UNIVERSITAT DE BARCELONA
Facultat de Matemàtiques i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

EL LASSO: REGULARITZACIÓ I SELECCIÓ DE PREDICTORS

Autor: Andrea Iglesias Munilla

Director: Dr. Josep Fortiana Gregori
Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, 20 de juny de 2021

Abstract

Lasso (Least Absolute Shrinkage and Selection Operator) is a regression method that performs both regularization and variable selection, improving prediction accuracy and interpretability of the resulting model.

In this project we follow evolution from the plain linear model, through Ridge regression, for many years the most popular technique to improve the precision of predictions, to Lasso. We delve into numerical procedures for calculating Lasso solutions: coordinate descent and LARS. We see some extensions of Lasso such as Elastic Net regression, a neat improvement when optimality fails.

We illustrate these methods with several real data examples using the R programming language (see notebooks and HTML files in appendices to the main text).

Resum

El Lasso (Least Absolute Shrinkage and Selection Operator) és un mètode de regressió que realitza selecció de variables i regularització per millorar la precisió en la predicció i la interpretabilitat del model resultant.

En aquest treball seguim l'evolució des del model lineal, passant per la regressió Ridge, que va ser durant molts anys la tècnica més popular per millorar l'exactitud en les prediccions, fins al Lasso. Aprofundim en procediments numèrics pel càlcul de solucions Lasso: descens seguint les coordenades i LARS. Veiem algunes extensions del Lasso com la regressió Elastic Net, que proporciona una millora en els casos en què el Lasso no funciona de manera òptima.

Il·lustrem aquests mètodes amb diversos exemples de dades reals utilitzant el llenguatge de programació R (vegeu fitxers HTML als annexos de la memòria).

2020 Mathematics Subject Classification. 62J07

Agraïments

Vull agrair al meu tutor Josep Fortiana per ajudar-me a escollir l'àmbit d'estudi, per solucionar tots els meus dubtes a distància degut al confinament, i guiar-me en tot moment al llarg d'aquests mesos de projecte.

A la meva família, amics i companys, per haver-me acompanyat en aquesta etapa acadèmica que arriba a la seva fi i obre altres de noves.

Índex

1 Introducció

Avui dia, la informació és un dels béns intangibles més valorats. Saber controlar-la és la millor manera d'estar per davant en qualsevol àmbit. A causa dels nous avanços tecnològics d'aquest segle, generem milions de dades per segon, les quals es poden fer servir per crear models amb els que predir o extreure informació per prendre decisions. Durant els últims anys s'han investigat diferents mètodes per crear aquests models aplicats en àrees molt diverses com la ciberseguretat, finances, viabilitat d'empreses, medicina, màrqueting, i un llarg etcètera.

Tal com veurem, per estimar els models, s'utilitza fonamentalment l'estadística, agafant com a base les relacions lineals entre les variables resposta i un o més predictors. Un dels models més senzills i que produeix bons resultats és calcular els coeficients mitjançant mínims quadrats ordinaris. Ara bé, aquest procediment té algunes limitacions que se solucionen amb una nova regressió anomenada Ridge (Hoerl, A.E., Kennard, R.W. 1970). La regressió Ridge però, no compleix exactament amb les condicions que ha de tenir un model lineal òptim.

En aquest treball tractem la regressió Lasso (Tibshirani, R. 1996), una tècnica usada actualment, que regularitza els coeficients estimats i alhora fa una selecció de variables, és a dir, tendeix a estimar alguns coeficients iguals a zero. L'objectiu principal és estudiar el funcionament del Lasso, les seves propietats i el càlcul de les solucions. Per tal de comprovar les seves característiques aplicarem el mètode a diferents bases de dades utilitzant el programari R.

Finalment veurem que el Lasso no és eficient en alguns escenaris i per això apareix un procediment recent que consisteix a utilitzar la regressió Elastic Net (Zou, H., Hastie, T. 2005), molt relacionada amb el Lasso, que resol aquestes limitacions.

