestro programa *PLS-VIP*. El código de las rutinas incluidas en *SIMFIT* ha sido desarrollado por W.G. Bardsley de la Universidad de Manchester (UK) y son de tipo "open source" (Bardsley (2013)). Una versión ejecutable del método PLS estándar, está también disponible en *SIMFIT* como parte de su opción "Multivariate Statistic". Esta opción proporciona al usuario todos los resultados habituales como puntuaciones, cargas, gráficos y predicción de nuevos casos.

El código de nuestro programa *PLS-VIP* ha sido escrito en FORTRAN 95 usando el compilador FTN95 de SilverFrost (http: //www.silverfrost.com), que incluye el depurador "Checkmate" que facilita la comprobación del código. Una vez compilado, el programa se puede ejecutar como un módulo interno del propio paquete *SIMFIT*.

El programa *PLS-VIP* presenta dos opciones basadas en la anterior filosofía:

- Una dedicada propiamente a analizar datos de *microarrays*.
- Otra para analizar en secuencia datos clínicos por una parte y datos génicos por otra; procediendo luego a unir ambos tipos de datos en su punto óptimo particular y volviéndolos a analizar para obtener un modelo clínico-genómico. Esta opción trataría de obtener mejores predicciones que con los dos tipos de datos por separado.

Los detalles de estos algoritmos se exponen a continuación.

Opción de análisis de datos de *microarrays*

Como ya se ha apuntado en previos apartados, en el presente trabajo se considera un problema de clasificación binaria donde cada sujeto pertenece a una de dos clases (control (A) y tumor (B)), indicadas por un vector $\mathbf{Y(n\times 1)}$ que usa codificación (0,1), asignando **0** a las muestras control y **1** a las muestras de tumor. La matriz $\mathbf{X(n\times m)}$ consiste de $\boldsymbol{m}$ columnas de expresión de genes medidos en el microarray para las $\boldsymbol{n}$ muestras analizadas. En todos los casos $\boldsymbol{n = n_A + n_B}$ **y** $\boldsymbol{n_A = n_B}$**.**

Cálculos matemáticos PLS en los que se basa esta opción del programa

Como ya se ha apuntado en el epígrafe 3.4.2, existen varios algoritmos para hacer los cálculos habituales de PLS. Nosotros hemos utilizado las rutinas del paquete *SIMFIT*, que sigue el método de la librería NAG (NAG (2012). En el presente trabajo estas etapas han sido las que aparecen en el siguiente recuadro:

1) Sean $\mathbf{X_1}$ y $\mathbf{Y_1}$ las matrices centradas por substracción de las medias de las columnas **X** e **Y** de partida a los valores originales. Las columnas también han sido escaladas a varianza unidad dividiendo los valores centrados por las desviaciones estándar de las columnas.

2) Desde $\boldsymbol{i = 1}$ a $\boldsymbol{k}$, siendo $\boldsymbol{k}$ el número de factores latentes deseados, se deben llevar a cabo los siguientes procedimientos:

a) Computar la dirección de máxima covarianza llevando a cabo una descomposición SVD de $\mathbf{(X_i)^T Y_i}$ y cogiendo el primer vector singular de la izquierda para definir los pesos **x** $\mathbf{(w_i)}$ de las variables predictoras.

b) Calcular el vector de puntuaciones de ***x*** como la combinación lineal $\mathbf{t_i = X_i w_i}$

c) Calcular las cargas de ***x*** por mínimos cuadrados: $\mathbf{(p_i)^T = (t_i)^T X_i}$

d) Calcular las cargas de ***y*** por mínimos cuadrados: $\mathbf{(q_i)^T = (t_i)^T Y_i}$

e) Calcular el vector de puntuaciones de ***y***: $\mathbf{u_i = Y_i q_i}$

f) Calcular las estimaciones de $\mathbf{X_i}$ e $\mathbf{Y_i}$:

$\mathbf{\hat{X}_i = t_i (p_i)^T}$; $\mathbf{\hat{Y} = t_i (q_i)^T}$

g) Deflactar las matrices previas:

$\mathbf{X_{i+1} = X_i - \hat{X}_i}$; $\mathbf{Y_{i+1} = Y_i - \hat{Y}}$

h) Se hace $\boldsymbol{i = i+1}$ y se vuelve al paso a).

3) finalmente, las puntuaciones **t** y **u** son usadas para calcular los parámetros de regresión usando los $\boldsymbol{k}$ factores latentes. Estos parámetros vienen dados por:

$$\beta = \mathbf{W(P^T W)^{-1} Q^T}$$

donde **W** es la matriz $\mathbf{m \times k}$ de los ***pesos-x***, **P** es la matriz $\mathbf{m \times k}$ de ***cargas-x***, y **Q** es la matriz $\boldsymbol{r \times k}$ de ***cargas-y***. Nótese que los valores β calculados de esta forma corresponden a valores internos correspondientes a las matrices centradas y escaladas de las **X** e **Y** originales. Deshaciendo el paso de centrado y escalado, el programa calcula finalmente los parámetros β referidos a los datos originales, que son más intuitivos.

Diagrama de flujo

El programa realiza automáticamente una eliminación de variables (genes) a lo largo de varias iteraciones mediante las puntuaciones del estadístico VIP. En cada iteración se varía en secuencia el número de factores PLS en el modelo y se ajustan las series **X-entrenamiento** y **Y-entrenamiento** y se mide el error de clasificación en cada paso usando unas series independientes **X-prueba** e **Y-prueba.** Las predicciones de los casos de prueba se calculan usando un corte ("cut-off") arbitrario elegido por el usuario (0.5 en el presente trabajo), asignando los valores ***y*** inferiores al "cut-off" a la clase "0" y los valores mayores o iguales al "cut-off" a la clase "1". La lista de las puntuaciones VIP correspondientes al número óptimo de factores en la iteración, se usa para eliminar los genes con poder discriminante bajo de acuerdo con un "cut-off" arbitrario elegido por el usuario (VIP < 1 en el presente trabajo), mientras que los genes remanentes son pasados a la siguiente iteración.

Después de varias iteraciones se elige la iteración y el número de factores óptimo siguiendo el siguiente criterio: el error de clasificación ha de ser el mínimo; si varias iteraciones tienen el mismo error mínimo la iteración seleccionada como óptima será la más alta y el número de factores PLS dentro de la iteración óptima será el más bajo. La iteración más alta con el mínimo error de clasificación se elige para filtrar los genes discriminantes de los genes ruido todo lo posible y el numero de factores PLS más bajo se selecciona por razones de parsimonia.

El diagrama de flujo de esta opción del programa puede verse en la Figura 19.

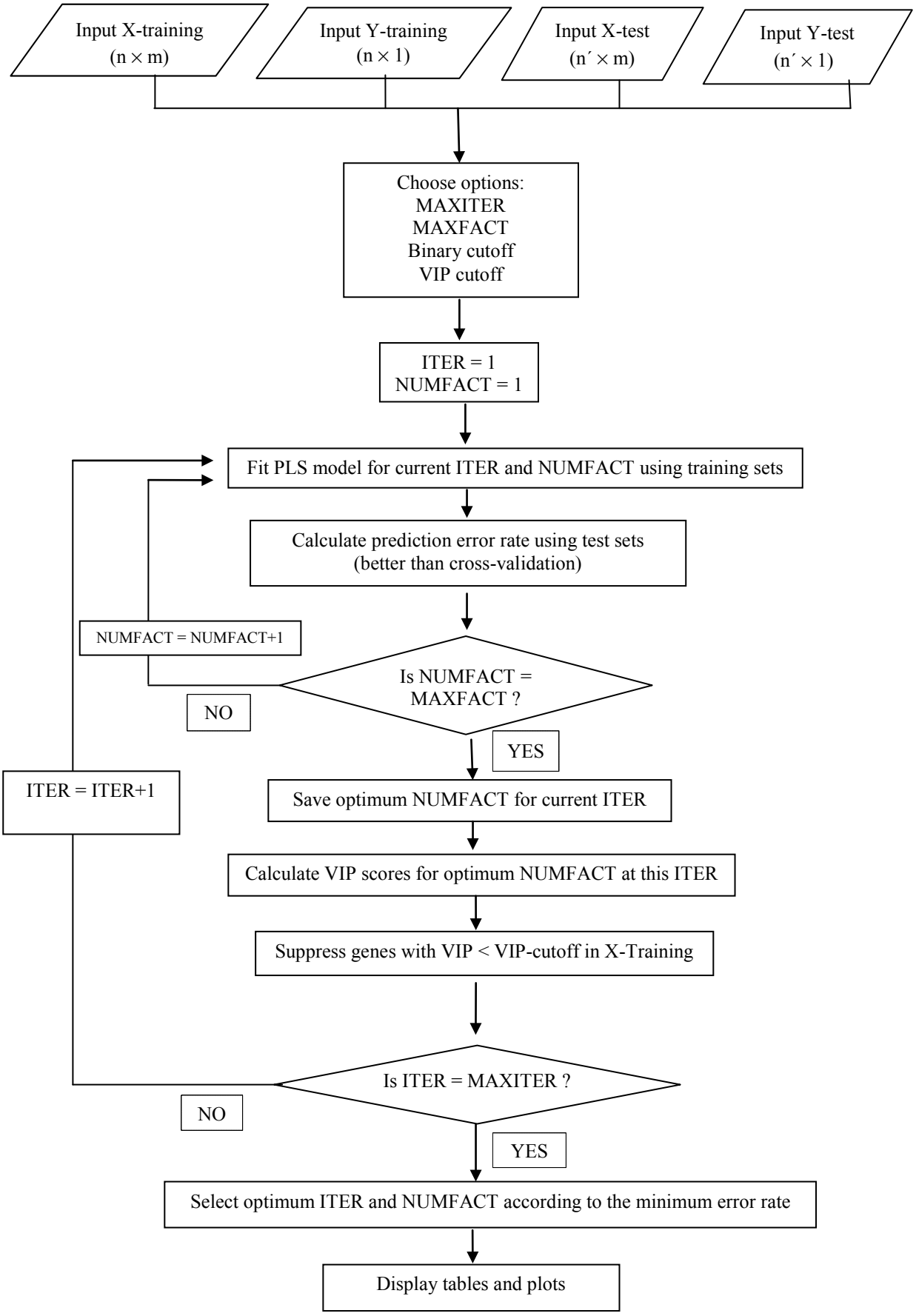

Figura 19. Diagrama de flujo de *PLS-VIP* en su opción de datos de *microarrays*. (Figura previamente publicada en Burguillo et al. (2014))

Como ya se ha mencionado en el apartado 3.4.4, se ha venido considerando hasta ahora que la validación cruzada era el método idóneo para obtener tanto la optimización del modelo PLS como evaluar el error de predicción. En el presente trabajo nosotros hemos usado la aproximación de la serie de prueba ("test-set") siguiendo la recomendación de Esbensen and Geladi (2010). Además, la validación por "test-set" resulta muy adecuada en estudios de simulación como el actual, ya que se pueden generar al azar series hermanas de las de entrenamiento, de manera que estas "test-sets" incluyan todas las causas de incertidumbre de las futuras muestras nuevas.

En la Figura 20 se recogen, en la parte superior la pantalla de selección de opciones y en la parte inferior el menú con las opciones de resultados y gráficos. El comportamiento de estas opciones se expondrá en el apartado 4.1.

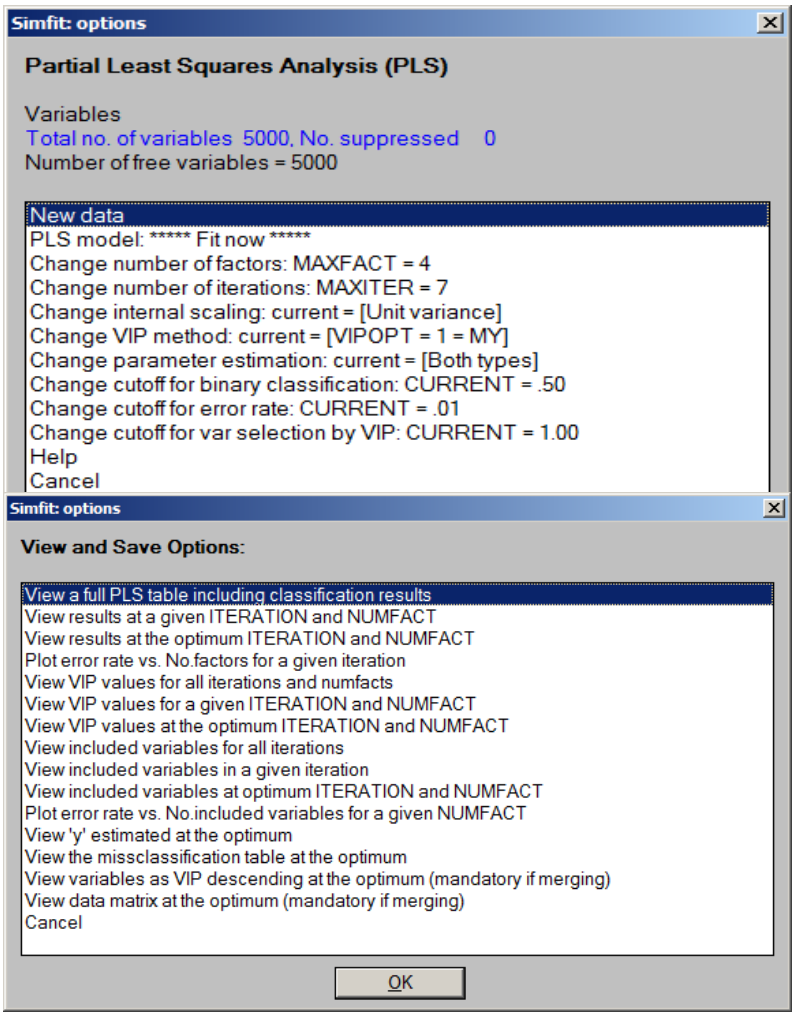

Figura 20. Superior: opciones de PLS. Inferior: Opciones de resultados.

Opción de análisis consecutivo de datos clínicos, génicos y clínicos + génicos

Una pregunta que ha estado siempre presente entre los clínicos y los investigadores era si los predictores clínicos (edad, género, tamaño del tumor, metástasis, linfocitos, Ca, etc.) no serían suficientes para asignar pacientes a grupos o si convendría añadir datos de expresión génica de *microarrays* para disminuir el error de clasificación. O viceversa, tal vez sería conveniente relegar los datos clínicos y centrarse en clasificar las patologías sólo en base a los marcadores génicos.

La pregunta tiene una sencilla respuesta cuando uno de los dos tipos de variables tiene más poder predictivo que el otro, ya que si el error de clasificación es prácticamente nulo con un tipo de las variables y grande con el otro, no tendría sentido el tratar de combinar ambas variables, ya que unas se impondrían sobre las otras en todos los métodos de clasificación y, además, se evitarían molestias al paciente y el coste económico sería menor. Análogamente, si los dos tipos de variables presentan predicciones excelentes, tampoco tiene interés combinarlas, bastaría con medir las más sencillas. El interés aparece cuando los dos tipos de variables presentan un poder predictivo intermedio y semejante, en este caso cabría preguntarse si la unión de ambas variables podría mejorar apreciablemente la predicción de nuevas muestras, situación que por otra parte es la más habitual en la práctica.

Para abordar estas situaciones, se ha desarrollado en el presente trabajo una nueva opción en el programa *PLS-VIP*, con el fin de aplicar la filosofía *PLS-VIP* consecutivamente a datos clínicos y génicos de la siguiente manera: En primer lugar aplicándolo a los datos clínicos por separado, luego a los datos génicos por separado y en tercer lugar fusionando las matrices de ambas variables correspondientes a la iteración y

número de factores óptimos en el óptimo de cada caso por separado; para finalmente aplicar de nuevo el tratamiento *PLS-VIP* a la matriz fusionada. De esta forma se podrá determinar si existe mejora en la predicción (menor error de clasificación) con la matriz fusionada que con las matrices individuales por separado. El algoritmo de este enfoque aparece recogido en la Figura 21.

En esencia, el algoritmo carga las matrices **X-training**, **Y-training**, **X-test** e **Y-test** para las variables clínicas y les aplica los pasos *PLS-VIP* descritos en la Figura 19 anterior. A continuación almacena internamente las matrices correspondientes a la iteración y número de factores óptimos. Seguidamente se cargan las matrices X-training, Y-training, X-test e Y-test de las variables génicas y se les aplica también el algoritmo *PLS-VIP* de dicha Figura 19. Finalmente se fusionan las correspondientes matrices en los óptimos de ambos tipos de variables y se procede a un último tratamiento *PLS-VIP* de la matriz combinada.

A lo largo de toda la secuencia existen diferentes pantallas de resultados (semejantes a las ya mostradas) en las que van apareciendo todos los datos de interés, como el error de clasificación, el número de variables utilizadas, las variables que se han mantenido (junto con su etiqueta) y las que se han eliminado. Con todos estos datos se comparan los diferentes ajustes PLS, especialmente su error de clasificación y tipo de las variables conservadas. Finalmente, se concluye si la combinación de las variables clínicas y génicas ha proporcionado una mejora en la clasificación de casos nuevos (series de prueba) que cada tipo de variables por separado.

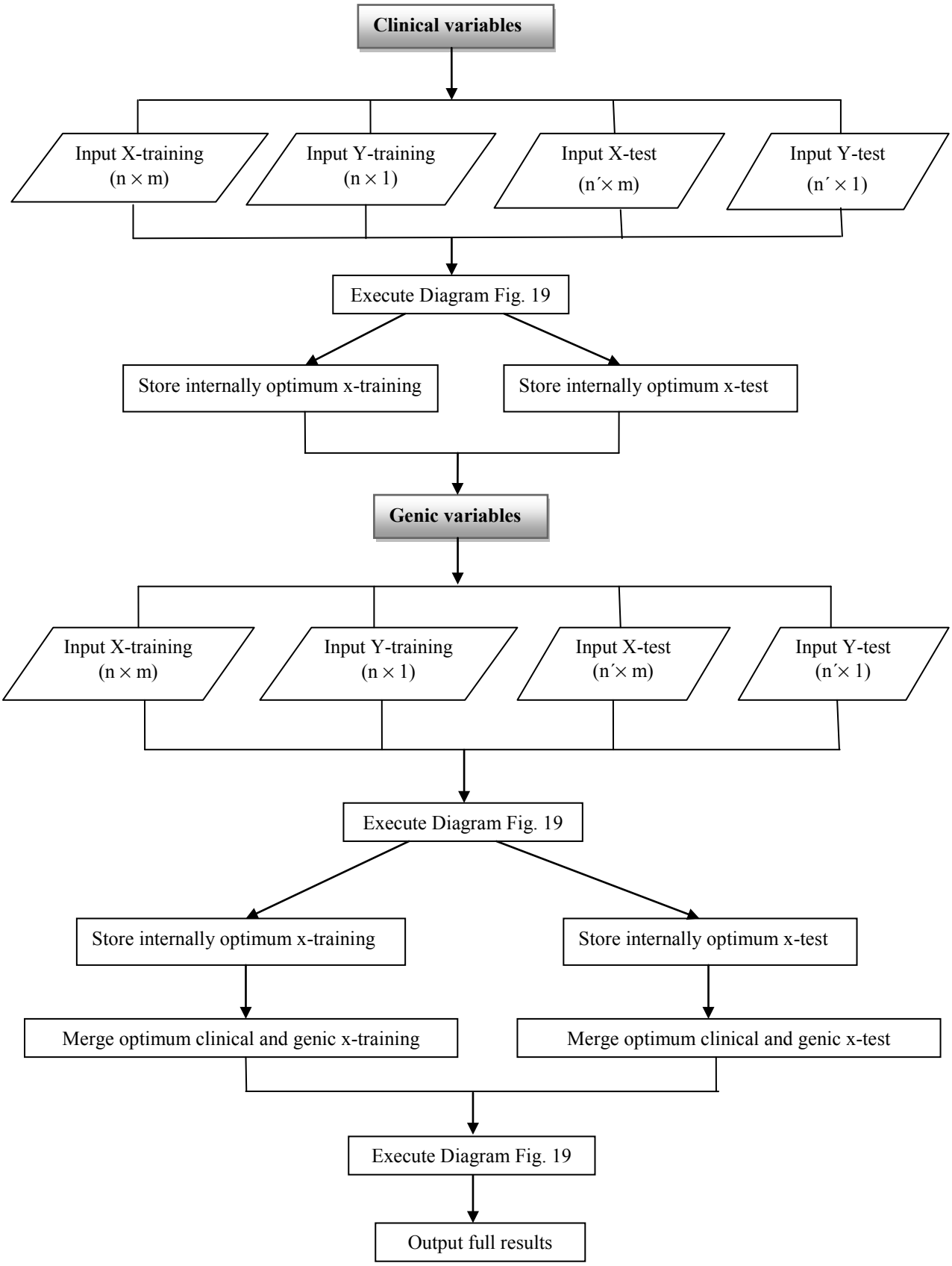

Figura 21. Diagrama de flujo del análisis en secuencia de datos clínicos y génicos y clínicos + génicos.

3.5. Simulaciones con el programa *SIMDATA* desarrollado en este trabajo

Para abordar el presente trabajo hubo necesidad de simular datos que mimetizaran lo más posible las variables clínicas y de expresión génica habituales en controles y pacientes, pero generando dichos valores de una forma aleatoria que incluyera también el error experimental esperado.

Para hacer las diferentes simulaciones se escribió el programa *SIMDATA*, encargado de generar al azar tanto las variables clínicas como las de expresión génica que se necesitarían a lo largo del trabajo.

También en este caso, se ha desarrollado el código fuente utilizando como base las rutinas de las "dlls" del paquete *SIMFIT* y empleando el compilador fortran FTN95 de SilverFrost. Una vez compilado el programa se ejecuta como un módulo interno del propio paquete *SIMFIT*.

Los detalles de las simulaciones se describen a continuación.

3.5.1. Simulación de variables clínicas categóricas y continuas

Se asumen siempre 2 clases: control (A) y tumor (B). El número de sujetos de cada grupo n_A y n_B es elegido por el usuario, pero con la condición de que los grupos estén balanceados ($n_A = n_B$). También se puede elegir entre simular sólo una serie de entrenamiento o generar a la vez una serie de prueba, hermana de la anterior, basada en las mismas semillas pero con una nueva simulación aleatoria. El programa permite como máximo generar 4 variables dicotómicas y 6 variables continuas. Las continuas pueden

ser de variación positiva o negativa en el tumor respecto al control, con el fin de mimetizar más la realidad. La estrategia de simulación varía según que las variables sean dicotómicas o continuas, pero en ambos casos ha tratado de acercarse a los valores de las variables clínicas reales. Estas estrategias se comentan a continuación.

Simulación de variables clínicas dicotómicas

Se empieza por generar un valor de probabilidad binomial al azar para el grupo control (**Prob-A**) a partir de una distribución uniforme entre los límites deseados (por ejemplo U(0.1,0.2)). Esta "**Prob-A**" sirve de semilla para una rutina que genera al azar una serie de valores de 0 y 1 basados en la distribución binomial de esa probabilidad, que constituye los datos de esa variable dicotómica en los sujetos control. Para los datos del grupo tumor se parte de "**Prob-A**" y se la desplaza una cuantía dada por el riesgo relativo (**RR**) del evento para esa variable elegido por el usuario: **"Prob-B" = "Prob-A" × RR**. Esta **"Prob-B"** actúa como una nueva semilla para la rutina de generación de valores 0 y 1 según la correspondiente distribución binomial, proporcionando los datos de los sujetos con tumor.

Los valores de "**Prob-A**" y "**Prob-B**" son almacenados y utilizados después para generar al azar los datos de la serie hermana de prueba. De nuevo se obtiene un vector de valores 0 y 1 desde la rutina de distribución binomial, de forma análoga a la ya expuesta.

A modo de ejemplo se muestra en la Figura 22 la pantalla de entrada de los valores para hacer una simulación de 2 variables dicotómicas con 5 controles y 5 tumores, con una distribución uniforme inicial para obtener "**Prob-A**" de U(0.08,0.12) y un riesgo relativo ("**proportion ratio"**) de 4. Nótese que los valores a la derecha de la pantalla son los valores por defecto y los que aparecen en las cajas son los elegidos por el usuario.

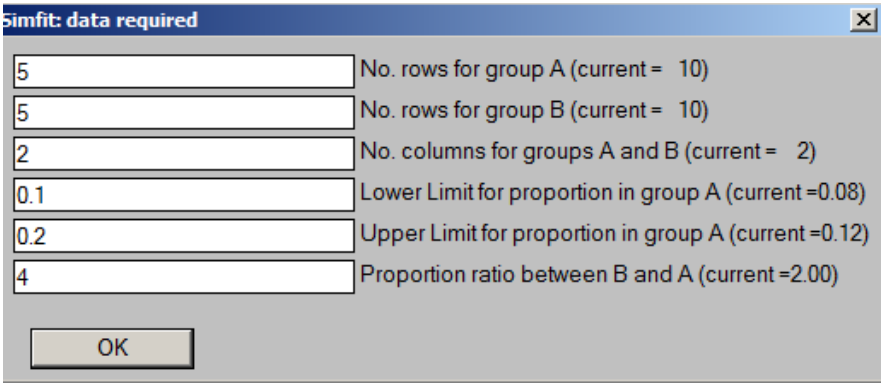

Figura 22. Opciones de entrada de datos en programa para variables dicotómicas.

Con estos valores iniciales se obtienen los valores de las 2 variables para los 10 sujetos en las 2 series de datos (Training y Test), según se muestran en la Figura 23.

Training matrix		Test-set matrix	
0.0000000E+00	0.0000000E+00	1.0000000E+00	0.0000000E+00
0.0000000E+00	0.0000000E+00	0.0000000E+00	0.0000000E+00
1.0000000E+00	0.0000000E+00	0.0000000E+00	0.0000000E+00
0.0000000E+00	1.0000000E+00	1.0000000E+00	1.0000000E+00
1.0000000E+00	0.0000000E+00	0.0000000E+00	0.0000000E+00
0.0000000E+00	1.0000000E+00	1.0000000E+00	1.0000000E+00
1.0000000E+00	1.0000000E+00	0.0000000E+00	1.0000000E+00
0.0000000E+00	0.0000000E+00	0.0000000E+00	1.0000000E+00
0.0000000E+00	1.0000000E+00	0.0000000E+00	1.0000000E+00
1.0000000E+00	0.0000000E+00	0.0000000E+00	1.0000000E+00

Figura 23.Ejemplo de simulación de variables dicotómicas.

Simulación de variables clínicas continuas

Los valores de las variables continuas reales varían aproximadamente entre 1 y 100, y este margen es el que se ha tomado como referencia. El número de sujetos control y tumor debiera ser el mismo que los utilizados para las variables clínicas si se desea fusionar luego ambas matrices de datos.

Para la serie de entrenamiento, lo primero que se hace es generar al azar para los controles (A) un valor semilla para cada variable desde una distribución uniforme (por

ejemplo U(1,80)). Esta semilla se toma como el valor medio de cada variable (**μ(A))** y se introduce en una rutina de generación al azar desde una distribución normal con media el valor de la semilla **μ(A)** y desviación estándar **σ(A),** es decir a partir de una distribución **N(μ(A),σ(A))).** La **σ(A)** se calcula como el producto de un coeficiente de variación (**CV**) elegido por el usuario y la media **μ(A)**, es decir **σ(A) = CV. μ(A)**. Los valores de los tumores se obtienen de la forma que se expone a continuación. La idea es desplazar el valor de cada celda control un cierto desplazamiento ("shift") al azar ("shift_random"). El "shift" puede elegirse positivo o negativo con el fin de contemplar variables con variación positiva o negativa en el tumor. Este "shift" se calcula en base a un tamaño de efecto (effect-size (**ES**)) elegido por el usuario y la desviación estándar de la variable en el control, es decir **shift = ES·σ(A)**. Seguidamente a partir de estos valores los datos del "shift" al azar (**"shift_random"**) se obtienen a partir de una distribución normal cuya media es "**shift"** y cuya desviación estándar se calcula como el producto de una desviación estándar del tamaño del efecto elegida por el usuario (σ_{ES} **)** y la desviación estándar de la variable en el control (**σ(A)),** es decir:

$$\textbf{shift\_random} = N(\textbf{shift}, \sigma_{ES}\sigma(A)).$$

Por último, el valor de la variable en el tumor se calcula como: **B = A + shift_ random.**

En la Figura 24 se muestran las dos pantallas de entrada de datos, con los valores que allí aparecen (recuérdese que los valores que se encuentran a la derecha como "current" son los valores por defecto y los elegidos son los que se teclean en las cajas).

