

LASSO.FREC: UN MODELO DE SELECCIÓN DE VARIABLES BASADO EN LAS SOLUCIONES DE LASSO EN TODA LA GRILLA

Verónica Moreno

Universidad Nacional de Tres de Febrero y Universidad de San Andrés, Argentina
veronicamoreno84@gmail.com

Lo más usual a la hora de descorrelacionar variables y reducir dimensión es usar el método LASSO, que consiste en minimizar

$$R(\beta_0, \beta) = \frac{1}{2n} \sum_{i=1}^n (y_i - \beta_0 - X_i \beta^T)^2 + \lambda \sum_{j=1}^p |\beta_j|.$$

La solución de este problema es esparza por lo cual para seleccionar variables se seleccionan aquellas con un valor de β igual a cero. Este método fue planteado por primera vez por [3], pero una solución numérica escalable se presentó recién en [1]. Muchas variantes de este modelo surgieron cómo métodos de selección de variables, para un estudio completo sobre todas las variantes de LASSO ver [2]. Hay numerosos trabajos que estudian la elección del parámetro de penalización λ . Lo más usual es considerar una grilla de valores de λ y quedarme con el que minimiza el error cuadrático medio al considerar los coeficientes de LASSO en un test data usando la técnica de cross validation (5 folds es lo más usual). Esta técnica de selección de variables se conoce como LASSO.MIN y es conocida por tener mucho poder predictivo, pero es muy conservadora a la hora de seleccionar las variables ya que suele tomar las verdaderas y muchas otras que no. En este trabajo proponemos un algoritmo de selección de variables llamado LASSO.FREC basado en resolver LASSO en una grilla de valores de λ . Lo que proponemos es considerar la frecuencia con que se selecciona cada variable teniendo en cuenta las soluciones de todos los λ 's. Vamos a seleccionar las variables que tienen mayor frecuencia.

El algoritmo LASSO.FREC tiene los siguientes pasos:

- 1- Elegir un threshold τ , este es un parámetro del algoritmo que tiene que ser elegido por el usuario.
- 2- Seleccionar una grilla de valores para λ .
- 3- Para cada valor de λ en la grilla considero los valores $(\beta_0, \beta)^\lambda$ que resuelven LASSO y armo el vector que me indica que variables selecciono con este λ de la siguiente manera: $S^\lambda \in \mathbb{R}^p$ tal que $S_j^\lambda = 1$ si seleccioné la variable j (o sea si β_j es distinto de cero) y $S_j^\lambda = 0$ en otro caso.
- 4- Armo un vector de frecuencias: $\text{Frec} \in \mathbb{R}^p$ con $\text{Frec}_j = \frac{1}{L} \sum_{\lambda} S_j^\lambda$, donde L es la cantidad de puntos que tiene la grilla.
- 5- Seleccionar las variables j que cumplan $\text{Frec}_j \geq \tau$.

Para un análisis de este algoritmo se seleccionaron los mismos tres escenarios que en [2] con el objetivo de poder comparar con otras variantes de LASSO. Para cada uno de estos tres escenarios, vamos a mostrar el gráfico de las frecuencias ordenadas de mayor a menor, con un color van a estar las verdaderas y con otros las falsas. Se puede observar que las verdaderas aparecen primeras y que las falsas al final, en algunos casos observando un salto entre estos dos grupos (verdaderas y falsas). En el escenario en que tenemos correlaciones muy marcadas, el algoritmo muy pocas veces confunde una variable verdadera con una de las falsas. Se realizó una simulación de monte carlo con 1000 simulaciones, promediando la cantidad de variables verdaderas y falsas que toma en cada selección. Se realizó una comparación de LASSO.FREC con diferentes thresholds (0.7, 0.8 y 0.9) y LASSO.MIN. Como resultado de esta comparación podemos ver que LASSO.MIN siempre selecciona las verdaderas pero selecciona muchas más falsas que LASSO.FREC mientras que LASSO.FREC selecciona muy pocas de las falsas, y muy pocas veces pierde una variable verdadera. Realizamos esta comparación para diferentes tamaños de muestras, observando que LASSO.FREC mejora en muchos escenarios cuando el tamaño de la muestra es más grande. Por último mostramos un ejemplo con datos reales, en donde comparamos las variables

seleccionadas por LASSO.FREC y LASSO.MIN, y con estas variables miramos el mse en un test data. Como conclusión se ve que seleccionando un threshold τ adecuado se puede lograr muchas menos variables que LASSO.MIN y mayor poder predictivo.

Trabajo en conjunto con Lucas Fernández Piana (Universidad de San Andrés, Argentina)..

Referencias

- [1] Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1), 1.
- [2] Freijeiro [U+2010] González, L., Febrero [U+2010] Bande, M., González [U+2010] Manteiga, W. (2022). A critical review of LASSO and its derivatives for variable selection under dependence among covariates. *International Statistical Review*, 90(1), 118-145.
- [3] Tibshirani, R. (1996). Regression shrinkage selection via the LASSO. *J. R. Stat. Soc. B. Methodol.*, 58(1):267–288.