2 El model lineal

Sigui $\{(x_i, y_i)\}_{i=1}^N$ una col·lecció de N mostres d'una regressió lineal, on cada $x_i = (x_{i1}, \cdots, x_{ip})$ és un vector p-dimensional de nombres constants que anomenarem predictors i cada $y_i \in \mathbb{R}$ forma part d'un vector de variables aleatòries que anomenarem resposta.

El model expressa l'esperança de la variable resposta y_i com a combinació lineal dels predictors x_i:

$$E(y_i) = \mu_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j. \tag{2.1}$$

En aquest cas el model està definit com

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i,$$

el qual està parametritzat pels coeficients de la regressió $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ i un biaix $\beta_0 \in \mathbb{R}$. El vector ϵ_i conté els errors aleatoris, que són variables centrades, i.i.d., amb llei $\epsilon_i \sim N(0, \sigma^2)$.

De manera equivalent, podem reescriure el model en forma matricial

$$y = X\beta + \epsilon$$

on y és el vector format per totes les variables resposta $y' = (y_1, \ldots, y_p)$, ϵ és el vector format pels errors aleatoris $\epsilon' = (\epsilon_1, \ldots, \epsilon_p)$, β és el vector format pels coeficients de la regressió $\beta' = (\beta_0, \ldots, \beta_p)$ i X és la matriu formada per la primera columna de uns, seguit per totes les variables explicatives tal que cada fila correspon a una observació de la mostra.

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}.$$

Per tant, el nostre objectiu és trobar una estimació $\hat{\beta}$ dels coeficients del model. Per poder calcular aquesta estimació necessitem que es compleixin una sèrie de condicions (Gauss-Markov): les observacions y_i són variables aleatòries amb segon moment finit, incorrelacionades i d'igual variància.

3 Mínims quadrats ordinaris

El mètode de mínims quadrats ordinaris (MQO) estima els coeficients β_i $(i = 0, 1, \cdots, p)$, a partir de trobar una solució al problema d'optimització:

$$\min_{\beta_0,\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\}, \tag{3.1}$$

o de manera equivalent en forma matricial

$$\min_{\beta} \left\{ \frac{1}{2N} \|y - X\beta\|_2^2 \right\}, \tag{3.2}$$

on $\|.\|$ és la norma euclidiana definida com $\|x\|_2 = \sqrt{\sum_{i=1}^{N} x_i^2}$.

Si $p < N$ i suposant que $\text{rang}(X) = p$, existeix una solució única del problema, calculant el mínim de la suma del quadrat dels errors, que ve donada per: $\hat{\beta} = (X'X)^{-1}X'y$.

En tal cas, el vector dels valors ajustats és: $\hat{y} = X\hat{\beta} = Hy$, on la matriu H s'anomena hat matrix. Aquesta matriu H és la projecció ortogonal sobre el subespai $\langle X \rangle \subset \mathbb{R}^N$; una matriu quadrada de dimensió $N \times N$, simètrica, idempotent, amb una traça igual a p, que indica els graus de llibertat del model, i que vé donada per: $H = X(X'X)^{-1}X'$.

Encara que $Q = X'X$ sigui no invertible, el subespai $\langle X \rangle \subset \mathbb{R}^N$ està ben definit i existeix sempre H, la projecció ortogonal sobre aquest subespai, definit de manera única.

Si, segons les condicions de Gauss-Màrkov, $\text{Var}(y) = \sigma^2 I$, obtenim que: $\text{Var}(\hat{y}) = \sigma^2 H$. Si $Q = X'X$ és no singular, existeix un $\hat{\beta}$ únic, i $\text{Var}(\hat{\beta}) = \sigma^2 Q^{-1}$.

En aquest cas, β_{MQO} és un estimador lineal, sense biaix, eficient i consistent.

3.1 Possible inestabilitat del model lineal

Però aquest model lineal pot ser 'millorat' reemplaçant l'ajust de mínims quadrats ordinaris per altres mètodes que poden tenir millor precisió en la predicció i millor interpretabilitat del model.

