REGULARIZACIÓN EN MÉTODOS DE REGRESIÓN

PRESENTA:

Enrique Santibañez Cortes
CIMAT
UNIDAD MONTERREY
enrique.santibanez@cimat.mx

10 de junio de 2021

ABSTRACT

En este trabajo desarrollamos la teoría de los principales métodos de regularización (*Ridge, LASSO y Elastic Net*), enfocándonos en regresión lineal múltiple para poder comprender estos conceptos en su forma más básica. Realizamos una análisis comparativo entre estos métodos, resaltando las ventajas y desventajas que tienen, además de los supuestos que se deben de cumplir para tener buen rendimiento. Posteriormente, realizamos una extensión del método de regularización aplicado en regresión multivariada múltiple presentando la teoría y un ejercicio práctico. Y por último, concluimos resaltando la importancia de estos métodos de regularización que tienen actualmente no solo en problemas de regresión.

1. Introducción

Considerando el planteamiento de regresión lineal,

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

donde $\beta, \mathbf{x}_i \in R^p$, y $\mathbf{X}$ es una matriz de tamaño $n \times p$ con renglones $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$. El método más frecuente para ajustar (1) es utilizar mínimos cuadrados ordinarios (OLS), el cual consiste en identifica como mejor modelo el hiperplanto que minimiza la suma de errores cuadrados, es decir,

$$\beta^{OLS} = \arg\min_{\beta} \left\{ ||\mathbf{y} - \mathbf{X}\beta||^2 \right\}. \tag{2}$$

Algunas de las limitaciones que aparecen en la práctica al tratar de emplear este tipo de modelos (ajustados por mínimos cuadrados ordinarios) son:

- Se ven perjudicados por la incorporación de predictores correlacionados.
- No pueden ajustarse cuando el número de predictores es superior al número de observaciones.

En problemas con muchas variables explicativas potenciales que pueden estar en parte altamente correlacionadas entre sí, el enfoque clásico de regresión por mínimos cuadrados puede sufrir el hecho de que los coeficientes de regresión estimados pueden llegar a estar bastante mal determinados, es decir, tener una alta varianza incluso cuando la superficie de regresión ajustada puede estar bien determinada [Boehmke and Greenwell, 2019]. Entonces, en problemas cuando tenemos $n > p$ y los datos presentan multicolinealidad, una alternativa a la regresión lineal usando mínimos cuadrados es utilizar la regresión regularizada (también conocida como modelos regularizados o métodos de shrinkage) para restringir el tamaño total de todas las estimaciones de coeficientes.

En este trabajo, se describen los fundamentos teóricos y aspectos prácticos de cómo combinar regresión lineal múltiple con métodos de regularización. Las siguientes consideraciones y métodos que aquí se presentan se pueden aplicar tanto para problemas de clasificación como a la regresión en principio. Sin embargo, nos centraremos en modelos de regresión lineal múltiple en la mayor parte de este trabajo.

2. Métodos de regularización o *shrinkage* en regresión lineal múltiple

Lo métodos de regularización son estrategias que incorporan penalizaciones en el ajuste por mínimos cuadrados ordinarios con el objetivo de evitar *overfitting*, reducir varianza, atenuar el efecto de la correlación entre predictores y minimizar la influencia en el modelo de los predictores menos relevantes.

La función objetivo en un modelo de regresión regularizado es similar al OLS, con un término de regularización $P(\beta)$

$$\arg\min_{\beta} \left\{ ||\mathbf{y} - \mathbf{X}\beta||^2 + P(\beta) \right\}. \tag{3}$$

Este parámetro de regularización restringe el tamaño de los coeficientes de modo que la única forma en que los coeficientes pueden crecer es si experimentamos una disminución comparable en la suma de errores cuadrados.

Este concepto se puede generalizar a todos los modelos GLM (ejemplo regresión logística y poisson) e incluso algunos modelos de supervivencias (ver [Tibshirani, 1997]). Con la única diferencia en que no necesariamente se deben de tomar la suma de errores cuadrados como función de perdida en los diferentes modelos de GLM o supervivencia. Pero el concepto es el mismo, podemos pensar en el parámetro de regularización que restringe el tamaño de los coeficientes de tal manera que la única forma en que puedan crecer es si experimentamos una disminución comparable en la función de pérdida del modelo.

Tres de los métodos de regularización más empleados son *Ridge, LASSO y Elastic net*. Dado que estos métodos de regularización actúan sobre la magnitud de los coeficientes del modelo, **todos deben de estar en la misma escala, por esta razón es necesario estandarizar o normalizar los predictores antes de entrenar el modelo**

2.1. Ridge

Cuando hay muchas variables correlacionadas en un modelo de regresión lineal, sus coeficientes pueden estar mal determinados y mostrar una alta varianza. Un coeficiente positivo tremendamente grande en una variable puede ser cancelado por un coeficiente negativo igualmente grande en su primo correlacionado. Los coeficientes de *ridge* minimizan una suma cuadrada del residual penalizada,

$$\hat{\beta}^{ridge}(\beta) = \arg\min_{\beta} \left\{ ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda ||\beta||_2^2 \right\} \tag{4}$$

Aquí $\lambda \geq 0$ es un parámetro de complejidad que controla la cantidad de contracción (shrinkage): cuanto mayor es el valor de λ, mayor es la cantidad de contracción. Los coeficientes se acercan a cero pero no necesariamente son iguales a cero. Una forma equivalente de escribir el problema de *ridge* es

$$\hat{\beta}^{ridge}(t) = \arg\min_{\beta} ||\mathbf{y} - \mathbf{X}\beta||_2^2 \tag{5}$$

$$s.a. \quad ||\beta||_2^2 \leq t, \tag{6}$$

lo que hace explícita la restricción de tamaño de los parámetros. Existe una correspondencia biunívoca entre los parámetros λ en (4) y t en (5). Los estimadores de regresión de *ridge* estan dados por

$$\hat{\beta}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$$

donde I es una matriz identidad de tamaño $p \times p$. La solución anterior, agrega una constantes positiva a la diagonal $\mathbf{X}^T\mathbf{X}$ antes de calcular la inversa. Esto hace que el problema no sea singular, incluso si $\mathbf{X}^T\mathbf{X}$ no es de rango completo. Esta fue la principal motivación por el cuál se introdujo por primera vez la regresión *ridge* [Hoerl and Kennard, 1970].

