

Aplicación de LASSO a modelos mixtos: un enfoque bayesiano

Rafael Ríos García, Elena Moreno Maestre,
y David Hervás Marín

Instituto de Investigación Sanitaria La Fe

<http://www.iislafe.es>

6 de Noviembre de 2015

Introducción

- Uno de los problemas que surgen con frecuencia en el ámbito bioestadístico es el de analizar datos donde:
 - 1 El número de variables predictoras supera, ampliamente, al número de observaciones disponibles.
 - 2 La observaciones no son independientes
- A partir de aquí, las opciones de análisis se reducen considerablemente: ya no podemos hacer inferencia con los modelos clásicos tipo

$$Y_i \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i + \epsilon$$

- Dos opciones:
 - 1 Darle la vuelta al problema y utilizar cada covariable como variable respuesta y aplicar test sencillos (test t, Chi-cuadrado, ANOVA) junto con FDR: problemas!
 - 2 Utilizar métodos que permitan construir modelos con $p \gg n$, y que admitan factores aleatorios: LASSO mixto.

El Lasso frecuentista

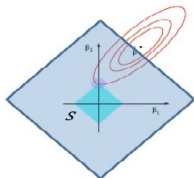
- La regresión LASSO contrae aquellos coeficientes que influyen poco o nada sobre la variable dependiente, haciendo que muchos de ellos converjan a cero.
- Las estimaciones de los coeficientes que proporciona LASSO son las soluciones al problema de optimización siguiente:

$$RSS_{\lambda}(\beta) = \sum_{i=1}^n \left(Y_i - \sum_{j=1}^{p-1} X_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

- λ es el parámetro que controla la penalización sobre la suma de los valores absolutos de las estimaciones, consiguiendo que algunas de las estimaciones de los parámetros sean cero.
- Si $\lambda = 0$, tenemos el problema clásico de mínimos cuadrados.

Intuición gráfica de LASSO

- En el siguiente figura podemos ver cómo LASSO consigue hacer ceros:



Penalización L_1 (Lasso)

- El área azul, $|\beta_1| + |\beta_2| \leq s$, representa la región de contracción de LASSO. Cuando aumentamos las dimensiones, aumentamos la probabilidad de que más estimaciones sean cero.
- El problema de LASSO: no tiene en cuenta la no independencia de la observaciones.

El Lasso frecuentista y su aproximación bayesiana

- LASSO puede ser interpretado desde el punto bayesiano, lo que facilita el análisis de las medidas correladas: tan solo tenemos que añadir un factor aleatorio
- Para conectar esta idea de LASSO con el mundo bayesiano, necesitamos definir, a priori, distribuciones de probabilidad sobre los parámetros que queremos estimar.
- Las estimaciones LASSO pueden ser interpretadas como una distribución doble exponencial a priori sobre los β_j :

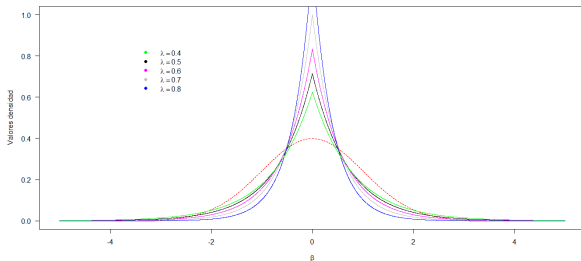
$$f(\beta|\mu, \lambda, b) = \frac{1}{2\lambda} \exp\left(-\frac{|\beta - \mu|}{\lambda}\right)$$

$$\beta_j \sim \text{DE}\left(0, \frac{1}{\lambda\tau}\right)$$

- λ es el parámetro de escala de la distribución de Laplace, y hace el papel de penalización en el LASSO bayesiano.

El Lasso frecuentista y su aproximación bayesiana

- Distribución de Laplace (o doble exponencial) para diferentes λ , junto con la Normal estándar.
- Valores altos de λ concentra mayor densidad de probabilidad cerca del cero.