- *Precisió en la predicció*: seguint amb el que s'ha dit a l'apartat 2, els estimadors calculats mitjançant MQO són sense biaix i de variància mínima. Això és cert sempre i quan es compleixin les condicions de Gauss-Markov i el nombre d'individus sigui major que el número de predictors. En cas contrari, ens podem trobar amb dues situacions: quan $N > p$ i existeixen relacions de dependència entre alguns predictors, o quan $N < p$. En ambdos casos, la matriu $Q = X'X$ no tindrà rang màxim, i per tant no es podrà calcular la seva inversa. Això provoca que no es pugui calcular l'estimador per MQO. També ens podem trobar que, encara que Q sigui invertible, a la pràctica, l'estimador $\hat{\beta}$ tingui males condicions de predicció ja que podem tenir problemes numèrics al calcular la seva inversa degut a la multicol-linealitat que veurem més endevant.

- *Interpretabilitat del model*: sovint algunes o moltes de les variables utilitzades en un model de regressió múltiple tenen poc poder predictiu, és a dir, no tenen una relació directa amb la resposta. La inclusió d'aquestes variables irrellevants comporta una complexitat perjudicial per al model resultant. Eliminant aquestes variables -posant un zero als corresponents estimadors- podem obtenir un model més fàcil d'interpretar. El procediment de MQO no produeix models 'sparse', en altres paraules, els models estimats no seran econòmics en predictors ja que és extremadament improbable que es produeixin estimacions de coeficients que siguin exactament iguals a zero.

3.2 Problemes de multicol·linealitat

La multicol·linealitat és un dels problemes principals que ens podem trobar a l'hora d'estimar un model. Si existeixen relacions lineals entre dos o més variables explicatives del model, no podem aplicar el mètode de MQO ja que la matriu X tindrà $\text{rang}(X) < p$ i conseqüentment serà singular. Llavors aquesta multicol·linealitat exacta provoca que la matriu $Q = X'X$ tingui determinant igual a zero, ja que per un resultat d'algebra lineal $\text{rang}(X'X) = \text{rang}(X)$. Aleshores no podem calcular Q^{-1} i, per tant, tampoc l'estimador $\hat{\beta}$.

A la pràctica aquesta multicol·linealitat exacta rarament succeeix. El que ens podem trobar és que dues variables estiguin fortament correlacionades, és a dir, alguna variable és 'quasi' combinació lineal d'una o més variables predictores.

En aquest cas la matriu Q és pròxima a singular, en altres paraules, el seu determinant és pròxim a zero. Això provoca que en calcular la inversa de Q, com hem de dividir entre el determinant, ens trobem amb problemes de precisió per calcular les estimacions dels paràmetres β_i donant lloc a un model amb molta variabilitat. Com més gran sigui el grau de multicol·linealitat, $|Q|$ serà més proper a zero, llavors major serà la variància de l'estimador β i, per consegüent, menor serà la seva precisió i estabilitat.

Detecció de la multicol·linealitat

Per poder identificar si el model de regressió amb el que treballem presenta problemes de multicol·linealitat, hi ha diversos instruments de detecció.

- **Matriu de correlacions** (R)

  La matriu de correlacions és una matriu quadrada i simètrica que té uns a la diagonal i fora d'ella els coeficients de correlació entre les variables ($r_{ij} = Corr(x_i, x_j)$).

  $$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1k} \\ \vdots & \vdots & & \vdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{pmatrix}.$$

  Considerem que dos predictors x_i i x_j, per $i \neq j$, estan fortament correlacionats si $r_{ij} > 0.8$. Tot i això, només podem detectar correlacions entre parells de variables i no possibles relacions entre més de dues variables explicatives.

  Per altra banda, en cas de no presentar multicol·linealitat elevada, el determinant de la matriu R tendeix a 1. En cas contrari, quan estem en presència de multicol·linealitat, la matriu de correlacions tindrà com a determinant un nombre proper a zero.