2.2. Least absolute shrinkage and selection operator (LASSO)

LASSO es un método de contracción como *ridge*, con diferencias sutiles pero importantes. La estimación de LASSO se define resolviendo [Tibshirani, 1996]

$$\hat{\beta}^{lasso}(\beta) = \arg\min_{\beta} \left\{ ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda ||\beta||_1 \right\} \tag{7}$$

Podemos reescribir el problema anterior como un problema de optimización de la siguiente forma:

$$\hat{\beta}^{lasso}(t) = \arg\min_{\beta} ||\mathbf{y} - \mathbf{X}\beta||^2 \tag{8}$$

$$s.a. \quad ||\beta||_1 \leq t, \tag{9}$$

Al igual que en *ridge*, el grado de penalización está controlado por $\lambda \geq 0$. Cuando $\lambda = 0$, el resultado es equivalente estimador por OLS. Y cuando mayor sea λ, mayor serpa la predicción.

Observe el parecido con el problema de regresión utilizando *ridge* (5) y (4), la penalización de *ridge* L_2 $||\beta||_2$ es remplazada por la penalizacipon $L_1||\beta||_1$. Este cambio hace que las soluciones no sean lineales en y, y por lo cuál no existe una expresión de forma cerrada como en el caso de regresión *ridge*, aunque se puede resolver de manera bastante eficiente utilizando algoritmos o métodos como el llamado regresión de ángulo mínimo [Efron et al., 2004] o utilizando gradiente descendiente [Hastie et al., 2001]. Estos algoritmos tienen el mismo costo computacional que en regresión *ridge*. [Hastie et al., 2001]

Debido a la naturaleza de la restricción, hacer t lo suficientemente pequeño hará que algunos de los coeficientes sean exactamente cero, por lo que LASSO hace una especie de selección continua de subconjuntos de variables.

2.2.1. Comparación entre Ridge y LASSO

La principal diferencia práctica entre *LASSO* y *ridge*, es que en *LASSO* es posible obtener coeficientes exactamente iguales a cero. Esto supone una ventaja notable de LASSO en escenarios donde no todos los predictores son importantes para el modelo y se desea que los menos influyentes queden excluidos. En consecuencia, cuando se tiene un problema de regresión con muchos predictores, LASSO se puede utilizar paraa identificar y extraer aquellas características *más importantes*.

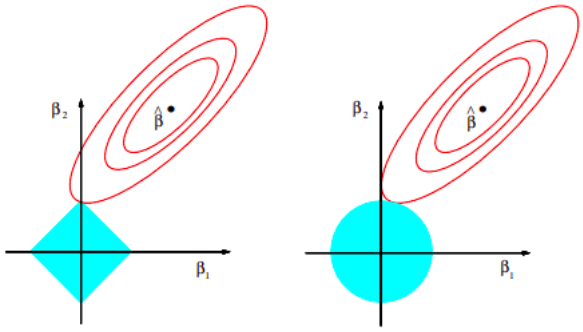

Figura 1: Regiones de los problemas de optimización, [Hastie et al., 2001]

Por otro lado, cuando existen predictores altamente correlacionados (linealmente), *ridge* reduce la influencia de todos ellos a la vez y de forma proporcional, mientras que LASSO tiende a seleccionar uno de ellos, dándole todo el peso y excluyendo al resto. En presencia de correlaciones, esta selección varía mucho con pequeñas perturbaciones (cambios en los datos de entrenamiento), por lo que, las soluciones de LASSO, son muy inestables si los predictores están altamente correlacionados. Por lo que en estos casos es recomendable utilizar regularización *ridge.*

Para conseguir un equilibrio óptimo entre estas dos propiedades (multicolinealidad y subgrupo de variables relevantes), se puede emplear lo que se conoce como penalización elastic net, que combina ambas estrategias.

2.3. Extensión LASSO

Hastie et al. [2001] mencionan que la regularización $L1$ de LASSO a tenido mucha relevancia, hasta el punto de desarrollar la detección de un campo comprimido en la literatura sobre procesamiento de señales. En esta subsección mencionaremos algunas de ellas, sin entrar mucho a detalle.

2.3.1. Elastic net

Zou and Hastie [2005] presentan por primera vez este enfoque de penalización, el cual es una generalización de las penalización *ridge* y *LASSO*, llamado *elastic net*. La estimación de *elastic net* se define

$$\hat{\beta} = \arg\min_{\beta}\{||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda_2||\beta||^2 + \lambda_1||\beta||\} \tag{10}$$

Sea $\alpha = \lambda_2/(\lambda_1 + \lambda 2)$, entonces resolver $\hat{\beta}$ en (10) es equivalente a el siguiente problema de optimización

$$\hat{\beta} = \arg\min_{\beta} \left\{ ||\mathbf{y} - \mathbf{X}\beta||^2 \right\} \tag{11}$$

$$s.a. \quad (1-\alpha)||\beta|| + \alpha||\beta||^2 \leq t. \tag{12}$$

Aunque los modelos de *LASSO* realizan la selección de características, cuando dos características fuertemente correlacionadas se empujan hacia cero, una se puede empujar completamente a cero mientras que la otra permanece en el modelo. Además, el proceso de uno dentro y otro fuera no es muy sistemático. Por el contrario, la penalización por regresión de la cresta es un poco más eficaz en el manejo sistemático de características correlacionadas juntas. En consecuencia, la ventaja de la penalización de la *elastic net* es que permite una regularización eficaz a través de la penalización de *ridge* con las características de selección de características de la penalización del *LASSO*.

2.3.2. LASSO adaptativo

Zou [2006] presenta una condición necesaria para que la selección de la variable *LASSO* sea consistente. En consecuencia, existen ciertos escenarios donde el *LASSO* es inconsistente para la selección de variables. Proponen una nueva versión del LASSO, llamado *LASSO* adaptativo. La función a optimizar en este caso es

$$\hat{\beta}^{lasso_ada}(\beta) = \arg\min_{\beta} \left\{ ||\mathbf{y} - \mathbf{X}\beta||^2 + \mathbf{w}\lambda||\beta||_1 \right\} \tag{13}$$

donde $\mathbf{w}$ es un vector de pesos conocidos. Se puede mostrar que el *LASSO* adaptativo disfruta de las propiedades del oráculo, es decir, funciona tan bien como si el verdadero modelo subyacente se diera de antemano. Además, el *LASSO* adaptativo se puede resolver con el mismo algoritmo eficiente para resolver el *LASSO*.

2.3.3. LASSO grupal

Cuando existe de variables predictores que pertenecen a grupos definidos, por ejemplo una colección de variables indicadores para representar los niveles de un predictor categórico. En estas situación, Hastie et al. [2001] menciona que es conveniente reducir y seleccionar a los miembro de un grupo juntos, lo cual es el concepto principal de *LASSO* agrupado.