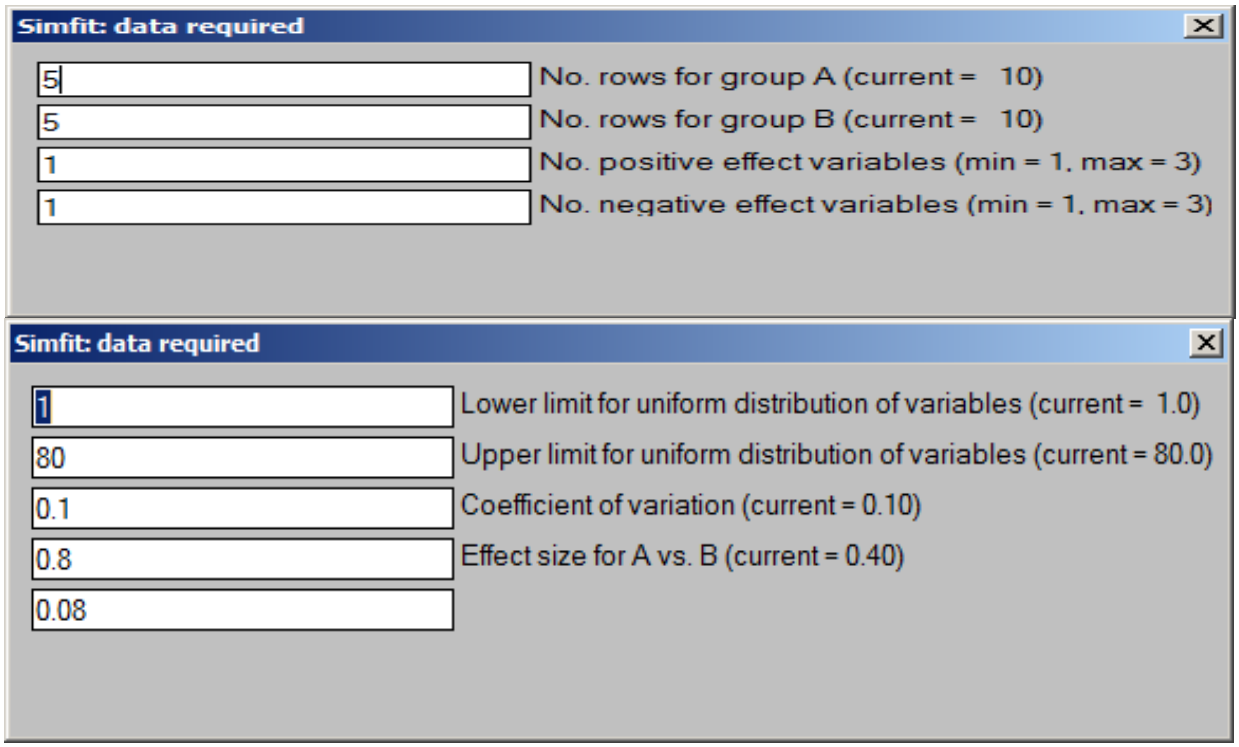

Fig. 24. Opciones de entrada de datos para variables clínicas continuas.

En la Figura 25 se recogen los valores simulados para la serie de entrenamiento y prueba.

Training-set		Test-set	
1.9499705E+01	2.7395864E+01	1.8608716E+01	2.3373093E+01
1.7373146E+01	2.6729273E+01	1.8562366E+01	2.4716795E+01
1.7633244E+01	2.6007631E+01	1.8142361E+01	2.8991918E+01
1.8384669E+01	2.6321353E+01	1.9138647E+01	2.7835410E+01
1.6501250E+01	2.7883857E+01	1.7933427E+01	1.9318440E+01
2.1225102E+01	2.5146964E+01	1.9709366E+01	2.1128289E+01
1.8859044E+01	2.4553412E+01	2.0098377E+01	2.2520645E+01
1.9226435E+01	2.4238167E+01	1.9795803E+01	2.6720035E+01
1.9622143E+01	2.3753286E+01	2.0594789E+01	2.5810493E+01
1.7712295E+01	2.5874795E+01	1.9362713E+01	1.7462954E+01

Fig. 25. Valores simulados de entrenamiento y prueba con variables continuas.

Se simularon diferentes escenarios combinando las distintas posibilidades comentadas más arriba, estos escenarios se recogen en la Tabla 1.

Table 1. Escenarios simulados con sólo variables clínicas			
Simulaciones y variables iniciales			
Escenario	**Casos**	**Dicotómicas iniciales**	**Continuas iniciales**
CLIN1	**10C y 10T**	**2 (RR = 3)**	**3 (TE=0.4)**
CLIN2	**10C y 10T**	**2 (RR = 6)**	**3 (TE=0.8)**
CLIN3	**25C y 25T**	**2 (RR = 3)**	**3 (TE=0.4)**
CLIN4	**25C y 25T**	**2 (RR = 6)**	**3 (TE=0.8)**
CLIN5	**10C y 10T**	**4 (RR = 3)**	**6 (TE=0.4)**
CLIN6	**10C y 10T**	**4 (RR = 6)**	**6 (TE=0.8)**
CLIN7	**25C y 25T**	**4 (RR = 3)**	**6 (TE=0.4)**
CLIN8	**25C y 25T**	**4 (RR = 6)**	**6 (TE=0.8)**
Las series "training" y "test" tienen el mismo número de controles (C) y tumores (T), según se indica. RR = riesgo relativo. TE = tamaño del efecto.			

Simulación de las variables respuestas (y)

Las correspondientes series **Y-training** e **Y-test** eran iguales y ambas consistían de vectores **"dummy"** con n_A ceros (controles) y n_B unos (tumores). Por ejemplo, para el caso de 5 sujetos A y 5 sujetos B, dichos vectores serían igual a: $(0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 1)^T$.

3.5.2. Simulación de variables de expresión génica en *microarrays*

Los estudios de simulación publicados hasta ahora para probar el funcionamiento de PLS, han recurrido a estrategias matemáticas con una extrapolación práctica limitada (Boulesteix (2004), Le Cao et al. (2008)). En su lugar, en el presente trabajo, se ha preferido simular los valores de la expresión génica en log(2), que es lo más habitual, a la vez que se han tratado de mimetizar las situaciones estándar de los *microarrays* reales. Previamente habíamos comprobado que los datos reales en la escala log (2) seguía aproximadamente distribuciones normales con un coeficiente de variación del 4-8 % (resultados no mostrados). Por tanto, se utilizaron distribuciones normales y un 6% de error relativo para simular las expresiones de los genes.

Se asumieron 2 clases, control (A) y tumor (B), con un total de ***n*** sujetos divididos por igual entre ambas clases. El número total de genes, ***m***, se reparte entre ***m**** genes diferencialmente expresados y ***m-m**** genes no informativos (ruido). A su vez, los ***m**** genes informativos se dividen por igual entre genes sobre-expresados e infra-expresados. Por su parte, los ***m-m**** genes no informativos se dividen por igual entre genes incorrelados y correlacionados. Bajo este diseño, la matriz total de expresión génica consiste de 5 bloques: A, B1, B2, N1 and N2, como se apunta en el siguiente esquema:

<table>
<tr><th></th><th colspan="2">m*</th><th colspan="2">m-m*</th></tr>
<tr><td>n/2</td><td colspan="2">A
Control</td><td rowspan="2">N1
Genes ruido
(incorrelados)</td><td rowspan="2">N2
Genes ruido
(correlacionados)</td></tr>
<tr><td>n/2</td><td>B1
Tumor
(sobre-
expresados)</td><td>B2
Tumor
(infra-
expresados)</td></tr>
</table>

Los valores de expresión génica para el bloque A fueron simulados con el siguiente patrón. Se eligió al azar una semilla para cada columna desde una distribución uniforme U(8,10), con el fin de minimizar la variabilidad entre los genes. Luego, los valores actuales para todas las celdas en cada columna fueron generados al azar a partir de una distribución normal $N(\mu,\sigma)$, siendo μ la semilla y siendo σ 0.54 (6 % de 9 que es el punto medio del intervalo 8-10 comentado con anterioridad). Para simular B1y B2, tel valor actual de cada celda A fue desplazada añadiendo un valor delta elegido aleatoriamente de una distribución normal (Δ,σ_Δ). El valor Δ fue añadido como positivo para los genes sobre-expresados (B1) y como negativo para los genes infra-expresados (B2), fijándose σ_Δ en el 6 % del valor simulado de Δ. Este valor Δ , a su vez, se varió convenientemente para similar escenarios con poder discriminatorio entre las clases más alto o más bajo.

Las expresiones para los genes ruido incorrelados en el bloque N1 se generaron en la misma forma que las muestras control en el bloque A. Para generar las expresiones correlacionadas de los genes ruido N2, se adoptó el siguiente procedimiento sencillo basado en la línea recta. La primera columna en N2 se rellenó aleatoriamente en la misma forma que cualquier columna en N1. Luego, los valores de las celdas de esa primera columna (valores guía) se extendieron hacia la derecha llenando las celdas de cada fila a partir de una distribución normal $N(\mu_x,0.54)$, siendo μ_x un valor desplazado a partir de una línea recta (μ_x= pendiente*(valor guía)), con dicha pendiente variando al azar según una distribución normal N(1,0.06) y mantenida constante a lo largo de cada fila para mimetizar una mejor correlación aleatoria.

Siguiendo el patrón anterior, se generaron dos series de datos. Primero, se creó una serie "X-entrenamiento", guardándose sus respectivas semilla para generar con ellas una X hermana llamada "X-prueba", la cual mantendrá todos los procedimientos aleatorios excepto las semillas que serán comunes a la de la serie de entrenamiento, con el fin de tener una serie de validación análoga a la de entrenamiento. En cuanto a las correspondientes series de la variable respuesta (Y-entrenamiento e Y-prueba), ambas fueron iguales y consistieron simplemente en una variable postiza que indica la clase, consistiendo de vectores de n/2 ceros (controles) y n/2 (tumores).

Como ejemplo, se muestra en la Figura 26 las pantallas de entrada de datos.

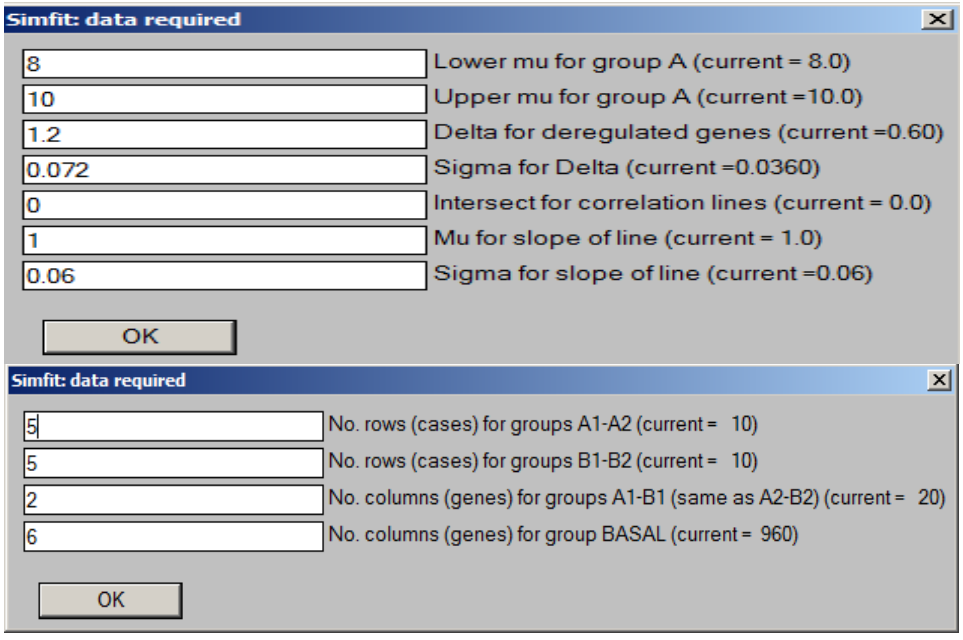

Fig. 26. Pantallas de entrada de datos para las expresiones génicas.

En la Figura 27 aparecen los datos, a modo de ejemplo, para la serie de entrenamiento con los valores de la Figura 26, es decir 10 casos (5 de control y 5 de tumor), 4 genes diferencialmente expresados (2 sobre-expresados y 2 infra-expresados) y 6 genes basales o de ruido (3 sin correlacionar y 3 correlacionados). La serie de prueba sería análoga y no se muestra por brevedad.

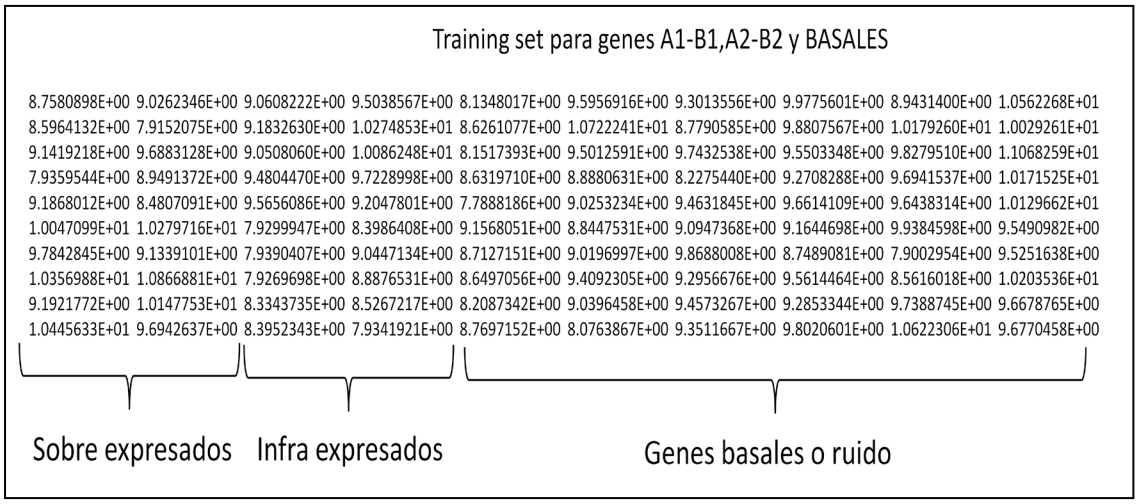

Figura 27. Serie de entrenamiento para las expresiones génicas.

Se analizaron diferentes escenarios de chips de ADN mediante combinaciones de los distintos tipos de genes arriba mencionados, junto con diferentes tamaños de muestras de control y tumor. Estos escenarios se recogen en la Tabla 2.

Tabla 2. Escenarios simulados con datos de microarrays

Condiciones comunes	Genes ruido desde N (μ,σ) (μ from U(8,10) and σ = 0.54)	Genes "sobre" e "infra" expresados desde N(Δ , σ_Δ)*	Escenario
Controles = 5 Tumores = 5 Genes "sobre expresados" = 10 Genes "infra expresados" = 10	980	(0 , 0.0054)	1
		(0.4 , 0.024)	2
		(0.8 , 0.048)	3
		(1.2 , 0.072)	4
	4980	(0 , 0.0054)	5
		(0.4 , 0.024)	6
		(0.8 , 0.048)	7
		(1.2 , 0.072)	8
	9980	(0 , 0.0054)	9
		(0.4 , 0.024)	10
		(0.8 , 0.048)	11
		(1.2 , 0.072)	12
Controles = 25 Tumores = 25 Genes "sobre expresados" = 40 Genes "infra expresados" = 40	920	(0 , 0.0054)	13
		(0.4 , 0.024)	14
		(0.8 , 0.048)	15
		(1.2 , 0.072)	16
	4920	(0 , 0.0054)	17
		(0.4 , 0.024)	18
		(0.8 , 0.048)	19
		(1.2 , 0.072)	20
	9920	(0 , 0.0054)	21
		(0.4 , 0.024)	22
		(0.8 , 0.048)	23
		(1.2 , 0.072)	24

* Los desplazamientos al azar se generan desde distribuciones normales con media Δ y desviación estándar σ_Δ .

Para los estudios en los que se combinaban variables clínicas y génicas, los escenarios de expresión génica se variaron ligeramente respecto a los de la tabla 2 anterior, eligiéndose otros un poco diferentes que son los que aparecen en la tabla 3.

Table 3. Escenarios simulados con sólo variables génicas				
Simulaciones y genes iniciales				
Escenario	**Casos**	**Genes dif. exp. Iniciales**	**Delta (σ_{delta})**	**Genes ruido iniciales**
GENIC1	**10C y 10T**	**20**	**0.4 (0.024)**	**980**
GENIC2	**10C y 10T**	**20**	**0.6 (0.036)**	**980**
GENIC3	**25C y 25T**	**20**	**0.4 (0.024)**	**980**
GENIC4	**25C y 25T**	**20**	**0.6 (0.036)**	**980**
GENIC5	**10C y 10T**	**40**	**0.4 (0.024)**	**960**
GENIC6	**10C y 10T**	**40**	**0.6 (0.036)**	**960**
GENIC7	**25C y 25T**	**40**	**0.4 (0.024)**	**960**
GENIC8	**25C y 25T**	**40**	**0.6 (0.036)**	**960**
Las series "training" y "test" tienen el mismo número de controles (C) y tumores (T), según se indica. Delta es el incremento del gen en su expresión diferencial.				

3.6. Programas estadísticos y de clasificación utilizados

El test t de Student y el método "Linear Discriminant Analysis (LDA)" se realizaron con el paquete estadístico *SIMFIT*. Los procedimientos "K-nearest neighbours (KNN)", "Random Forest (RF)" y "Support Vector Machines (SVM)" se llevaron a cabo con la "suite" Babelomics en su version 4.3.0 (http://babelomics.bioinfo.cipf.es), y el método de "Prediction Analysis for Microarrays (PAM)" se aplicó mediante una macro de Excel de Stanford University (http://www-stat.stanford.edu/~tibs/PAM).

Las multicomparaciones realizadas con el test t de Student fueron tenidas en cuenta usando el criterio de "False Discovery Rate (FDR)" (Benjamini and Hochberg, 1995) ya expuesto anteriormente.

El procedimiento LDA se llevó a cabo usando la distancia de Mahalanobis y aplicando las opciones *SIMFIT* de "estimative" para las asignaciones Bayesianas, "equal" para las matrices de covarianzas, y "equal" para las probabilidades a priori ("prior").

El número K elegido en el procedimiento KNN fue aquel que daba la mejor clasificación con la serie de entrenamiento.

El "threshold" (corte) usado para construir el modelo PAM fue aquel que proporcionaba con la serie de entrenamiento la mejor clasificación y el número más bajo de genes seleccionados.

El número de árboles y genes usados en la predicción con RF se eligieron en el punto de mejor clasificación con la serie de entrenamiento.

Finalmente el corte denominado "cost" en SVM fue el que proporcionaba mejor porcentaje de clasificación también con la serie de entrenamiento.

3. RESULTADOS Y DISCUSION

4.1. Resultados con datos simulados de *microarrays*

En este apartado, el objetivo es analizar la bondad del procedimiento *PLS-VIP* bajo muy diferentes escenarios que mimeticen, todo lo que sea posible, la gran variedad de aspectos experimentales que se pueden dar en el análisis de datos de *microarrays*. Todo ello encaminado a obtener las adecuadas conclusiones acerca del uso de la metodología PLS con dichos datos de *microarrays*.

4.1.1. Escenarios simulados para probar el programa *PLS-VIP*

Para analizar la potencia del programa propuesto, se simularon diferentes situaciones de *microarrays*, para ello se variaron el tamaño de muestra, el número de genes sobre-expresados e infra-expresados, el valor de desplazamiento empleado ("delta (Δ)") y el número de genes basales o de "ruido".

Como se ha descrito en el apartado 3.5 de "Metodología", estas simulaciones intentan abarcar los valores observados en experimentos con muestras reales (ver la Tabla 2 en el apartado 3.5). Se simularon dos condiciones: una con tamaño de muestra de 5 controles y 5 tumores y a su vez con 10 genes sobre-expresados y 10 infra-expresados; la otra con 25 controles y 25 tumores, 40 genes sobre-expresados y 40 infra-expresados. Para cada una de estas dos condiciones, se simularon a su vez diferente número de genes basales o de ruido, siendo éstos 980, 4980, 9980 y 920, 4920 y 9920, respectivamente. Esto forma un total de 6 bloques. Así mismo, dentro de cada uno de estos bloques se simularon 4 escenarios de distinta potencia: potencia cero, baja, media y alta (valores Δ de 0, 0.4, 0.8 y 1.2). En total se simularon, pues, 24 escenarios (ver Tabla 2 en el apartado 3.5).

Para cada escenario, se generaron aleatoriamente 50 series **entrenamiento-X** y sus correspondientes 50 series "hermanas" de **prueba-X**. Las respectivas **entrenamiento-Y** y **prueba-Y** fueron unos vectores columna idénticos que simplemente incluían valores **0** y **1** para las muestras control y tumor, respectivamente, de acuerdo con el tamaño de muestra considerado en cada escenario.

4.1.2. Funcionamiento del algoritmo *PLS-VIP*

Aunque se analizarán el resto de escenarios más adelante, se muestra a continuación el comportamiento del algoritmo *PLS-VIP* bajo dos escenarios particulares, con el fin de exponer su comportamiento. Uno se ha elegido con genes que tienen un bajo poder discriminante y el otro con genes de una potencia discriminatoria media.

Escenario con genes de bajo poder discriminante

Para este escenario, las condiciones de la serie **entrenamiento-X** fueron las siguientes: 25 muestras control y 25 muestras tumor; los genes discriminantes fueron 80, de los que 40 estaban sobre-expresados y 40 infra-expresados, desplazados de los genes basales (ruido) mediante unos valores Δ elegidos de una distribución normal (Δ,σ_Δ), con $\Delta = 0.4$ y $\sigma_\Delta = 0.024$ (6% de error al azar) y adicionando Δ como positivo o negativo según se requiriese. Los genes ruido fueron 4920; estando la mitad de ellos no correlacionados y la otra mitad correlacionados (ver el apartado 3.5.2 de "Metodología"). A la vez, fue simulada bajo condiciones similares una serie **prueba-X** tipo "hermana" de la anterior. Los correspondientes vectores **entrenamiento-Y** y **prueba-Y** fueron idénticos y estaban formados por una columna con 25 ceros (muestras control) y 25 unos (muestras tumor).

El algoritmo comienza en la primera iteración con los 5000 genes de partida, y usa las series de entrenamiento para estimar los factores PLS en secuencia (1 a 4 factores). Simultáneamente se calcula el error de clasificación usando las series de prueba en cada factor. Luego, en el número óptimo de factores (cuando el error de clasificación es el más bajo), los genes con VIP < 1 son eliminados, mientras que los genes remanentes se pasan a la segunda iteración. Los correspondientes resultados para 8 iteraciones y factores de 1 a 4 aparecen recogidos en la Tabla 4.

Tabla 4. Funcionamiento del propuesto algoritmo *PLS-VIP* basado en una eliminación iterativa de variables hacia atrás mediante puntuaciones VIP. Los datos X son una matriz simulada de expresión génica en log(2) de genes con bajo poder de discriminación. El algoritmo encuentra un modelo óptimo con pocos genes y un mínimo error de clasificación. (Tabla previamente publicada en Burguillo et al. (2014)).

Iteración	Genes Usados en predicción	Número de factores	Varianza acumulada (X)	Varianza acumulada (Y)	Error de clasificación
1	5000	1	25.0	23.0	0.50
		2	29.5	99.4	0.28
		3	31.0	100	0.30
		4	32.4	100	0.30
2	2201	1	37.3	26.5	0.48
		2	46.8	98.9	0.16
		3	47.9	100	0.16
		4	49.1	100	0.16
3	739	1	6.2	94.9	0.18
		2	42.3	98.4	0.06
		3	43.8	99.9	0.06
		4	45.2	100	0.06
4	298	1	12.4	91.6	0.18
		2	19.6	98.6	0.04
		3	21.9	99.8	0.04
		4	24.2	100	0.04
5	**132**	1	22.5	72.2	0.24
		2	**31.3**	**96.9**	**0.00**
		3	34.2	99.0	0.02
		4	36.5	99.6	0.02
6	57	1	36.0	52.4	0.42
		2	47.2	89.6	0.18
		3	50.2	94.9	0.22
		4	52.6	97.1	0.18
7	19	1	21.6	84.0	0.20
		2	30.2	88.8	0.22
		3	36.6	91.0	0.14
		4	41.4	92.3	0.18
8	9	1	26.6	78.3	0.22
		2	35.9	82.2	0.24
		3	46.2	82.7	0.22
		4	55.7	82.8	0.20

Como se observa en la Tabla 4, la varianza capturada aumenta apreciablemente al pasar de 1 a 2 factores PLS, tanto para las variables **X** como la variable **Y**, estabilizándose a continuación o creciendo ligeramente. También puede apreciarse que la varianza capturada es baja para las **X**, debido al elevado número de variables **X** (genes), mientras que es alta para la variable **Y** que es única (un vector de ceros y unos). Por otra parte, puede verse en dicha tabla cómo el error de clasificación disminuye generalmente al pasar de 1 factor a 2 factores PLS y luego se estabiliza o aumenta ligeramente. Este comportamiento se ha representado en la Figura 28(A) para la iteración 1 y 5.

Asimismo, puede observarse en la Tabla 4, cómo el error de clasificación disminuye gradualmente con las iteraciones, hasta alcanzar un mínimo (error 0) en la iteración 5, donde se usan 2 factores y 132 genes, siendo la varianza capturada acumulada para las **X** del 31% y para las **Y** del 97%. A partir de este mínimo, el error de clasificación va aumentando hasta la última iteración. Estos resultados se pueden visualizar en la Figura 28(B). La interpretación de este comportamiento sería que los genes ruido son suprimidos hasta un cierto punto, donde se alcanza el óptimo del funcionamiento del algoritmo, pero a partir del cual comienza también la eliminación de algún gen informativo y la predicción desde una serie independiente de prueba empeora, debido a una mayor influencia de los genes ruido. Un comportamiento similar ha sido también descrito en Cho et al. (2002). Algunas mejoras en el presente trabajo en comparación con otras investigaciones previas (Cho et al. (2002), Perez-Enciso and Tenenhaus (2003)) serían las siguientes: no es necesaria una preselección de genes (por ejemplo con *t-test*), el número de factores PLS no es un valor fijo para todas las iteraciones sino que es optimizado dentro de cada iteración, y las puntuaciones del

estadístico VIP, usadas para seleccionar los genes que pasarán a la siguiente iteración, son las correspondientes al número de factores óptimo dentro de cada iteración.

