REGULARIZACIÓN EN UN MODELO CON MULTIPLES RESPUESTAS

PRESENTA:

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

Victor Manuel Martinez Santiago
CIMAT
UNIDAD MONTERREY
victor.santiago@cimat.mx

3 de junio de 2021

ABSTRACT

El objetivo general de este trabajo es determinar una función $f(\mathbf{x})$ que prediga las q salidas de un vector de entrada x, mediante un problema de optimización a partir de datos de entrada $X_{n\times p}$ y $Y_{n\times q}$, más un parámetro de regularización.

Para resolver ese problema, consideramos el enfoque estadístico de regresión multivariada. El problema ha sido abordado inicialmente considerando un enfoque donde las variables respuestas no están correlacionadas visto como un modelo de regresión multivariado múltiple, también se puede interpretar que los errores no están correlacionados. Un segundo análisis es realizado por Rothman et al. [2010], en el cual considera que los errores están correlacionados, dando lugar a la métodología de regresión multivariada estimando la covarianca (MRCE).

En la primera parte del presente se discute la importancia del presente trabajo, posteriormente en la sección de Metodología se realiza el planteamiento de ambos enfoques así como sus consideraciones. En la parte de Resultados se presentan un análisis aplicando ambos enfoques a un conjunto de datos sintéticos. Y por último, Concluimos la relevancia de estos dos enfoques.

Keywords Regularización · Regresión · Multivariado · LASSO · Optimización

1. Introducción

1.1. Definición del problema a resolver.

Sea $X \in \mathbb{R}^{n\times p}$ una matriz de $n \times p$ con n ejemplos de entrenamiento y p características. $Y \in \mathbb{R}^{n\times q}$ tal que cada fila representa q respuestas. El objetivo principal es estimar una función $f(\mathbf{x})$ que prediga las k salidas de un vector de entrada x, mediante un problema de optimización. Y posteriormente programar el algoritmo en Python para determinar la función apartir de un conjunto de datos X, Y y un parametro de regularización λ.

1.2. Enfoque general de la solución.

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

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \tag{1}$$

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

$$g(\mathbf{B}, \Omega) = \left[\frac{1}{n}(\mathbf{Y} - \mathbf{XB})^T(\mathbf{Y} - \mathbf{XB})\Omega\right] - \log(\det(\Omega)) \tag{2}$$

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

$$\hat{\mathbf{B}}^{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{3}$$

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.

2. Metodología

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.

2.1. Análisis del planteamiento del problema

El problema definido (2) también se puede modificar para agregar un parámetro de regularización. Denotando a $C(\mathbf{B})$ como el parámetro de regularización en función de $\mathbf{B}$, es decir,

$$\hat{\mathbf{B}} = \underset{\mathbf{B}}{argmin}\ ||\mathbf{Y} - \mathbf{XB}||_2^2 = \underset{\mathbf{B}}{argmin}\ [(\mathbf{Y} - \mathbf{XB})^T(\mathbf{Y} - \mathbf{XB})] \tag{4}$$

$$\text{sujeto a:}\quad C(\mathbf{B}) \leq t$$

Y se puede mostrar que el problema anterior es equivalente a resolver el problema (ver [Peng et al., 2010]).

$$\hat{\mathbf{B}} = \underset{\mathbf{B}}{argmin}\ \{\ \ ||\mathbf{Y} - \mathbf{XB}||_2^2 - \lambda(C(\mathbf{B}))\ \} \tag{5}$$

Argumentaremos un poco mas al respecto, para ello, observamos que el lagrangiano del problema (4) es:

$$L(\mathbf{B}, \mu) = ||\mathbf{Y} - \mathbf{XB}||_2^2 + \mu(C(\mathbf{B}) - t)$$
Mas las condiciones de KKT

Observaremos que el gradiente del lagrangiano que se obtiene a partir de (4) es el mismo gradiente del problema formulado en (5).