Suponga que los p predictores se dividen en L grupos con p_l el número del grupo l. Para facilitar la notación, usamos una matriz X_l para representar los predictores correspondientes en el $l-$ésimo grupo, con el vector de coeficientes correspondiente β_l. Entonces la función a minimiza con este plantamento es

$$\hat{\beta}^{lasso_grupo}(\beta) = \arg\min_{\beta} \left\{ ||\mathbf{y} - \mathbf{X}\beta||^2 + \sqrt{\mathbf{p}}\lambda||\beta||_2 \right\} \tag{14}$$

donde $\sqrt{p} = (p_1 \quad \cdots \quad p_l)$ son los diferentes tamaños de los grupos.

3. Resultados numéricos

Para comparar los efectos entre los distintos métodos de regularización y el efecto del parámetro de regularización, realizamos dos problemas de regresión lineal múltiple en dónde existen evidencia de multicolinealidad en donde el enfoque de mínimos cuadrados no tiene buen rendimiento.

3.1. Ejemplo: Datos de delitos (diferencias de los métodos de regularización y el parámetro de regularización)

El objetivo de este análisis es poder predecir el el número de crímenes violentos a partir de un conjunto de aspectos poblacionales: porcentaje de la población considerada urbana y el ingreso familiar medio, y la participación de las fuerzas del orden público, como el número per cápita de agentes de policía y el porcentaje de agentes asignados a las unidades de drogas, etc. Estos datos fueron obtenidos del repositorio de Aprendizaje Automático de UCI [Dua and Graff, 2017]. Se tiene 122 variables explicativas y una variable respuesta crímenes violentos, los cuales están normalizaron en el rango decimal $0{,}00 - 1{,}00$ [Faraway, 2006].

Las correlación de estas variables es alta (ver Figura 2), lo que supone un problema a la hora de emplear modelos de regresión lineal. Por lo que incita a utilizar una alternativa de regularización. Para medir el impacto de la regularización primero implementamos la estimación por OLS como y *baseline* de este ejercicio, utilizando como métrica el error cuadrático medio (RMSE).

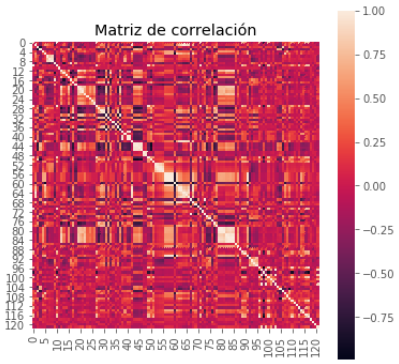

Figura 2: Heatmap de la matriz de correlación entre las 122 variables predictoras.

Para este ejercicio utilizamos un conjunto de prueba y de entrenamiento. El modelo de regresión por OLS arroja un RMSE de 0.3097 en el conjunto de prueba. Posteriormente ajustamos modelos de regresión con regularización de *ridge*, *LASSO* y *elastic net*, y para determinar el mejor valor de λ utilizamos validación cruzada utilizando 10 folds, los valores probados fueron considerando los números 200 espaciados uniformemente en una escala logarítmica entre -4 y 1.

Podemos notar claramente, que cuando se utilizan la regularización *ridge* que entre más grande sea el valor de λ hace que los coeficientes se acerquen a cero, pero no necesariamente sean cero. Que a diferencia con la regularización $LASSO$ vemos que entre más grande sea λ de igual manera los coeficientes se van acercando a cero hasta llegar a cero.

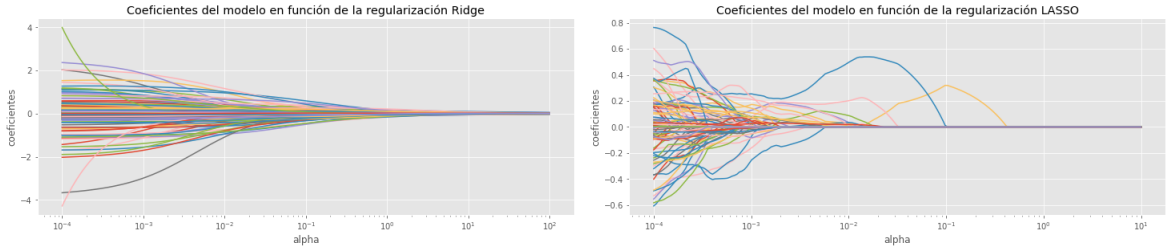

Figura 3: Comparación de los coeficientes entre rigde y *LASSO* utilizando diferentes valores de λ

Posteriormente utilizando los mejores valores de λ obtenidos por validación cruzada para cada uno de las regularizaciones calculamos el RMSE en el conjunto de prueba (ver Figura 4). Podemos notar que los tres métodos de regularización tienen mejores RMSE en comparación con el ajuste de OLS, estos resultados tienen sentido debido a que los datos tenías multicolinealidad.

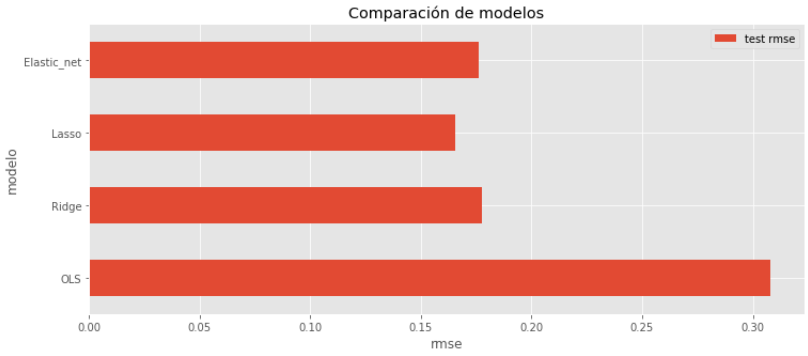

Figura 4: Comparación del RMSE para los distintos ajustes

El mejor modelo fue utilizando regularización *LASSO*, una gran diferencias con los dos métodos de regularización es que este método considera solamente 22 variables predictoras y el resto de los coeficientes son valores iguales a 0

(ver Figura 5). De igual manera, vemos que cuando λ crece esto implica que el número de coeficientes iguales a cero aumenta.

