Selección de variables en modelos de regresión lineal controlando la tasa de falsos positivos

Leonardo Alfonso Martínez González.
Addy Margarita Bolívar Cimé.
Rogelio Ramos Quiroga (CIMAT)
Universidad Juárez Autónoma de Tabasco.
email: Mat.LAMG@outlook.com

1. Introducción

En muchos campos de la ciencia, comúnmente observamos una variable de respuesta junto con un gran número de variables explicativas potenciales, y deseamos ser capaces de descubrir cuales son las variables que están verdaderamente asociadas con la respuesta. Al mismo tiempo, necesitamos saber que la *tasa de falsos positivos* (*false discovery rate*)–la fracción esperada de los falsos positivos entre todas las variables seleccionadas– no es demasiado alta, para asegurarle al científico que la mayoría de las variables seleccionadas son verdaderas y replicables.

Supongamos que tenemos registros de una variable de respuesta de interés y y muchas variables explicativas potenciales X_j, $j = 1, 2, \ldots, p$, en n unidades de observación. Nuestras observaciones obedecen un modelo de regresión lineal clásico

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \tag{1}$$

donde $\mathbf{y} \in \mathbb{R}^n$ es un vector de respuestas, $\mathbf{X} \in \mathbb{R}^{n\times p}$ es una matriz conocida, $\beta \in \mathbb{R}^p$ es un vector de coeficientes desconocido, y $\epsilon \sim N(0, \sigma^2\mathbf{I})$ es un vector aleatorio de ruido. Considerando el caso en que $p > n$, en este cartel se verá como seleccionar las variables de un modelo de regresión lineal utilizando la metodología de knockoff.

2. False discovery rate (FDR)

El FDR es la proporción esperada de variables falsamente seleccionadas, el FDR de un procedimiento de selección de variables en el modelo (1), en el que se obtiene un subconjunto $\widehat{S} \subset \{1, 2, \ldots, p\}$ de variables, se define como

$$FDR = E\left[\frac{\#\left\{j : \beta_j = 0 \text{ y } j \in \widehat{S}\right\}}{\#\left\{j : j \in \widehat{S}\right\} \vee 1}\right].$$

Por otro lado, también podemos considerar el error de tipo S (de equivocarse de signo) para definir la tasa de error direccional

$$FDR_{dir} = E\left[\frac{|\left\{j \in \widehat{S} : \widehat{sign}(\beta_j) \neq sign(\beta_j)\right\}|}{|\widehat{S}| \vee 1}\right].$$

3. Selección de variables mediante Knockoff

Se presenta un procedimiento de control general de FDR que está garantizado para funcionar bajo cualquier diseño fijo $X \in \mathbb{R}^{n\times p}$ con $n \geq p$ y la respuesta y sigue un modelo lineal Gaussiano como en (1).

3.1 Variables knockoff

Para cada característica X_j en el modelo, es decir, las columnas de X, construiremos una característica "Knockoff" $\widehat{X}_j$.

- Una construcción explícita, para el caso $n \geq p$

$$\tilde{X} = X\left(I - \Sigma^{-1}\text{diag}(s)\right) + \tilde{U}C,$$

donde Σ^{-1} es la inversa de la matriz de Gram, $\tilde{U}$ es una matriz ortonormal de $n \times p$ que es ortogonal a la matriz de características X, $C^TC = 2\text{diag}(s) - \text{diag}(s)\Sigma^{-1}\text{diag}(s)$ es una descomposición de Cholesky y s es un vector p-dimensional no negativo.

- Con estas nuevas variables se ajusta el modelo, usando lasso

$$y = [X, \tilde{X}]\,\beta + \epsilon.$$

3.2 Selección

- Se construyen estadísiticos para las $2p$ variables

$$Z_j = \sup\left\{\lambda : \ \widehat{\beta}_j(\lambda) \neq 0\right\},$$

para cada $j = 1, \ldots, p$

$$W_j = (Z_j \vee \tilde{Z}_j) \cdot \begin{cases} +1 & Z_j > \tilde{Z}_j \\ -1 & Z_j < \tilde{Z}_j \end{cases};$$

- Se selecciona la variable X_j si $W_j \geq T$, donde

$$T = \min\left\{t : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q\right\}.$$

- Este criterio tiene la propiedad: para alguna $q \in [0, 1]$

$$\text{E}\left[\frac{\#\{j : \beta_j = 0,\ j \in \widehat{S}\}}{\#\{j : j \in \widehat{S}\} + \frac{1}{q}}\right] \leq q.$$

- T es el primer tiempo para el cual cierto cociente cae por abajo de q, entonces T puede considerarse como un tiempo de paro de un proceso estocástico, con la propiedad de ser (super)martingala. Entonces, por el teorema de tiempo de paro "opcional":

$$E\left[\frac{V^+(T)}{1+V^-(T)}\right] \leq E\left[\frac{V^+(0)}{1+V^-(0)}\right] = E\left[\frac{V^+(0)}{1+N_0-V^+(0)}\right] \leq 1,$$

donde N_0 es el numero total de casos nulos, $V^+(T) = \#\{nullj : 1 \leq j \leq T; p_j \leq q\}$ y $V^-(T) = \#\{nullj : 1 \leq j \leq T; p_j > q\}$, con p_j el correspondiente p-valor de β_j.