Ahora bien, durante el presente trabajo, los dos modelos que se abarcaran, consideran esencialmente una restricción L1 (Norma L1).

$$C(B) = \sum_{j=1}^{p}\sum_{k=1}^{q}|b_{jk}| = \sum_{k=1}^{q}|B_k|$$

2.2. REMMAP(REgularized Multivariate regression for identifying Master Predictors)

En un modelo de regresión dado como :

$$y_k = \mathbf{X}_i\mathbf{B}_k + \epsilon \qquad i = 1, ..., n \;\; j = 1, ..., p \;\; k = 1, ..., q$$

$$\mathbf{y}_k = \mathbf{x}\mathbf{B}_k + \epsilon \qquad i = 1, ..., n \;\; j = 1, ..., p \;\; k = 1, ..., q$$

Donde n es el numero de observaciones, p es numero de regresores y q es el numero de respuestas.

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:

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

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

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

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

La actualización de los parametros se realiza considerando un descenso coordinado para cada elemento en b_{jk} respuesta, este proceso se podria realizar hasta converger o bien una cantidad de iteraciones t, lo cual nos llevara a la conclusión de que el algoritmo es de orden O(TPQ).

De acuerdo con (Peng et al, 2010) la actualización de cada elemento $b_{j.k}$ se realiza tal que:

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

$$Donde \qquad \tilde{\mathbf{Y}}_k = \mathbf{Y}_k - \sum_{j \neq j_0} \mathbf{X}_j \mathbf{B}_{jk}$$

En palabras $\tilde{\mathbf{Y}}_k$ es el residual que se obtiene de ajustar los pesos sin considerar la j-esima variable, la cual se esta optimizando, esta es la escencia del gradiente por descenso coordinado.

La metodologia de regresion LASSO y la busqueda del resultado utilizando un algoritmo de descenso coordinado, fue tratado para el caso de la regresión multiple especificamente en [Rothman et al., 2010], donde la actualización de la función se realiza de manera identica considerando una unica variable respuesta k, ademas observemos que dado que los datos estan estandarizados la norma del vector seria igual a n, es decir $||\mathbf{X}_{j_0}||_2^2 = n$

Utilizando un algoritmo sencillo de optimización, el algoritmo en el caso de regresión multiple tendria un costo computacional de orden O(TP).

De esto ultimo observamos que el resultado de aplicar el algoritmo seria k regresiones LASSO .

El pseudocodigo de la función es:

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

Input: $Y_{n\times q}$,$X_{n\times p}$
Result: $B_{p\times q}$
Inicializamos parametros, $[B = 0_{p\times q}, ...]$;
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);

Caso cuando $\lambda_2 \neq 0$

Mencionamos brevemente que cuando $\lambda_2 \neq 0$, (Peng et al, 2010) muestran que la actualización se realiza tal que la $B_{j0,k}$ ya calculada que llamaremos $B_{j0,k}^{lasso}$, es evaluada nuevamente en una función donde y toma los siguientes valores dependiendo de dos casos

$$\hat{B}_{j0,k} = \begin{cases} 0 & si \quad ||\mathbf{B}_j^{lasso}||_2 = 0 \\ (1 - \frac{\lambda_2}{||\mathbf{B}_j^{lasso}||_2 \quad ||X_k||_2^2})_+ (B_{j0,k}^{lasso}) & \text{en otro caso} \end{cases}$$

Hacemos la observación de que en esta situación el orden del algoritmo no cambia, el algoritmo es de orden O(TPQ)

2.3. 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 (2) para construir un estimador disperso $\mathbf{B}$ que dependa de $\Omega = \{\omega_{j'j}\}$,

$$(\hat{\mathbf{B}}, \hat{\Omega}) = \arg\min_{\mathbf{B},\Omega} \left\{ g(\mathbf{B}, \Omega) + \lambda_1 \sum_{j' \neq j} |\omega_{j'j}| + \lambda_2 \sum_{j=1}^{p} \sum_{k=1}^{q} |b_{jk}| \right\} \tag{6}$$

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 $\hat{\mathbf{B}}^{OLS}$ [3].

El problema de optimización en (6) 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

$$\hat{\Omega}(\mathbf{B}_0) = \arg\min \left\{ tr\left(\hat{\Sigma}_R \Omega\right) - \log(\det \Omega) + \lambda_1 \sum_{j' \neq j} |\omega_{j'j}| \right\}, \tag{7}$$

donde $\hat{\Sigma}_R = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\mathbf{B}_0)^T(\mathbf{Y} - \mathbf{X}\mathbf{B}_0)$. 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 7. Se abordará con más detalle este algoritmo en las secciones posteriores.