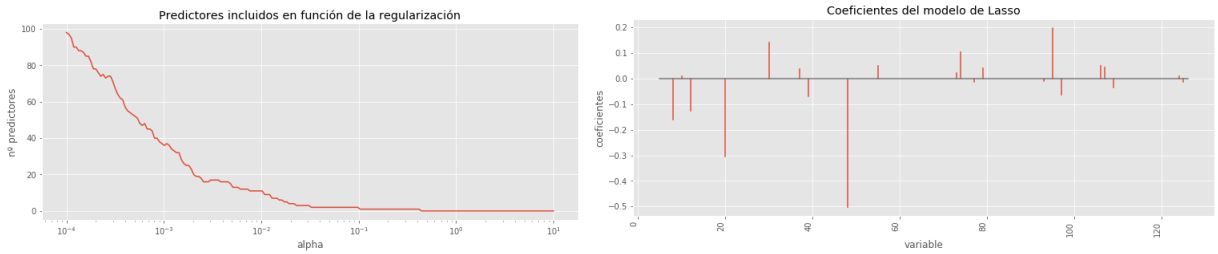

Figura 5: Número de coeficientes diferentes de cero y coeficientes de la mejor configuración en *LASSO*

Este resultado es muy importante cuando se desea considerar el comportamiento de los datos a analizar. Es decir, considerando que en regresión hay dos objetivos en un análisis (predictor y descriptivo), en ocasiones es preferible modelos en donde se tenga menos variables debido a que esto tiene más interpretación que modelos que tienen más variables. Obviamente esta ganancia en la interpretación esta relacionada con el aumente en el error de predicción.

3.2. Ejemplo: Contenido de grasa (efecto de escalas)

Este ejercicio y los datos fue propuesto por Boehmke and Greenwell [2019], con una pequeña modificación para observar el impacto de estandarizar los datos en los métodos de regularización. Se tiene un departamento de calidad de una empresa de alimentación se encarga de medir el contenido en grasa de la carne que comercializa. Este estudio se realiza mediante técnicas de analítica química, un proceso relativamente costoso en tiempo y recursos. Una alternativa que permitiría reducir costes y optimizar tiempo es emplear un espectrofotómetro (instrumento capaz de detectar la absorbancia que tiene un material a diferentes tipos de luz en función de sus características) e inferir el contenido en grasa a partir de sus medidas.

Antes de dar por válida esta nueva técnica, la empresa necesita comprobar qué margen de error tiene respecto al análisis químico. Para ello, se mide el espectro de absorbancia a 100 longitudes de onda (variables predictoras) en 215 muestras de carne (número de registros), cuyo contenido en grasa se obtiene también por análisis químico. Entonces el objetivo es predecir el contenido en grasa a partir de los valores dados por el espectrofotómetro.

Entonces para medir el efecto que tiene estandarizar los datos en los métodos de regularización, multiplicamos los datos de los últimas 20 variables predictoras por un factor de expansión de 10000. Posteriormente, veamos que la matriz de correlación de nuestras diferentes longitudes de ondas (ver Figura 6) presenta un efecto muy fuerte de multicolinealidad, por lo que en teoría el método de regularización se espera que tenga mejores resultados con los otros métodos de regularización.

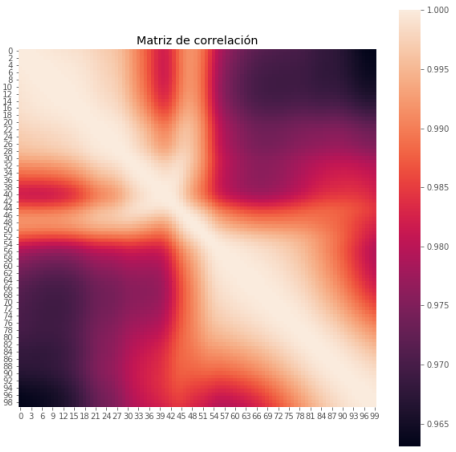

Figura 6: Heapmap de la matriz de correlación de nuestras longitudes de onda

Utilizamsos validación cruzada para obtener el mejor valor de λ de los métodos de regularización con los datos originales y a los datos estandarizados. Posteriormente, calculamos el RMSE en nuestro conjunto de prueba (ver Figura 7). Se observa claramente que el efecto de ajustar los métodos de regularización de los datos originales y datos estandarizados. Además, se observa que la regularización *ridge* presenta los mejores resultados.

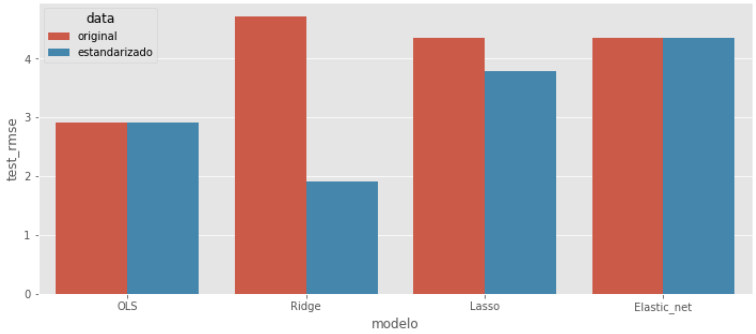

Figura 7: Comparación del RMSE en estandarizados y originales.

Este resultado demuestra la importancia de utilizar los datos estandarizados, este supuesto debe de ser considerado a la hora de considerara métodos de regularización.

4. Extensión de regularización para el caso multivariado

En esta sección tienen el objetivo de extender el método de regularización para el caso de regresión lineal multivariada, con el fin de ver la facilidad de implementación en otros modelos distintos a la regresión lineal. Además se presenta un ejemplo numérico en donde se observa la importancia que tiene los métodos de regularización en contra del enfoque clásico.

4.1. Regresión lineal multivariada.

La regresión multivariada es una generalización del modelo de regresión clásico pero considerando $q > 1$ variables respuestas. Es decir, sea $\mathbf{X}$ la matriz de las variables independientes $n \times p$, $\mathbf{Y}$ la matriz de las variables independientes $n \times q$ y sea $\mathbf{E}$ la matriz de error aleatorio $n \times q$. Entonces el modelo de regresión multivariada es

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \tag{15}$$

donde $\mathbf{B}$ es la matriz de coeficientes de regresión $p \times q$. Si $q = 1$ el modelo se simplifica al problema de regresión clásico donde $\mathbf{B}$ es el vector de coeficientes de regresión $p-$dimensional. Consideremos que las $\mathbf{X}$ y $\mathbf{Y}$ están centradas para facilitar los cálculos.

La función de verosimilitud logarítmica negativa de $(\mathbf{B}, \Omega)$, donde $\Omega = \Sigma^{-1}$ se puede expresar como

$$g(\mathbf{B}, \Omega) = \left[\frac{1}{n}(\mathbf{Y} - \mathbf{XB})^T(\mathbf{Y} - \mathbf{XB})\Omega\right] - \log(\det(\Omega)) \tag{16}$$

Es fácil ver (derivando con respecto a $\mathbf{B}$ e igualando a 0, y simplificando), que el estimador de máxima verosimilitud de $\mathbf{B}$ es

$$\hat{\mathbf{B}}^{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{17}$$

Lo anterior es equivalente a realizar las estimaciones de $\mathbf{B}$ utilizando mínimos cuadrados ordinarios de forma separada para cada una de las q variables de respuestas y no este implica que no dependan de Ω.