Definición del Modelo 1

```
model{  
  
  #LIKELIHOOD  
  for(i in 1:N){  
    Y[i] ~ dnorm(mu[i], tau)  
    mu[i] <-eta[i]  
    eta[i] <-inprod(X[i,], bgamma[])+u[G[i]] #bgamma es beta*gamma(0/1)  
  }  
  
  #PREVIA EFECTO ALEATORIO (random intercept)  
  for (j in 1:Nre) {  
    u[j] ~ dnorm(0,tau.re)  
  }  
  tau.re ~ dgamma(0.001,0.001)
```

Definición del Modelo 2

```

#PREVIA PARA b
t1 <- lambda*tau
for(j in 1:P){
  b[j] ~ ddexp(0, t1)
}
tau ~ dgamma(0.0001, 0.0001)
lambda ~ dunif(0.001,10)

#DESV. TIPICA (Fixed & Random Effect)
sz.sq <- 1/tau; sz <- sqrt(sz.sq)
sre.sq<-1/tau.re; sre<-sqrt(sre.sq)

#ZERO TRICK
for(j in 1:P){
  bgamma[j] <- b[j]*gamma[j] #b*gamma
  gamma[j] ~ dbern(0.5) #PREVIA PARA gamma
}

```


Validación

- ¿Qué pretendemos con la validación del método?
 - 1 Controlar qué variables se seleccionan y cuáles no.
 - 2 Controlar la variabilidad debida al factor aleatorio.
 - 3 Controlar el error residual.
 - 4 Controlar la precisión de las estimaciones (en menor medida)
- Construcción de la variable respuesta Y

$$Y_i = \alpha_k X_{ik} + u_i + \epsilon_i, \quad i \in \{1, \dots, n\}, k \in \{1, \dots, 10\}, \alpha_k \in \mathbb{R}$$

$X_{ik} \sim \mathbf{N}(\mu, \sigma)$, matriz de predictores influyentes

$u_i \sim \mathbf{N}(0, \sigma_{re})$, factor aleatorio

$\epsilon_i \sim \mathbf{N}(0, \sigma_{sz})$, error residual

- Construcción de la matriz de predictores no influyentes, $\hat{\mathbf{X}}$, de dimensión $n \times 200$:

$$\hat{\mathbf{X}} \sim \mathbf{N}(\mu, \sigma)$$

- Lo combinamos todo para validarlo utilizando nuestro modelo

$$\mathbf{Y} = \beta \mathbf{X} + \epsilon$$

Conclusiones de la Validación

- Resultados para $n = 9$, $\sigma_{re} = 2$, y $\sigma_{sz} = 1$:
 - ① Se selecciona el 50% de las covariables influyentes, y el 1.05% de las no influyentes (variables espurias)
 - ② Para un error debido al factor aleatorio de 2, hemos obtenido una estimación de 1.91
 - ③ Para un error residual de 1, hemos obtenido un estimación de 0.69
 - ④ La estimaciones de los parámetros para esta configuración de la validación son las siguientes:

Variable	Efecto	Estimación
X1	10,6	5,41
X5	9,8	6,99
X7	11,34	5,36
X8	-14,34	-10,13
X9	12,1	3,89

- Como era de esperar, estos resultados mejoran a medida que se incrementa el tamaño muestral

Aplicación del modelo a un problema biomédico

- Hemos aplicado nuestro modelo a un problema biomédico con un doble objetivo:
 - 1 Seleccionar aquellas covariables potencialmente influyentes
 - 2 Controlar la no independencia de las observaciones
- Para el estudio hemos contado con $n = 18$ pacientes y $p = 188$ covariables que provienen de analizar determinados compuestos químicos del hígado en diferentes momentos del tiempo.
- La variable dependiente es continua, y lo que mide es la capacidad de regeneración hepática.






Resultados

- De las 188 covariables, se seleccionan tan solo 5 como potencialmente influyentes en la regeneración hepática.
- La mayoría las covariables seleccionadas intervienen en procesos bioquímicos relacionados con el hígado.
- Dado que el objetivo es también seleccionar un número contenido de covariables relacionadas con la regeneración hepática, podemos considerar que los resultados son positivos en este sentido.

Conclusión

- Fácil de implementar en R.
- Da buenos resultados, tanto en la validación como en el análisis.
- Parece posible la generalización a más casos: glm, random slope.
- Desde el punto de vista biomédico, se cumple el objetivo previsto: se selecciona un número contenido de metabolitos
- Es aplicable a problemas similares, donde $p \gg n$ y las muestras no son independientes.

Bibliografía

-  Anastasia Lykou, Ioannis Ntzoufras
On Bayesian lasso variable selection and the specification of the Shrinkage Parameter Stat Comput (2013)
-  Trevor Park and George Casella
The Bayesian Lasso Journal of the American Statistical Association (2008)
-  Anastasia Lykou, Ioannis Ntzoufras
WinBUGS: a tutorial, John Wiley and Sons (2011)
-  Tibshirani, R.
Regression shrinkage and selection via the lasso, J. Royal. Statist (1996)
-  THui Zou and Trevor Hastie
Regularization and variable selection via the elastic net, J. Royal. Statist (2005)