Resumen

En las últimas décadas los avances tecnológicos han tenido como consecuencia la generación de una creciente cantidad de datos en el campo de la biología y la biomedicina. A día de hoy, las así llamadas tecnologías "ómicas", como la genómica, epigenómica, transcriptómica o metabolómica entre otras, producen bases de datos con cientos, miles o incluso millones de variables.

El análisis de datos ómicos presenta una serie de complejidades tanto metodológicas como computacionales que han llevado a una revolución en el desarrollo de nuevos métodos estadísticos específicamente diseñados para tratar con este tipo de datos.

A estas complejidades metodológicas hay que añadir que, en la mayor parte de los casos, las restricciones logísticas y/o económicas de los proyectos de investigación suelen conllevar que los tamaños muestrales en estas bases de datos con tantas variables sean muy bajos, lo cual no hace sino empeorar las dificultades de análisis, ya que se tienen muchísimas más variables que observaciones.

Entre las técnicas desarrolladas para tratar con este tipo de datos podemos encontrar algunas basadas en la penalización de los coeficientes, como lasso o elastic net, otras basadas en técnicas de proyección sobre estructuras latentes como PCA o PLS y otras basadas en árboles o combinaciones de árboles como random forest.

Todas estas técnicas funcionan muy bien sobre distintos datos *ómicos* presentados en forma de matriz (IxJ). Sin embargo, en ocasiones los datos ómicos

pueden estar expandidos, por ejemplo, al tomar medidas repetidas en el tiempo sobre los mismos individuos, encontrándonos con estructuras de datos que ya no son matrices, sino arrays tridimensionales o *three-way* ($IxJxK$). En estos casos, la mayoría de las técnicas citadas pierden parte de su aplicabilidad, quedando muy pocas opciones viables para el análisis de este tipo de estructuras de datos.

Una de las técnicas que sí es útil para el análisis de estructuras *three-way* es N-PLS, que permite ajustar modelos predictivos razonablemente precisos, así como interpretarlos mediante distintos gráficos.

Sin embargo, relacionado con el problema de la escasez de tamaño muestral relativa al desorbitado número de variables, aparece la necesidad de realizar una selección de variables relacionadas con la variable respuesta. Esto es especialmente cierto en el ámbito de la biología y la biomedicina, ya que no solo se quiere poder predecir lo que va a suceder, sino entender por qué sucede, qué variables están implicadas y, a poder ser, no tener que volver a recoger los cientos de miles de variables para realizar una nueva predicción, sino utilizar unas cuantas, las más importantes, para poder diseñar kits predictivos coste/efectivos de utilidad real. Por ello, el objetivo principal de esta tesis es mejorar las técnicas existentes para el análisis de datos ómicos, específicamente las encaminadas a analizar datos *three-way*, incorporando la capacidad de selección de variables, mejorando la capacidad predictiva y mejorando la interpretabilidad de los resultados obtenidos. Todo ello se implementará además en un paquete de R completamente documentado, que incluirá todas las funciones necesarias para llevar a cabo análisis completos de datos *three-way*.

El trabajo incluido en esta tesis por tanto, consta de una primera parte teórico-conceptual de desarrollo de la idea del algoritmo, así como su puesta a punto, validación y comprobación de su eficacia; de una segunda parte empírico-práctica de comparación de los resultados del algoritmo con otras metodologías de selección de variables existentes, y de una parte adicional de programación y desarrollo de *software* en la que se presenta todo el desarrollo del paquete de R, su funcionalidad y capacidades de análisis.

El desarrollo y validación de la técnica, así como la publicación del paquete de R, que ya cuenta con varios cientos de usuarios, ha permitido ampliar las opciones actuales para el análisis de datos ómicos *three-way* abriendo un gran número de líneas futuras de investigación.