4.2. Estimación

De lo anterior podemos observar dos enfoques distintos cuando se considera una regresión multivariada. Lo primero es considerar que los datos no están correlacionados, es decir, que no dependan de Ω y el otro enfoque es considerar la matriz de covarianzas de los errores. Pero en ambos métodos agregamos un parámetro de regularización.

4.2.1. Regularized Multivariate regression for identifying Master Predictors (REMMAP)

El problema de minimización con restricciones propuesto por [Peng et al., 2010], considera una optimización L1 y L2, considera tambien que las q respuestas observadas no estan correlacionadas, la función a optimizar es representada como:

$$L(\hat{\mathbf{B}}, \mathbf{X}, \mathbf{Y}) = \underset{B}{argmin}\{||(\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB})||^2 + \lambda_1 \sum_j \sum_k |b_{jk}| + \lambda_2 \sum_j \sum_k (b_{jk})^2\}$$

$$L(\hat{\mathbf{B}}, \mathbf{X}, \mathbf{Y}) = \underset{\mathbf{B}}{argmin}\{\frac{1}{2}\sum_{k=1}^{q}(\mathbf{y}_k - \mathbf{xB}_k)^2\} + \lambda_1 \sum_{k=1}^{q} |\mathbf{B}_k| + \lambda_2 \sqrt{\sum_{k=1}^{q}(\mathbf{B}_k)^2}\}$$

Considerando $\lambda_2 = 0$, pues solo trabajaremos con la restricción de norma L1, tenemos:

$$L(\hat{\mathbf{B}}, \mathbf{X}, \mathbf{Y}) = \underset{\mathbf{B}}{argmin}\{\frac{1}{2}\sum_{k=1}^{q}(\mathbf{y}_k - \mathbf{xB}_k)^2\} + \lambda_1 \sum_{k=1}^{q} |\mathbf{B}_k|$$

La metodología de regresión *LASSO* y la búsqueda del resultado utilizando un algoritmo de descenso coordinado, fue tratado para el caso de la regresión múltiple específicamente en [Rothman et al., 2010], donde la actualización de la función se realiza de manera idéntica considerando una única variable respuesta k, además observemos que dado que los datos están estandarizados la norma del vector seria igual a n, es decir $||\mathbf{X}_{j_0}||_2^2 = n$. De esto ultimo observamos que el resultado de aplicar el algoritmo seria k regresiones *LASSO* (ver **Algoritmo 1**).

4.2.2. Regresión multivariada con estimación de covarianza (MRCE)

Rothman et al. [2010] plantea un procedimiento para construir un estimador de una matriz de coeficientes de regresión multivariada que tenga en cuenta la correlación de las variables de respuesta. Básicamente propone un estimador para $\mathbf{B}$ que considera los errores correlacionados utilizando la verosimilitud normal. Considera dos penalizaciones a la verosimilitud logarítmica negativa (16) para construir un estimador disperso $\mathbf{B}$ que dependa de $\Omega = \{\omega_{j'j}\}$,

$$(\hat{\mathbf{B}}, \hat{\Omega}) = \arg\min_{\mathbf{B},\Omega} \left\{ g(\mathbf{B}, \Omega) + \lambda_1 \sum_{j' \neq j} |\omega_{j'j}| + \lambda_2 \sum_{j=1}^{p} \sum_{k=1}^{q} |b_{jk}| \right\} \tag{18}$$

donde $\lambda_1 \geq 0$ y $\lambda_2 \geq 0$ son los parámetros de regularización. Se considera una penalización del *LASSO* en las entradas fuera de la diagonal de la covarianza del error inverso Ω por dos razones.

1. Se asegura una solución óptima para Ω tenga un valor finito cuando hay más respuestas que muestras ($q > n$).
2. Tiene un efecto de reducir el número de parámetros en la covarianza del error inverso, lo cuál es útil cuando q es grande. [Rothman et al., 2008].

Y la penalización *LASSO* en $\mathbf{B}$ introduce escases en $\hat{\mathbf{B}}$, que reduce el número de parámetros en el modelo y proporciona una interpretación a los coeficientes. Además, esta penalización implica una solución óptima para $\mathbf{B}$ en función de Ω. Cabe resaltar, que sin una penalización en $\mathbf{B}$ (es decir, $\lambda_2 = 0$) la solución óptima para $\mathbf{B}$ es siempre $\hat{\mathbf{B}}^{OLS}$ [17].

El problema de optimización en (18) no es convexo, sin embargo, resolver $\mathbf{B}$ o Ω con el otro parámetro fijo hace al problema convexo. Entonces, si dejamos fijo $\mathbf{B}$ en un punto $\mathbf{B}_0$ el problema de optimización para Ω se convierte a

$$\hat{\Omega}(\mathbf{B}_0) = \arg\min \left\{ tr\left(\hat{\Sigma}_R \Omega\right) - \log(\det \Omega) + \lambda_1 \sum_{j' \neq j} |\omega_{j'j}|) \right\}, \tag{19}$$

donde $\hat{\Sigma}_R = \frac{1}{n}(\mathbf{Y} - \mathbf{XB}_0)^T(\mathbf{Y} - \mathbf{XB}_0)$. Este problema es conocido como el problema de estimación de covarianza considerando una penalización L_1. Friedman et al. [2008] plantea el algoritmo de *LASSO* gráfico para resolver el problema de optimización 19. Se abordará con más detalle este algoritmo en las secciones posteriores.

Por otro lado, resolver 18 fijando Ω en un punto elegido Ω_0 transforma el problema a optimizar

$$\hat{\mathbf{B}}(\Omega_0) = \arg\min \left\{ tr\left(\frac{1}{n}(\mathbf{Y} - \mathbf{XB})^T(\mathbf{Y} - \mathbf{XB})\Omega_0 + \lambda_2 \sum_{j=1}^{p} \sum_{k=1}^{q} |b_{jk}| \right) \right\} \tag{20}$$

Una solución para el problema anterior es utilizar un descenso de coordenadas cíclicas. Rothman et al. [2010] resume en el procedimiento de optimización como se describe en el **Algoritmo 2** (ver Anexos). Se utiliza la estimación de mínimos cuadrados penalizados por rigde $\hat{\mathbf{B}}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda_2 I)^{-1}\mathbf{X}^T\mathbf{Y}$ para escalar nuestra prueba de convergencia de párametroas, ya que siempre está bien definida (incluso cuando $p > n$). La derivación completa del algoritmo se puede ver en la **Sección 6.4**.

