Variable selection and predictive models for Big Data environments

Álvaro Méndez-Civieta
UC3M - Santander Big Data institute
alvaro.mendez@uc3m.es

Directoras:
M. Carmen Aguilera-Morillo
Department of Applied Statistics and Operational Research, and Quality
Universitat Politècnica deValència
mdagumor@eio.upv.es

Rosa E. Lillo
Department of Statistics
Universidad Carlos III de Madrid
UC3M - Santander Big Data institute
lillo@est-econ.uc3m.es

Keywords: Dimension-reduction, Functional-data, High-dimension, Quantile-regression, Variable-selection.

MSC Subject classifications: 62J07, 62H25.

En los últimos años, los avances en las tecnologías de recopilación de datos han planteado un difícil reto al extraer conjuntos de datos cada vez más complejos y de mayor tamaño. Tradicionalmente, las metodologías estadísticas trataban con conjuntos de datos en los que el número de variables no superaba el número de observaciones, sin embargo, enfrentarse a problemas en los que el número de variables es mayor que el número de observaciones se ha convertido en algo cada vez más común, y puede verse en áreas como la economía, la genética, los datos climáticos, la visión por ordenador, etc. Este problema ha exigido el desarrollo de nuevas metodologías adecuadas para un marco de alta dimensión.

La mayoría de las metodologías estadísticas se limitan al estudio de los promedios. La regresión por mínimos cuadrados, el análisis de componentes principales, los mínimos cuadrados parciales (PLS), etc. Todas estas técnicas proporcionan estimaciones basadas en la media y se basan en la idea clave

de que los datos se distribuyen normalmente. Pero esta es una suposición que no suele verificarse en conjuntos de datos reales, en los que es fácil encontrar asimetría, heterocedasticidad y valores atípicos. La estimación de métricas alternativas más robustas, como los cuantiles, puede ayudar a resolver estos problemas, proporcionando una imagen más completa de la distribución de los datos.

Esta tesis se construye en torno a estas dos ideas centrales. Buscamos desarrollar metodologías más robustas, basadas en cuantiles, y extenderlas a problemas de alta dimensión donde el número de variables es posiblemente mayor que el número de observaciones.

Una solución cuando se trata de problemas de alta dimensión en el campo de la regresión es el uso de técnicas de selección de variables. En este sentido, Friedman, Hastie y Tibshirani (2010) propuso el *sparse group lasso* (SGL), una combinación lineal de *lasso* y *group lasso* que ha demostrado ser una alternativa muy eficaz. Sin embargo, este tipo de penalizaciones se basan en el concepto del equilibrio entre sesgo y varianza, y buscan reducir la variabilidad de las estimaciones introduciendo cierto sesgo en el modelo, lo que en el contexto de selección de variables significa que es posible que las variables seleccionadas por el modelo no sean las verdaderamente significativas. La primera contribución de esta tesis estudia la formulación de un *adaptive sparse group lasso* (ASGL) para la regresión cuantílica, una formulación más flexible del SGL que hace uso de la idea adaptiva, es decir, el uso de pesos adaptativos en la penalización para ayudar a corregir el sesgo, mejorando así la selección de variables y la precisión de la predicción. Sin embargo, la idea adaptiva se ha limitado tradicionalmente a escenarios de baja dimensión, ya que requiere resolver un modelo no penalizado (lo cual es inviable en alta dimensión). En esta tesis se estudian una serie de alternativas para el cálculo de pesos basadas en componentes principales y PLS que extienden de forma efectiva los estimadores basados en la idea adaptiva a problemas de alta dimensión, y también se muestran los beneficios de esta propuesta en un conjunto de datos de genética. Estos resultados han sido publicados en Mendez-Civieta, Aguilera-Morillo y Lillo (2021).

Una solución alternativa al problema de alta dimensión es el uso de una técnica de reducción de dimensión como los PLS Wold (1973). Los PLS son una metodología propuesta inicialmente en el campo de la quimiometría como alternativa a la regresión tradicional por mínimos cuadrados en casos en los que los datos son de alta dimensión o son colineales. Los PLS funcionan proyectando la matriz de datos independientes en un subespacio de variables no correlacionadas que maximizan la covarianza con la matriz de respuesta. Sin embargo, el hecho de ser un proceso iterativo basado en mínimos cuadrados implica que esta metodología proporciona estimaciones basadas en la media, y la hace extremadamente sensible a la presencia de valores atípicos, asimetría o heterocedasticidad. La segunda contribución de esta tesis define la *fast partial quantile regression* (fPQR), una metodología que realiza una proyección en un subespacio donde se maximiza una métrica de covarianza cuantílica, extendiendo de forma efectiva los PLS al marco de la regresión cuantílica. A diferencia de la covarianza tradicional, no existe una definición única de lo que debe ser una covarianza cuantílica. Por ello, en este trabajo se estudian tres alternativas diferentes para esta métrica mediante una serie de conjuntos de datos sintéticos y se proporciona una implementación eficiente del algoritmo de fPQR. Finalmente, se compara el algoritmo fPQR frente al PLS y una versión robusta de PLS en un conjunto de datos de quimiometría. Estos resultados se han publicado en Méndez Civieta, Aguilera-Morillo y Lillo (2022).