Por otro lado, resolver 6 fijando Ω en un punto elegido Ω_0 transforma el problema a optimizar

$$\hat{\mathbf{B}}(\Omega_0) = \arg\min \left\{ tr\left(\frac{1}{n}(\mathbf{Y} - \mathbf{X}\mathbf{B})^T(\mathbf{Y} - \mathbf{X}\mathbf{B})\Omega_0 + \lambda_2 \sum_{j=1}^{p}\sum_{k=1}^{q} |b_{jk}| \right) \right\} \tag{8}$$

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**. 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 2.5**.

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

Input: $\mathbf{Y}_{n\times q}$,$\mathbf{X}_{n\times p}$, $\Omega_{p\times p}$, λ_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 $\left(\hat{\mathbf{B}}^{(m)}\right)$

El costo computacional para el cálculo de μ_{rc} es $O(pq)$ y el costo total de todo el algoritmo es $O(p^2q^2)$.

Considerando lo anterior, podemos resumir la resolución del problema de optimización 6 usando el descenso de coordenadas en bloque, es decir, iteramso minimizando con respecto a $\mathbf{B}$ y miniminzando con respecto a Ω. El **Algoritmo 3** usa el descenso de coordenadas por bloques para calcular una solución loca para 6.

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

Input: $\mathbf{Y}_{n\times q}$,$\mathbf{X}_{n\times p}$, λ_1, λ_2, ϵ.

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 8 utilizando el **Algoritmo 2**.
2. Calcular $\hat{\Omega}^{(m+1)} = \hat{\Omega}(\hat{\mathbf{B}}^{(m+1)})$ resolviendo 7 usando el algoritmo de LASSO gráfico.

return $\left(\hat{\mathbf{B}}^{(m)}\right)$

2.4. Algoritmo de LASSO gráfico

Friedman et al. [2008] describe este método para máximizar el problema de optimización **(7)**. 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

$$w_{12} = \arg\min_{y} \left\{ y^T W_{11}^{-1} y : ||y - s12||_\infty \geq p \right\}. \tag{9}$$

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 (9) es equivalente a resolver el problema dual

$$\min_{\beta} \left\{ \frac{1}{2} |W_{11}^{-1/2} \beta - b|^2 + \lambda |\beta|_1 \right\} \tag{10}$$

donde $b = W_{11}^{-1/2} s_{12}$. Si β resuelve (10) entonces $w_{12} = W_{11}\beta$ resuelve (9). Además es sencillo ver que las soluciones en (7) son equivalentes a resolver (10). Para resolver (10) 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 (10). 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

$$\hat{\beta}_j = S(u_j - \sum_{k \neq j} V_{jk}\beta_k, \lambda)/V_{jj} \tag{11}$$

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.

2.5. 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

$$f(\mathbf{B}) = g(\mathbf{B}, \Omega) + \lambda_2 + \sum_{j=1}^{p} \sum_{k=1}^{q} |b_{jk}| \tag{12}$$

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

$$\frac{\partial f+}{\partial \mathbf{B}} = \frac{2}{n}\mathbf{X}^T\mathbf{X}\mathbf{B}\Omega - \frac{2}{n}\mathbf{X}^T\mathbf{Y}\Omega + \lambda_2 1_{(b_{ij}>0)} - \lambda_2 1_{(b_{ij}<0)} \tag{13}$$

$$\frac{\partial f-}{\partial \mathbf{B}} = -\frac{2}{n}\mathbf{X}^T\mathbf{X}\mathbf{B}\Omega + \frac{2}{n}\mathbf{X}^T\mathbf{Y}\Omega - \lambda_2 1_{(b_{ij}>0)} + \lambda_2 1_{(b_{ij}<0)} \tag{14}$$

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

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

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

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

3. Resultados

Para verificar el rendimientos de los modelos descritos en las secciones anteriores, consideramos probar las estimaciones de los algoritmos REMMAP y MRCE, utilizando MSE como métrica para comparar los errores por la facilidad de comprender la misma. Comparando con las estomaticones obtenidas con máxima verosimilitud. Estos algoritmos REMMAP y MRCE se implementaron en el lenguaje de programación Python, en el sistema x86_64, Ubuntu.