Considerando lo anterior, podemos resumir la resolución del problema de optimización (18) usando el descenso de coordenadas en bloque, es decir, iteramos minimizando con respecto a $\mathbf{B}$ y minimizando con respecto a Ω. El **Algoritmo 3** (ver Anexos) usa el descenso de coordenadas por bloques para calcular una solución loca para (18).

4.3. Ejemplos: Datos sintéticos (efecto $n > p$)

Recordando que los estimadores de OLS tienen problemas cuando existen un número mayor de predictores que de observaciones, entonces para verificar este hecho lo probamos en esta extensión de regularización.

4.4. Evaluación de los modelos con datos sintéticos

El conjunto de datos sintéticos fue generado con la función $make_regression()$ de la librería de Scikit-learn [Pedregosa et al., 2011]. Consideramos diferentes parámetros de la función anterior: $n\_samples(n) = [100, 20]$, $n\_features(p) = [20, 100]$, y $n\_targets(q) = [2, 5]$. Esto con el objetivo de observar el efecto que tiene las dimensiones de diferentes datos en nuestros modelos. Consideramos partir el conjunto de datos original, en dos conjuntos uno de prueba y otro de entrenamiento. Además de que nuestro conjunto de datos, consideramos una estandarización debido a los supuestos que se tienen en los modelos.

Para el caso en el que $n > p$, los rendimientos de regularización no son buenos en comparación con el enfoque clásico de OLS (ver Figura 10). Esto tiene completamente sentido debido a que nuestros datos sintéticos son creados sin considerar correlaciones entres las variables y por lo tanto no existe ninguna incumplimiento en los supuestos de OLS. Además, vemos que MRCE es muy parecido a OLS esto debido a la relación que existe cuando $\lambda = 0$.

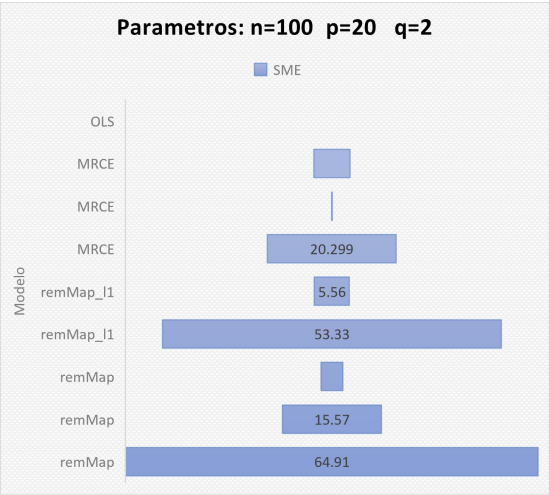

Figura 8

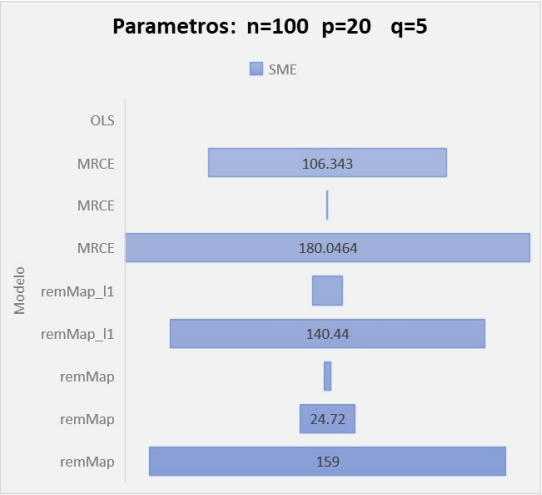

Figura 9

Figura 10: MSE considerando distintos modelos, con $n > p$.

Ahora, si consideramos cuando $n < p$ (ver Figura 13) notamos que los mejores predictores son ocupando algún método de regularización lo cual tienen sentido debido a que como existe más variables predictoras que registros se esta incumpliendo el supuesto de OLS y por ende los resultados no serían adecuados.

Es decir, pudimos observar el efecto que tiene utilizar los métodos de regularización cuando se incumple algún supuesto de mínimos cuadrados.

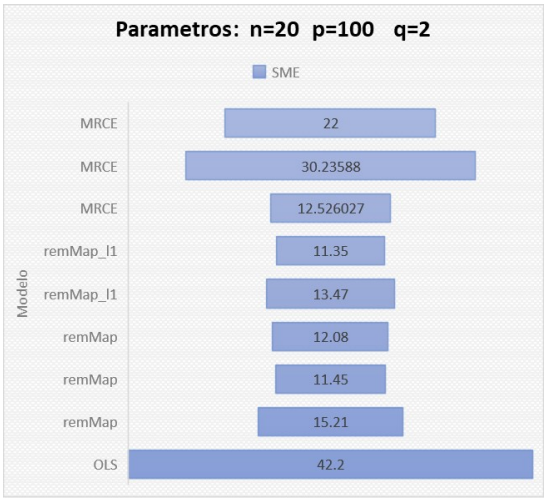

Figura 11

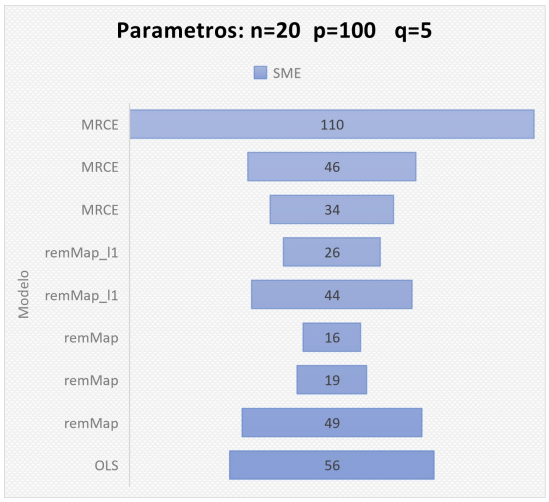

Figura 12

Figura 13: MSE considerando distintos modelos, con $n < p$.

5. Conclusiones

Se presentaron de forma concreta los principales métodos de regularización (*ridge, LASSO y elastic net*) en regresión lineal multiple , y la importancia que tienen en comparación del enfoque clásico de mínimos cuadrados. Además, estos enfoques se pueden extender fácilmente a otro tipo de modelos: modelos lineales generalizados (regresión logística o poisson), modelos de supervicencia, etc.