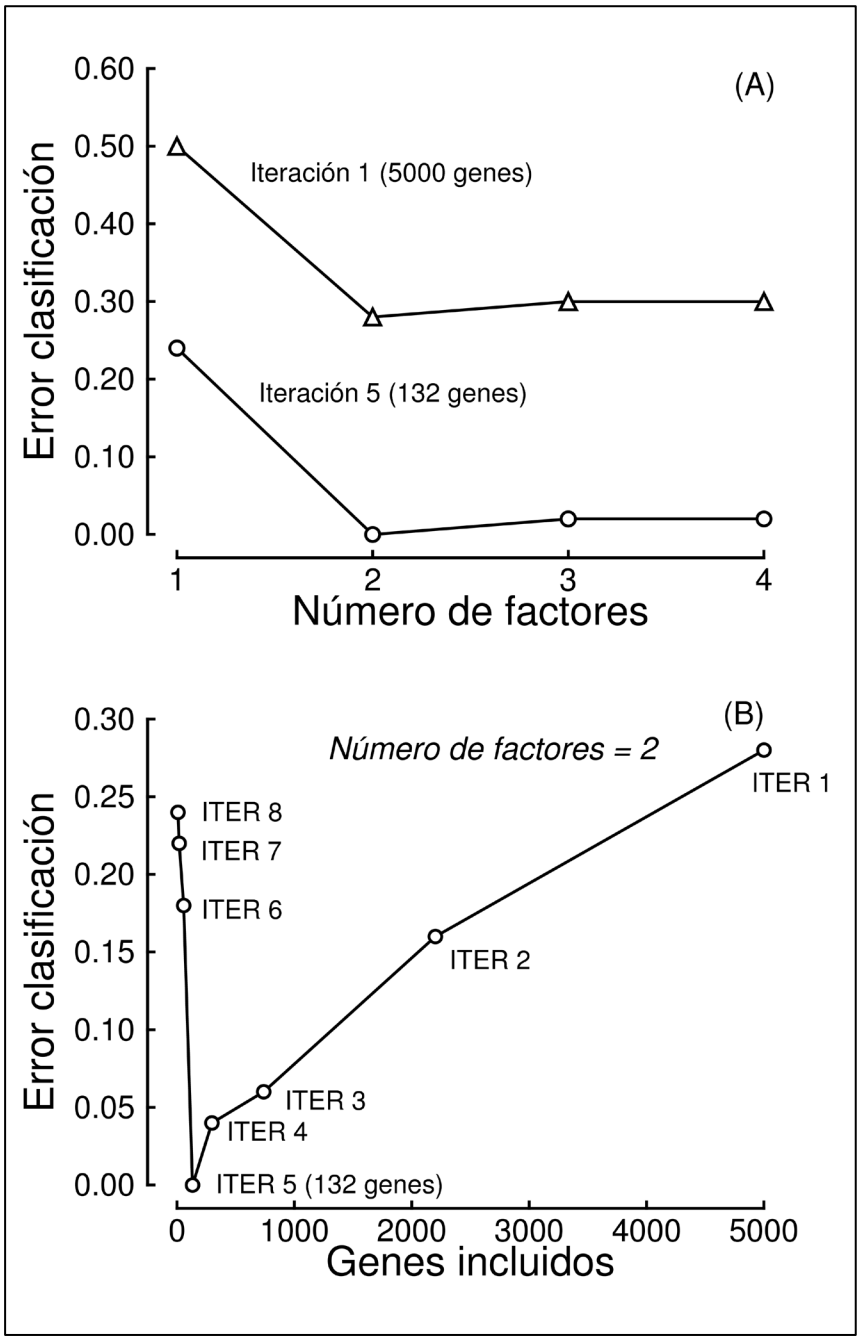

Figura 28. Funcionamiento del propuesto algoritmo *PLS-VIP*. (A) Variación del nº de factores en dos iteraciones. (B) Variación del nº de genes con las iteraciones. (Figura previamente publicada en Burguillo et al. (2014))

Una vez que el número óptimo de iteraciones y de factores ha sido encontrado, se puede analizar posteriormente la correspondiente matriz **entrenamiento-X** (que tiene 132 genes en este caso), usando las opciones PLS que existen en el paquete estadístico *SIMFIT* dentro de su apartado "Multivariate statistics". Así, en la Figura 29(A) se muestran las ***puntuaciones-X***, pudiéndose apreciar que, usando los dos primeros factores PLS, las 25 muestras control y las 25 tumorales se separan correctamente. Sin embargo, una de las muestras tumor cae alejada de su grupo, lo que se podría interpretar como una muestra atípica debida al azar.

En la Fig. 29(B) se han representado las ***cargas-X*** para los 132 genes. Debido a que tales genes provienen de simulaciones, se puede deducir que esta lista incluye exactamente 51 genes diferencialmente expresados, mientras que 81 son genes ruido incluidos por azar. La presencia de estos genes ruido no es de extrañar en este escenario concreto, ya que los 80 genes diferencialmente expresados fueron simulados con una potencia discriminatoria baja y los 4920 restantes fueron generados de forma aleatoria, permitiendo por tanto la aparición de algunos genes falsos positivos. A pesar de esto, el algoritmo fue capaz de acometer cierta selección de genes de entre los 5000 genes de partida y de alcanzar una buena proporción de muestras bien clasificadas. Este funcionamiento puede ser muy útil cuando se enfrentan clases muy similares, como ocurre con pacientes que responden o no responden a un fármaco, donde las diferencias en la expresión de los genes suelen ser muy pequeñas.

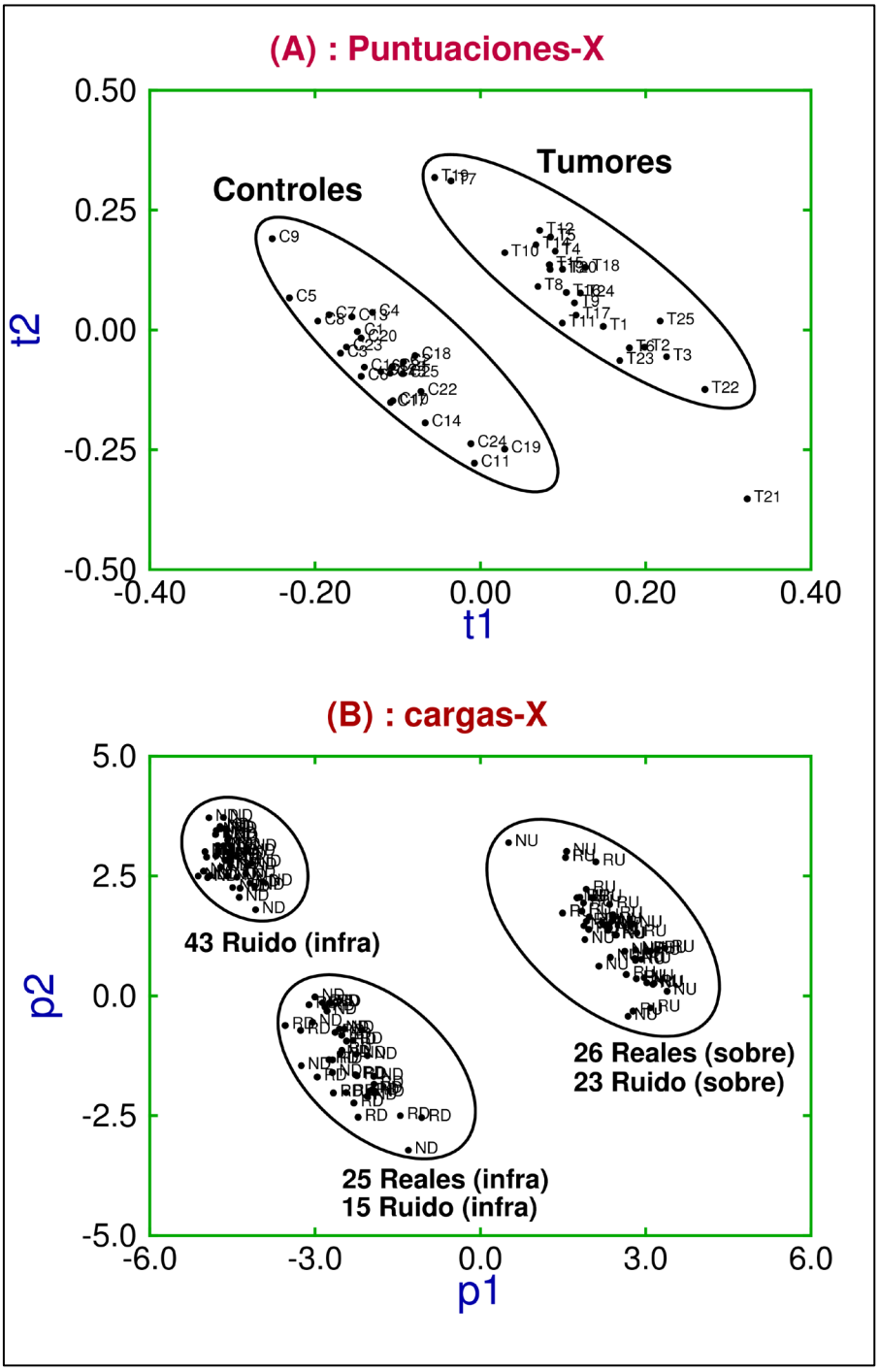

Figura 29. Puntuaciones y cargas de las variables *X* para la iteración 5 y 2 factores PLS con la serie de entrenamiento y un escenario potencia baja. Reales = genes verdaderamente discriminantes. Ruido = genes basales. Sobre = sobre-expresados. Infra = infra-expresados. RU = real sobre-expresado (up), RD = real infra-expresado (down). NU = noise sobre-expresado y ND = Noise infra-expresado. (Figura previamente publicada en Burguillo et al. (2014))

Finalmente se han representado en la Figura 30 las puntuaciones de la variable **Y**, observándose cómo los grupos de control y tumor se separan claramente. No es posible para la variable **Y** representar las cargas, ya que se trata de una variable única y no es posible el calcular cargas.

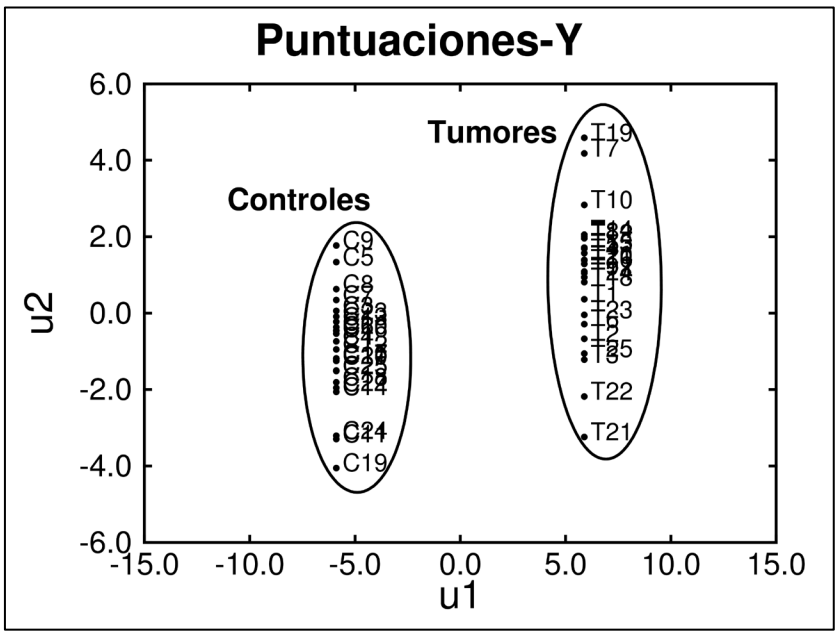

Figura 30. Puntuaciones de las variables *Y* para la iteración 5 y 2 factores PLS con la serie de entrenamiento y un escenario potencia baja. C# =Control(nº paciente). T# = Tumor(nº paciente).

Quedaría una última representación de interés en PLS, la de las puntuaciones ***u*** frente a las puntuaciones ***t***, pero este tipo de gráfica se analizará en el apartado siguiente con un escenario de mayor potencia discriminante.

Escenario con genes de potencia discriminante media

En este caso, las condiciones fueron las mismas que en el escenario anterior, pero ahora los genes discriminantes fueron simulados con $\Delta = 0.8$ y $\sigma_\Delta = 0.048$. En esta situación, el algoritmo encuentra el óptimo en la iteración 6 con 2 factores PLS, siendo la varianza acumulada capturada de 58% para las **X** y de 95% para las **Y**. En este punto, el error de clasificación vale cero y el número de genes usados para la predicción es de 15,

siendo todos ellos verdaderos genes diferencialmente expresados sin la presencia de ningún falso positivo. En la Figura 31(A) se han representado las ***puntuaciones-X*** de la serie de entrenamiento, observándose dos "clusters" bien separados, uno para las muestras control y el otro para las tumorales.

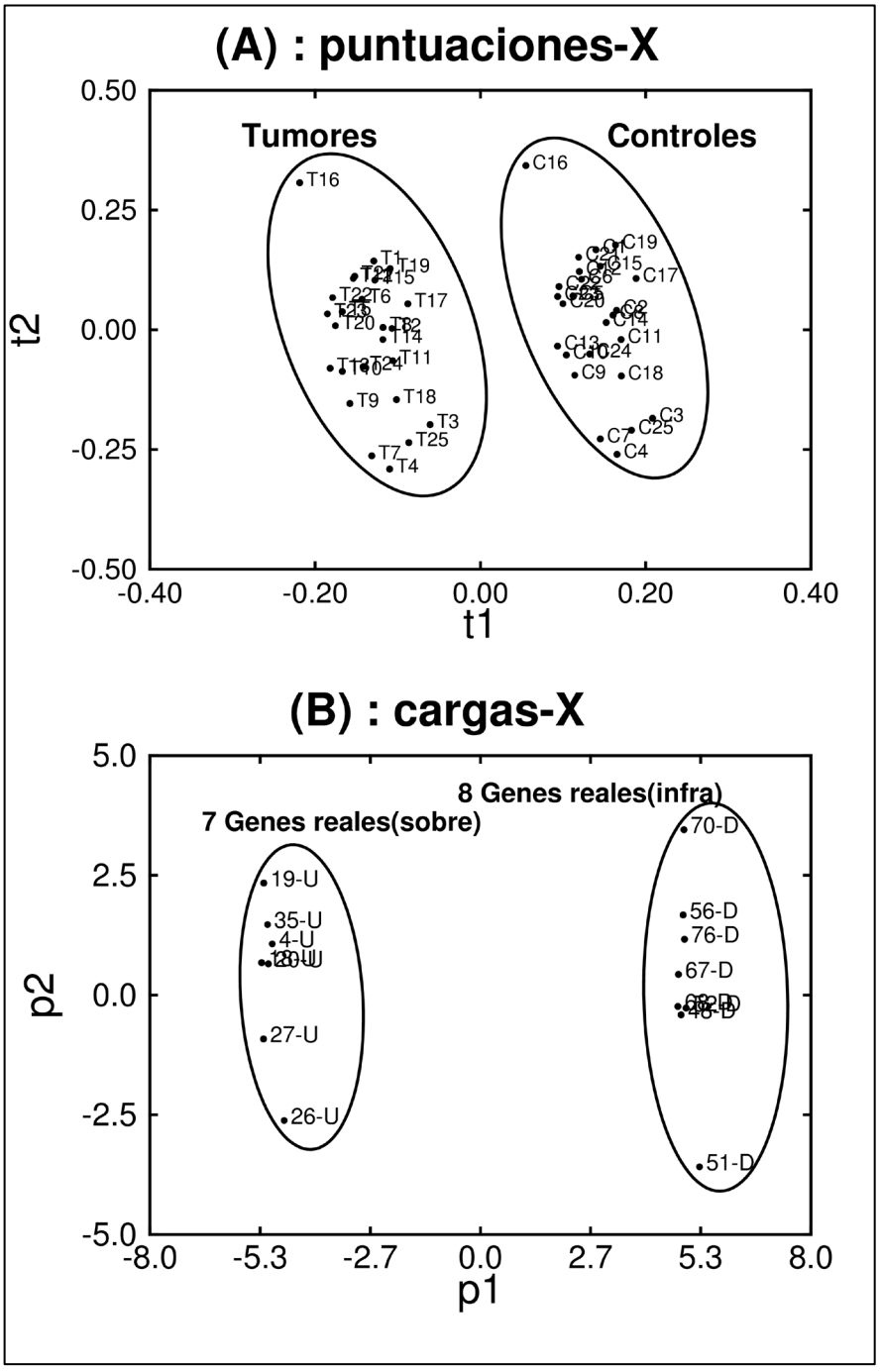

Figura 31. Puntuaciones y cargas de las variables *X* para la iteración 6 y 2 factores PLS con la serie de entrenamiento y el escenario potencia media. Reales = genes diferencialmente expresados: U(up) = sobre-expresados,D(down) = infraexpresados. (Figura previamente publicada en Burguillo et al. (2014))

Por su parte, en la Figura 31(B) se muestran las ***cargas-X,*** donde se puede apreciar 2 "clusters" agrupando los genes en 7 genes discriminantes sobre-expresados y 8 genes discriminantes infra-expresados. En las dos figuras anteriores; al tratarse de un número más pequeño de datos, se han podido incluir las etiquetas para facilitar la identificación de pacientes y de genes.

Las ***puntuaciones-Y*** son análogas a las de la Figura 30 y se han omitido por brevedad. Por último, como una medida de la bondad del ajuste de la serie de entrenamiento en su punto óptimo, se han representado en las Figuras 32(A) y 32(B) las correlaciones sucesivas entre las puntuaciones de **X** e **Y**, denominadas $\boldsymbol{t}$ y $\boldsymbol{u}$, para los 2 factores del óptimo. Cada gráfica muestra el ajuste de regresión lineal de los valores de $\boldsymbol{u_i}$ sobre $\boldsymbol{t_i}$, así como el ajuste de regresión lineal de $\boldsymbol{t_i}$ sobre $\boldsymbol{u_i}$, junto con los coeficientes de correlación $\boldsymbol{r}$ y los niveles de significancia $\boldsymbol{p}$. Claramente las puntuaciones del primer factor ($\boldsymbol{u_1}$ y $\boldsymbol{t_1}$) se encuentran bien correlacionadas, como indica la buena superposición de ambas líneas de regresión y los valores de $\boldsymbol{r}$ y $\boldsymbol{p}$. La correlación entre las puntuaciones del segundo factor ($\boldsymbol{u_2}$ and $\boldsymbol{t_2}$) todavía es significativa pero resulta más débil.

En resumen, se puede concluir, que las diferentes representaciones mostradas más arriba suponen un valioso valor añadido de PLS frente a otros métodos de clasificación, ya que permite diferentes representaciones de puntuaciones y cargas que pueden sugerir aspectos nuevos tanto de los casos como de las variables del estudio en cuestión, como ya ha sido apuntado por otros autores (Boulesteix (2004), Perez-Enciso and Tenenhaus (2003)).

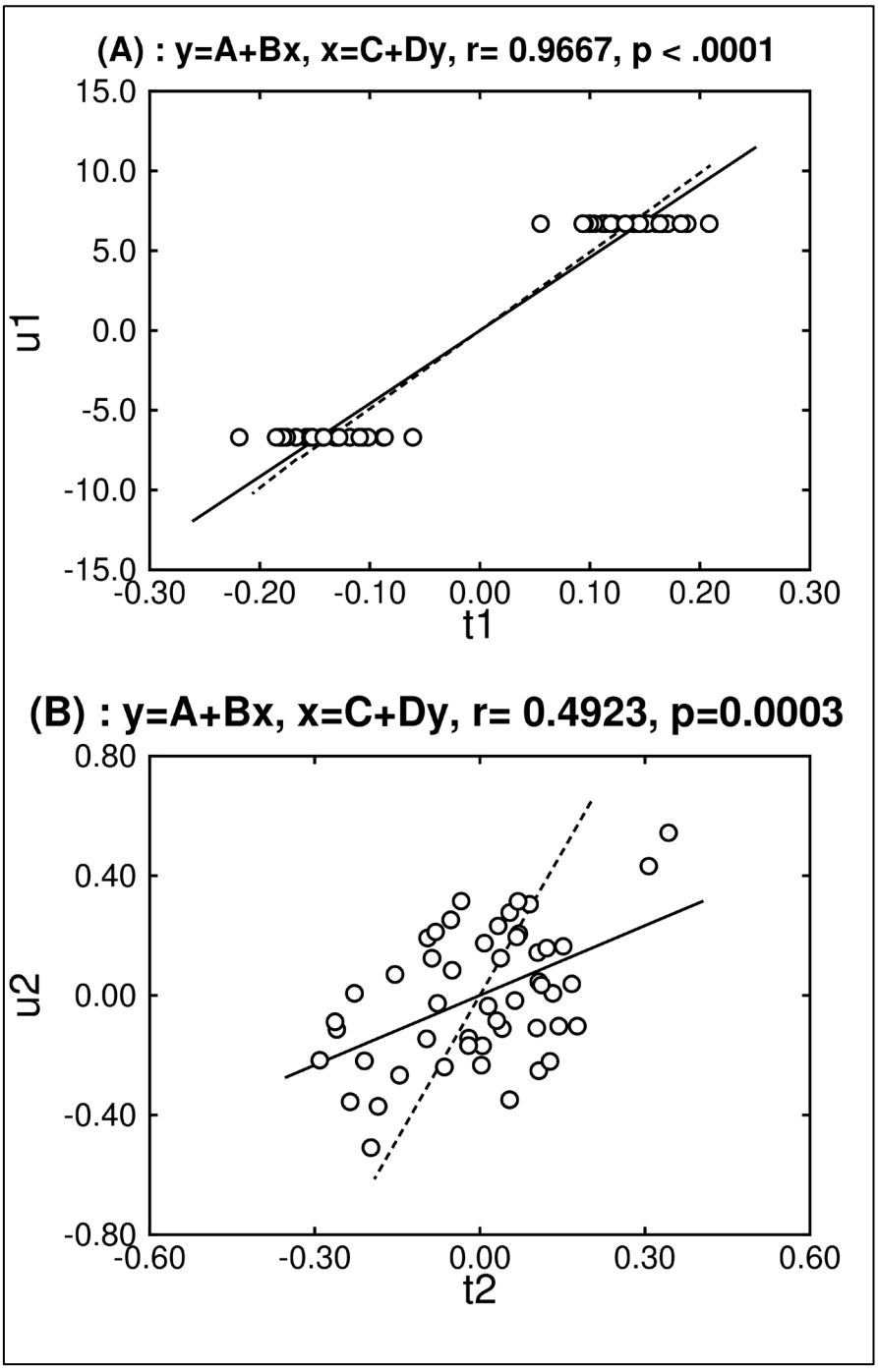

Figura 32. Representaciones de puntuaciones u_i frente a t_i . (A) para el factor 1 y (B) para el factor 2. Linea continua: regresión de *y* sobre *x*. Línea discontinua = regresión de *x* sobre *y*. (Figura previamente publicada en Burguillo et al. (2014))

4.1.3. Análisis de *PLS-VIP* bajo diferentes escenarios

Para analizar la potencia del algoritmo propuesto, se simularon diferentes situaciones de *microarrays*, para ello se variaron el tamaño de muestra, el número de genes sobre-expresados e infra-expresados, el valor de desplazamiento empleado en los genes (Δ) y el número de genes basales o de ruido.

Como se ha descrito en el apartado 3.5.2 de "Metodología", estas simulaciones intentan mimetizar los valores observados en experimentos con muestras reales. Se simularon 6 bloques con 4 escenarios cada uno de distinta potencia, incluyendo para los 4 escenarios de cada bloque potencia cero, baja, media y alta. En total se simularon 24 escenarios. Para cada escenario se generaron aleatoriamente 50 series **entrenamiento-X** y sus correspondientes 50 series "hermanas" de **prueba-X**. Las respectivas **entrenamiento-Y** y **prueba-Y** fueron unos vectores columna iguales que simplemente incluían valores **0** y **1** para las muestras control y tumorales, respectivamente, de acuerdo con las muestras consideradas en cada escenario. Todas estas series fueron analizadas con el algoritmo *PLS-VIP* y los resultados aparecen recogidos en la Tabla 5, donde los valores promedio de cada 50 series se valoran con la mediana, a la que se acompaña de la información correspondiente al primer cuartil y tercer cuartil con objeto de valor su dispersión, y poder así minimizar la influencia de algunos valores atípicos que aparecen normalmente.

Los escenarios 1, 5, 9, 13, 17 y 21 son de control, sin genes diferencialmente expresados ($\Delta = 0$), con el fin de tener una referencia (no efecto) en cada bloque de escenarios.

La primera observación en la Tabla 5, es que el algoritmo *PLS-VIP* progresa en todos los escenarios a través de varias iteraciones hasta alcanzar una iteración óptima con

una mínima "proporción de error de clasificación (PEC)". También se puede observar que 1 o 2 factores PLS fueron suficientes, en el óptimo, para capturar un porcentaje alto de variabilidad de la variable **Y**, como cabe esperar para un vector de respuesta **Y** de tipo binario.

Con respecto a los tres primeros bloques (escenarios 1 a 12), el tamaño de muestra fue pequeño (5 controles y 5 tumores). Los genes discriminantes fueron pocos (10 sobre-expresados y 10 infra-expresados) y su potencia discriminante en base al desplazamiento Δ varió de 0 a 1.2 en la escala de log(2), asumiendo que la expresión génica está centrada entre valores de 8 y 10. Los genes ruido variaron de 920 a 9980 y tuvieron expresiones génicas aleatorias en todos los escenarios a partir de una distribución normal $N(\mu,0.54)$, procediendo μ de una distribución uniforme $U(8,10)$.

Dejando aparte los escenarios de referencia (1, 5 y 9), se observa que los errores de clasificación disminuyen gradualmente a medida que aumenta el valor de Δ, y que este efecto resulta más aparente cuando el número de genes ruido es más pequeño (es decir escenarios 2 a 4 mejores que escenarios 6 a 8 y 10 a 12). También se puede ver en cada escenario que, en la iteración óptima, el número de genes usados en la predicción es pequeño (alrededor de 15), pero que estos genes seleccionados están algo contaminados por genes ruido, lo que conlleva un "proporción de genes falsos positivos (PGFP)" relativamente alta. Esta contaminación disminuye gradualmente de nuevo al aumentar Δ dentro de cada grupo, y este efecto se hace más evidente, de nuevo, cuando el número de genes ruido disminuye (escenarios 2 a 4 mejores que escenarios 6 a 8 y 10 a 12). Así, cuando el tamaño de muestra y el número de genes discriminantes es escaso, la estrategia *PLS-VIP* ayuda a mejorar el error de clasificación y disminuye el número de genes seleccionados para la predicción, aunque se debe tener en cuenta que, en estos escenarios, una proporción relevante de genes ruido es incluida por azar en la lista final de los genes.

En los 3 bloques siguientes (escenarios 13 a 24), el tamaño de muestra se aumentó a 50 muestras (25 de control y 25 tumorales), los genes fueron ahora 80 (40 sobre-expresados y 40 infra-expresados) y el poder discriminante fue el mismo que anteriormente (Δ variando de 0 a 1.2). Igual que antes, se simularon también genes ruido, variando de 920 a 9920.

Dejando a un lado los escenarios de referencia (13, 17 y 21), los resultaros fueron ahora mucho mejores que en los escenarios anteriores, como era de esperar. Así el valor de PEC fue ya 0 incluso para el valor más bajo de Δ de 0.4 y 920 genes ruido (escenario 14), pero mostrando todavía un valor de PGFP de 0.23. Asimismo, y como también cabía esperar, los valores de PEC y PGFP fueron ambos de 0 cuando el poder de los genes discriminantes (Δ) subió a valores de 0.8 y 1.2, independientemente del nº de genes ruido. En resumen, se puede concluir, que el algoritmo *PLS-VIP* reduce con éxito el número de genes predictores mediante la eliminación de los genes ruido, lo cual mejora el funcionamiento de la clasificación.

De las simulaciones anteriores se sugieren algunas recomendaciones prácticas:

a) Se recomienda realizar una simple exploración de datos de la matriz de datos del *microarray* en unidades de log(2) con el fin de calcular las diferencias entre las medias para las expresiones de los genes de los dos grupos de muestras.

b) Si unas pocas diferencias de medias son ≥ 0.8, pero el tamaño de muestra es escaso, el método *PLS-VIP* podría servir para fines de clasificación, pero la lista final de genes seleccionados debiera ser tomada con precaución ya que habrá genes falsos positivos en la lista.

c) Si varias diferencias de medias son ≥ 0.8 y el tamaño de muestra es grande, el algoritmo *PLS-VIP* construirá un buen clasificador y la lista final de genes tendrá, probablemente, sólo genes verdaderos positivos.