La paquetería de Python Scikit-learn [Pedregosa et al., 2011], tiene una función en donde esta implementado el **Algoritmo** 3. Entonces esta función nos ayudo a determinar si nuestra implementación estaba bien.

3.1. Evaluación de los modelos con datos sintéticos

El conjunto de datos sinteticos fue generado con la función $make_regression()$ de la librería de Scikit-learn. 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.

Observando la **Figura 3**, podemos notar que cuando se consideran tamaños de $n < p$ notamos que los mejores predictores son ocupando la metodología de REMMAP. Además, observamos que entre mayor sea el tamaño de nuestras variables respuestas nuestro error aumenta. Si comparamos los mejores modelos se observa claramente que en general los la metodologia de MRCE tiene rendimientos similares que los estimadores de máxima verosimilitud. Pero, en general la metodología REMMAP es mejor.

Modelo	MSE	Parametros	Numero de variables
MRCE	12.52	$\lambda_1 = 0{,}1\lambda_2 = 0{,}5$	21
$REMMAP\_l1$	11.35	$\lambda_1 = 0{,}1\lambda_2 = 0$	15
REMMAP	11.45	$\lambda_1 = 0{,}1\lambda_2 = 0$	8
OLS	42.2	–	–

Cuadro 1: Parametros: n=20 , p=100 , q=2

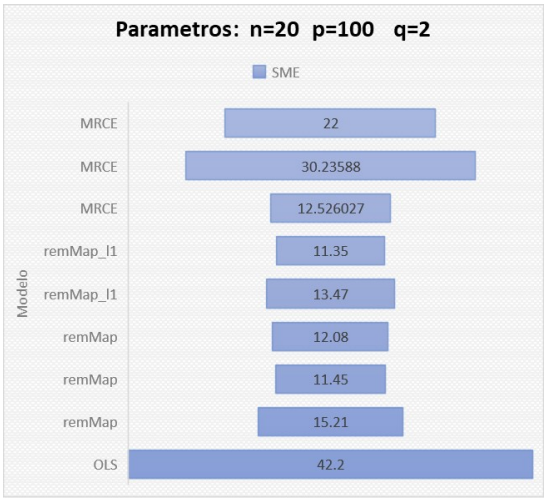

Figura 1

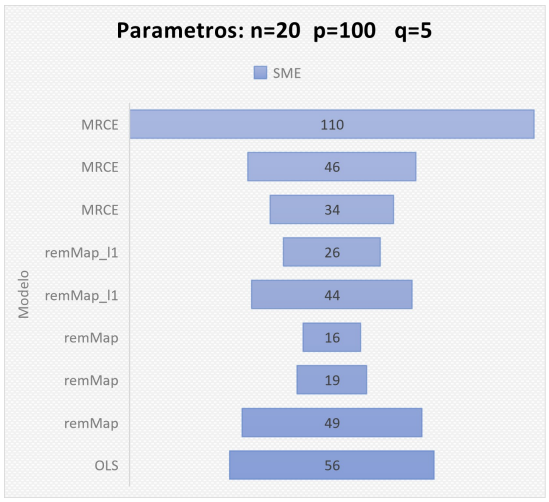

Figura 2

Figura 3: MSE considerando distintos modelos, con $n < p$.

Modelo	MSE	Parametros	Numero de variables
MRCE	34	$\lambda_1 = 0{,}1 \lambda_2 = 0{,}5$	10
$REMMAP\_l1$	26.47	$\lambda_1 = 0{,}1 \lambda_2 = 0$	40
REMMAP	15.64	$\lambda_1 = 1e-20 \lambda_2 = 0{,}1$	25
OLS	55.85	–	–

Cuadro 2: Parametros: n=20 , p=100 , q=5

Por otro lado, si consideramos $n > p$ la metodología MRCE tiene mejor rendimiento y es muy cercano al estimador de (3) cuando se considera $\lambda_2 = 0$.