3.3 Knockoff en dimensiona alta

En dimensiones altas, donde $p > n$, la construcción de knockoff ya no es posible. En esta situacion, proponemos la siguiente estrategia general

- Paso de proyección: en un primer paso, filtrar todas las características $X_j, j = 1, 2, \ldots, p$ para identificar un conjunto $S_0 \subset \{1, 2, \ldots, p\}$ de características potencialmente relevantes con $|S_0| < n$. Consideraremos este paso como bastante liberal en el sentido de que típicamente produce una lista larga, que con suerte contiene la mayoría de las características importantes, pero también posiblemente muchas características con efectos nulos ($\beta_j = 0$) o efectos de casi desaparición ($\beta_j \approx 0$).
- Paso de inferencia: el paso anterior produce un modelo reducido

$$\mathbf{y} = \mathbf{X}\beta^{partial} + \epsilon$$

donde $\beta^{partial}$ indican que tanto la definición como el significado de los coeficientes de regresión han cambiado. Luego, pruebe las asociaciones en este modelo reducido controlando el FDR_{dir}.

- Ahora, considere una característica X_j que no aparece en el modelo completo original, es decir, $\beta_j = 0$. En la regresión parcial, podemos encontrar uno de dos escenarios:

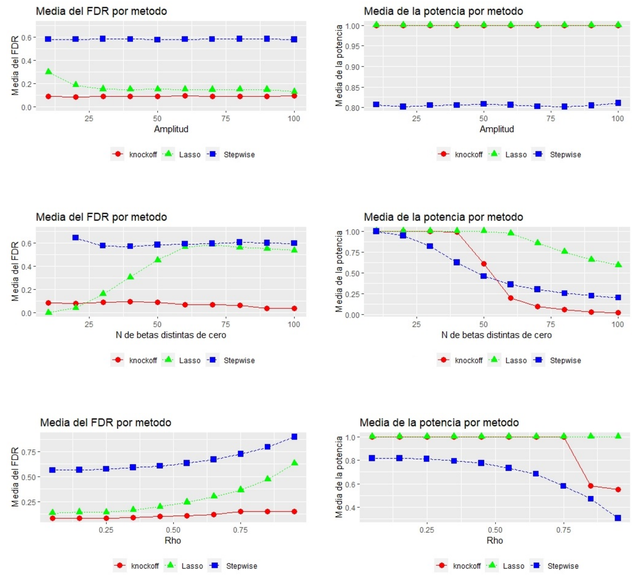
1) El coeficiente en la regresión parcial puede ser grande, es decir, $\beta_j^{parcial}$ no se aproxima a 0. Esto ocurre típicamente si X_j está altamente correlacionado con alguna señal fuerte X_k que se perdió en el paso de proyección. Aquí, en general preferiríamos incluir X_j en el modelo seleccionado en cualquier caso, ya que es una buena proximacion para la característica relevante perdida X_k.
2) Alternativamente, el coeficiente podría permanecer cerca de cero, $\beta_j^{parcial} \approx 0$. Esto es probable siempre que X_j no sea una aproximación de ninguna característica perdida. Aquí, el signo de $\beta_j^{parcial}$ no se puede estimar con mucha certeza, por lo que con el control de FDR_{dir}, es probable que excluyamos X_j de nuestro modelo final.

Estas consideraciones claramente dejan en claro que una selección cuidadosa seguida de un control FDR_{dir} podría producir un procedimiento de selección valioso.

4. Simulación de procedimientos de selección

Se realizaron simulaciones para comparar la metodología de knockoff contra los métodos de Lasso y stepwise. En estas simulaciones tenemos fija a $p = 1200$ y $n = 250$.
Cuando variamos la amplitud (el tamaño de la señal) fijamos $k = 30$ (donde k es el número de betas distintas de cero) y $\rho = 0.3$ (donde ρ es la correlación); cuando variamos k fijamos la amplitud a 30 y $\rho = 0.3$; y cuando variamos ρ, fijamos la amplitud a 30 y $k = 30$.
De estas simulaciones medimos el FDR y la potencia (P):

$$FDR = E\left[\frac{\#\left\{j : \beta_j = 0 \text{ y } j \in \widehat{S}\right\}}{\#\left\{j : j \in \widehat{S}\right\} \vee 1}\right],\ P = E\left[\frac{\#\left\{j : \beta_j \neq 0 \text{ y } j \in \widehat{S}\right\}}{K}\right].$$



5. Conclusiones

- Observamos de las gráficas que la metodología de knockoff tiene una tasa de descubrimientos falsos pequeña en comparación a los otros dos métodos.
- Esta metodología controla el FDR, pero no el error de tipo I.
- Observamos en la gráfica que la potencia de la metodología knockoof no es muy buena pero tampoco es tan mala.

Referencias

[Barber y Candes, 2015] Barber, R. F., Candes, E. (2015). Controlling the false discovery rate via Knockoffs. Ann. Statist. 43 (2015), no. 5, 2055–2085. doi:10.1214/15-AOS1337.

[Barber, R. F., y Candès, E. J. (2019)] knockoff filter for high-dimensional selective inference. Annals of Statistics, 47(5), 2504-2537.

[Efron, 2010] Efron B. (2010). Large-scale inference: empirical bayes methods for estimation, testing and prediction. Cambridge University Press.