Tabla 5. Probando el algoritmo PLS-VIP bajo diferentes escenarios simulados con 50 repeticiones cada uno

Condiciones comunes	Genes ruido desde N (μ,σ) (μ from U(8,10) and σ = 0.54)	Genes "sobre" e "infra" expresados desde N(Δ , σ_Δ)*	Escenario	Iteración óptima (Q1 , Q3)**	Nº factores optimos (Q1 , Q3)	Genes totales en el óptimo (Q1 , Q3)	Proporción genes falsos positivos (Q1 , Q3)	Proporción errores de clasificación (Q1 , Q3)
Controles = 5 Tumores = 5 Genes "sobre-expresados" = 10 Genes "infra-expresados" = 10	980	(0 , 0.0054)	1	**6** (4 , 7)	**2** (1 , 2)	**11** (7 , 51)	**1** (1 , 1)	**0.30** (0.20 , 0.40)
		(0.4 , 0.024)	2	**5** (3 , 6)	**2** (1 , 2)	**25** (9 , 125)	**0.96** (0.90 , 0.97)	**0.30** (0.20 , 0.40)
		(0.8 , 0.048)	3	**6** (5 , 6)	**1** (1 , 2)	**14** (8 , 24)	**0.59** (0.47 , 0.72)	**0** (0 , 0.10)
		(1.2 , 0.072)	4	**7** (6 , 7)	**1** (1 , 1)	**7** (4 , 9)	**0.13** (0 , 0.23)	**0** (0 , 0)
	4980	(0 , 0.0054)	5	**7** (6 , 8)	**1** (1 , 2)	**15** (7 , 63)	**1** (1 , 1)	**0.30** (0.20 , 0.40)
		(0.4 , 0.024)	6	**8** (5 , 8)	**2** (1 , 2)	**13** (7 , 94)	**1** (0.99 , 1)	**0.30** (0.20 , 0.40)
		(0.8 , 0.048)	7	**7** (6 , 8)	**1** (1 , 2)	**17** (7 , 41)	**0.83** (0.63 , 0.91)	**0.10** (0.05 , 0.20)
		(1.2 , 0.072)	8	**8** (8 , 8)	**1** (1 , 2)	**8** (5 , 11)	**0.45** (0.24 , 0.59)	**0** (0 , 0)
	9980	(0 , 0.0054)	9	**8** (6 , 9)	**2** (1 , 3)	**14** (7 , 63)	**1** (1 , 1)	**0.30** (0.20 , 0.40)
		(0.4 , 0.024)	10	**9** (8 , 9)	**2** (1 , 3)	**10** (7 , 25)	**1** (1 , 1)	**0.30** (0.20 , 0.40)
		(0.8 , 0.048)	11	**8** (7 , 9)	**2** (1 , 2)	**13** (7 , 35)	**0.89** (0.76 , 0.98)	**0.20** (0.10 , 0.30)
		(1.2 , 0.072)	12	**9** (4 , 10)	**1** (1 , 2)	**7** (5 , 13)	**0.48** (0.33 , 0.67)	**0** (0 , 0.10)
Controles = 25 Tumores = 25 Genes "sobre-expresados" = 40 Genes "infra-expresados" = 40	920	(0 , 0.0054)	13	**4** (3 , 6)	**2** (1 , 3)	**38** (9 , 163)	**1** (1 , 1)	**0.39** (0.36 , 0.44)
		(0.4 , 0.024)	14	**4** (3 , 4)	**1** (1 , 2)	**83** (49 , 106)	**0.23** (0.13 , 0.28)	**0** (0 , 0)
		(0.8 , 0.048)	15	**5** (5 , 5)	**1** (1 , 1)	**15** (13 , 17)	**0** (0 , 0)	**0** (0 , 0)
		(1.2 , 0.072)	16	**6** (6 , 6)	**1** (1 , 1)	**7** (5 , 8)	**0** (0 , 0)	**0** (0 , 0)
	4920	(0 , 0.0054)	17	**5** (4 , 7)	**2** (1 , 3)	**73** (13 , 216)	**1** (1 , 1)	**0.38** (0.36 , 0.42)
		(0.4 , 0.024)	18	**4** (4 , 5)	**2** (1 , 3)	**179** (120 , 283)	**0.61** (0.53 , 0.73)	**0.02** (0 , 0.04)
		(0.8 , 0.048)	19	**7** (6 , 7)	**1** (1 , 1)	**15** (13 , 17)	**0** (0 , 0)	**0** (0 , 0)
		(1.2 , 0.072)	20	**7** (7 , 8)	**1** (1 , 1)	**7** (6 , 9)	**0** (0 , 0)	**0** (0 , 0)
	9920	(0 , 0.0054)	21	**7** (5 , 8)	**2** (1 , 3)	**28** (14 , 209)	**1** (1 , 1)	**0.40** (0.36 , 0.42)
		(0.4 , 0.024)	22	**5** (5 , 5)	**2** (1 , 3)	**224** (133 , 259)	**0.71** (0.61 , 0.78)	**0.04** (0.02 , 0.07)
		(0.8 , 0.048)	23	**7** (7 , 8)	**1** (1 , 1)	**13** (11 , 24)	**0** (0 , 0)	**0** (0 , 0)
		(1.2 , 0.072)	24	**8** (8 , 8)	**1** (1 , 1)	**7** (5 , 10)	**0** (0 , 0)	**0** (0 , 0)

* Los desplazamientos al azar se generan desde distribuciones normales con media Δ y desviación estándar σ_Δ .

** Los valores se refieren a la mediana (primer cuartil - tercer cuartil) a partir de 50 repeticiones de datos simulados para cada escenario.

(Tabla publicada previamente en Burguillo et al. (2014))

4.1.4. Comparación de *PLS-VIP* con otros métodos de predicción

Para comparar la bondad del algoritmo propuesto en este trabajo frente a otros métodos de clasificación habituales en Genómica, se procesaron las 50 series aleatorias del escenario 18 con distintos procedimientos de clasificación. Se seleccionó este escenario ya que, debido a su bajo poder discriminatorio, podría permitir la detección de algunas diferencias entre los métodos probados. Las condiciones fueron 25 muestras control y 25 tumorales, 80 genes diferencialmente expresados que se simularon mediante un desplazamiento aleatorio de los controles de tipo (N (Δ,σ_Δ), con $\Delta = 0.4$ y $\sigma_\Delta = 0.024$), otros 4920 genes fueron genes ruido generados aleatoriamente. Bajo estas condiciones, se simularon 50 series de entrenamiento que se analizaban con sus 50 series "hermanas" de prueba, con el fin de promediar el funcionamiento de los distintos métodos ensayados. Se analizaron dos estrategias, una sin realizar una selección preliminar de genes sino manejando un alto número de variables directamente, la otra incluyendo un primer paso de selección de genes con algún método apropiado. Los resultados se muestran en la Tabla 6, en la que los valores de tendencia central y dispersión se expresan como: mediana (primer cuartil, tercer cuartil), para minimizar los valores atípicos.

En la primera estrategia (parte superior de la Tabla 6), los métodos funcionaban sin ninguna selección de genes (es decir con los 5000 genes originales). Se puede observar cómo los métodos KNN y PAM presentan un valor de "proporción de error de clasificación (PEC)" bastante pobre, con valores alrededor de 0.4-0.5. Este hecho parece lógico debido al ruido de 4920 genes frente a 80 genes con poder discriminatorio bajo. Como contraste, SVM funcionó extremadamente bien bajo estas condiciones desfavorables, con un valor de PEC de 0.04 (0.02,0.07). Por su parte, PLS se probó en su

formulación original, es decir sin ninguna selección de genes (PLS-Estándar), mostrando un mal valor de PEC de 0.19 (0.12,0.29), lo que significa que los genes ruido también impiden una buena clasificación por PLS. El valor de "proporción de genes falsos positivos (PGFP)" para estos métodos fue muy alto (0.98), como cabía esperar, teniendo en cuenta que los genes verdaderamente discriminantes fueron 80 y los genes ruido 4920. Sin embargo, el funcionamiento de *PLS-VIP* sobre los 5000 genes de partida fue excelente, con un error de clasificación de 0.02 (0,0.04), mejorando a todos los métodos anteriores, incluido SVM (0.04 (0.02,0.07)).

En una segunda opción (parte intermedia de la Tabla 6), se realizó una selección previa de genes para cada serie de ***entrenamiento-X*** aplicando el test t de Student a los 5000 genes iniciales. Para cada serie de entrenamiento, se seleccionaron los 20 genes que presentaban el valor más bajo de FDR (genes "top"), algunos de ellos eran verdaderos genes diferencialmente expresados entre clases (control, tumor) y el resto eran falsos positivos que aparecían aleatoriamente como "ruido", encontrándose aproximadamente el 50% de cada tipo como promedio (PGFP $\approx$ 0.50). Cada una de las 50 series de ***entrenamiento-X*** y ***prueba-X*** de partida, fueron recortadas para incluir solamente los 20 genes "top" seleccionados por el test-t. Finalmente, las series así recortadas se utilizaron para ensayar el funcionamiento de tres métodos de clasificación: LDA, KNN y PAM. Como puede observarse en la Tabla 6, el comportamiento de LDA, KNN y PAM es muy similar, siendo sus valores de PEC de 0.24 (0.19,0.32), 0.22 (0.16,0.29) y 0.26 (0.22,0.44), respectivamente, por lo que puede considerarse que el funcionamiento de tales métodos en estas condiciones es deficiente, debido principalmente a la presencia de genes ruido introducidos por azar.

En la tercera aproximación (parte inferior de la Tabla 6), se acometía una selección previa de genes para cada serie de ***entrenamiento-X***. En todos los casos se observó que las listas de los genes seleccionados estaban formadas por verdaderos genes diferencialmente expresados y por genes falsos positivos (ruido) que aparecen aleatoriamente durante el proceso de simulación. Con los métodos KNN, SVM y RF, la selección previa de genes se hizo con el método "correlation feature selection o CFS" (Hall (1999)) a partir de los 5000 genes iniciales, seleccionándose 47 (43,50) genes con un valor de PGFP de 0.68 (0.64,0.71). Con el método PAM, se usó una validación cruzada con 3-reparticiones conduciendo a una selección de 316 (249,470) genes, presentando un valor de PGFP de 0.79 (0.71,0.89). En el método *PLS-VIP*, su propio procedimiento iterativo hizo una selección de genes a partir de las puntuaciones VIP, seleccionando 179 (120,283) genes con un valor PGFP de 0.61 (0.53,0.73). Con respecto al error de clasificación, se puede apreciar que KNN y PAM se comportan modestamente, con valores de PEC de 0.22 (0.18,0.28) and 0.26 (0.10,0.48), respectivamente. El método RF presento una clasificación moderada, con un valor PEC de 0.14 (0.06,0.19). Como contraste, el funcionamiento de SVM fue excelente con una predicción en promedio totalmente correcta (PEC = 0 (0,0)). Análogamente, el algoritmo *PLS-VIP* funcionó muy bien consiguiendo un valor de PEC de 0 (0, 0.04). De todo lo anterior, se puede concluir que el algoritmo propuesto de *PLS-VIP* tiene un funcionamiento comparable al método de SVM (sin selección y con selección previa de genes por CFS). La ventaja de la aproximación *PLS-VIP* frente a SVM reside en su simplicidad, ya que usa solamente un procedimiento matemático sobre la serie original completa, no necesita una selección previa de genes, y su ejecución en el ordenador es muy rápida.

Tabla 6. Comparación de métodos habituales de clasificación frente a los algoritmos PLS-Estándar y PLS-VIP usando 50 repeticiones de datos simulados (escenario 18) . Los modelos óptimos en todos los métodos, excepto en los algoritmos PLS, se construyeron con la serie de entrenamiento pero los errores de clasificación se calcularon con la serie de prueba. En los métodos PLS, tanto la dimensionalidad óptima del modelo como el error de predicción se calcularon en cada paso con la serie de entrenamiento seguida de la serie de pureba

Método	Genes de partida (Nº dif. exp.)	Selección previa de genes	Genes totales seleccionados Mediana (Q1 , Q3)*	Proporción genes falsos positivos (PGFP) Mediana (Q1 , Q3)*	Proporción errores de clasificación (PEC) Mediana (Q1 ,Q3)*
KNN	5000 (80)	Ninguna	**5000**	**0.98**	**0.40** (0.36 , 0.46)
PAM					**0.48** (0.41 , 0.52)
SVM					**0.04** (0.02 , 0.07)
PLS-Estándar					**0.19** (0.12 , 0.29)
PLS-VIP			**179** (120 , 283)	**0.61** (0.53 , 0.73)	**0.02** (0 , 0.04)
LDA	5000 (80)	test-t	**20 top****	**0.50** (7 , 11)*	**0.24** (0.19 , 0.32)
KNN				**0.50** (0.43 , 0.67)	**0.22** (0.16 , 0.29)
PAM				**0.49** (0.40 , 0.67)	**0.26** (0.22 , 0.44)
KNN	5000 (80)	CFS, LOO	**47** (43 , 50)	**0.68** (0.64 , 0.71)	**0.22** (0.18 , 0.28)
PAM		CV (3 fold)	**316** (249 , 470)	**0.79** (0.71 , 0.89)	**0.26** (0.10 , 0.48)
SVM		CFS, LOO	**47** (43 , 50)	**0.68** (0.64 , 0.71)	**0** (0 , 0)
RF		CFS, LOO	**47** (43 , 50)	**0.68** (0.64 , 0.71)	**0.14** (0.06 , 0.19)

* mediana (primer cuartil , tercer cuartil) a partir de 50 simulaciones. Los datos son los del escenario 18, donde las series de entrenamiento y de prueba tienen 25 muestras control y 25 muestras tumor, 5000 genes en total, con 80 genes discriminantes de baja potencia (Δ=0.4) y 4920 genes ruido . ** top = genes más significativos (test-t). CFS = "Correlation feature selection. LOO = "Leave One Out"". CV (3 fold) = "Cross validation (3 particiones)". VIP = "Variable Influence on Projection".

(Tabla publicada previamente en Burguillo et al. (2014))

4.2. Resultados de *PLS-VIP* con datos reales de *microarrays*

4.2.1 *"Macroglobulemia de Waldenström frente a Leucemia Linfocítica Crónica"* y *"Mieloma Múltiple sin ganancia 1q frente a Mieloma Múltiple con ganancia 1q"*-

Las muestras de los pacientes hematológicos se eligieron de estudios ya publicados. Como primer ejemplo se compararon 10 muestras de linfocitos B (BL) de pacientes con Macroglobulemia de Waldenström frente a 11 muestras de BL de pacientes con Leucemia Linfocítica Crónica (CLL). Todas las muestras habían sido hibridadas al "Microarray Human Genome U133 A" de Affymetrix, como se describe en la publicación original (Gutierrez et al. (2007)). Los archivos primarios sin procesar (.CEL) fueron descargados de la base de datos GEO ("accession number: GSE6691"). Estos datos fueron primero normalizados usando el algoritmo RMA incluido en la "Affymetrix Expression Console". Los datos se filtraron a continuación eliminando las sondas control y aquellas sondas que no tenían anotación "gene-symbol". Por último, las sondas restantes se filtraron eliminando aquellas que en todas las muestras presentaban una expresión el $\log(2) < 6.6$, por considerar que dicha expresión era despreciable. Este procedimiento dejó 5797 sondas.

En cuanto a las muestras, éstas se repartieron en dos series, una de entrenamiento con 8 WM y 9 CLL muestras, y otra de prueba o validación con 2 WM y 2 CLL muestras. Esta estrategia permite construir modelos de predicción menos sesgados que si sólo se utilizara una serie de entrenamiento y validación cruzada.

El segundo ejemplo consistió en una comparación entre muestras de células plasmáticas (PC) de 20 mielomas múltiples (MM) con ganancia cromosómica 1q (1q) frente a 36 muestras MM PC sin esta aberración (non-1q). Las muestras fueron hibridadas

al chip "Human Gene 1.0 st" de Affymetrix según se describe en la publicación original (Gutierrez et al. (2010). Los archivos de datos se obtuvieron de la base de datos GEO con la identificación GSE16558. El preprocesado y el filtrado se realizaron como se ha indicado más arriba, pero en este caso las "probesets" con un valor de log(2) < 6.7 en todas las muestras fueron eliminadas, por considerarlas no diferencialmente expresadas. Asimismo, las "probesets" que presentaban un "gene-symbol" duplicado o que han sido consideradas en la bibliografía como genes concomitantes (Broyl et al., 2010) también fueron eliminadas. Tras estos filtrados, se obtuvo un total de 9789 "probesets" que son las que se utilizaron para el análisis. Los modelos de predicción se construyeron con una serie de entrenamiento formada por 15 "1q" y 27 "non-1q" muestras, mientras que la serie de prueba consistió de 5 "1q" y 9 "non-1q" muestras.

En ambos ejemplos, las muestras que formarían parte de la serie de entrenamiento y de la serie de prueba fueron extraídas mediante permutaciones al azar de las correspondientes series completas.

Los datos finales de los *microarrays* fueron analizados mediante *PLS-Estándar* y *PLS-VIP* y sus resultados fueron comparados con otros métodos de clasificación. En el primer ejemplo se compararon BL de WM frente a BL de CLL usando las series de entrenamiento y de prueba previamente descritas. Los resultados se muestran en la mitad superior de la Tabla 7, donde se puede observar cómo todos los métodos probados dieron el mismo error de clasificación de cero. La explicación para esta excelente predicción por cualquier método podría ser la gran diferencia que existe entre las dos patologías, lo que trae consigo el que muchos genes sean altamente discriminantes y eso hace que la clasificación sea fácil con cualquier método. Por otra parte, convendría destacar que en

ciertos casos el uso de *PLS-VIP* podría resultar útil, pues, como puede verse en este ejemplo, reduce el número de sondas discriminantes de 5797 a 11.

El segundo ejemplo es una comparación entre muestras de PC de MM con ganancia 1q (1q) frente a muestras sin ganancia 1q (non-1q). Ahora, la diferencia entre muestras es mucho más pequeña y cabe esperar algunas discrepancias entre los métodos de clasificación. Los resultados se muestran en la mitad inferior de la Tabla 7, donde se puede ver que *PLS-VIP* presenta un buen comportamiento, teniendo uno de los errores de clasificación más bajos (0.07) y reduciendo las "probesets" de 9789 a 10. Cabe destacar que en este ejemplo real, donde las diferencias de expresión génica eran menores entre las dos clases, *PLS-VIP* se ha comportado como el mejor método, por lo que podría ser un método recomendable para comparaciones con clases no muy diferentes, como ocurre en la respuesta o no respuesta a los fármacos. Pero habrá que hacer más estudios que validen esta propuesta.

Resulta extraño que en este ejemplo el método "t-test-SVM" se haya portado tan pobremente (PEC = 0.56), ya que suele dar buenos resultados. Tal vez la razón podría ser que la selección previa de genes por t-test no es la mejor opción para SVM, que suele utilizar una selección por CFS, pero este aspecto no se pudo comprobar, ya que el algoritmo no pudo manejar el alto número de genes (9789) usando CFS, dando "time out error".

Tabla 7. Comparación de diferentes métodos de clasificación frente a PLS usando datos reales

Comparación	Método de clasificación	Sondas de partida	Selección previa de sondas	Sondas entran método	Parámetros óptimos con "training set"	Sondas totales en el óptimo	Proporción errores de clasificación con "test set" (PEC)
Waldenström Macroglobulinemia (WM) (B limfocitos) ***frente a*** Leucemia linfocítica crónica (CLL) (B limfocitos)[a]	KNN	5797	Ninguna	5797	KNN = 3 , LOO	5797	**0**
	PAM				Threshold = 0, CV (3 fold)	5797	**0**
	SVM				Cost = 0.6, LOO	5797	**0**
	PLS-Estándar				ITER = 1, NUMFACT = 1	5797	**0**
	PLS-VIP				ITER = 8, NUMFACT = 1	11	**0**
	LDA	5797	test-t*	7 top‡	Covarianzas iguales	7	**0**
	KNN		test-t	433	KNN = 7, LOO	433	**0**
	PAM				Threshold = 0, CV(3 fold)	433	**0**
	SVM				Cost = 0.2, LOO	433	**0**
Mieloma Múltiple sin ganancia en 1q (células plasmáticas) ***frente a*** Mieloma Múltiple con ganancia en 1q (células plasmáticas)[b]	KNN	9789	Ninguna	9789	KNN = 5, LOO	9789	**0.21**
	PAM				Threshold = 0, CV(3 fold)	9789	**0.36**
	SVM				Cost = 2.4, LOO	9789	**0.29**
	PLS-Estándar				ITER = 1, NUMFACT = 2	9789	**0.14**
	PLS-VIP				ITER = 8, NUMFACT = 1	10	**0.07**
	LDA	9789	test-t**	14 top‡	Covarianzas iguales	14	**0.14**
	KNN		test-t	272	KNN = 8, LOO	272	**0.14**
	PAM				Threshold = 0, CV (3 fold)	272	**0.29**
	SVM				Cost = 0.2, LOO	272	**0.36**
	KNN		CFS	9789	Time out error	–	–
	PAM		Theshold		Threshold = 2.0, CV (3 fold)	171	**0.14**
	SVM		CFS		Time out error	–	–

[a] Serie entrenamiento (8 WM, 9 CLL), serie prueba (2 WM, 2 CLL). [b] Serie entrenamiento (15 1q, 27 non-1q), serie prueba (5 1q, 9 non-1q). *Test-t con FDR < 0.05 dejó 433 sondas top. ** test-t con FDR < 0.05 dejó 272 sondas top. CFS = Correlation feature selection. ‡ LDA requiere menos variables que casos en una de las clases (7 en un caso y 14 en otro)). ITER = Nº de iteracciones y NUMFACT= Nº de factores (en los algoritmos PLS). LOO = "Leave One Out", CV (3 fold) = Cross Validation

4.3. Construcción de modelos predictores que combinan variables clínicas y génicas usando el algoritmo *PLS-VIP-Consecutivo*.

Con el fin de estudiar si los modelos predictores serían más exactos combinando variables clínicas clásicas con variables génicas de *microarrays*, se procedió a ajustar con *PLS-VIP* las variables clínicas por un lado y las génicas por otro; luego se adicionaban las variables óptimas obtenidas (clínicas + génicas), con el fin de analizar en qué escenarios dicha combinación de variables resultaba ser mejor predictor con *PLS-VIP* que los dos tipos de variables por separado.

La metodología seguida fue la ya expuesta en el diagrama de la Figura 21, que viene a ser un *PLS-VIP-Consecutivo*, en el que se ajustan sucesivamente las variables clínicas y génicas, cada una por su lado, y las variables retenidas se adicionan y se ajustan de nuevo, midiendo en todos los casos el error de clasificación.

4.3.1. Predictores clínicos: Escenarios simulados con variables clínicas

Según los procedimientos aleatorios descritos en el 3.5.1 de "Metodología", relativo al programa *SIMDATA*, se simularon diferentes escenarios con variables clínicas, tanto dicotómicas como continuas. El objetivo era mimetizar, dentro de lo posible, diferentes situaciones clínicas en las que tales variables suelen cambiar en número y en potencia discriminante entre clases. En las variables dicotómicas la potencia se simulaba variando el riesgo relativo (RR) en la dirección tumor a control y en las continuas variando el tamaño del efecto (TE) en la misma dirección. Se simularon 8 escenarios en lo que se varió el tamaño de muestra y el número de variables tanto dicotómicas como continuas. Los detalles de estos escenarios se mostraron ya en la Tabla 1, pero

brevemente se puede recordar aquí que los escenarios comprenden dos bloques y cada uno a su vez se desdobla en otros cuatro escenarios. Se variaba el tamaño de muestra (10 controles y 10 tumores o 25 controles y 25 tumores), el número de variables dicotómicas (2 ó 4) y el número de variables continuas (3 ó 6) y el poder discriminatorio de las variables dicotómicas (RR = 3 y RR = 6) y continuas (TE = 0.4 y TE = 0.8).

De cada escenario se simulaban aleatoriamente 50 series usando las mismas condiciones, con la intención de obtener unos valores estadísticos promedio, y en todos ellos se simularon tanto series de "entrenamiento" como las respectivas series de "prueba". Como medida de tendencia central se eligió la mediana y como indicadores de dispersión el primer y tercer cuartil: M (Q1,Q3). Esta elección trataba de evitar el efecto de sesgo de los valores atípicos que se observaron en alguna de las 50 repeticiones analizadas.

Resultados de los escenarios con solo variables clínicas

En la Tabla 8, se han recogido los resultados de los 8 escenarios simulados (CLIN1 a CLIN8). En la primera columna se cita el escenario, en la segunda se incluyen los casos de dicho escenario (nº de controles (C) y nº de tumores (T)), en la tercera se anota el número de variables dicotómicas iniciales y el RR simulado para las mismas, en la cuarta se incluyen el número de variables continuas iniciales junto con el TE asociado a cada variable. Las cinco columnas restantes se refieren ya al resultado de ajustar con *PLS-VIP* las variables iniciales y su significado es intuitivo. Recuérdese que el algoritmo *PLS-VIP* se "entrena" con las series de entrenamiento, pero que la bondad del modelo se calcula a partir de las series de prueba simuladas aleatoriamente en las mismas condiciones que las de entrenamiento.

En la Tabla 8 se muestran 2 bloques bien diferenciados, el de los escenarios CLIN1 a CLIN4 (parte superior de la tabla), en el que el número de variables es de 5 (2 dicotómicas y 3 continuas), y el de los escenarios CLIN5 a CLIN8, que fueron simulados con 10 variables (4 dicotómicas y 6 continuas). En ambos bloques se ha seguido la misma sistemática en cuanto a la potencia de las variables, variando el RR de las dicotómicas de 3 a 6 y el TE de las continuas de 0.4 a 0.8.

La observación detenida de esta tabla 8 sugiere varias interpretaciones y conclusiones:

1) En la aplicación de *PLS-VIP*, normalmente ha bastado con 1 iteración y 1 factor, apreciándose una mediana de 1 para el número de iteraciones y de factores en todos los escenarios, si bien el tercer cuartil alcanza en algunos casos el valor de 2. Este comportamiento parece lógico si se considera que el número de variables es muy pequeño para la técnica *PLS-VIP*.

2) El número de variables dicotómicas y continuas conservadas en el óptimo obtenido por *PLS-VIP*, presenta medianas que coinciden con el número de variables de partida en todos los escenarios. No obstante las horquillas del primer y tercer cuartil (Q1,Q3) varían ahora ligeramente, sobre todo en las variables continuas de algún escenario. La interpretación sería que, al existir pocas variables iniciales, el algoritmo *PLS-VIP* no tiene capacidad para eliminar muchas variables mediante iteraciones, aunque algunas veces si puede hacerlo, como se aprecia viendo las horquillas de Q1 a Q3 en algunos escenarios, como por ejemplo el CLIN1 y el CLIN5.

3) La comparación de CLIN1 con CLIN3 y de CLIN2 con CLIN4, donde se ha aumentado el tamaño de muestra (de 20 a 50 casos), manteniendo constante el resto de las condiciones, pone de manifiesto que el tamaño de muestra influye poco en los errores de clasificación, cambiando solamente de 0.30 a 0.32 y de 0.10 a 0.14, respectivamente. Un comportamiento semejante puede observarse al contrastar CLIN5 con CLIN7 y CLIN6 con CLIN8, cambiando ahora el error de clasificación de 0.25 a 0.26 y de 0.05 a 0.06, respectivamente. Parece, por tanto, que el tamaño de muestra, en estas condiciones, no tiene especial influencia sobre la bondad de predicción, posiblemente porque se alcanzado digamos una "meseta de potencia", de forma que a partir de 20 muestras (10C y 10T) el error de predicción no disminuye y un número mayor de muestras no parecería necesario en la práctica.

4) Dentro de cada bloque (CLIN1 a CLIN4 y CLIN5 a CLIN8), se aprecia, como cabría esperar, que los errores de clasificación mejoran (disminuyen) cuando el RR pasa de 3 a 6 y el TE pasa de 0.4 a 0.8. En concordancia con lo comentado en el punto anterior, esta mejora no se ve afectada apreciablemente por el tamaño de muestra (CLIN5 *vs.* CLIN7 y CLIN6 *vs.* CLIN8)

5) Comparando el bloque superior (CLIN1 a CLIN4) con el inferior (CLIN5 a CLIN8), se observa que el error de clasificación disminuye al pasar de 5 variables a 10. Este hecho resulta lógico, ya que al aumentar el número de variables, la información de la que dispone el algoritmo para la clasificación es

mayor. Precisamente el cuantificar estos efectos es una de las ventajas de los métodos de simulación.

6) En resumen, para *PLS-VIP*, las simulaciones han puesto de manifiesto que un tamaño de muestra de 10 controles y 10 tumores, junto con 4 variables dicotómicas con RR = 6 y también 6 variables clínicas con TE = 0.8 (es decir CLIN6 y CLIN8), podría ser una buena opción para obtener unos errores de predicción en torno al 0.05 (5% de mal clasificados). El TE de 0.8 para una variable continua es considerado un efecto de tipo medio y no parece difícil de alcanzar en variables clínicas continuas. Más difícil, sin duda, es disponer en la práctica de variables dicotómicas con RR = 6, ya que supone una diferencia clínica bastante grande. En base a esto, se podría concluir que las variables dicotómicas son menos precisas y por tanto de mayor confusión que las variables continuas, lo cual podría considerarse lógico.