- **Càlcul del Factor d'Inflació de la Variància (FIV_j)**

  El Factor d'inflació de la variància per la variable x_j (FIV_j) compara la variància real del coeficient β_j amb la que s'obtindria en cas que hi hagués absència total de multicol·linealitat. Així, per cadascuna de les variables explicatives del model podem calcular el FIV de la manera següent:

$$FIV_j = \frac{1}{1 - R_j^2}$$

on R_j^2 és el coeficient de determinació de la regressió auxiliar per la variable x_j. Considerem que existeixen problemes de multicol·linealitat si algun FIV_j és superior a 10 ($R_j^2 > 0,9$). Encara que pot existir multicol·linealitat amb factors d'inflació de la variància baixos, i a més també pot haver-hi multicol·linealitats que no impliquin a totes les variables predictores i per tant no siguin detectades amb el FIV_j.

- **Número de condició**

  En general, el número de condició $\kappa(A)$ d'una matriu A mesura la inexactitud numèrica introduïda resolent l'equació $Ax = b$ per a un vector columna b donat. Tècnicament és la relació entre els valors singulars màxim i mínim de A. Quan $\kappa \approx 1$ no hi ha pèrdues de precisió significatives i quan $\kappa = 10^k$, la pèrdua de precisió és de k dígits decimals significatius. Es pot calcular mitjançant la fórmula

  $$\kappa(A) = \|A\| \cdot \|A^{-1}\|$$

  per qualsevol norma matricial.

4 La regressió Ridge

Compromís biaix-variància

Aquest és un principi general que s'utilitza en situacions on diferents models estadístics es poden aplicar a unes mateixes dades. S'observa que, en general, si fem el model més gran, utilitzant el nombre màxim de predictors amb la intenció d'ajustar millor (menys biaix) les dades observades, el model obtingut resulta inestable (amb més variància). Un model amb més variància ajusta pitjor altres dades procedents de la mateixa població, i produeix prediccions menys fiables.

Restringint o reduint els coeficients estimats, podem reduir substancialment la variància a costa d'un augment negligible del biaix. Això pot conduir a millores substancials en la precisió amb la qual podem predir les respostes de les observacions no utilitzades per predir el model.

Seguint el compromís biaix-variància, sorgeix el model de regressió Ridge.

La regressió Ridge

La regressió Ridge va ser introduïda per primera vegada el 1970 per Hoerl i Kennard a la revista *Technometrics* [1].

El model de la regressió resulta de cercar un estimador $\hat{\beta}_\lambda$ de β, intencionadament amb biaix, però més estable que l'estimador calculat per MQO, amb una menor variància. S'obté com la solució al problema d'optimització:

$$\min_{\beta_0,\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\} \quad \text{tal que} \quad \sum_{j=1}^{p} {\beta_j}^2 \leq t. \tag{4.1}$$

De manera equivalent, el mateix problema es pot reescriure utilitzant normes,

$$\min_{\beta} \left\{ \frac{1}{2N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\}, \tag{4.2}$$

on $\beta = (\beta_1, \cdots, \beta_p)$, X és la matriu formada pels predictors i $\lambda > 0$ és un paràmetre, el valor del qual es tria adequadament, usualment amb algun sistema de validació encreuada. Aquesta penalització amb la norma l_2 restringeix o regularitza els coeficients estimats, és a dir, que els comprimeix cap a zero a mesura que λ augmenta.

A diferència dels mínims quadrats ordinaris, que només generen un conjunt d'estimacions dels coeficients, la regressió ridge produirà diferents conjunts d'estimacions dels coeficients, $\hat{\beta}_\lambda$, per a cada valor de λ.

La penalització per contracció s'aplica només a $\beta_1, \cdots, \beta_p$, però no a la intersecció β_0. Volem reduir el coeficient associat a cada variable, no obstant això, no volem reduir la intercepció, que és simplement una mesura del valor mitjà de la resposta quan $x_{i1} = x_{i2} = \cdots = x_{ip} = 0$.

Més en general, la regressió Ridge és una regularització de Tikhònov[5] d'un problema mal posat. El concepte de problema ben posat es deu a Jacques Hadamard i es defineix com un problema per al qual existeix una solució, que és única i depèn de manera contínua de les condicions inicials.