Resaltamos que la regularización *ridge* presenta mejores resultados cuando existe multicolinealidad en los datos esto se observa claramente en el ejemplo de *contenido de grasa*, en donde la mayoría de las variables estaba correlaciones. Por otro lado, la regularización *LASSO* presenta mejores resultados cuando se presenta un grupo de variables relevantes, y por consecuente es muy útil para poder interpretar los modelos ya que podemos este método restringe a los coeficientes a cero. Además, de que este método es ampliamente utilizada en *feature selection* para un modelo de aprendizaje aútomatico más complejo. Y por último, la principal ventaja de la penalización de la red elástica es que permite una regularización eficaz a través de la penalización de *ridge* con las características de selección de características de la penalización del *LASSO*. Debido a que los métodos de regularización son aplicados a el tamaño del coeficiente, es de suma importancia estandarizar los datos antes de aplicar un método de regularización, esto se observo en el ejemplo de *contenido de grasa* expuesto.

Además, desarrollamos la extensión de regularización para un modelo de regresión multivariada. En este caso presentamos el efecto que tiene los métodos de regularización cuando se incumple el supuesto de que existe un número de variables predictoras que número de registros, lo cuál es muy común en problemas de genoma o medicina.

Es claro que este tipo de enfoque es muy importante no solo en métodos de regresión, si no en otros modelos. Ya que además de que es una alternativa cuando se violan cierto tipos de supuestos, también sirve como métodos para evitar *overfitting* en problemas de aprendizaje automático.

6. Anexos

6.1. Códigos

Todos las implementaciones se realizaron en el lenguaje de programación Python, en el sistema x86_64, Ubuntu. Todos los códigos utilizados para estos resultados se pueden encontrar en mi página personal de Git- gub: Enriquesec. En el repositorio Ciencia de Datos/Tareas/Proyecto_final/.

6.2. Pseudo-algoritmos

Aquí se presentan los pseudo-algoritmos de las metodologías propuestas de REMMAP y MCRE, en la sección 4.

Algorithm 1: REMMAP [Peng et al., 2010]

Input: $Y_{n\times q}$,$X_{n\times p}$
Result: $B_{p\times q}$

Inicializamos parámetros, $[B = 0_{p\times q}, ...]$;
while *True* **do**
  Para j=1,...p ; k=1,...,q

$$B_{j_0,k} = (|\mathbf{X}_{j_0}^T \tilde{\mathbf{Y}}_k| - \lambda_1)_+ \frac{sign(\mathbf{X}_{j_0}^T \tilde{\mathbf{Y}}_k)}{||\mathbf{X}_{j0}||_2^2}$$

  if *B no cambia* **then**
    break;
    return(B);

Algorithm 2: Descenso de coordenadas cíclicas [Rothman et al., 2010].

Input: $\mathbf{Y}_{n\times q}$,$\mathbf{X}_{n\times p}, \Omega_{p\times p}, \lambda_2$ y ϵ
Result: $\hat{B}_{p\times q}$

$S = \mathbf{X}^T\mathbf{X}$
$H = \mathbf{X}^T\mathbf{Y}\Omega$
$\hat{\mathbf{B}}^{rigde} = (\mathbf{X}^T\mathbf{X} + \lambda_2 I)^{-1}\mathbf{X}^T\mathbf{Y}$.
while $\sum |\hat{\mathbf{B}}^{(m)} - \hat{\mathbf{B}}^{m-1}| > \epsilon \sum |\hat{\mathbf{B}}^{rigde}|$ **do**
  for *r=1,...p* **do**
    for *c=1,...,q* **do**
      $\mu_{rc} = \sum_{j=1}^{p}\sum_{k=1}^{q} \hat{b}_{jk}^{(m)} s_{rj} w_{kc}$
      $\hat{b}_{rc}^{(m)} = sign\left(\hat{b}_{rc}^{(m)} + \frac{h_{rc}-\mu_{rc}}{s_{rr}\omega_{cc}}\right)\left(\left|\hat{b}_{rc}^{(m)} + \frac{h_{rc}-\mu_{rc}}{s_{rr}\omega_{cc}}\right| - \frac{n\lambda_2}{s_{rr}\omega_{cc}}\right)_+$
return$\left(\hat{\mathbf{B}}^{(m)}\right)$

Algorithm 3: MRCE [Rothman et al., 2010].

Input: $\mathbf{Y}_{n\times q}$,$\mathbf{X}_{n\times p}, \lambda_1,\ \lambda_2, \epsilon$.
Result: $\hat{B}_{p\times q}$

Inicializamos
$\hat{\mathbf{B}}^{(0)} = 0$
$\hat{\Omega}^{(0)} = \hat{\Omega}(\hat{\mathbf{B}}_{(0)})$
while $\sum |\hat{\mathbf{B}}^{(m)} - \hat{\mathbf{B}}^{m-1}| > \epsilon \sum |\hat{\mathbf{B}}^{rigde}|$ **do**
  1. Calcular $\hat{\mathbf{B}}^{m+1} = \hat{\mathbf{B}}(\hat{\Omega}^{(m)}$ resolviendo 20 utilizando el **Algoritmo 2**.
  2. Calcular $\hat{\Omega}^{(m+1)} = \hat{\Omega}(\hat{\mathbf{B}}^{(m+1)})$ resolviendo 19 usando el algoritmo de LASSO gráfico.
return$\left(\hat{\mathbf{B}}^{(m)}\right)$

6.3. Algoritmo de LASSO gráfico

Friedman et al. [2008] describe este método para máximizar el problema de optimización **(19)**. Sea W el estimador para Σ (matriz de covarianza poblacional). Se puede mostrar que se puede resolver el problema optimizando cada fila y la columna correspondiente a W en una forma de descenso de coordenadas de bloque. Partimos W y S,

$$W = \begin{pmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{pmatrix}, \quad S = W = \begin{pmatrix} S_{11} & s_{12} \\ s_{12}^T & s_{22} \end{pmatrix}$$

donde S es la matriz de correlación empírica. Entonces se puede mostrar que

$$w_{12} = \arg\min_{y} \left\{ y^T W_{11}^{-1} y : ||y - s12||_\infty \geq p \right\}. \tag{21}$$

Lo anterior es un programa cuadrático con restricciones de caja que resuelven usando un procedimiento de punto interior. Pero de igual manera se puede mostrar usando dualidad convexa que el problema (21) es equivalente a resolver el problema dual

$$\min_{\beta} \left\{ \frac{1}{2} |W_{11}^{-1/2}\beta - b|^2 + \lambda|\beta|_1 \right\} \tag{22}$$

donde $b = W_{11}^{-1/2} s_{12}$. Si β resuelve (22) entonces $w_{12} = W_{11}\beta$ resuelve (21). Además es sencillo ver que las soluciones en (19) son equivalentes a resolver (22). Para resolver (22) usamos W_{11} y s. Luego actualizamos w y corremos todas las variables hasta la convergencia. Consideramos que la solución de $w_{ii} = s_{ii} + \lambda$ para todo i. Este algoritmo se le conoce como algoritmo *LASSO* gráfico (ver **Algoritmo** 4).