Table 8. Aplicación de PLS-VIP a escenarios simulados con sólo variables clínicas y 50 repeticiones cada uno

Simulaciones y variables iniciales				Resultado del ajuste PLS-VIP y variables óptimas finales				
Escenario	**Casos**	**Dicotómicas iniciales**	**Continuas iniciales**	**Iteración óptima (Q1 , Q3)***	**Nº factores óptimo (Q1 , Q3)**	**Dicotómicas conservadas en óptimo (Q1 , Q3)**	**Continuas conservadas en óptimo (Q1 , Q3)**	**Proporción errores de clasificación (Q1 , Q3)**
CLIN1	10C y 10T	2 (RR = 3)	3 (TE=0.4)	1 (1 , 2)	1 (1 , 2)	2 (2 , 2)	3 (1 , 3)	0.30 (0.25 , 0.40)
CLIN2	10C y 10T	2 (RR = 6)	3 (TE=0.8)	1 (1 , 2)	1 (1 , 1)	2 (2 , 2)	3 (1 , 3)	0.10 (0.05 , 0.15)
CLIN3	25C y 25T	2 (RR = 3)	3 (TE=0.4)	1 (1 , 2)	1 (1 , 1)	2 (2 , 2)	3 (1 , 3)	0.32 (0.28 , 0.38)
CLIN4	25C y 25T	2 (RR = 6)	3 (TE=0.8)	1 (1 , 1)	1 (1 , 1)	2 (2 , 2)	3 (3 , 3)	0.14 (0.10 , 0.17)
CLIN5	10C y 10T	4 (RR = 3)	6 (TE=0.4)	1 (1 , 2)	1 (1 , 2)	4 (2 , 4)	6 (2 , 6)	0.25 (0.20 , 0.30)
CLIN6	10C y 10T	4 (RR = 6)	6 (TE=0.8)	1 (1 , 2)	1 (1 , 1)	4 (4 , 4)	6 (3 , 6)	0.05 (0.05 , 0.10)
CLIN7	25C y 25T	4 (RR = 3)	6 (TE=0.4)	1 (1 , 1)	1 (1 , 2)	4 (4 , 4)	6 (6 , 6)	0.26 (0.22 , 0.30)
CLIN8	25C y 25T	4 (RR = 6)	6 (TE=0.8)	1 (1 , 1)	1 (1 , 1)	4 (4 , 4)	6 (6 , 6)	0.06 (0.04 , 0.08)

*Los valores se refieren a la mediana (primer cuartil , tercer cuartil) a partir de 50 repeticiones para cada escenario.
Las series "training" y "test" tienen el mismo número de controles (C) y tumores (T), según se indica.
RR = riesgo relativo. TE = tamaño del efecto.

Representaciones PLS de una serie con solo variables clínicas

Como ya se ha comentado en otros apartados, la ventaja adicional que presenta PLS frente a otros métodos de clasificación es que dispone de múltiples representaciones gráficas, las cuales permiten visualizar los resultados y facilitar su interpretación. Así, a modo de ejemplo, se muestran a continuación diferentes gráficas de la serie de datos 17 de las 50 que forman el escenario CLIN4. Se ha elegido esta serie por tratarse de unos datos con potencia discriminante intermedia, por lo que puede resultar más ilustrativa. Concretamente, se trata de una simulación con 25 muestras control y 25 de tumor, con 2 variables dicotómicas con RR = 6 y 3 variables continuas con TE = 0.8.

Al aplicar *PLS-VIP*, se obtuvo el óptimo con nº de iteraciones = 1, nº de factores = 2, teniendo como variables finales las 5 de partida y un error de clasificación con la serie "hermana" de prueba (test17) de 0.18. La varianza acumulada para las ***X*** con los 2 factores y la serie de entrenamiento (training17) fue de 52% (38% + 14%), que resulta un poco baja, lo que indicaría que 5 variables parecen escasas para el buen funcionamiento de PLS. La varianza acumulada para las ***Y*** fue mayor, del 64% (62% + 2%).

En la Figura 33 se muestra las puntuaciones y las cargas de las variables ***X*** (clínicas) para la serie de entrenamiento en el óptimo arriba mencionado. En las puntuaciones puede apreciarse una cierta separación entre las muestras control que quedan a la derecha y las muestras tumor que se sitúan a la izquierda, si bien los puntos no están muy agrupados. En cuanto a las cargas, se observa que, con los 2 factores PLS del óptimo, se separan principalmente las variables dicotómicas de las continuas. El error de clasificación usando la serie "hermana" de prueba (test17) fue de 0.18, que supone 9 casos mal clasificados de 50, apareciendo 3 falsos positivos y 5 falsos negativos.

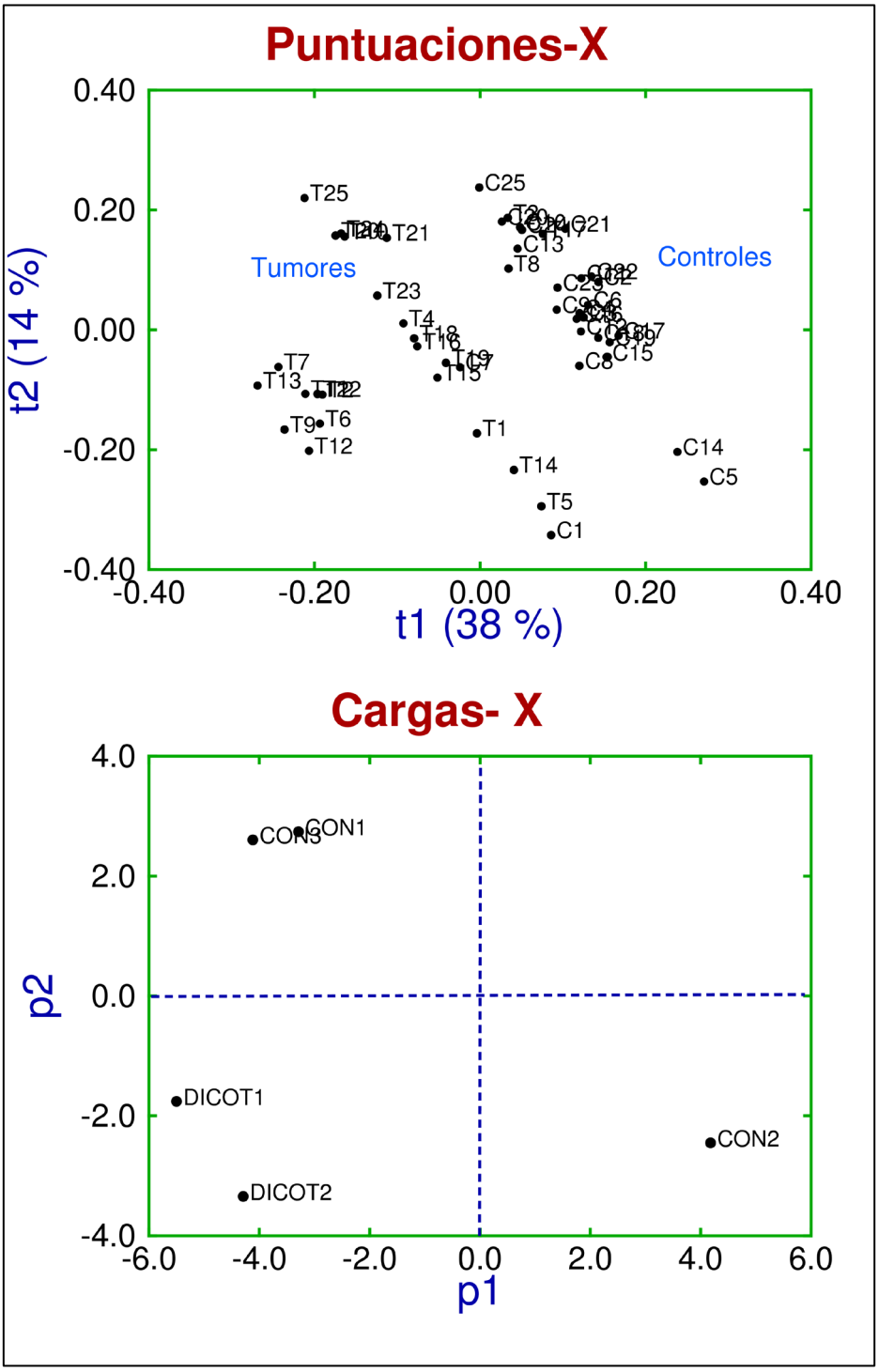

Figura 33. Puntuaciones y cargas de las variables clínicas X de la serie 17 de CLIN4. DICOT# = variable dicotómica-nº. CON# = variable continua-nº.

En la Figura 34 se han representado las puntuaciones de la variable ***Y*** (***u*** $_1$ *vs.* ***u*** $_2$). El factor 1 (***u*** $_1$) se emplea en separar los controles a la derecha y los tumores a la izquierda, como era de esperar para una variable dicotómica (0 y 1), ocupándose el factor 2 (***u*** $_2$) en matizar las diferencias entre las muestras.

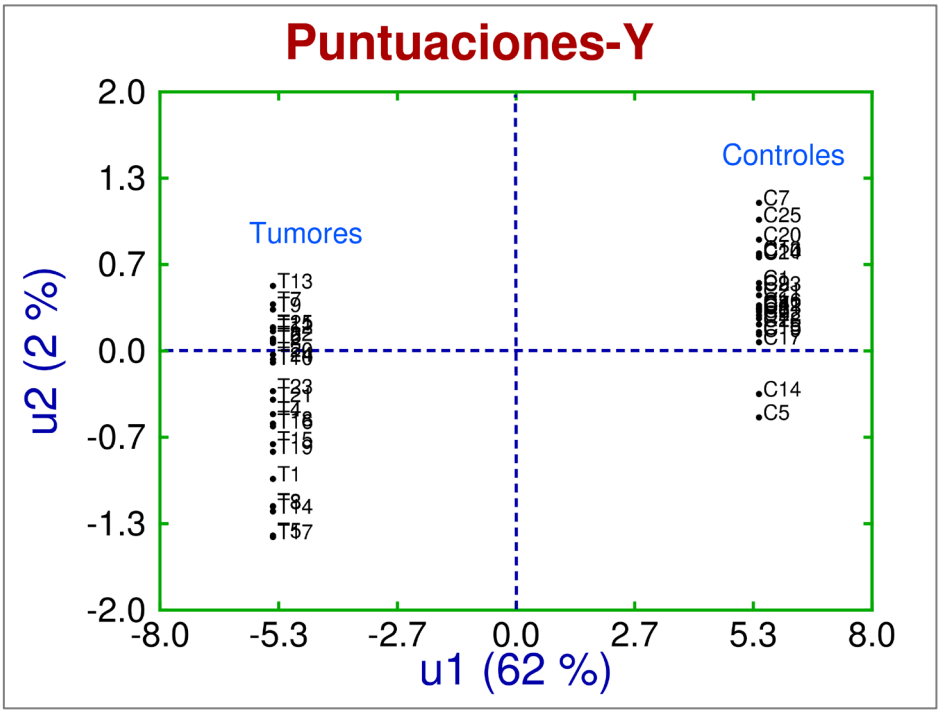

Figura 34. Puntuaciones de las variables respuesta Y de la serie 17 de CLIN4.

Por último, en la Figura 35 se recogen 2 gráficas importantes referidas a la bondad del ajuste PLS con la serie de entrenamiento en su óptimo de *PLS-VIP*. Estas gráficas son las que corresponden a las sucesivas correlaciones entre las puntuaciones de las ***X*** y de las ***Y*** , llamadas ***t*** y ***u***. Cada gráfica muestra el ajuste de regresión lineal de los valores $\boldsymbol{u}_i$ sobre los valores de $\boldsymbol{t}_i$ (línea continua) y también el ajuste de regresión lineal de $\boldsymbol{t}_i$ sobre $\boldsymbol{u}_i$ (línea discontinua), junto con los coeficientes de correlación, ***r***, y los niveles de significancia ***p***. Como puede observarse, las puntuaciones del primer factor ($\boldsymbol{u}_1$ y $\boldsymbol{t}_1$) están aceptablemente correlacionadas como indica el buen solapamiento de ambas líneas de

regresión y los valores de ***r*** y ***p***. Sin embargo la correlación entre las puntuaciones del segundo factor (***u₂*** y ***t₂***) aunque ligeramente apreciable no resulta significativa, posiblemente debido al escaso número de variables de la serie que es de 5.

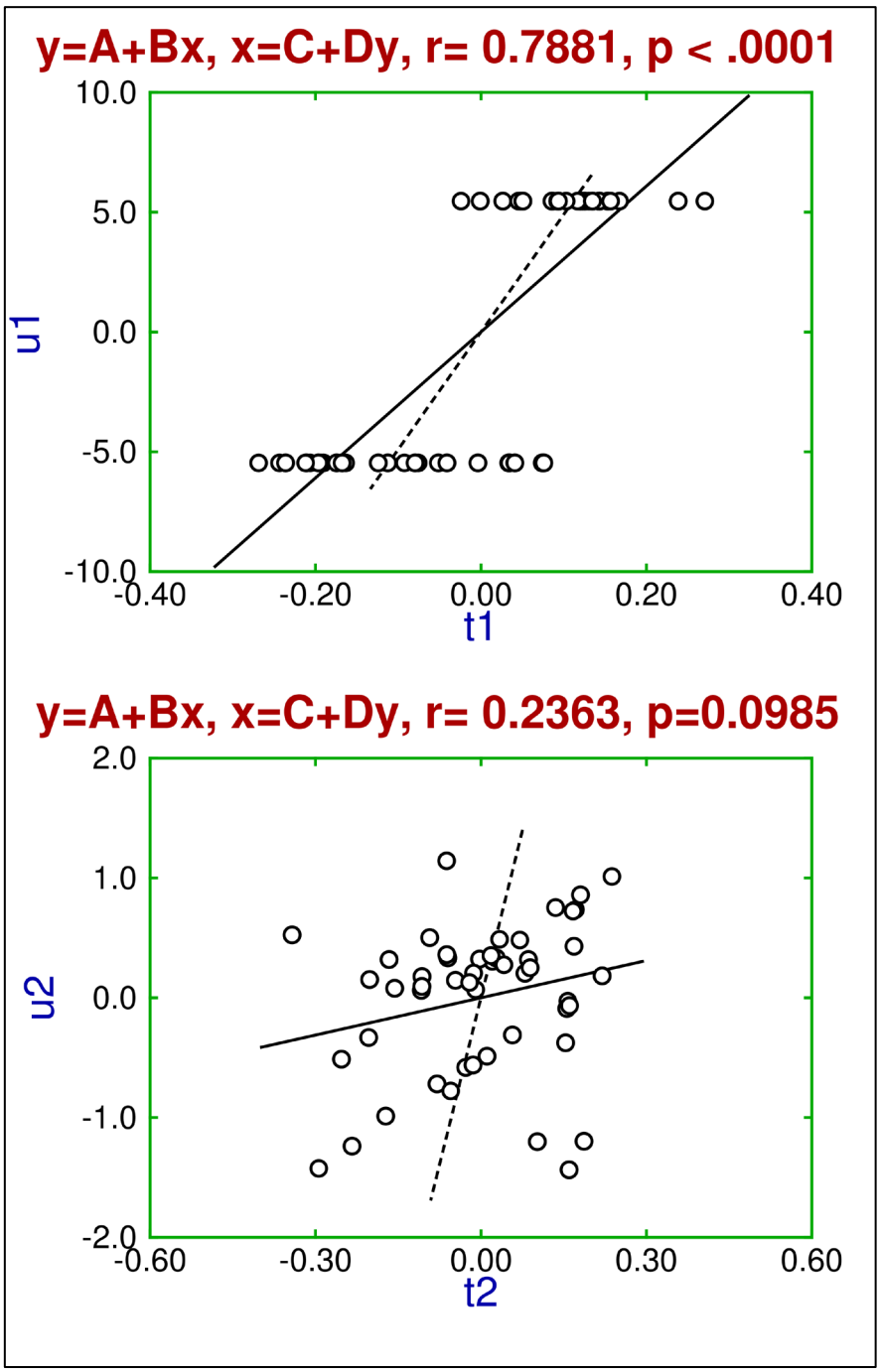

Figura 35. Representaciones u_i frente a t_i para dos factores de la serie 17 de CLIN4. (A) Factor 1 y (B) factor 2. Línea continua = regresión de *y* sobre *x*. Línea discontinua = regresión de *x* sobre *y*.

4.3.2 Predictores genómicos: Escenarios simulados con sólo variables de expresión génica de *microarrays*

Las variables génicas se simularon aleatoriamente con el programa *SIMDATA* descrito en el apartado 3.5.2 de "Metodología". En esencia, se simularon 8 escenarios en los que se iba variando el tamaño de muestra, el número de genes diferencialmente expresados (dif. exp.) y el número de genes basales o genes ruido. La potencia discriminante de los genes dif. exp. se variaba mediante un desplazamiento Δ del valor de la expresión génica en las muestras de tumor respecto a las de control; el cual podía ser positivo (genes sobre-expresados) o negativo (genes infra-expresados). Estos escenarios se han mostrado ya en la Tabla 3 y son análogos a los de las variables clínicas de la Tabla 1. Brevemente, se puede recordar aquí que comprenden dos bloques y cada uno a su vez se desdobla en cuatro escenarios. Se variaba el tamaño de muestra (10 controles y 10 tumores o 25 controles y 25 tumores), el número de genes dif. exp. (20 ó 40) y el número de genes ruido (980 ó 960) y el poder discriminatorio de los genes ($\Delta = 0.4$ y $\Delta = 0.6$).

De cada escenario se simularon 50 series de datos de "entrenamiento" y otras 50 series de datos de "prueba", con el objeto de obtener al final unos valores estadísticos promedio. Al igual que en el caso de las variables clínicas se optó por recopilar para cada escenario el valor de la mediana y del primer y tercer cuartil: M (Q1,Q3).

Resultados de los escenarios con solo variables génicas

En la Tabla 9 se han recopilado los resultados obtenidos al aplicar el método *PLS-VIP* a los 8 escenarios simulados. La primera columna muestra el nombre del escenario (GENIC1 a GENIC8), en la segunda se incluye el tamaño de muestra (nº de controles (C) y nº de tumores (T)), en la tercera se indica el nº de genes dif. exp., en la cuarta el valor del desplazamiento Δ y en la quinta el número de genes ruido. Las 5 columnas restantes

muestran los resultados obtenidos al aplicar *PLS-VIP* a las variables de partida. En dicha tabla aparecen de nuevo dos bloques, el de la parte superior que abarca los escenarios GENIC1 a GENIC4 con 20 genes dif. exp. y 980 genes ruido, y el de la parte inferior que incluye los escenarios GENIC5 a GENIC8 simulados con 40 genes dif. exp. y 960 genes ruido. En ambos bloques se variaba el tamaño de muestra (20 y 50, respectivamente) y el valor del desplazamiento Δ para los genes (0.4 y 0.6, respectivamente).

A la vista de la Tabla 9 se pueden hacer las siguientes interpretaciones:

1) Debido a que todos los escenarios tienen ahora 1000 variables (genes) el algoritmo *PLS-VIP* evoluciona haciendo uso de mayor número de iteraciones que en el caso anterior de las variables clínicas, llegando a alcanzar óptimos con un promedio de 3 a 5 iteraciones. El número de factores PLS en los óptimos es ahora de 1 ó 2, lo cual es razonable si se piensa que la variable respuesta sigue siendo dicotómica (Control =0, Tumor = 1).
2) El número de genes totales conservados en el óptimo varía según los escenarios, con medianas que oscilan entre 18 y 133 genes, de los cuales no todos son verdaderos genes dif. exp. sino que en las listas aparecen por azar genes falsos positivos (FP) en casi todos los escenarios. Esta proporción de genes FP disminuye siempre al aumentar el tamaños de muestra (GENIC1 vs. GENIC3, GENIC2 vs. GENIC4, GENIC5 vs. GENIC7, GENIC6 vs. GENIC8). También disminuyen estos FP al aumentar el valor del desplazamiento Δ en los genes dif. exp., por ejemplo cuando se pasa de GENIC3 a GENIC4 o de GENIC7 a GENIC8. Este comportamiento parece lógico y confirma que *PLS-VIP* está funcionando correctamente, encontrando menos genes FP cuando las condiciones de los genes dif. exp. se van haciendo más discriminantes en los escenarios.

3) A partir de los errores de clasificación para los distintos escenarios (última columna de la Tabla), se puede concluir lo siguiente:
   a) El tamaño de muestra juega un papel más importante en las variables génicas que en las anteriores clínicas, por ejemplo disminuyendo el error de 0.25 a 0.12 cuando se pasa de GENIC1 (20 muestras) a GENIC3 (50 muestras) o de 0.20 a 0.02 al pasar de GENIC5 (20 muestras) a GENIC7 (50 muestras).
   b) La disminución del error es aún mayor cuando se aumenta el desplazamiento Δ de 0.4 a 0.6 en los genes dif. exp., como puede verse comparando GENIC1 frente a GENIC2 (el error pasa de 0.25 a 0.05) o GENIC5 frente a GENIC6 (el error pasa de 0.20 a 0).
   c) El aumento de los genes dif. exp. de 20 a 40 disminuye también el error de clasificación, como cabía esperar, y que puede observarse comparando por ejemplo GENIC3 con GENIC7 (el error pasa de 0.12 a 0.02).
   d) Los menores errores de clasificación se alcanzan, como también era previsible, cuando las condiciones simuladas son las más discriminantes dentro de cada bloque (GENIC4 en el primero y GENIC8 en el segundo), donde sus medianas para el error de predicción llegan a valer 0, es decir se consigue el 100% de aciertos en los 50 casos (25C y 25T) de las series de prueba.
4) A la vista de las simulaciones, se puede sugerir desde el punto de vista práctico que, usando *PLS-VIP*, las condiciones más favorables para datos de microarrays con 1000 genes, se dan cuando el número de muestras es en torno a 25 controles y 25 tumores y el número de genes dif. exp. es $\geq$ 20 con un desplazamiento en log(2) en torno a 0.6 en valor absoluto (sobre-expresados o infra-expresados).

Table 9. PLS-VIP con distintos escenarios simulados con sólo variables génicas y 50 repeticiones cada uno

Simulaciones y genes iniciales					Resultado del ajuste PLS-VIP y genes en el óptimo *				
Escenario	**Casos**	**Genes dif. exp. Iniciales**	**Delta (σ_{delta})**	**Genes ruido iniciales**	**Iteración óptima (Q1 , Q3)**	**Nº factores óptimo (Q1 , Q3)**	**Genes totales conservados en óptimo (Q1 , Q3)**	**Proporción de genes falsos positivos (Q1 , Q3)**	**Proporción errores de clasificación (Q1 , Q3)**
GENIC1	10C y 10T	20	0.4 (0.024)	980	3.5 (3 , 5)	2 (1 , 3)	96 (24 , 169)	0.91 (0.86 , 0.94)	0.25 (0.20 , 0.30)
GENIC2	10C y 10T	20	0.6 (0.036)	980	5 (4 , 6)	2 (1 , 3)	18 (8 , 42)	0.55 (0.39 , 0.69)	0.05 (0.05 , 0.10)
GENIC3	25C y 25T	20	0.4 (0.024)	980	4 (4 , 5)	2 (1 , 2)	39 (29 , 56)	0.55 (0.47 , 0.67)	0.12 (0.10 , 0.16)
GENIC4	25C y 25T	20	0.6 (0.036)	980	5 (4 , 5)	1 (1 , 2)	21 (13 , 25)	0.09 (0 , 0.23)	0 (0 , 0.02)
GENIC5	10C y 10T	40	0.4 (0.024)	960	3 (2 , 5)	2 (2 , 2)	113 (40 , 347)	0.85 (0.79 , 0.93)	0.20 (0.10 , 0.25)
GENIC6	10C y 10T	40	0.6 (0.036)	960	5 (4 , 5)	1 (1 , 2)	35 (15 , 47)	0.36 (0.26 , 0.51)	0 (0 , 0.03)
GENIC7	25C y 25T	40	0.4 (0.024)	960	4 (3 , 4)	2 (1 , 2)	69 (41 , 126)	0.47 (0.29 , 0.69)	0.02 (0.02 , 0.06)
GENIC8	25C y 25T	40	0.6 (0.036)	960	5 (4 , 5)	1 (1 , 1)	25 (17 , 33)	0 (0 , 0.06)	0 (0 , 0)

*Los valores se refieren a: mediana (primer cuartil - tercer cuartil), a partir de 50 series simuladas al azar.
Las series "training" y "test" tienen el mismo número de controles (C) y tumores (T), según se indica.
Delta es el incremento del gen en su expresión diferencial.

Representaciones PLS de una serie con variables génicas

Como se hizo con las variables clínicas, se muestran a continuación, a modo de ejemplo, diferentes gráficas de la misma serie 17 pero ahora en variables génicas (escenario GENIC3). Concretamente, se trata de una simulación con 25 muestras control y 25 de tumor, con 20 genes dif. exp. con $\Delta = 0.4$ y 980 genes ruido.

Al aplicar *PLS-VIP* a esta serie, se obtuvo el óptimo en nº de iteraciones = 6 y nº de factores = 2, teniendo 11 genes finales (7 dif. exp. y 4 ruido). La varianza acumulada para las ***X*** con los 2 factores y la serie de entrenamiento fue de 36% (26% + 10%), que resulta baja, lo que parece significar que, aunque ahora el nº de variables es mayor (11), su bajo poder discriminante y la presencia de variables ruido conlleva una extracción de varianza escasa. La varianza acumulada para las ***Y*** fue, sin embargo mayor, siendo del 64% (62% + 2%), con una escasa participación del factor 2 frente al 1. El error de clasificación usando la serie de prueba fue de 0.20, presentando 10 mal clasificados de 50 casos, habiendo 6 casos falsos positivos y 4 casos falsos negativos.

En la Figura 36 se muestra las puntuaciones y las cargas de las variables ***X*** (génicas ahora) para la serie de entrenamiento en el óptimo arriba mencionado. En las puntuaciones puede apreciarse una cierta separación entre las muestras control y las muestras tumor, si bien los puntos no están muy agrupados. En cuanto a las cargas, se observa que con los 2 factores PLS se separan aceptablemente las variables génicas, siendo el factor 1 el que separa los genes infra-expresadas (D en la gráfica (del inglés "Down")) de los sobre-expresados (U de "Up" en la gráfica). Los genes ruido ("Noise" en la gráfica) se adaptan también a esta separación de genes sobre expresados (genes 129,

703 y 394) y un gen infra-expresado (el 385), como se comprobó en la matriz de datos comparando las medias de los controles frente a los tumores.

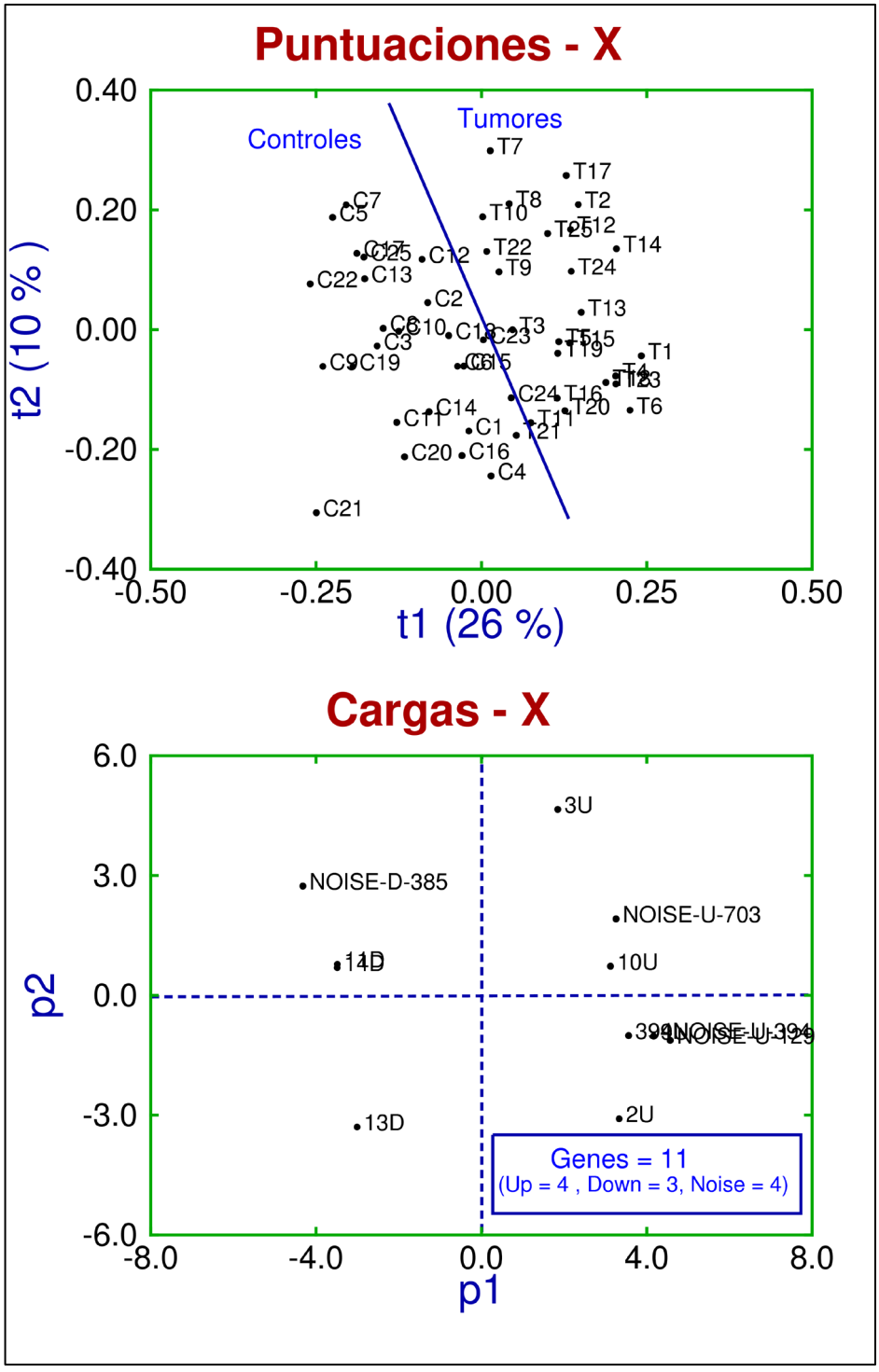

Figura 36. Puntuaciones y cargas para las variables *X* de la serie 17 de GENIC3. U = up (gen sobre-expresado), D = down (gen infra-expresado), NOISE-U = gen ruido sobre-expresado, NOISE-D = gen ruido infra-expresado.