La tercera contribución de esta tesis se engloba en el campo del análisis de datos funcionales (FDA), y está motivada por un conjunto de datos reales que estudian el nivel de actividad física en 420 niños, medida mediante *wearables*. En el campo del FDA, es habitual tratar cada observación de proporcionada por un *wearable* como una función, una curva de actividad, normalmente registrada durante un periodo de 24 horas. Para estudiar este tipo de datos se pueden utilizar diferentes

metodologías, siendo el análisis funcional de componentes principales (FPCA) una de las alternativas más utilizadas. El FPCA puede descomponer los datos en un conjunto de funciones *loading* que identifican y describen la variación en las curvas muestrales. Un inconveniente del FPCA es que se centra en la reconstrucción del valor esperado para cada sujeto, y no capta aspectos ocultos que pueden afectar a la escala, desplazando los cuantiles. Este problema es especialmente importante en escenarios donde los datos están sesgados o muestran una gran variabilidad. En estas situaciones, comprender los patrones no sólo en el centro, sino también en las colas de la distribución puede ser muy útil. Esta tesis introduce el *functional quantile factor model* (FQFM), una metodología que extiende el concepto de FPCA a la regresión cuantílica, obteniendo un modelo que puede explicar los cuantiles de los datos condicionados a un conjunto de funciones *loading*. Así mismo se propone un algoritmo iterativo para el cálculo del estimador FQFM. Este algoritmo es adecuado para tratar con datos ausentes, y con observaciones medidas en mallas de tiempo irregulares.

La última contribución de esta tesis es *asgl*, un paquete de Python que resuelve modelos de mínimos cuadrados y de regresión cuantílica penalizados en espacios de baja y alta dimensión. Este paquete llena un vacío en las metodologías existentes en diferentes lenguajes de programación como R, Matlab o Python, haciendo posible el uso de penalizaciones basadas en la idea adaptiva. Además proporciona diferentes alternativas para el cálculo de los pesos, y está programado de forma que pueda ejecutarse en paralelo, reduciendo potencialmente el tiempo de cálculo. El paquete ha sido muy bien recibido, consiguiendo en el momento de escribir este documento más de 11000 descargas, y su documentación completa se puede encontrar en Méndez-Civieta, Aguilera-Morillo y Lillo (2021).

Finalmente, el último capítulo de la tesis presenta las conclusiones de este trabajo, e incluye posibles líneas de investigación futuras.

Agradecimientos

Los autores desean agradecer el apoyo financiero recibido por las becas y proyectos de investigación PIPF UC3M, ECO2015-66593-P (Ministerio de Economía y Competitividad, España) y PID2020-113961GB-I00 (Agencia Estatal de Investigación España).

Acerca del autor

Álvaro Méndez-Civieta es Doctor en Estadística por la Universidad Carlos III de Madrid, Máster en Big Data por la Universidad Carlos III de Madrid y Licenciado en Matemáticas por la Universidad de Oviedo. Su investigación se centra en el desarrollo de alternativas robustas basadas en cuantiles para el análisis de conjuntos de datos de alta dimensión, con un enfoque en la selección de variables, la reducción de la dimensión y los datos funcionales.

Referencias

Friedman, J., T. Hastie y R. Tibshirani (2010). «A note on the group lasso and a sparse group lasso». En: *ArXiv:1001.0736*, págs. 1-8. ISSN: 15410420. DOI: 10.1111/biom.12292. URL: http://arxiv.org/abs/1001.0736.

Méndez Civieta, Álvaro, M. Carmen Aguilera-Morillo y Rosa E. Lillo (mar. de 2022). «Fast partial quantile regression». En: *Chemometrics and Intelligent Laboratory Systems*, pág. 104533. ISSN: 01697439. DOI: 10.1016/j.chemolab.2022.104533. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169743922000442.

Mendez-Civieta, Alvaro, M. Carmen Aguilera-Morillo y Rosa E. Lillo (2021). «Adaptive sparse group LASSO in quantile regression». En: *Advances in Data Analysis and Classification* 15.3, págs. 547-573. ISSN: 18625355. DOI: 10.1007/s11634-020-00413-8.

Méndez-Civieta, Álvaro, M. Carmen Aguilera-Morillo y Rosa E. Lillo (2021). «Asgl: A Python Package for Penalized Linear and Quantile Regression». En: págs. 1-31. URL: http://arxiv.org/abs/2111.00472.

Wold, H (1973). «Nonlinear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments». En: *Multivariate Analysis-III*. Ed. por Paruchuri R Krishnaiah. Academic Press, págs. 383-407. ISBN: 978-0-12-426653-7.