La regularització de Tikhònov minimitza

$$\|y - X\beta\|_2^2 + \|\Gamma\beta\|_2^2$$

per alguna matriu de Tikhònov Γ convenientment escollida. Aleshores la solució és

$$\hat{\beta} = (X'X + \Gamma'\Gamma)^{-1}X'y.$$

Per $\Gamma = 0$ el problema es redueix a mínims quadrats ordinaris. El problema Ridge és un cas especial de la regularització Tikhònov on $\Gamma = \alpha I$ ($\alpha^2 = \lambda$).

Fent calculs, resulta que: $\hat{\beta}_\lambda = (X'X + \lambda I)^{-1}X'y$. Triant un λ suficientment gran, es pot aconseguir que $Q_\lambda = X'X + \lambda I$ sigui no singular i que la variància de l'estimador sigui acceptable, al preu d'afegir biaix.

En tal cas, el vector dels valors ajustats és: $\hat{y} = X\hat{\beta} = X(X'X + \lambda I)^{-1}X'y = H_\lambda y$. Per analogia amb el cas de MQO, la matriu H_λ s'anomena hat-matrix de la regressió Ridge, tot i que aquesta no és idempotent (no és una projecció ortogonal). Seguint aquesta analogia, la seva traça és, en general, un nombre no enter, que per definició és equivalent als graus de llibertat del model: $\mathrm{df}(\lambda) = \mathrm{tr}(H_\lambda)$.

La regressió ridge només regularitza, és a dir, restringeix els coeficients cap a zero i, per la seva naturalesa, no produeix cap estimació que sigui exactament igual a zero, cosa que veurem si fa la regressió Lasso a continuació.

Per tant, no es produeix selecció de variables i conseqüentment tots els predictors apareixen en el model estimat final. Aquest fet resulta un problema pel que fa a la interpretabilitat del model quan, per exemple, tenim unes dades amb un elevat número de predictors.

5 El Lasso

Una nova tècnica anomenada Lasso (**L**east **A**bsolute **S**hrinkage and **S**election **O**perator) va ser proposada per Robert Tibshirani el 1996 [3].

El Lasso és un mètode que combina l'ajust per MQO amb una regularització dels coeficients mitjançant la norma l_1, que, a diferència de la regressió ridge (restricció amb norma l_2) proporciona una nova propietat als models estimats: la selecció de variables.

Donat $\{(x_i, y_i)\}_{i=1}^N$ una col·lecció de N parells de variables predictors-resposta, el Lasso troba una solució $(\hat{\beta}_0, \hat{\beta})$ al problema d'optimització

$$\min_{\beta_0,\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\} \quad \text{tal que} \quad \sum_{j=1}^{p} |\beta_j| \leq t. \tag{5.1}$$

La restricció $\sum_{j=1}^{p} |\beta_j| \leq t$ es pot escriure de forma més compacta amb la norma l_1 com la restricció $\|\beta\|_1 \leq t$. Normalment, es representa utilitzant notació de matriu-vector. Sigui $y = (y_1, ..., y_N)$ que denota el vector de dimensió N de respostes, i X la matriu $N \times p$ amb $x_i \in \mathbb{R}^p$ la fila i-essima, llavors el problema d'optimització anterior es pot reescriure com

$$\min_{\beta_0,\beta} \left\{ \frac{1}{2N} \|y_i - \beta_0 \mathbf{1} - \mathbf{X}\beta\|_2^2 \right\} \quad \text{tal que} \quad \|\beta\|_1 \leq t, \tag{5.2}$$

on $\mathbf{1}$ és el vector de N uns, i $\|.\|_2$ representa la usual norma Euclidiana per a vectors.

La penalització t limita la suma dels valors absoluts dels paràmetres estimats. Atès que una estimació de paràmetres reduïts correspon a un model més restringit, aquesta t limita la forma en què podem ajustar les dades. S'ha d'especificar mitjançant un procediment extern com la validació encreuada.