Algorithm 4: LASSO gráfico [Friedman et al., 2008]

Input: S, λ y ϵ.
Result: W

1 *Inicializamos*
2 $W = S + \lambda\rho$
3 **while** $|W - \{diagonal\}| > \epsilon|S - \{diagonal\}|$ **do**
4 **for** *j=1,2,*$\cdots, p, 1, 2, \cdots, p, \cdots$ **do**
5 *Resolver el problema de LASSO en (22). Esto regresa un vector solución* $\hat{\beta}$ *de tamaño* $p-1$*, por lo que imputamos el renglón y la columna de* W *usando* $w_{12} = W_{11}\hat{\beta}$.
6 return$\left(W^{-1}\right)$

Para resolver el paso 5 del (**Algoritmo** 4) consideramos un descenso coordinado. Sea $V = W_{11}$ y $u = s_{12}$, entonces actualizamos β_j de la forma

$$\hat{\beta}_j = S(u_j - \sum_{k \neq j} V_{jk}\beta_k, \lambda)/V_{jj} \tag{23}$$

para $j = 1, 2, \cdots, p, 1, 2, \cdots, p, \cdots$. Donde S es el operador soft-threshold:

$$S(x, y) = sign(x)(|x| - t)_+.$$

Para más detalle de este algoritmo consulte Friedman et al. [2008], ahí se presentan las demostraciones más a detalle y más referencias sobre problemas similares.

6.4. Descenso de coordenadas cíclicas

El descenso de coordenadas es un algoritmo de optimización que minimiza sucesivamente a lo largo de las direcciones de las coordenadas para encontrar el mínimo de una función. En nuestro problema, tenemos la función objetivo para Ω fija en Ω_0 es

$$f(\mathbf{B}) = g(\mathbf{B}, \Omega) + \lambda_2 + \sum_{j=1}^{p}\sum_{k=1}^{q} |b_{jk}| \tag{24}$$

Se puede resolver para $\mathbf{B}$ utilizando un descenso de coordenadas cíclicas. Expresamos las derivadas direccionales como

$$\frac{\partial f+}{\partial \mathbf{B}} = \frac{2}{n}\mathbf{X}^T\mathbf{X}\mathbf{B}\Omega - \frac{2}{n}\mathbf{X}^T\mathbf{Y}\Omega + \lambda_2 1_{(b_{ij}>0)} - \lambda_2 1_{(b_{ij}<0)} \tag{25}$$

$$\frac{\partial f-}{\partial \mathbf{B}} = -\frac{2}{n}\mathbf{X}^T\mathbf{X}\mathbf{B}\Omega + \frac{2}{n}\mathbf{X}^T\mathbf{Y}\Omega - \lambda_2 1_{(b_{ij}>0)} + \lambda_2 1_{(b_{ij}<0)} \tag{26}$$

donde $1_{(.)}$ es un indicador. Si definimos a $S = \mathbf{X}^T\mathbf{X}$ y $H = \mathbf{X}^T\mathbf{Y}\Omega$ y $\mu_{rc} = \sum_{j=1}^{p}\sum_{k=1}^{q} b_{jk}s_{rj}w_{kc}$, entonces considerando un solo parametro b_{rc} tenemos que las derivadas direccionales son

$$\frac{\partial f+}{\partial b_{rc}} = \mu_{rc} - h_{rc} + n\lambda_2 1_{(b_{ij}>0)} - n\lambda_2 1_{(b_{ij})<0},$$

$$\frac{\partial f-}{\partial b_{rc}} = -\mu_{rc} + h_{rc} - n\lambda_2 1_{(b_{ij}>0)} + n\lambda_2 1_{(b_{ij})<0}.$$

Sea b_{rc}^0 nuestra iteración actual, entonces míminizar lo anterior es equivalente a resolver $\hat{b}_{rc}^*$

$$\hat{b}_{rc}^* s_{rr}\omega_{cc} - b_{rc}^0 s_{rr}\omega_{cc} + \mu_{rc} - h_{rc} = 0.$$

Por lo anterior, es sencillo ver que implica que

$$\hat{b}_{rc} = sign\left(\hat{b}_{rc}^*\right)\left(\left|\hat{b}_{rc}^*\right| - \frac{n\lambda_2}{s_{rr}\omega_{cc}}\right)_+.$$

Si $\hat{b}_{rc}^* = 0$ tiene un valor de cero, entonces tanto la parte de pérdida como la de penalización de la función objetivo se minimizan y el párametro permance en 0. Por lo que podemos escribir esta solución como

$$\hat{b}_{rc} = sign\left(\hat{b}_{rc}^0 + \frac{h_{rc} - \mu_{rc}}{s_{rr}\omega_{cc}}\right)\left(\left|\hat{b}_{rc}^0 + \frac{h_{rc} - \mu_{rc}}{s_{rr}\omega_{cc}}\right| - \frac{n\lambda_2}{s_{rr}\omega_{cc}}\right)_+.$$

Referencias

Bradley C. Boehmke and Brandon M. Greenwell. Hands-on machine learning with r. 2019.

Robert Tibshirani. The lasso method for variable selection in the cox model. In *Statistics in Medicine*, pages 385–395, 1997.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12: 55–67, 1970.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58: 267–288, 1996.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–451, 2004. ISSN 00905364.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006. URL `https://EconPapers.repec.org/RePEc:bes:jnlasa:v:101:y:2006:p:1418-1429`.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

J.J. Faraway. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC Press, 2006.

Jie Peng, Ji Zhu, Anna Bergamaschi, Wonshik Han, Dong-Young Noh, Jonathan R. Pollack, and Pei Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics*, 4(1):53 – 77, 2010. doi:10.1214/09-AOAS271. URL `https://doi.org/10.1214/09-AOAS271`.

Adam J. Rothman, Elizaveta Levina, and Ji Zhu. Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962, 2010. doi:10.1198/jcgs.2010.09188. URL `https://doi.org/10.1198/jcgs.2010.09188`. PMID: 24963268.

Adam J. Rothman, Peter J. Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2(none), Jan 2008. ISSN 1935-7524. doi:10.1214/08-ejs176. URL `http://dx.doi.org/10.1214/08-EJS176`.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, July 2008. ISSN 1465-4644, 1468-4357. doi:10.1093/biostatistics/kxm045. URL `http://biostatistics.oxfordjournals.org/content/9/3/432`.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.