El comportamiento para las puntuaciones de la variable ***Y*** es análogo al de la Figura 34 con variables clínicas, por lo que se omite por brevedad.

Por último, en la Figura 37 se recogen las 2 gráficas importantes acerca de la bondad del ajuste PLS con la serie de entrenamiento en el óptimo de PLS-VIP, que son, como ya se ha comentado más arriba, las que corresponden a las sucesivas correlaciones entre las puntuaciones de las ***X*** y de las ***Y*** , llamadas ***t*** y ***u***. Cada gráfica muestra el ajuste de regresión lineal de los valores u_i sobre los valores de t_i (línea continua) y también el ajuste de regresión lineal de t_i sobre u_i (línea discontinua), junto con los coeficientes de correlación, ***r***, y niveles de significancia ***p***. Como puede observarse, las puntuaciones del primer factor (u_1 y t_1) están aceptablemente correlacionadas como indica el buen solapamiento de ambas líneas de regresión y los valores de ***r*** y ***p***. La correlación entre las puntuaciones del segundo factor (u_2 y t_2) resulta también significativo (p =0.001), a diferencia de lo observado en la Figura 35 para las variables clínicas (p = 0.10).

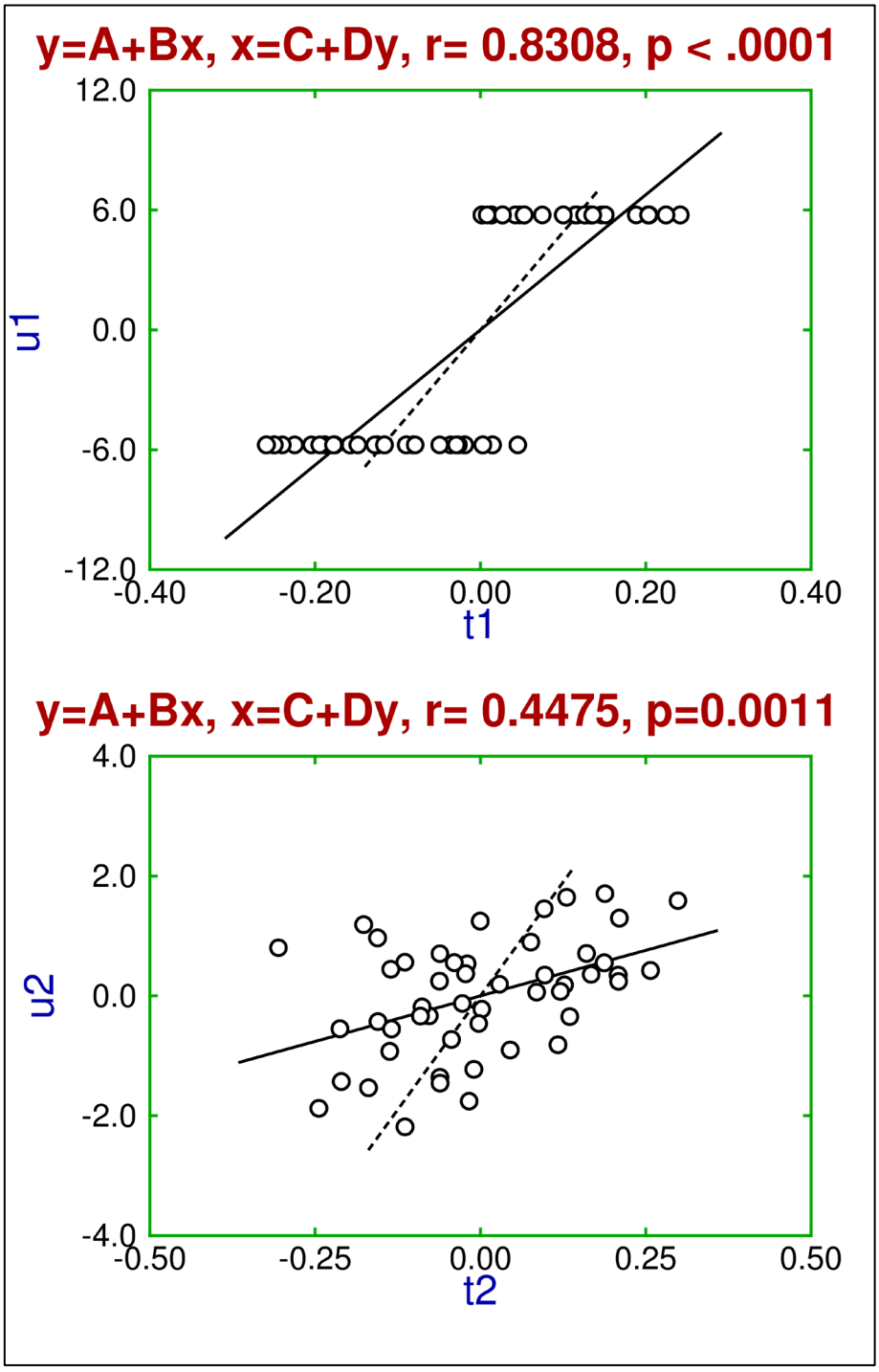

Figura 37. Representaciones de u_i frente a t_i para los factores la serie 17 de GEN3. (A) Factor 1 y (B) factor 2. Línea continua = regresión de *y* sobre *x*. Línea discontinua = regresión de *x* sobre *y*.

4.3.3 Predictores clínico-genómicos: Combinación de los modelos óptimos obtenidos en los apartados 4.3.1 y 4.3.2

En los apartados anteriores se ha aplicado *PLS-VIP* a las variables clínicas y génicas por separado usando diferentes escenarios simulados aleatoriamente. En todos los casos el algoritmo alcanza un óptimo en el que el error de predicción, usando la correspondiente serie de prueba, es mínimo. Ahora la pregunta clave es si se llegaría a un menor error de clasificación si se unieran ambos tipos de variables, fusionando sus matrices de datos en sus respectivos óptimos. A continuación se tratará de responder a esta pregunta combinando los datos de los escenarios clínicos con los génicos en unos nuevos escenarios que llamaremos clínico-genómicos.

Resultados de los escenarios clínico-genómicos

Combinando los datos de los 8 escenarios clínicos con los 8 escenarios génicos se diseñaron 16 escenarios clínico-genómicos que se denominaron sucesivamente CLIGEN1-1, CLIGEN1-2, ….CLIGEN8-7, CLIGEN8-8; donde el primer dígito se refiere al escenario clínico y el segundo al génico, según los nombres utilizados en los apartados anteriores.

En todos los casos la matriz de datos clínicos de cada escenario, en su óptimo de *PLS-VIP*, se fusionaba con la matriz de datos óptima del escenario génico correspondiente. Esta fusión de matrices se hacía tanto para la serie de entrenamiento como la de prueba.

Los 16 escenarios considerados abarcaban diferentes tipos de tamaño de muestra, de variables clínicas y génicas, de parámetros discriminantes (ER, TE, genes dif. exp., etc.), y cada uno consistía de 50 matrices de entrenamiento y 50 de prueba, con el fin de

obtener un promedio estadístico. Los resultados de aplicar *PLS-VIP* a dichos escenarios se han reunido en la Tabla 10. Esta Tabla consta de 3 paneles verticales, el de la izquierda se refiere a un recordatorio de las condiciones de los escenarios clínicos ya analizados en el apartado 4.3.1, el panel del medio es análogo pero relativo a los escenarios génicos expuestos en el apartado 4.3.2. El panel más ancho de la derecha muestra los nuevos resultados obtenidos con *PLS-VIP* sobre las matrices clínico-genómicas de cada escenario en sus respectivos óptimos, y las columnas recogen el nº de iteraciones, el nº de factores, las variables dicotómicas y continuas retenidas, los genes totales conservados, la proporción de genes FP y la proporción de casos mal clasificados (error de clasificación).

Para leer adecuadamente la Tabla 10 uno debe moverse horizontalmente de izquierda a derecha con el fin de recordar los resultados anteriores clínicos y génicos, deteniéndose luego en el panel de la derecha que es el que contiene los nuevos resultados clínico-genómicos. También debe repararse en que la Tabla 10 tiene horizontalmente 2 bloques semejantes en cuanto al tamaño de muestra, pero el superior tiene 5 variables clínicas, 20 genes dif. exp. y 980 genes ruido y el inferior tiene 10 variables clínicas, 40 genes dif. exp. y 960 genes ruido.

Observando en detalle la Tabla 10 se pueden hacer las siguientes interpretaciones:

1) Al combinar las matrices clínicas y génicas en sus óptimos *PLS-VIP* resulta claro que el número de variables de partida no es muy elevado, ya que muchas variables poco discriminantes fueron eliminadas por el algoritmo. Por tanto, al aplicar ahora *PLS-VIP* al escenario clínico-genómico, el algoritmo puede hacer pocas iteraciones, normalmente 1 ó 2 y en ciertos casos 3. En cuanto al nº de factores PLS, basta normalmente con 1 ó 2 y en algunas condiciones 3.

2) Cuando las variables clínicas y génicas son de potencia semejante en el correspondiente escenario clínico + génico, el nº de iteraciones suele ser 1, y por tanto se conservan todas las variables de partida. Sin embargo, cuando las variables génicas tienen mayor potencia discriminante que las clínicas, entonces el nº de iteraciones suele ser de 2, rechazando el algoritmo las variables clínicas y reteniéndose las génicas, bien las de partida o filtradas a un número menor. Esto es lo que se observa al comparar CLIGEN1-1 con CLIGEN1-2 donde se eliminan todas las variables clínicas, o comparando CLIGEN3-3 con CLIGEN3-4, o también CLIGEN7-7 con CLIGEN7-8.

3) Los genes totales conservados en el óptimo suelen ser ligeramente inferiores a los genes de partida proporcionados por el escenario génico, bien porque el algoritmo opera ahora con alguna variable más como son las clínicas (cuando la iteración es 1) o porque el propio algoritmo avanza hasta una iteración 2. Y respecto a la proporción de genes FP, se observa un patrón semejante al observado con sólo las variables génicas (Tabla 9), es decir disminuyen siempre al aumentar el tamaño de muestra o al aumentar el valor del desplazamiento Δ en los genes dif. exp., como parece lógico.

4) Analizando los errores de clasificación para los distintos escenarios, es decir clínicos, genómicos y clínico-genómicos (columnas sombreadas en verde), se puede destacar lo siguiente:

    a) Cuando el error es semejante en los escenarios clínico y génico, la adición de los dos puede traer consigo un error igual al más pequeño de los 2 (por ej. CLIGEN1-1 vs. CLIN1 y GENIC1) o más usualmente un

error intermedio entre los dos (por ej. CLIGE6-5 *vs.* CLIN6 y GENIC5).

b) Cuando el error es mucho mayor en un escenario que en otro, la adición de los dos conduce estrictamente a que el algoritmo se decanta por el de menor error, simplemente eliminando alguna variable, sin que haya producido ninguna ventaja el haber adicionado los dos escenarios, tal es el caso de CLIGEN7-8 vs. CLIN7 y GENIC8 o de CLIGEN8-8 vs. CLIN8 y GENIC8.

c) Aquí estamos llegando al punto capital de si tiene algún interés práctico el adicionar variables clínicas y génicas. A la vista de los resultados, parece que la respuesta no es clara, ya que si unas variables tienen mucha más potencia que las otras el algoritmo PLS-VIP, razonablemente, sólo se queda con una de ellas y el trabajo habría sido en vano desde un punto de vista experimental al haber medido los dos tipos de variables. Parece que, solamente en el caso de que los dos tipos de variables tengan potencias intermedias parecidas, su adición daría un error de clasificación inferior a las dos, al sumarse la potencia de ambas. Pero quedaría por determinar si estas diferencias son estadísticamente significativas. De tratar de contestar a este dilema nos ocuparemos en el apartado siguiente al aplicar el Test U de Mann Whitney a todos los escenarios usando todas sus repeticiones.

Table 10. PLS-VIP sobre escenarios simulados (50 repeticiones) con variables clínicas y génicas por separado y luego unidas en sus modelos óptimos

Escenario clínico					Escenario génico					Escenario clínico + génico unidos en sus óptimos							
Escenario Inicial (variables)	Óptimos desde PLS-VIP				Escenario Inicial (variables)	Optimos desde PLS-VIP				Escenario partida (con variables óptimas)	Valores óptimos después de PLS-VIP						
	Cas.	Var. Dicot	Var. Cont.	Error clasif.		Casos	Genes Totales	Proporción Fals. Pos.	Error clasific.		NITER	NFACT	Var. Dicot.	Var. Cont.	Genes totales	Proporción genes FP	Error clasif.
CLIN1 (Dicot. = 2 , RR=3, Cont. = 3, TE=0.4)	10C 10T	2	3	0.30	***GENIC1*** (20 dif. exp. (Δ=0.4), 980 ruido)	10C 10T	96	0.91	0.25	***CLIGEN1-1*** (CLIN1 + GENIC1)	1 (1 , 1)	2 (1 , 3)	2 (0.5 , 2)	3 (0 , 3)	77 (16 , 159)	0.89 (0.85 , 0.94)	0.25 (0.20 , 0.30)
					GENIC2 (20 dif. exp. (Δ=0.6), 980 ruido)		18	0.55	0.05	***CLIGEN1-2*** (CLIN1 + GENIC2)	2 (1 , 2)	2 (1 , 3)	0 (0 , 1)	0 (0 , 1)	15 (8 , 35)	0.51 (0.31 , 0.69)	0.05 (0.03 , 0.10)
CLIN2 (Dicot. = 2 , RR=6, Cont. = 3, TE=0.8)		2	3	0.10	***GENIC1*** (20 dif. exp. (Δ=0.4), 980 ruido)		96	0.91	0.25	***CLIGEN2-1*** (CLIN2 + GENIC1)	1 (1 , 3)	2 (1 , 3)	2 (1 , 2)	1 (0 , 3)	29 (6 , 117)	0.92 (0.86 ,1.00)	0.20 (0.13 , 0.25)
					GENIC2 (20 dif.exp. (Δ=0.6), 980 ruido)		18	0.55	0.05	***CLIGEN2-2*** (CLIN2 + GENIC2)	1 (1 , 2)	2 (1 , 2)	2 (1 , 2)	1 (0 , 3)	15 (7 , 34)	0.50 (0.29 , 0.69)	0.05 (0 , 0.10)
CLIN3 (Dicot. = 2 , RR=3, Cont. = 3, TE=0.4)	25C 25T	2	3	0.32	***GENIC3*** (20 dif.exp. (Δ=0.4), 980 ruido)	25C 25T	39	0.55	0.12	***CLIGEN3-3*** (CLIN3 + GENIC3)	1 (1 , 1)	2 (1 , 2)	2 (2 , 2)	3 (0 , 3)	39 (27 , 53)	0.56 (0.47 , 0.66)	0.12 (0.08 , 0.14)
					GENIC4 (20 dif.exp. (Δ=0.6), 980 ruido)		21	0.09	0	***CLIGEN3-4*** (CLIN3 + GENIC4)	2 (1 , 2)	1 (1 , 2)	1 (0 , 2)	0 (0 , 1)	17 (13 , 21)	0.05 (0 , 0.15)	0 (0 , 0.02)
CLIN4 (Dicot. = 2 , RR=6, Cont. = 3, TE=0.8)		2	3	0.14	***GENIC3*** (20 dif.exp. (Δ=0.4), 980 ruido)		39	0.55	0.12	***CLIGEN4-3*** (CLIN4 + GENIC3)	1 (1 , 2)	2 (1 , 2)	2 (2 , 2)	3 (2 , 3)	32 (15 , 47)	0.53 (0.36 , 0.64)	0.06 (0.04 , 0.10)
					GENIC4 (20 dif. exp. (Δ=0.6), 980 ruido)		21	0.09	0	***CLIGEN4-4*** (CLIN4 + GENIC4)	1 (1 , 2)	1 (1 , 1)	2 (2 , 2)	3 (0 , 3)	19 (10 ,23)	0.05 (0 , 0.17)	0 (0 , 0.02)
CLIN5 (Dicot. = 4 , RR=3, Cont. = 6, TE=0.4)	10C 10T	4	6	0.25	***GENIC5*** (40 dif.exp . (Δ=0.4), 960 ruido)	10C 10T	113	0.85	0.20	***CLIGEN5-5*** (CLIN5 + GENIC5)	1 (1 , 2)	2 (2 , 3)	3 (1 , 4)	3 (0 , 6)	91 (22 ,243)	0.83 (0.79 , 0.91)	0.15 (0.10 , 0.23)
					GENIC6 (40 dif. exp. (Δ=0.6), 960 ruido)		35	0.36	0	***CLIGEN5-6*** (CLIN5 + GENIC6)	2 (1 , 2)	1 (1 , 2)	0 (0 , 2)	0 (0 , 1)	19 (13 , 35)	0.31 (0.17 , 0.47)	0 (0 , 0)
CLIN6 (Dicot. = 4, RR=6, Cont. = 6, TE=0.8)		4	6	0.05	***GENIC5*** (40 dif. exp. (Δ=0.4), 960 ruido)		113	0.85	0.20	***CLIGEN6-5*** (CLIN6 + GENIC5)	1 (1 , 3)	2 (1 , 2)	4 (2 , 4)	3 (1 , 6)	41 (11 , 115)	0.83 (0.77 , 0.91)	0.10 (0.05 , 0.15)
					GENIC6 (40 dif. exp. (Δ=0.6), 960 ruido)		35	0.36	0	***CLIGEN6-6*** (CLIN6 + GENIC6)	2 (1 , 2)	1 (1 , 2)	2 (1 , 3)	1 (0 , 6)	17 (8 , 33)	0.29 (0.15 , 0.46)	0 (0 , 0)
CLIN7 (Dicot. = 4 , RR=3) (Cont. = 6, TE=0.4)	25C 25T	4	6	0.26	***GENIC7*** (40 dif. exp. (Δ=0.4), 960 ruido)	25C 25T	69	0.47	0.02	***CLIGEN7-7*** (CLIN7 + GENIC7)	1 (1 , 2)	2 (1 , 2)	4 (2 , 4)	6 (0 , 6)	47 (35 , 85)	0.39 (0.27 , 0.56)	0.04 (0.02 , 0.04)
					GENIC8 (40 dif. exp. (Δ=0.6), 960 ruido)		25	0	0	***CLIGEN7-8*** (CLIN7 + GENIC8)	2 (2 , 2)	1 (1 , 1)	0 (0 , 1)	0 (0 , 0)	25 (17 , 30)	0 (0 , 0.01)	0 (0 , 0)
CLIN8 (Dicot. = 4 , RR=6, Cont. = 6, TE=0.8)		4	6	0.06	***GENIC7*** (40 dif. exp. (Δ=0.4), 960 ruido)		69	0.47	0.02	***CLIGEN8-7*** (CLIN8 + GENIC7)	1 (1 , 2)	2 (1 , 2)	4 (4 , 4)	6 (4 , 6)	40 (25 , 65)	0.37 (0.23 , 0.48)	0 (0 , 0.02)
					GENIC8 (40 dif. exp. (Δ=0.6), 960 ruido)		25	0	0	***CLIGEN8-8*** (CLIN8 + GENIC8)	2 (1 , 2)	1 (1 , 1)	3 (2 , 4)	0 (0 , 6)	19 (14 , 25)	0 (0 , 0)	0 (0 , 0)

* Los valores se refieren a mediana (Q1 , Q3) a partir de 50 repeticiones. Las series "training" y "test" tienen igual número de controles (C) y tumores (T), según se indica.

Comparación de los errores de clasificación de los distintos escenarios usando la prueba U de Mann Whitney

El parámetro más importante en cualquier modelo predictivo es la proporción de casos mal clasificados (o error de clasificación). En los apartados anteriores, se han comparado estos errores de clasificación entre los distintos escenarios simulados (clínicos, génicos y clínico-genómicos) desde un punto semi-cuantitativo, pero sin abordar su significancia estadística. Para analizar este aspecto se ha recurrido a los errores de clasificación de todos los escenarios entre sí, aplicando la prueba U de Mann Whitney a los valores de las respectivas 50 repeticiones de cada escenario. Los resultados de esta prueba, junto con sus p-valores se han recopilado en la Tabla 11. Esta Tabla consta de 3 paneles, el primero y el segundo de la izquierda recuerdan las condiciones y errores de clasificación de los escenarios clínicos y génicos en sus respectivos óptimos de *PLS-VIP*. El tercer panel, situado a la derecha, recoge en sus primeras dos columnas un recordatorio semejante para los escenarios clínicos + génicos, mientras que las 4 columnas restantes muestran los resultados de aplicar la prueba U de Mann Whitney. Estas cuatro columnas del extremo de la derecha son ahora las más importantes, las tres primeras porque nos proporcionan los p-valores de las comparaciones de interés: CLINICAS *vs.* GENICAS, : CLINICAS + GENICAS *vs.* CLINICAS y CLINICAS + GENICAS *vs.* GENICAS; por su parte la última columna recoge el interesante aspecto de si la adición de variables correspondiente tendría o no aplicación práctica.

Observando detenidamente las 3 columnas con los errores de clasificación de cada tipo de variables (en verde) y la última columna de la derecha relativa a la aplicación práctica, se podrían proponer las siguientes conclusiones:

1) La condición necesaria para que la adición de variables pueda tener interés práctico es que el escenario tenga un p-valor < 0.05 en las comparaciones CLI+GEN *vs.* CLI y CLI+GEN *vs.* GEN, ya que de lo contrario bastaría sólo con un tipo de variables (las clínicas o las génicas). Sin embargo, no es necesario que el p-valor de CLI vs. GEN sea significativo. Este es el caso de CLIGEN4-3 donde el error fue de 0.14 en las clínicas, 0.12 en las génicas y bajó hasta 0.06 al unir las clínicas con las génicas. Los respectivos p-valores de CLIN vs. GENIC, CLIN+GENIC vs. CLIN y CLIN+GENIC *vs.* GENIC fueron 0.44, < 0.001 y < 0.001. Un escenario en el que también ocurre una mejora al combinar las variables es CLIGEN8-7.

2) La anterior es condición necesaria, pero no suficiente, ya que puede ocurrir lo que se observa en el escenario CLIGEN2-1, donde los 3 p-valores de las tres comparaciones entre escenarios son < 0.001, pero sin embargo este escenario no tendría interés práctico ya que los errores de las clínicas, génicas y clínicas + génicas valen 0.10, 0.25 y 0.20, respectivamente, por lo que bastaría con medir las clínicas solas, que tienen más potencia, y ahorrar la medición de las génicas. Una situación semejante se da en CLIGEN6-5.

3) Alternativamente, puede ocurrir que sean las génicas las que presente mayor potencia frente a las clínicas, en cuyo caso tampoco habría un p-valor significativo para la adición de ambas variables frente a las génicas sólas, y bastaría con medir simplemente las variables génicas. Tal es el caso de varios escenarios como: CLIGEN3-3, CLIGEN3-4, CLIGEN7-7 o CLIGEN7-8, entre otros.

4) En resumen, para que la adición de variables clínicas y génicas pueda resultar de interés práctico, ninguna de las dos debe predominar en potencia respecto a la otra, sino que deben tener errores de clasificación moderados y parecidos, con el fin de que su unión pueda resultar en una disminución práctica del error de clasificación. Es lógico que, si unas de las variables tiene un error alto, digamos las clínicas con error

0.25, y la otra un error muy bajo, digamos las génicas con error 0, su adición no tiene ya interés, ya que PLS-VIP eliminará las variables de baja potencia y se mantendrá en el error 0 ya observado con las de alta potencia (ver por ejemplo CLIGEN5-6 y CLIGEN7-8).

Table 11. Proporciones de error de clasificación en diferentes escenarios y su comparación con la prueba U de Mann-Whitney

Escenario clínico inicial (CLI)	Casos	Error clasif. en óptimo de PLS-VIP	Escenario génico inicial (GEN)	Casos	Error clasif. en óptimo de PLS-VIP	Escenario CLI + GEN fusionando sus var. óptimas	Error clasif. en óptimo de PLS-VIP	p-valor CLI vs. GEN	p-valor CLI + GEN vs. CLI	p-valor CLI + GEN vs. GEN	Aplicación práctica
CLIN1 (Dicot. = 2 , RR=3) (Cont. = 3, TE=0.4)	10C 10 T	**0.30 ***	***GENIC1*** (20 dif. exp. (Δ=0.4), 980 ruido)	10C 10 T	**0.25**	***CLIGEN1-1 (CLIN1 + GENIC1)***	**0.25 (0.20 , 0.30)**	< 0.01	< 0.001	0.31	**NO (GEN solas)**
			GENIC2 (20 dif. exp. (Δ=0.6), 980 ruido)	10C 10 T	**0.05**	***CLIGEN1-2 (CLIN1 + GENIC2)***	**0.05 (0.03 , 0.10)**	< 0.001	< 0.001	0.54	**NO (GEN solas)**
CLIN2 (Dicot. = 2 , RR=6) (Cont. = 3, TE=0.8)	10C 10 T	**0.10**	***GENIC1*** (20 dif. exp. (Δ=0.4), 980 ruido)	10C 10 T	**0.25**	***CLIGEN2-1 (CLIN2 + GENIC1)***	**0.20 (0.13 , 0.25)**	< 0.001	< 0.001	< 0.001	**NO (CLIN solas)**
			GENIC2 (20 dif. exp.(Δ=0.6), 980 ruido)	10C 10 T	**0.05**	***CLIGEN2-2 (CLIN2 + GENIC2)***	**0.05 (0 , 0.10)**	0.02	< 0.001	0.054	**SI (CLIN + GEN)**
CLIN3 (Dicot. = 2 , RR=3 Cont. = 3, TE=0.4)	25C 25 T	**0.32**	***GENIC3*** (20 dif. exp. (Δ=0.4), 980 ruido)	25C 25 T	**0.12**	***CLIGEN3-3 (CLIN3 + GENIC3)***	**0.12 (0.08 , 0.14)**	< 0.001	< 0.001	0.15	**NO (GEN solas)**
			GENIC4 (20 dif. exp. (Δ=0.6), 980 ruido)	25C 25 T	**0**	***CLIGEN3-4 (CLIN3 + GENIC4)***	**0 (0 , 0.02)**	< 0.001	< 0.001	0.30	**NO (GEN solas)**
CLIN4 (Dicot. = 2 , RR=6) (Cont. = 3, TE=0.8)	25C 25 T	**0.14**	***GENIC3*** (20 dif. exp. (Δ=0.4), 980 ruido)	25C 25 T	**0.12**	***CLIGEN4-3 (CLIN4 + GENIC3)***	**0.06 (0.04 , 0.10)**	0.44	< 0.001	< 0.001	**SI (CLIN + GEN)**
			GENIC4 (20 dif. exp. (Δ=0.6), 980 ruido)	25C 25 T	**0**	***CLIGEN4-4 (CLIN4 + GENIC4)***	**0 (0 , 0.02)**	< 0.001	< 0.001	0.37	**NO (GEN solas)**
CLIN5 (Dicot. = 4 , RR=3) (Cont. = 6, TE=0.4)	10C 10 T	**0.25**	***GENIC5*** (40 dif. exp. (Δ=0.4), 960 ruido)	10C 10 T	**0.2**	***CLIGEN5-5 (CLIN5+ GENIC5)***	**0.15 (0.10 , 0.23)**	< 0.001	< 0.001	0.56	**NO (GEN solas)**
			GENIC6 (40 dif. exp. (Δ=0.6), 960 ruido)	10C 10 T	**0**	***CLIGEN5-6 (CLIN5 + GENIC6)***	**0 (0 , 0)**	< 0.001	< 0.001	0.66	**NO (GEN solas)**
CLIN6 (Dicot. = 4 , RR=6) (Cont. = 6, TE=0.8)	10C 10 T	**0.05**	***GENIC5*** (40 dif. exp. (Δ=0.4), 960 ruido)	10C 10 T	**0.2**	***CLIGEN6-5 (CLIN6 + GENIC5)***	**0.10 (0.05 , 0.15)**	< 0.001	< 0.001	< 0.001	**NO (CLIN solas)**
			GENIC6 (40 dif. exp. (Δ=0.6), 960 ruido)	10C 10 T	**0**	***CLIGEN6-6 (CLIN6 + GENIC6)***	**0 (0 , 0)**	< 0.001	< 0.001	0.07	**NO (GEN solas)**
CLIN7 (Dicot. = 4 , RR=3) (Cont. = 6, TE=0.4)	25C 25 T	**0.26**	***GENIC7*** (40 dif. exp. (Δ=0.4), 960 ruido)	25C 25 T	**0.02**	***CLIGEN7-7 (CLIN7+ GENIC7)***	**0.04 (0.02 , 0.04)**	< 0.001	< 0.001	0.99	**NO (GEN solas)**
			GENIC8 (40 dif. exp. (Δ=0.6), 960 ruido)	25C 25 T	**0**	***CLIGEN7-8 (CLIN7+ GENIC8)***	**0 (0 , 0)**	< 0.001	< 0.001	0.57	**NO (GEN solas)**
CLIN8 (Dicot. = 4 , RR=6) (Cont. = 6, TE=0.8)	25C 25 T	**0.06**	***GENIC7*** (40 dif. exp.(Δ=0.4), 960 ruido)	25C 25 T	**0.02**	***CLIGEN8-7 (CLIN8+ GENIC7)***	**0 (0 , 0.02)**	< 0.001	< 0.001	< 0.001	**SI (CLIN + GEN)**
			GENIC8 (40 dif. exp. (Δ=0.6), 960 ruido)	25C 25 T	**0**	***CLIGEN8-8 (CLIN8 + GENIC8)***	**0 (0 , 0)**	< 0.001	< 0.001	0.99	**NO (GEN solas)**

* Los valores son mediana o mediana (primer cuartil , tercer quartil) de 50 repeticiones. Series "training" y "test" tienen mismo nº controles (C) y tumores (T).