Com que la restricció del lasso tracta tots els coeficients per igual, normalment té sentit que tots els elements de X estiguin en les mateixes unitats. Si no, sense cap pèrdua de generalitat, centrarem i estandarditzarem els predictors per tal que cadascuna de les columnes de X tingui mitjana 0 i variància 1, és a dir

$$\frac{1}{N} \sum_{i=1}^{N} x_{ij} = 0 \quad \text{i} \quad \frac{1}{N} \sum_{i=1}^{N} x_{ij}^2 = 1.$$

A partir d'ara considerarem que tant les variables de X com els valors de y_i estan centrades, és a dir $\frac{1}{N} \sum_{i=1}^{N} y_i = 0$. La condició de centrar les variables és convenient, ja que podem ometre el terme β_0 en l'optimització del lasso.

Donada una solució òptima del Lasso $\hat{\beta}$ amb les dades centrades, podem recuperar les solucions lasso per a les dades descentrades: $\hat{\beta}$ és el mateix i $\hat{\beta}_0$ ve donada per

$$\hat{\beta}_0 = \overline{y} - \sum_{j=1}^{p} \overline{x}_j \hat{\beta}_j, \tag{5.3}$$

on $\overline{y}$ i $\{\overline{x}_j\}_1^p$ són les mitjanes de les dades originals.

També pot ser útil reescriure el problema del lasso

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2N} \|y - X\beta\|_2^2 \right\} \quad \text{tal que} \quad \|\beta\|_1 \leq t, \tag{5.4}$$

en la seva forma Lagrangiana

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}, \tag{5.5}$$

per alguna $\lambda \geq 0$, on t i λ són els paràmetres de regularització o penalització.

La forma lagrangiana i el problema de la restricció són problemes equivalents: per a cada valor de t en el rang on la restricció $\|\beta\|_1 \leq t$ es compleix, hi ha una λ corresponent que dona com a resultat la mateixa solució que amb la fórmula lagrangiana. En canvi la solució $\hat{\beta}_\lambda$ del problema en forma lagrangiana resol el problema de la restricció amb $t = \|\hat{\beta}_\lambda\|_1$.

5.1 Selecció de predictors

El Lasso juga un doble paper: per una banda soluciona el problema de deficiència de rang (quan hi ha més variables predictores que individus o quan hi ha multicol·linealitat) reduint els coeficients estimats cap a zero. Per altra banda té l'efecte de forçar que algunes de les estimacions dels coeficients siguin exactament iguals a zero quan el paràmetre t és prou petit, dit d'una altra manera, realitza selecció de variables.

Com a resultat, els models generats a partir del Lasso són generalment molt més fàcils d'interpretar. Diem que el Lasso tendeix a produir models 'sparse', és a dir, models econòmics en predictors en els quals una proporció elevada de coeficients són iguals a zero.

Això és degut a les propietats de la norma l_1 i la restricció $\|\beta\|_1 \leq t$.

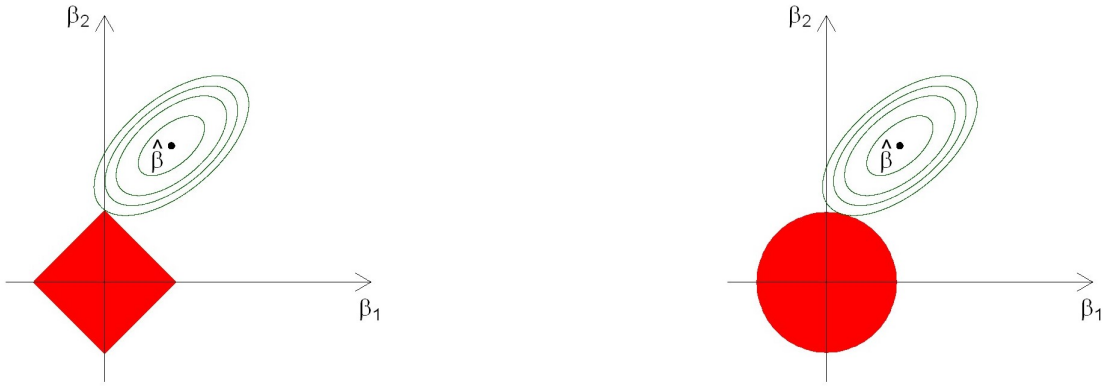

Figura 1: *Respresentació gràfica de les funcions d'error i restricció per a la regressió Lasso (esquerra) i per a la regressió Ridge (dreta). Les zones vermelles són les regions de restricció,* $|\beta_1| + |\beta_2| \leq t$ *i* $\beta_1^2 + \beta_2^2 \leq t$*, mentre que les el·lipses verdes representen la SQR.*