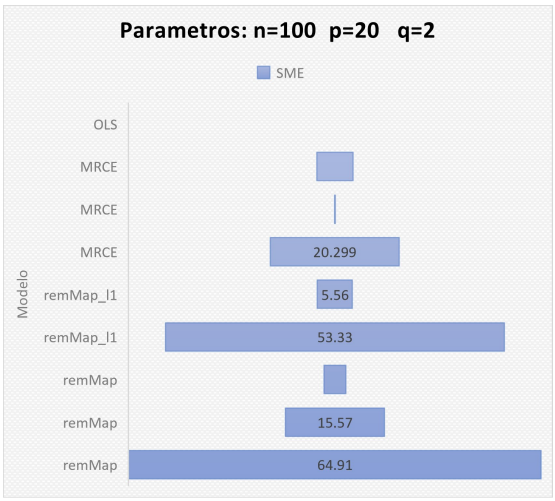

Figura 4

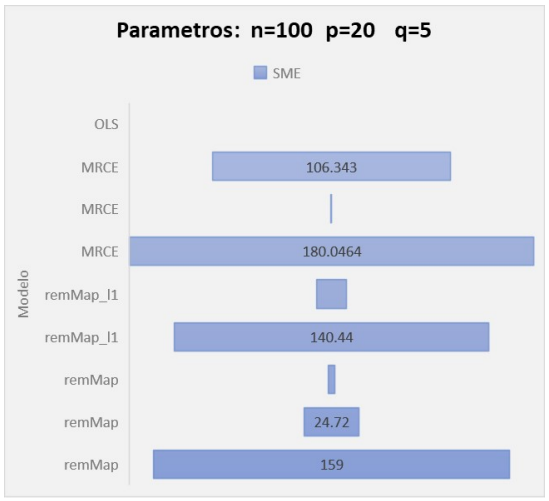

Figura 5

Figura 6: MSE considerando distintos modelos, con $n > p$.

Para este caso observamos que MRCE tiene mejores rendimientos que los otros modelos. Además, vemos claramente que existe una relación con el estimador de máxima verosimilitud. En general, preferíamos las estimadores de MRCE debido a que tiene una mayor interpretación que los estimadores OLS. Ya que tener estimadores iguales a ceros, nos permiter considerar que esas variables no son importantes en el modelo.

Modelo	MSE	Parametros	Numero de variables
MRCE	0.1227	$\lambda_1 = 0{,}1 \lambda_2 = 0{,}1$	14
REMMAP_$l1$	5.56	$\lambda_1 = 0{,}1 \lambda_2 = 0$	10
REMMAP	3.43	$\lambda_1 = 1e-20 \lambda_2 = 0{,}1$	8
OLS	0.1e-16	–	–

Cuadro 3: Parametros: n=100 , p=20 , q=2

Modelo	MSE	Parametros	Numero de variables
MRCE	0.3224	$\lambda_1 = 0{,}1 \lambda_2 = 0{,}1$	33
REMMAP_$l1$	13.15	$\lambda_1 = 0{,}1 \lambda_2 = 0$	16
REMMAP	2.89	$\lambda_1 = 1e-20 \lambda_2 = 0{,}1$	20
OLS	0.1e-16	–	–

Cuadro 4: Parametros: n=100 , p=20 , q=5

En general observamos que si aumentamos el tamaño de las variables (q), los modelos presentan un rendimiento menor.

4. Conclusiones

Primeramente podemos observar que se cumplió el objetivo principal de este trabajo, determinar una función f tal que sirva como función predictora usando un conjunto de datos X y Y planteándolo como un problema de optimización. Para resolver este problema, utilizamos distintos algoritmos de optimización: **LASSO gráfico (4) (descenso de coordenadas por bloque y descenso coordinado), y descenso de coordenadas cíclicas (2), descenso de coordenadas por bloques (3)**.

Por otro lado, planteamos dos metodología para resolver el problema de optimización de regresión multivariada con regularización. En general REMMAP parece desempeñarse mejor cuando n<p, mientras MRCE tiene un buen desempeño cuando $n > p$. MRCE presento la particularidad de requerir una parametrización mas cuidadosa por lo que en ese sentido es mas exigente, además de que involucra un costo computacional mas alto comparándolo con REMMAP.

Referencias

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.

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, 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.