Representaciones PLS de una serie con variables clínicas y génicas

Como se hizo en apartados anteriores con las variables clínicas y génicas, a modo de ejemplo, se muestran a continuación algunas gráficas de interés de la misma serie 17 pero ahora adicionando ambas variables en sus óptimos PLS-VIP (escenario CLIGEN4-3). Concretamente, se trata de unas condiciones de partida que tienen 25 muestras de control y 25 de tumor, con 2 variables dicotómicas (RR = 6), 3 variables continuas (TE = 0.8) con 7 genes dif. exp. con $\Delta = 0.4$ y 4 genes ruido, es decir 16 variables en total.

Al aplicar PLS-VIP, se obtuvo el óptimo en nº de iteraciones = 1, nº de factores = 3, manteniéndose todas las variables de partida antes citadas, es decir 16, ya que el nº de iteraciones = 1. La varianza acumulada para las $\boldsymbol{X}$ con los 3 factores y la serie de entrenamiento fue de 39% (24% + 8% + 7%), que resulta baja, lo que parece significar que, aunque ahora el nº de variables es grande (16), su bajo poder discriminante y la presencia de variables ruido conlleva una extracción de varianza escasa. La varianza acumulada para las *Y* fue, sin embargo mayor, siendo del 86.8% (81% + 4.4% + 1.4%), con una escasa participación del factor 2 y el 3 frente al 1. Utilizando la correspondiente serie de prueba obtenida también por adición de las mismas variables, el error de clasificación fue ahora menor, de 0.06, en comparación con el 0.18 de las variables clínicas solas y el 0.20 de las génicas solas. El pequeño error de 0.06 presenta solamente 1 falso positivo y 2 falsos negativos de los 50 casos a predecir que incluye la serie de prueba. Este mejor comportamiento del modelo clínico-genómico confirma que las condiciones idóneas para *PLS-VIP-Consecutivo*, en su estrategia de adicionar variables clínicas y génicas, sería que ambas variables presentaran errores de clasificación moderados y parecidos con ambas variables. Por último conviene señalar que aquí se

necesitaron 3 factores en el óptimo, a diferencia de los 2 factores utilizados para esta serie en las variables clínicas y génicas por separado.

En cuanto a visualizar estos resultados, nos vamos a limitar, por brevedad, a las gráficas de ***puntuaciones*** y ***cargas*** de las ***X***, ya que las ***puntuaciones*** de ***Y*** muestran una representación análoga a la ya incluida en la Figura 34, y en cuanto a las gráficas de $\boldsymbol{u_1}$ vs. $\boldsymbol{t_1}$, $\boldsymbol{u_2}$ vs. $\boldsymbol{t_2}$ y $\boldsymbol{u_3}$ vs. $\boldsymbol{t_3}$ (ahora hay 3 factores), también son análogas a las representadas en apartados anteriores (Figuras 35 y 37), sólo que ahora las correlaciones para los 3 factores son mejores, con unos p-valores de < 0.0001, < 0.001 y 0.03, respectivamente, lo que indica una mejor bondad del modelo clínico-genómico frente al de las variables por separado.

En la Figura 38 se representan las ***puntuaciones-X*** con los datos de la serie de entrenamiento y los dos primeros factores . En las puntuaciones puede apreciarse ahora una mejor separación entre las muestras control y las muestras tumor, que cuando se usaron las variables por separado, no observándose casos entremezclados.

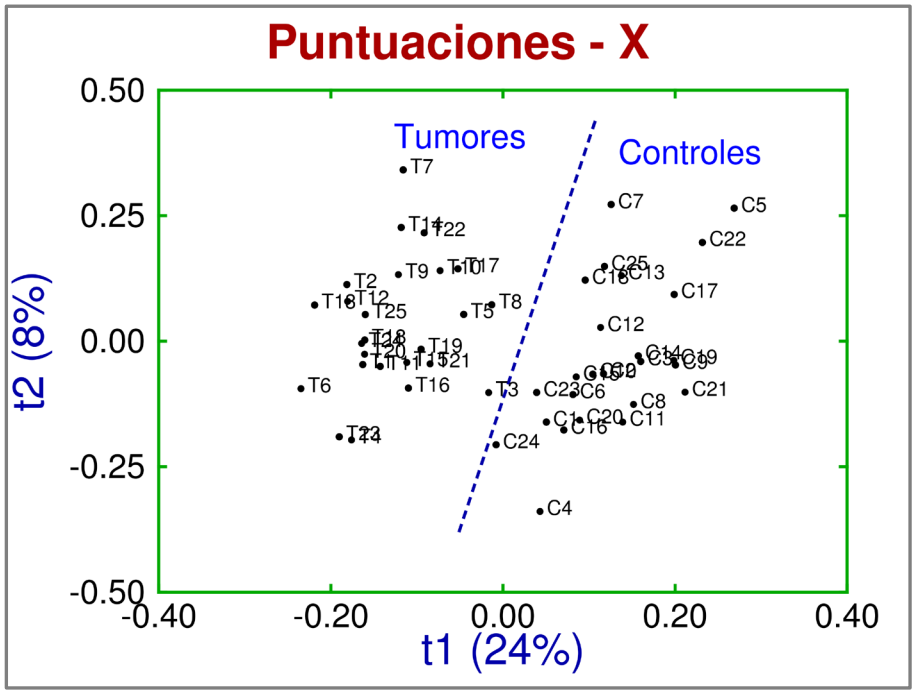

Figura 38. Puntuaciones de las variables X para la serie 17 de CLIGEN4-3.

Las ***cargas-X*** se han representado en la Figura 39. Allí se observa que con los 2 primeros factores PLS se separan aceptablemente todas las variables, no situándose ninguna en el centro de coordenadas. Las variables clínicas no se agrupan del todo, mientras que las variables génicas están bien separadas, principalmente por el factor 1, que las distribuye en genes infra-expresados (D en la gráfica (del inglés "Down")) de los sobre-expresados (U de "Up" en la gráfica). Los genes ruido ("Noise" en la gráfica) se adaptan también a esta separación de genes ruido sobre expresados (129, 703 y 394) y el gen ruido infra-expresado (385), como se comprobó en la matriz de datos comparando las medias de los controles frente a los tumores.

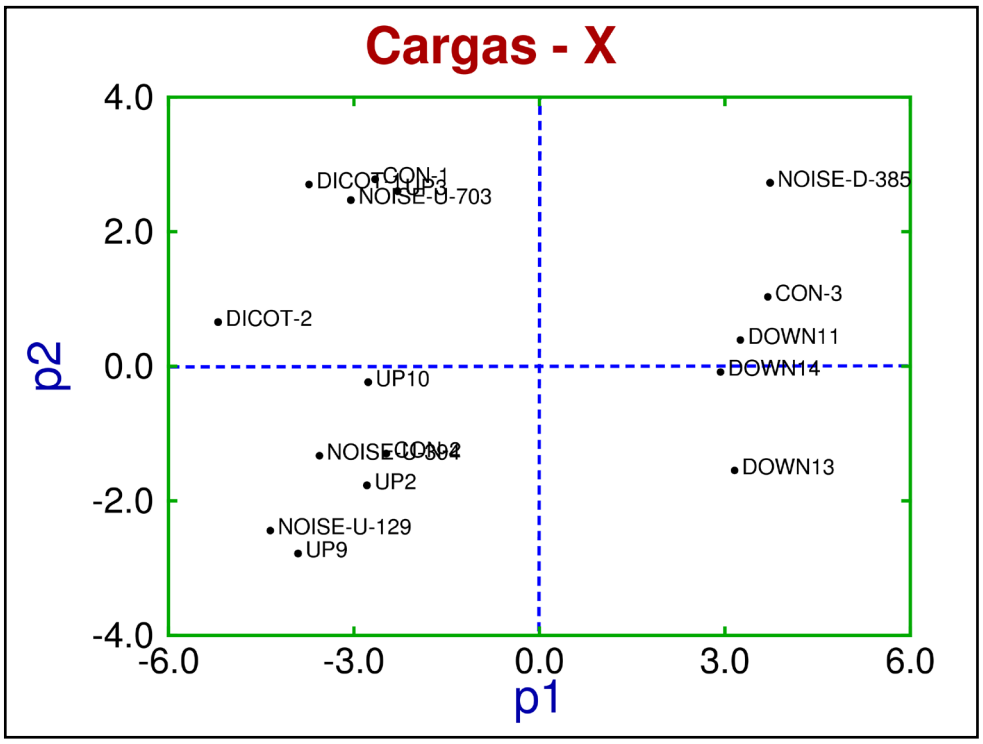

Fig. 39. Cargas de las variables X para la serie 17 de CLIGEN4-3. Las etiquetas de los puntos significan: DICOT = dicotómica, CON = continua, UP = gen sobre expresado, DOWN = gen infra-expresado, NOISE-U = gen ruido sobre-expresado, NOISE-D = gen ruido infra-expresado.

4.3.4 Estudio comparativo de *PLS-VIP-Consecutivo* frente a otros métodos predictivos usando variables clínicas más génicas

El procedimiento propuesto en este trabajo para analizar variables clínicas con génicas consiste, como ya se ha comentado más arriba, en aplicar consecutivamente *PLS-VIP*, es decir primero a las variables clínicas, luego a las variables génicas, y finalmente adicionar las variables clínicas y génicas en sus óptimos y aplicar de nuevo *PLS-VIP* a esos datos.

Llegados a este punto, lo interesante sería comparar el método *PLS-VIP-Consecutivo*, propuesto en este trabajo, con otros métodos predictivos habituales en Genómica, como PAM, KNN, SVM y RF. Para ello se eligió el escenario CLIGEN4-3, por tratarse de una situación de potencia intermedia con la que poder observar las posibles diferencias entre los distintos métodos. Como ya nos es familiar, este escenario consta de 50 series de entrenamiento junto con sus 50 series "hermanas" de prueba. En los métodos PAM, KNN, SVM y RF los modelos óptimos se construyeron solamente con las series de entrenamiento pero los errores de clasificación se calcularon a con las series de prueba, mientras que en el caso de *PLS-VIP-Consecutivo*, tanto los modelos óptimos como los errores de clasificación se obtuvieron con las series de prueba. Las condiciones de todas las series simuladas fueron: 25 controles y 25 tumores, siendo 2 las variables dicotómicas, 3 las variables continuas y 1000 las variables génicas (20 genes dif. exp. y 980 genes ruido). Los resultados obtenidos aparecen recogidos en la Tabla 12.

Las tres primeras columnas de la Tabla 12 muestran los métodos analizados y los casos y variables de partida. La cuarta, más importante, cita el procedimiento de selección de variables seguido en cada método con las 1005 variables de partida, que en esencia han sido tres: a) "Correlation Feature Selection (CFS)" con validación por "Leave One Out

(LOO)" para los métodos KNN, SVM y RF, b) "Cross validation" con 3 particiones (*CV* (*3 fold*)" en el caso de PAM y c) estadístico VIP para *PLS-VIP-Consecutivo*. La columna quinta recoge las variables finales seleccionadas por cada método que son las que entrarán propiamente en el proceso de predicción con la serie de prueba, y donde los valores están expresados como mediana (primer cuartil , tercer cuartil), abreviados como M(Q1,Q3). La columna sexta presenta la proporción de genes falsos positivos (PGFP), expresados también como M(Q1,Q3), que se encuentran en las variables finales seleccionadas por cada método, pudiéndose apreciar que dicha proporción es muy semejante en todos los métodos (en torno a 0.6). Nótese que las variables clínicas se han simulado sin variables ruido y no pueden tener falsos positivos. Por último se encuentra la columna siete, la más importante para nuestro objetivo de comparación de los métodos, pues recoge la proporción de sus respectivos errores de clasificación. Como puede apreciarse, el menor error, citado como M(Q1,Q3), lo ha conseguido el método SVM con un valor de 0 (0,0.04), seguido de cerca por RF con 0.04 (0.02-0.08) y *PLS-VIP-Consecutivo* con 0.06 (0.04,0.10), y ya con errores más grandes aparecen KNN (0.16 (0.11,0.21) y PAM (0.22 (0.15,0.30)).

En resumen, se puede concluir que, en esta ocasión, SVM presenta el mejor comportamiento de los métodos probados, pero *PLS-VIP* se le acerca apreciablemente y presenta una menor proporción de genes falsos positivos (0.53 frente a 0.67 de SVM), a la vez que dispone de numerosas posibilidades gráficas, como ha quedado ya expuesto en otros apartados. En base a estos resultados, se puede considerar que el método *PLS-VIP consecutivo* sería una alternativa interesante a los procedimientos de predicción habituales, con la ventaja de que al analizar las variables clínicas por un lado, las génicas por otro y luego la combinación de ambas en sus óptimos, puede proporcionar al investigador algún tipo de información valiosa adicional de ambos tipos de variables.

Table 12. Comparación de métodos habituales de clasificación frente al algoritmo PLS-VIP-Consecutivo promediando 50 repeticiones de datos simulados según el escenario CLIGEN4-3. Los modelos óptimos en todos los métodos (excepto en el algoritmo PLS), se construyeron solamente con la serie de entrenamiento pero los errores de clasificación se calcularon con la serie de prueba en todos los casos. Con PLS, tanto la dimensionalidad óptima del modelo como el error de predicción fueron calculados usando en cada paso del algoritmo la serie de entrenamiento seguida de la serie de prueba

Método	Casos	Variables de partida	Selección previa de variables	Variables finales entran al método para predicción	Proporción genes falsos positivos (PGFP)	Proporción errores de clasificación (PEC)
KNN	25 C y 25 T	Clínicas = 5 (2 dicotómicas y 3 continuas) Génicas = 1000 (20 Dif. Exp. y 980 ruido)	CFS, LOO	Clínicas Dicot. = 2 (1 , 2) Clínicas Cont. = 1 (0 , 1) Génicas = 13 (10 , 16)	0.67 (0.56 , 0.71)	0.16 (0.11 , 0.21)
PAM			CV (3 fold)	Clínicas Dicot. = 2 (1 , 2) Clínicas Cont. = 3 (3 , 3) Génicas = 26 (1 , 73)	0.64 (0.14 , 0.87)	0.22 (0.15 , 0.30)
SVM			CFS, LOO	Clínicas Dicot. = 2 (1 , 2) Clínicas Cont. = 1 (0 , 1) Génicas = 13 (10, 16)	0.67 (0.56 , 0.71)	0 (0 , 0.04)
RF			CFS, LOO	Clínicas Dicot. = 2 (1 , 2) Clínicas Cont. = 1 (0 , 1) Génicas = 13 (10 , 16)	0.67 (0.56 , 0.71)	0.04 (0.02 , 0.08)
PLS-VIP-Consecutivo*			VIP	Clínicas Dicot. = 2 (2 , 2) Clínicas Cont. = 3 (2 , 3) Génicas = 32 (15 , 47)	0.53 (0.36 , 0.64)	0.06 (0.04 , 0.10)

* Consecutivo significa que PLS-VIP actua primero sobre las variables clínicas, luego sobre las génicas y finalmente se fusionan las variables óptimas y aplica de nuevo PLS-VIP. ** Los valores se refieren a la mediana (primer cuartil - tercer cuartil) a partir de 50 simulaciones. CFS = "Correlation feature selection". LOO = "Leave One Out". CV (3 fold) = "Cross validation con 3 particiones. VIP = "Variable Influence on Projection".

4.4. Comportamiento de *PLS-VIP-Consecutivo* con datos reales clínicos y génicos combinados en un predictor

Conviene empezar diciendo que no es fácil recopilar pacientes de los que se disponga a la vez de datos clínicos y génicos de microarrays. Es práctica común en medicina registrar variables clínicas de los pacientes, pero es poco frecuente hacerles un estudio de expresión génica, salvo en contextos de investigación, por lo que en la práctica el tamaño de muestra de que se dispone con ambos tipos de variables es normalmente escaso. Una opción aceptable ha sido la de una muestra de pacientes de Mieloma Múltiple facilitados por el servicio de Hematología del Hospital Universitario de Salamanca. Los resultados obtenidos con *PLS-VIP* se compararán con otros métodos de clasificación habituales en Genómica (KNN, PAM, SVM y RF). Con todos los métodos se utilizó la misma estrategia: primero se analizaban las variables clínicas sólas, luego las génicas sólas y por último la combinación de ambas en una matriz de variables clínicas + génicas.

4.4.1. *Pacientes con Mieloma Múltiple tratados con 6 ciclos de quimioterapia categorizados como Respuesta Incompleta (RI) frente a Respuesta Completa (RC).*

Se partió de 2 series de datos de mieloma (Gutierrez et al. (2010), Lopez-Corral et al., (2014)), que se encuentra depositados en GEO con los números GSE16558 y GSE47552, respectivamente, cuyos archivos se descargan y se procesan. Se tienen así 43 muestras de *microarrays* medidos al diagnóstico que fueron normalizadas con el algoritmo RMA de la consola de Affymetrix. Para corregir el efecto de diferentes "batchs" se utilizó el paquete "COMBAT" (http://www.bu.edu/jlab/wpassets/ComBat/Abstract.html). Seguidamente se hizo un filtrado para eliminar los controles de Affymetrix, las sondas que no tenían gen asociado (en la versión NA33 de anotaciones de

Affymetrix) y también aquellas sondas que presentaban en todas las muestras una expresión menor de 6.7 en log(2) por considerarlas de expresión despreciable. Al final se pudo disponer de 8728 sondas génicas para 35 pacientes que pasaron a ser nuestras variables génicas.

En cuanto a las variables clínicas se disponía de varias de ellas recogidas también al diagnóstico, al igual que los datos de microarrays. Se seleccionaron las siguientes:

Variables clínicas categóricas:

- Tipo de inmunoglobulina de cadena pesada
- Tipo de inmunoglobulina de cadena ligera
- Categoría en la escala de Dury Salmon
- Grado de lesiones óseas

Para tratar estas variables se recurrió al uso de variables postizas (*dummys*) en la forma convencional de asignar ceros y unos, variables postizas que son las que entraron propiamente a los métodos de clasificación.

Variables clínicas de tipo continuo:

Se refieren a las concentraciones o proporciones de diferentes marcadores. Se han incluido los valores promedio de la respuesta, RI y RC, en la notación M(Q1,Q3):

- Componente monoclonal en suero [RI: 4(1 , 6) ; RC: 1.5(0 , 2)]
- Componente monoclonal en orina [RI: 0.26(0 , 2.7) ; RC: 1.1(0 , 4.2)]
- β_2 microglobulina [RI: 3.9(3.2 , 7.9) ; RC: 2.3(1.3 , 4.1)]
- Hemoglobina [RI: 9.8(8.6 , 10.4) ; RC: 11.2(9.7 , 12.3)]
- Leucocitos [RI: 5.4(4.3 , 8.9) ; RC: 7.1(5.7 , 8.5)]
- Plaquetas [RI: 222(181 , 288) ; RC: 213(189 , 271)]
- Creatinina [RI: 1.20(0.91 , 2.78) ; RC: 1.12(0.86 , 1.65)]
- Calcio [RI: 9.5(8.9 , 10.6) ; RC: 9.6(9.2 , 10)]
- Albúmina [RI: 3.4(3.1 , 4.0) ; RC: 3.7(3.4 , 3.9)]
- % de plasmacitosis celular [RI: 36(24 , 56) ; RC: 48(34 , 79)]

Todas las variables clínicas entraron en los diferentes métodos de clasificación con sus valores originales, es decir sin transformación alguna, como podría ser la estandarización ($\bar{x} = 0, s = 1$). La razón es que PLS realiza dicha transformación en el propio algoritmo y en los otros métodos la estadarización es un tema controvertido que excede el analizarlo en esta investigación. Se disponía de todas las variables arriba mencionadas para los 35 pacientes de los que se tenían también sus datos de *microarrays*. Los pacientes se categorizaron en dos clases: respuesta incompleta (RI) y respuesta completa (RC). Finalmente los pacientes se distribuyeron al azar mediante permutaciones aleatorias en dos series, una de entrenamiento compuesta por 12 RI y 6 RC y otra de prueba formada por 11 RI y 6 RC.

Resultados usando sólo variables clínicas

El comportamiento observado con los distintos métodos de predicción se muestra en la Tabla 13. Allí puede verse que hay 3 métodos (KNN, SVM y RF) que eliminan todas las variables menos una, de modo que no se ha considerado incluir los resultados que se obtienen por considerarlos poco fiables al tener sólo una variable (se denotan como "NA"). Los dos métodos restantes son PAM y PLS-VIP. El método PAM selecciona 5 variables continuas mientras que *PLS-VIP* selecciona 9 variables (3 categóricas y 6 continuas), sin embargo ambos métodos hacen predicciones con un mismo error de mal clasificados del 29%. La única ventaja a favor de PLS-VIP es que los mal clasificados están más balanceados, mientras que en PAM están sesgados a RC. En resumen, se aprecian dos características: Una, que las variables clínicas no parecen ser bien aceptadas por algunos métodos como KNN, SVM y RF, tal vez por la presencia de variables categóricas formateadas en forma de variables postizas de ceros y unos; y la otra, que los

métodos PAM y *PLS-VIP*, aunque parecen funcionar correctamente, presentan valores moderados de predicción con errores del 29%, tal vez porque la respuesta a los 6 ciclos de quimioterapia pudiera estar poco relacionada con unas variables clínicas al diagnóstico.

Tabla 13. Comparación de métodos de clasificación usando datos de Mieloma con sólo variables clínicas

Clases	Método clasificación	Variables de partida	Previa selección de variables	Variables optimas usadas	Parámetros óptimos	Proporción error de clasificación	Fallos en RI	Fallos en RC
Multiple Mieloma (Respuesta incompleta después de 6 ciclos de quimioterapia (RI) ***frente a*** Multiple Mieloma (Respuesta completa después de 6 ciclos de quimioterapia (RC)[a]	KNN	Clínicas = 14 (4 categóricas y 10 continuas)	CFS, LOO	Clínicas = 1 (categórica)	KNN = 17	NA	NA	NA
	PAM		CV (3 fold)	Clínicas = 5 (continuas).	Threshold = 1.3123	0.29	0/11	5/6
	SVM		CFS, LOO	Clínicas = 1 (categórica)	Cost = 0.8	NA	NA	NA
	RF				Trees = 35	NA	NA	NA
	PLS-VIP		No	Clínicas = 9 (3 categóricas y 6 continuas)	ITER = 3 NUMFACT = 1	0.29	2/11	3/6

[a] Serie entrenamiento (12 RI, 6 RC), serie prueba (11 RI, 6 RC). CFS = "correlation based feature selection". LOO = "CV with leave one out". ITER es Nº de iteraciones y NUMFACT es Nº de factores PLS. NA = no diponible actualmente debido a que el método sólo selecciona 1 variable y el resultado se ha considerado poco fiable.

Resultados usando solo variables génicas

Al comparar los diferentes métodos de clasificación con los datos génicos de microarrays se obtuvieron los resultados que se recogen en la Tabla 14. Ahora todos los métodos han funcionado correctamente y han seleccionado un número de variables coherente. KNN y PAM presentaron los errores de clasificación más altos frente a la serie de prueba (35% en ambos casos), con la diferencia de que los mal clasificados están sesgados a RC en KNN y bien balanceados en PAM. Los algoritmos de RF y *PLS-VIP* dieron ambos errores de clasificación del 29%, estando los fallos de clasificación ligeramente sesgados a RC en el método RF.

El método de SVM es el que mejor se ha comportado con la actual estructura de datos reales, obteniendo un porcentaje de mal clasificados del 12%, si bien los mal clasificados están sesgados hacia el lado de los pacientes con RC. Cabe concluir que, con estos datos, SVM es el método con mejor potencia predictiva de los aquí analizados. Este hecho contrasta con el observado en la Tabla 7, cuando se analizaron los datos reales de mielomas "sin" y "con" ganancia 1q, donde *PLS-VIP* presentó un error de clasificación del 7% frente al de 36% de t-test+SVM, si bien es posible que, con el método CFS+SVM utilizado en la presente comparación, se hubiera también alcanzado un resultado semejante al de *PLS-VIP*, sin embargo este punto no pudo ser comprobado porque el algoritmo utilizado no pudo procesar el número tan elevado de genes que tenía la serie de la Tabla 7 (9789). Por otra parte, si se observa de nuevo la Tabla 6, con 50 series de datos simulados, se aprecia que los errores de clasificación para SVM, *PLS-VIP* y RF fueron de 0%, 2% y 14%, respectivamente, situándose *PLS-VIP* muy cerca de SVM.

En conclusión, la clasificación con datos reales parece depender en parte de la estructura de los datos, si bien SVM, *PLS-VIP* y tal vez RF resultan las opciones más prometedoras de las ensayadas, aunque harían falta más estudios con datos reales y con diferentes estructuras de datos para confirmar esta interpretación.

Tabla 14. Comparación de métodos de clasificación usando datos de Mieloma con sólo variables génicas

Clases	Método clasificación	Variables génicas de partida	Previa selección de variables	Variables optimas usadas	Parámetros óptimos	Proporción error de clasificación	Fallos en RI	Fallos en RC
Multiple Mieloma (Respuesta incompleta después de 6 ciclos de quimioterapia (RI) ***frente a*** Multiple Mieloma (Respuesta completa después de 6 ciclos de quimioterapia (RC)[a]	**KNN**	**8728**	**CFS, LOO**	**18**	**KNN = 15**	**0.35**	**0/11**	**6/6**
	PAM		**CV (3 fold)**	**343**	**Threshold = 1.0829**	**0.35**	**3/11**	**3/6**
	SVM		**CFS, LOO**	**18**	**Cost = 0.6**	**0.12**	**0/11**	**2/6**
	RF		**CFS, LOO**	**18**	**Trees = 65**	**0.29**	**1/11**	**4/6**
	PLS-VIP		**No**	**62**	**ITER = 6 NUMFACT = 3**	**0.29**	**2/11**	**3/6**

[a] Serie entrenamiento (12 RI, 6 RC), serie prueba (11 RI, 6 RC). CFS significa "correlation based feature selection". LOO = "CV with leave one out". ITER es Nº de iteraciones y NUMFACT es Nº de factores PLS.