La figura 1 contrasta les dues restriccions utilitzades en el lasso i en el ridge, en el cas $p = 2$. La solució obtiguda per mínims quadrats ordinaris es denota com a $\hat{\beta}$, mentre que el diamant i el disc vermells representen les restriccions de la regressió Lasso i de la regressió Ridge, respectivament.

Si t és prou gran, les regions de restricció contindran $\hat{\beta}$, de manera que les estimacions de la regressió ridge i de la regressió lasso seran les mateixes que les estimacions per MQO. No obstant això, a la figura 1 les estimacions per MQO es troben fora del diamant i del disc, de manera que les estimacions per MQO no són les mateixes que les de la regressió ridge o lasso.

Les el·lipses centrades al voltant de $\hat{\beta}$ representen regions amb la suma dels residus quadrats constant. Dit d'una altra manera, tots els punts d'una el·lipse determinada comparteixen un valor comú de la SQR. A mesura que les el·lipses s'allunyen de les estimacions del coeficient per MQO, la SQR augmenta.

Els problemes d'optimització (4.1) i (5.1) indiquen que les estimacions del coeficient de regressió Lasso i Ridge són causades pel primer punt en què una el·lipse entra en contacte amb la regió de restricció. Com que la regressió Ridge té una restricció circular, sense

puntes, aquesta intersecció no es produirà generalment en un eix, de manera que els coeficients estimats seran exclusivament diferents de zero. No obstant això, la restricció del Lasso té cantonades a cadascun dels eixos, de manera que l'el·lipse sovint tallarà la regió de restricció en un eix. Quan això passa, un dels coeficients serà igual a zero. En dimensions superiors, moltes de les estimacions del coeficient poden ser iguals a zero. A la figura 1, la intersecció es produeix a $\beta_1 = 0$, de manera que el model resultant només inclourà β_2.

A la figura 1, hem considerat el cas simple de $p = 2$. Quan $p = 3$, llavors la regió de restricció per a la regressió Ridge es converteix en una esfera i la regió de restricció per al Lasso es converteix en un poliedre. Quan $p > 3$, la restricció per a la regressió Ridge es converteix en una hiperesfera i la restricció per al Lasso es converteix en un polítop. Tot i això, les idees clau que es mostren a la figura 1 encara es mantenen. En particular, el Lasso porta a la selecció de variables quan $p > 2$ a causa de les cantonades del poliedre o polítop.

5.2 Graus de llibertat

Suposem que tenim p predictors i que estimem un model de regressió lineal utilitzant només un subconjunt k d'aquests predictors. Llavors, si triem aquests k predictors sense tenir en compte la variable resposta, la regressió té k graus de llibertat, mentre que els residus tenen $N - k$ graus de llibertat. Aquesta és una manera de dir que l'estadístic de prova per justificar la hipòtesi que tots els k coeficients són nuls segueix una distribució Chi-quadrada amb k graus de llibertat.

Tanmateix, si els k predictors els escollim utilitzant la variable resposta, per exemple, els predictors que tinguin l'error de predicció més petit entre tots els subconjunts de mida k, esperaríem que el procediment d'ajust tingui més de k graus de llibertat. Anomenarem adaptatiu a un procediment en el qual escollim un subconjunt de predictors tenint en compte la variable resposta. Clarament, el lasso n'és un exemple.

Com veurem a continuació, en el cas del lasso, els graus de llibertat es poden calcular comptant els coeficients diferents del zero en el model estimat.