Resultados usando variables clínicas + génicas

Por último, se procedió a comparar los distintos métodos de clasificación, combinando en una sola matriz los datos clínicos y génicos, que ya se han analizado por separado en los apartados anteriores. El objetivo era ver si la combinación de ambos tipos de variables producía una mejora en el error de clasificación respecto a las variables por separado.

En la Tabla 15 se recogen los resultados de este estudio. Allí puede apreciarse que todos los métodos han funcionado adecuadamente, observándose también que KNN, RF y *PLS-VIP-consecutivo* presentan el mismo error de clasificación del 35%, PAM tiene un error del 41% y de nuevo SVM se revela como el mejor método de predicción con un 12% de error. Un hecho destacable es que los tres métodos que utilizan CFS para la selección de variables, el algoritmo ha excluido todas las variables clínicas y seleccionado sólo 18 variables génicas, mientras que *PLS-VIP-Consecutivo* selecciona algunas variables clínicas (9 de 14), además de las génicas (62 de 8728), al igual que hace PAM (selecciona 7 clínicas de 14 y 450 génicas de 8728). Esta propiedad de *PLS-VIP-consecutivo*, de retener alguna variable clínica, coincide con lo observado en la Tabla 7 con algunos de los escenarios simulados.

Conviene hacer notar, por último, que *PLS-VIP-Consecutivo* es el método que tiene los fallos mejor balanceados entre RI y RC, mientras que los otros métodos se encuentran sesgados y fallan más con pacientes con RC.

Tabla 15. Comparación de métodos de clasificación usando datosde Mieloma con variables clínicas + génicas

Clases	Método clasificación	Variables génicas de partida	Previa selección de variables	Variables optimas usadas	Parámetros óptimos	Proporción error de clasificación	Fallos en RI	Fallos en RC
Multiple Mieloma (Respuesta incompleta después de 6 ciclos de quimioterapia (RI) ***frente a*** Multiple Mieloma (Respuesta completa después de 6 ciclos de quimioterapia (RC)[a]	**KNN**	**Clínicas = 14 génicas = 8728**	**CFS, LOO**	**18 (clínicas = 0, génicas = 18)**	**KNN = 15**	**0.35**	**0/11**	**6/6**
	PAM		**CV (3 fold)**	**457 (clínicas = 7, génicas = 450)**	**Threshold = 0.9924**	**0.41**	**3/11**	**4/6**
	SVM		**CFS, LOO**	**18 (clínicas = 0, génicas = 18)**	**Cost = 0.6**	**0.12**	**0/11**	**2/6**
	RF			**18 (clínicas = 0, génicas = 18)**	**Trees = 65**	**0.35**	**1/11**	**5/6**
	PLS-VIP (Consecutivo)		**No**	**71 (clínicas = 9, génicas = 62)**	**ITER = 1 NUMFACT =3**	**0.35**	**3/11**	**3/6**

[a] Serie entrenamiento (12 RI, 6 RC), serie prueba (11 RI, 6 RC). CFS = "correlation based feature selection". LOO = "CV with leave one out". ITER es Nº de iteraciones y NUMFACT es Nº de factores PLS.

Comparación de las tres estrategias: clínicas, génicas y clínicas + génicas

Con el fin de comparar más directamente las tres estrategias analizadas, se han recopilado en la tabla 16 los valores de errores de clasificación de todas ellas. La pregunta a la que hay que contestar sería esta: ¿El uso de una combinación de variables clínicas y génicas mejora la potencia predictiva de un método de clasificación? Con los datos reales de mieloma que se acaban de exponer la respuesta es claramente "no". Así, puede observarse en la Tabla 16, de forma que algunos métodos se han mantenido en su error de clasificación (KNN con un 35% y SVM con un 12%), y otros lo han empeorado al combinar las variables clínicas y génicas frente a ellas mismas por separado: PAM con 41% frente a 29% en clínicas y 35% en génicas, RF con 35% frente a 29% en génicas y *PLS-VIP-consecutivo* con 35% frente al 29% de clínicas y 29% de génicas. De nuevo, SVM es el que presenta el error más bajo (12%), pero sin aprovecharse de la información adicional de las variables clínicas para mejorar su resultado con las génicas sólas (también 12%).

¿Cómo explicar este comportamiento? Cabrían tres interpretaciones. Una sería el que los pacientes con mieloma tienen unas variables clínicas al diagnóstico que no son indicadoras después de un tratamiento agresivo como son 6 ciclos de quimioterapia. La segunda explicación sería que las variables clínicas en mieloma tengan un tamaño de efecto pequeño comparado con el de las variables génicas, por lo que al combinarlas parecen actuar de ruido y además acaban perdiéndose entre las numerosas variables génicas. Esta última explicación estaría de acuerdo con los resultados de simulación para los distintos escenarios de la Tabla 11, donde sólo en 3 casos de 16 se apreció una mejora estadísticamente significativa en la combinación de variables, y estos casos coincidieron con variables de tamaño de efecto alto, como son dicotómicas con un RR = 6 y continuas con un TE = 0.8. La tercera y última explicación sería que los métodos probados están diseñados para un elevado número de variables y con variables clínicas podrían funcionar mejor otros métodos como regresión logística binaria o análisis discriminante.

Tabla 16. Recopilación de los errores de clasificación con las 3 estrategias de tipos de variables usadas

		Clínicas solas		Génicas solas		Clínicas + Génicas	
Clases	**Método clasificación**	**Variables optimas usadas**	**Proporción error de clasificación**	**Variables optimas usadas**	**Proporción error de clasificación**	**Variables optimas usadas**	**Proporción error de clasificación**
Multiple Mieloma (Respuesta incompleta después de 6 ciclos de quimioterapia (RI) ***frente a*** Multiple Mieloma (Respuesta completa después de 6 ciclos de quimioterapia (RC)[a]	**KNN**	Clínicas = 1 (categórica)	**NA**	Génicas = 18	**0.35**	18 (clínicas = 0, génicas = 18)	**0.35**
	PAM	Clínicas = 5 (continuas).	**0.29**	Génicas =343	**0.35**	457 (clínicas = 7, génicas = 450)	**0.41**
	SVM	Clínicas = 1 (categórica)	**NA**	Génicas = 18	**0.12**	18 (clínicas = 0, génicas = 18)	**0.12**
	RF		**NA**	Génicas = 18	**0.29**	18 (clínicas = 0, génicas = 18)	**0.35**
	PLS-VIP (Consecutivo)	Clínicas = 9 (3 categóricas y 6 continuas)	**0.29**	Génicas = 62	**0.29**	71 (clínicas = 9, génicas = 62)	**0.35**

[a] Serie entrenamiento (12 RI, 6 RC), serie prueba (11 RI, 6 RC).

5. CONCLUSIONES

CONCLUSIONES

1) Se ha desarrollado un programa basado en Mínimos Cuadrados Parciales (*Partial Least Squares, PLS*), con el fin de adaptar las estrategias de esta técnica al análisis de datos de expresión génica en chips de ADN o *microarrays*. El algoritmo utiliza el estadístico VIP (*Variable Influence on Projection*) para seleccionar iterativamente los genes más discriminantes entre clases de entre los miles existentes en un chip, y encontrar, a su vez, el número óptimo de iteraciones y de factores latentes en los que se alcanza el mínimo error de clasificación. El código ha sido escrito en Fortran 95 y utiliza las rutinas de las "dlls" del Paquete Estadístico *SIMFIT*. Se le ha denominado *PLS-VIP* y está desarrollado para clasificación binaria, si bien se podría extender fácilmente a clasificación multicategórica. Se realizó también una variante de dicho programa, denominada *PLS-VIP-Consecutivo*, que aplica *PLS-VIP* a variables clínicas clásicas por un lado y a génicas de *microarrays* por otro, adiciona luego las variables seleccionadas y aplica de nuevo *PLS-VIP* a dicha combinación para crear un modelo mixto clínico-genómico.

2) Se ha escrito otro programa, llamado *SIMDATA*, que simula datos aleatorios a partir de diferentes distribuciones de probabilidad. Permite generar datos de expresión génica de microarrays usando las unidades habituales de log(2), así como variables clínicas categóricas con diferente riesgo relativo entre clases (RR) y variables continuas con distinto tamaño de efecto (TE). El programa simula simultáneamente, para clases binarias, una serie de datos de "entrenamiento" y otra serie "hermana" de "prueba", la primera para optimizar el modelo PLS y la segunda para validarlo, ya que esta validación

con serie independiente de prueba presenta menos sesgo que los métodos de validación cruzada. Se ha escrito también usando Fortran 95 y las "dlls" de *SIMFIT*.

3) Para analizar la potencia de *PLS-VIP*, se simularon 24 escenarios de microarrays con 50 repeticiones cada uno, en los que se variaba el tamaño de muestra, el número de genes sobre e infra expresados, el valor del desplazamiento de estos genes respecto al control (Δ), y el número de genes basales o de "ruido". El análisis de estos resultados puso de manifiesto que el algoritmo *PLS-VIP* construye un modelo clasificador con poco error de clasificación, tanto si el tamaño de muestra y el número de genes discriminatorios es escaso respecto a los genes ruido, como cuando se aumenta el tamaño de muestra y el número de genes discriminantes.

4) Se llevó a cabo un estudio comparativo, para un escenario concreto, con el fin de valorar el error de predicción de *PLS-VIP* frente a otros métodos de clasificación usuales en Genómica, como KNN, PAM, SVM y RF. Se encontró que *PLS-VIP* mejoraba a KNN, PAM y RF y era comparable a SVM con selección previa de genes por CFS. La ventaja de *PLS-VIP* frente a SVM reside en su simplicidad, ya que solamente usa un procedimiento matemático sobre la serie completa de datos, no necesita una selección previa de los genes y su ejecución en tiempo de ordenador es más rápida.

5) Un estudio comparativo semejante, realizado con datos reales de Mieloma Múltiple estratificados en dos categorías, demostró que *PLS-VIP* obtenía también mejores resultados que KNN, PAM y SVM.

6) Un estudio de simulación y análisis de variables clínicas, génicas y clínicas + génicas, ha llevado a las siguientes conclusiones:

a) Predictores con sólo variables clínicas

- El número de variables dicotómicas y continuas conservadas en el óptimo de *PLS-VIP* coincide normalmente con el número de variables de partida en todos los escenarios.
- Cuando se mantienen las condiciones discriminantes, el aumento del tamaño de muestra no ejerce, paradójicamente, gran influencia en la bondad de las clasificaciones.
- Las variables dicotómicas resultaron ser menos precisas y de mayor capacidad de confusión que las variables continuas.

b) Predictores con sólo variables génicas

- El tamaño de muestra tiene más influencia en las variables génicas que el que tenía antes en las clínicas, disminuyendo apreciablemente el error de clasificación cuando el tamaño de muestra aumenta.
- Las condiciones más favorables se dan cuando el número de muestras es en torno a 25 controles y 25 tumores y el número de genes discriminantes es ≥ 20 con un desplazamiento Δ en log(2) en torno a 0.6, y siendo el nº de genes ruido del orden de 1000.

c) Predictores con variables clínicas + génicas

- Para que la adición de variables clínicas y génicas pueda resultar de interés práctico, ninguna de las dos debe predominar en potencia sobre la otra, sino que deben presentar errores de clasificación moderados y semejantes, con el fin de que su unión pueda resultar en una disminución del error de clasificación.

7) Se realizó un estudio totalmente análogo al anterior con datos reales de Mieloma Multiple. Se estudiaron 35 pacientes a los que se habían dado 6 ciclos de quimioterapia y se había medido su repuesta en dos categorías, incompleta (RI) y completa (RC), y de los que se disponía de 14 variables clínicas (4 categóricas, 10 continuas) y de 8728 variables génicas. Al haber suficientes variables clínicas fue posible hacer también una comparación con otros algoritmos predictores, llegándose a las siguientes conclusiones:

a) Con sólo variables clínicas

- Algunos métodos como KNN, SVM y RF no se ejecutaban correctamente, tal vez por tratarse de un número escaso de variables para estos métodos. Por su parte PAM y *PLS-VIP*, aunque funcionan adecuadamente, presentan valores pobres de clasificación, posiblemente porque la respuesta a la quimioterapia en mieloma esté poco relacionada con las variables clínicas al diagnóstico utilizadas.

b) Con sólo variables génicas

- KNN y PAM presentaron los errores más altos, seguidos de RF y *PLS-VIP*, siendo SVM en esta ocasión el de menor error. La clasificación parece depender de la estructura de los datos y harían falta más estudios con datos reales para esclarecer estos comportamientos.

c) Con variables clínicas + génicas

- Al combinar los dos tipos de variables, se comprueba que algunos métodos no han mejorado su error de clasificación (KNN y SVM) y otros lo han empeorado frente a sus respectivas variables por separado (PAM, RF y *PLS-VIP-Consecutivo*).

- SVM fue el método que presentó en este caso el error más bajo de clasificación, pero sin aprovecharse de la información adicional de las variables clínicas, ya que no las incluyó en el modelo y alcanzó su óptimo con 18 variables génicas. En este sentido *PLS-VIP-Consecutivo*, aunque con mayor error de clasificación, resulta muy prometedor ya que en el óptimo había retenido 9 variables clínicas y 62 génicas. *PLS-VIP Consecutivo* presenta también la ventaja de que sus fallos de clasificación están más balanceados que en SVM. No obstante, las conclusiones con datos reales siempre deben ser consideradas con cautela, ya que suelen depender del tipo de comparación binaria tratada y de la severidad de la patología.

8) Se ha comprobado a lo largo del trabajo que, en general, las propiedades de PLS son excelentes para clasificación en Genómica, ya que permiten un número de variables mucho mayor que el de casos y las variables pueden estar correlacionadas como ocurre con las expresiones génicas. Además PLS presenta diferentes posibilidades de visualización de los resultados que no presentan otros métodos, como son las representaciones de las puntuaciones de los casos, las cargas de las variables o los gráficos t-u para los factores, como ha quedado patente en nuestro estudio.

Salamanca, Marzo de 2015

6. BIBLIOGRAFÍA

6. BIBLIOGRAFÍA

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). *DOI 10.1002/WICS.051/ WIREs Comp Stat/ www.wiley.com/wires/compstats*, 6.

Abdi, H. and Williams, L. J. (2013). Partial least squares methods: partial least squares correlation and partial least square regression. *Methods in molecular biology* **930**, 549-579.

Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 10101-10106.

Bardsley, W. G (2013). The SIMFIT Statistical Package. University of Manchester (U.K.). http://www.simfit.org.uk.

Barker, M. and Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics* **17**, 166-173.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple testing **57**, 289-300.

Boulesteix, A. L. (2004). PLS dimension reduction for classification with microarray data. *Statistical applications in genetics and molecular biology* **3**, Article33.

Boulesteix, A. L. and Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in bioinformatics* **8**, 32-44.

Boulesteix, A. L., Strobl, C., Augustin, T., and Daumer, M. (2008a). Evaluating microarray-based classifiers: an overview. *Cancer informatics* **6**, 77-97.

Boulesteix, A. L., Porzelius, C., and Daumer, M. (2008b). Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics* **24**, 1698-1706.

Boulesteix, A. L. and Hothorn, T. (2010). Testing the additional predictive value of high-dimensional molecular data. *BMC bioinformatics* **11**, 78-2105-11-78.

Boulesteix, A. L. and Sauerbrei, W. (2011). Added predictive value of high-throughput molecular data to clinical data and its validation. *Briefings in bioinformatics* **12**, 215-229.

Breiman, L. (1996). Bagging predictors. *Machine Learning* **24**.

Breiman, L. (2001). Random forests. *Machine Learning* **45**.

Broyl, A., Hose, D., Lokhorst, H., de Knegt, Y., Peeters, J., Jauch, A., Bertsch, U., Buijs, A., Stevens-Kroef, M., Beverloo, H. B., Vellenga, E., Zweegman, S., Kersten, M. J., van der Holt, B., el Jarari, L., Mulligan, G., Goldschmidt, H., van Duin, M., and Sonneveld, P. (2010). Gene expression profiling for molecular classification of multiple myeloma in newly diagnosed patients. *Blood* **116**, 2543-2553.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* **2**, 121-167.

Burguillo, F.J., Corchete, L.A., Martin, J., Barrera, I. and Bardsley, W.G. (2014). A Partial Least Squares Algortim for Microarray Data Analysis using the VIP Statistic for gene selection and Binary Classification. *Current Bioinformatics* **9**, 348-359.

Celis, J. E., Kruhoffer, M., Gromova, I., Frederiksen, C., Ostergaard, M., Thykjaer, T., Gromov, P., Yu, J., Palsdottir, H., Magnusson, N., and Orntoft, T. F. (2000). Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. *FEBS letters* **480**, 2-16.

Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. *Machine Learning* **20**, 273-297.

Cho, J. H., Lee, D., Park, J. H., Kim, K., and Lee, I. B. (2002). Optimal approach for classification of acute leukemia subtypes based on gene expression data. *Biotechnology progress* **18**, 847-854.

Chong, I. G. and Jun, C. H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems* **78**, 103-112.

Chun, H. and Keles, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society.Series B, Statistical methodology* **72**, 3-25.

Chung, D. and Keles, S. (2010). Sparse partial least squares classification for high dimensional data. *Statistical applications in genetics and molecular biology* **9**, Article17.

De Bin, R., Herold, T., and Boulesteix, A. L. (2014). Added predictive value of omics data: specific issues related to validation illustrated by two case studies. *BMC medical research methodology* **14**, 117.

De Bin, R., Sauerbrei, W., and Boulesteix, A. L. (2014). Investigating the prediction ability of survival models based on both clinical and omics data: two case studies. *Statistics in medicine* **33**, 5310-5329.

Diaz-Uriarte, R. (2005). Supervised methods with genomic data: a review and cautionary view. In *Data Analysis and visualisation in genomics and proteomics*, Azuaje, F. and Dopazo, J. (eds), 193-210. Chichester (England): John Wiley and Sons.

Diaz-Uriarte, R. and Alvarez de Andres, S. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics* **7**, 3.

Ding, S., Xu, Y., Hao, T., and Ma, P. (2014). Partial least squares based gene expression analysis in renal failure. *Diagnostic pathology* **9**, 137.

Doledec, S. and Chessel, D. (1994). Co-Inertia Analysis - an Alternative Method for Studying Species Environment Relationships. *Freshwater Biology* **31**, 277-294.

Dudoit, S. and Fridlyan, J. (2003). Classification in microarray experiments. In *Interdiciplinary Statistics:Statistical Analysis of Gene Expression Microarray Data*, Speed, T. (ed), 93-157. Boca Raton: Chapman and Hall/CRC.

Dudoit, S., Fridlyand, J., and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97**, 77-87.

Eckart, C. and Young, G. (1936). The Approximation of One Matrix by another of Lower Rank. *Psychometrika* **1**, 211-218.

Eden, P., Ritz, C., Rose, C., Ferno, M., and Peterson, C. (2004). "Good Old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. *European journal of cancer* **40**, 1837-1841.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14863-14868.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikström, C., and Wold, S. (2006). *Multi- and Megavariate Data Analysis. Part I: Basic Principles and Applications*. Umeà (Sweden): UMETRICS AB.

Esbensen, K. H. (2010). *Multivariate Data Analysis In Practice*. Oslo: CAMO software.

Esbensen, K. H. and Geladi, P. (2010). Principles of Proper Validation: use and abuse of re-sampling for validation. *Journal of Chemometrics* **24**, 168-187.

Everitt, B. S. (1993). *Cluster analysis*. London: Edward Arnold.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics (Oxford, England)* **16**, 906-914.

Gabriel, K. (1971). Biplot Graphic Display of Matrices with Application to Principal Component Analysis. *Biometrika* **58**, 453-&.

Galindo, M. P. (1986). Una Alternativa de representación Simultánea: HK-Biplot. *Qüestiió* **10**, 13-23.

Gauchi, J. P. and Chagnon, P. (2001). Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. *Chemometrics and Intelligent Laboratory Systems* **58**.

Gevaert, O., De Smet, F., Timmerman, D., Moreau, Y., and De Moor, B. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics* **22**, e184-90.

Gutierrez, N. C., Ocio, E. M., de Las Rivas, J., Maiso, P., Delgado, M., Ferminan, E., Arcos, M. J., Sanchez, M. L., Hernandez, J. M., and San Miguel, J. F. (2007). Gene expression profiling of B lymphocytes and plasma cells from Waldenstrom's macroglobulinemia: comparison with expression patterns of the same cell counterparts from chronic lymphocytic leukemia, multiple myeloma and normal individuals. *Leukemia* **21**, 541-549.

Gutierrez, N. C., Sarasquete, M. E., Misiewicz-Krzeminska, I., Delgado, M., De Las Rivas, J., Ticona, F. V., Ferminan, E., Martin-Jimenez, P., Chillon, C., Risueno, A., Hernandez, J. M., Garcia-Sanz, R., Gonzalez, M., and San Miguel, J. F. (2010). Deregulation of microRNA expression in the different genetic subtypes of multiple myeloma and correlation with gene expression profiling. *Leukemia* **24**, 629-637.

Hall, M.A. (1999).Correlation based Feature Selection for Machine Learning. Hamilton (New Zeland): The University of Waikato.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer-Verlag.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* **28**, 321-377.

Huang, X. and Pan, W. (2003). Linear regression and two-class classification with gene expression data. *Bioinformatics* **19**, 2072-2078.

Huang, X., Pan, W., Park, S., Han, X., Miller, L. W., and Hall, J. (2004). Modeling the relationship between LVAD support time and gene expression changes in the human heart by penalized partial least squares. *Bioinformatics* **20**, 888-894.

Indahl, U. G., Liland, K. H., and Naes, T. (2009). Canonical partial least squares-a unified PLS approach to classification and regression problems. *Journal of Chemometrics* **23**, 495-504.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis* **5**, 248-264.

Karlsson, M. K., Lonneborg, A., and Saebo, S. (2012). Microarray-based prediction of Parkinson's disease using clinical data as additional response variables. *Statistics in medicine* **31**, 4369-4381.

Kruskal, J. B. and Wish, M. (1978). *Multidimensional scaling (Vol. 11)*: Sage Publications.

Lazraq, A., Cleroux, R., and Gauchi, J. P. (2003). Selecting both latent and explanatory variables in the PLS1 regression model. *Chemometrics and Intelligent Laboratory Systems* **66,** 117-126.

Le Cao, K. A., Boitard, S., and Besse, P. (2011). Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC bioinformatics* **12**, 253.

Le Cao, K. A., Rossouw, D., Robert-Granie, C., and Besse, P. (2008). A sparse PLS for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology* **7**, Article 35.

Lebart, L., Morineau , A., and Fénelon J.P. (1982). *Traitement des données statistiques : (méthodes et programmes)[Note(s) : XIII, 510 p.,] (bibl.: 8 p.)*. Paris: Dunod.

Lindgren, F., Geladi, P., and Wold, S. (1993). The Kernel Algorithm for Pls. *Journal of Chemometrics* **7**, 45-59.

Lopez-Corral, L., Corchete, L. A., Sarasquete, M. E., Mateos, M. V., Garcia-Sanz, R., Ferminan, E., Lahuerta, J. J., Blade, J., Oriol, A., Teruel, A. I., Martino, M. L., Hernandez, J., Hernandez-Rivas, J. M., Burguillo, F. J., San Miguel, J. F., and Gutierrez, N. C. (2014). Transcriptome analysis reveals molecular profiles associated with evolving steps of monoclonal gammopathies. *Haematologica* **99**, 1365-1372.

Man, M. Z., Dyson, G., Johnson, K., and Liao, B. (2004). Evaluating methods for classifying expression data. *Journal of Biopharmaceutical Statistics* **14**, 1065-1084.

Martens, H. and Naes, T. (1989). *Multivariate calibration*. Chichester: Wiley.

Mehmood, T., Martens, H., Saebo, S., Warringer, J., and Snipen, L. (2011). A Partial Least Squares based algorithm for parsimonious variable selection. *Algorithms for molecular biology* **6**, 27-38.

Mehmood, T., Liland, K. H., Snipen, L., and Sæbø, S. (2012). A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems* **118**, 62-69.

NAG. Numerical Algorithms Group (2012). Partial Least Squares Documentation, routines G02LAF, G02LCF and G02LDF.

Nguyen, D. V. and Rocke, D. M. (2002a). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **18**, 39-50.

Nguyen, D. V. and Rocke, D. M. (2002b). Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics* **18**, 1216-1226.

Olson Hunt, M. J., Weissfeld, L., Boudreau, R. M., Aizenstein, H., Newman, A. B., Simonsick, E. M., Van Domelen, D. R., Thomas, F., Yaffe, K., and Rosano, C. (2014). A variant of sparse partial least squares for variable selection and data exploration. *Frontiers in neuroinformatics* **8**, article18.

Osten, D. W. (1988). Selection of optimal regression models via cross-validation. *Journal of Chemometrics* **2**, 39-48.

Perez-Enciso, M. and Tenenhaus, M. (2003). Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Human genetics* **112**, 581-592.

Perez-Lopez, C. (2005). *Metodos Estadisticos Avanzados con SPSS*. Madrid: Thomson.

Rosipal, R. and Kraemer, N. (2006). Overview and recent advances in partial least squares. *Subspace, Latent Structure and Feature Selection* **3940**, 34-51.

Sampson, D. L., Parker, T. J., Upton, Z., and Hurst, C. P. (2011). A comparison of methods for classifying clinical samples based on proteomics data: a case study for statistical and machine learning approaches. *PloS one* **6**, e24973.

Simon, R., Korn, E., McShane, L., Radmacher, M., Wright, G., and Zhao, Y. (2003). *Design and Analysis of DNA Microarray Investigations*. New York: Springer.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* **3**, Article 3.

Stekel, D. (2003). *Microarray Bioinformatics*. New YorK: Cambridge University Press.

Sturn, A., Quackenbush, J., and Trajanoski, Z. (2002). Genesis: cluster analysis of microarray data. *Bioinformatics* **18**, 207-208.

Sun, Y., Goodison, S., Li, J., Liu, L., and Farmerie, W. (2007). Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics* **23**, 30-37.

Takane, Y., Young, F. W., and Deleeuw, J. (1977). Nonmetric Individual-Differences Multidimensional-Scaling - Alternating Least-Squares Method with Optimal Scaling Features. *Psychometrika* **42**, 7-67.

Ter Braak, C. J. (1986). Canonical correspondence analysis: a new eigenvector tehcnique for multivariate gradient analysis. *Ecology* **67(5)**, 1167-1179.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 6567-6572.

Tobias, R. D. (1995). An introduction to partial least squares regression. In *Proc. Ann. SAS Users Group Int. Conf.* 2-5. Orlando.

Torgerson, W. S. (1952). Multidimensional Scaling: I. Theory and Method. *Psychometrika* **17**, 401-419.

Tucker, L. R. (1958). An Inter-Battery Method of Factor-Analysis. *Psychometrika* **23**, 111-136.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5116-5121.

Van Den Wollenberg, A. L. (1977). Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika* **42(2)**, 207-219.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-536.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.

Viala, M., Bhakar, A. L., de la Loge, C., van de Velde, H., Esseltine, D., Chang, M., Dhawan, R., and Dubois, D. (2007). Patient-reported outcomes helped predict survival in multiple myeloma using partial least squares analysis. *Journal of clinical epidemiology* **60**, 670-679.

Wang, D., Song, X., Wang, Y., Li, X., Jia, S., and Wang, Z. (2014). Gene expression profile analysis in epilepsy by using the partial least squares method. *The Scientific World Journal* **2014**, 731091.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis*, Krishnaiach, P. R. (ed), 391-420. New York: Academic Press.

Wold, H. (1975). Path models with latent variables: The NIPALS approach. In *Quantitative Sociology: International perspectives on mathematical and statistical model building.*, Blalock, H. (ed), 307-357. New York: Academic Press.

Wold, S. (1994). PLS for Multivariate Linear Modeling QSAR: Chemometric Methods in Molecular Design. In *Methods and Principles in Medicinal Chemistry*, van de Waterbeemd, H. (ed), : Verlag-Chemie.

Yeniay, O. and Goktas, A. (2002). A comparison of partial least squares regression with other predicción methods. *Hacettepe Journal of Mathematics and Statistic.* **31**, 91-111.

Zhu, J. and Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics* **5**, 427-443.