En primer lloc, hem de definir amb precisió què volem dir amb els graus de llibertat d'un model adaptatiu. Suposem que tenim un model amb

$$y_i = f(x_i) + \epsilon_i \quad \text{per a} \quad i = 1, \cdots, N,$$

per alguna f desconeguda i amb els errors ϵ_i iid $(0, \sigma^2)$. Si les prediccions de la mostra es denoten per $\hat{y}$, aleshores podem definir, de manera alternativa, els graus de llibertat com:

$$\mathrm{df}(\hat{y}) := \frac{1}{\sigma^2} \sum_{i=1}^{N} Cov(\hat{y}_i, y_i). \tag{5.6}$$

Així, els graus de llibertat corresponen a la quantitat total d'autoinfluència que cada mesura de la resposta té en la seva predicció. Com més s'adapta el model a les dades, més gran són els graus de llibertat.

En el cas d'un model lineal fix, utilitzant k predictors escollits independentment de la variable resposta, es pot demostrar que $\mathrm{df}(\hat{y}) = k$. Tanmateix, sota un ajust adaptatiu, normalment es dona el cas que els graus de llibertat són més grans que k.

Es pot demostrar que per al lasso, amb un paràmetre de penalització fix λ, el nombre de coeficients diferents de zero k_λ és una estimació no esbiaixada dels graus de llibertat. (Zou, Hastie i Tibshirani 2007, Tibshirani i Taylor 2012).

La raó és que el lasso no només selecciona predictors (que fan augmentar els graus de llibertat), sinó que també redueix els seus coeficients cap a zero, en relació amb les estimacions per mínims quadrats. Aquesta contracció resulta ser la quantitat adequada per a reduir els graus de llibertat a k. Aquest resultat és útil perquè ens proporciona una mesura qualitativa de la quantitat d'ajusts que hem fet en qualsevol punt del procediment del lasso.

5.3 Unicitat de les solucions

Primerament observem que la teoria de la dualitat convexa es pot utilitzar per demostrar que quan les columnes de X estan en posició general, llavors per $\lambda > 0$ la solució al problema del lasso (5.5) és única. Això es manté fins i tot quan $p \geq N$, tot i que llavors el nombre de coeficients diferents de zero de qualsevol solució lasso serà com a màxim N. (Rosset, Zhu i Hastie 2004, Tibshirani 2013).

Ara, quan la matriu de predictors X no té rang màxim, els valors ajustats per mínims quadrats són únics, però els paràmetres estimats no ho són. Aquesta situació es pot produir quan $p \leq N$ a causa de la multicol·linealitat, i sempre es produeix quan $p > N$. En aquest últim escenari, hi ha un nombre infinit de solucions $\hat{\beta}$ que ofereixen un ajust perfecte amb un error d'entrenament igual a zero.

Considerem el problema del lasso en forma Lagrangiana per $\lambda > 0$. Els valors ajustats $X\hat{\beta}$ són únics, però resulta que la solució $\hat{\beta}$ pot no ser única.

Considerem un exemple senzill amb dos predictors x_1, x_2, la resposta y, i suposem que els coeficients de la solució lasso $\hat{\beta}$ són $(\hat{\beta}_1, \hat{\beta}_2)$. Si ara incloem un tercer predictor $x_3 = x_2$ a la mostra, una còpia idèntica del segon predictor, llavors per a qualsevol $\alpha \in [0, 1]$, el vector $\tilde{\beta}(\alpha) = (\hat{\beta}_1, \alpha\hat{\beta}_2, (1-\alpha)\hat{\beta}_2)$ produeix un ajust idèntic amb norma l_1 $|\tilde{\beta}(\alpha)|_1 = |\hat{\beta}|_1$. En conseqüència, per a aquest model (en el qual podríem tenir $p \leq N$ o $p > N$), hi ha una família infinita de solucions.

En general, quan $\lambda > 0$, si les columnes de la matriu del model X estan en posició general, les solucions del lasso són úniques. Per ser precisos, diem que les columnes $\{x_j\}_{j=1}^p$ estan en posició general si algun subespai afí $L \subset \mathbb{R}^N$ de dimensió $k < N$ conté com a màxim $k+1$ elements del conjunt $\{\pm x_1, \pm x_2, \cdots, \pm x_p\}$, excloent els parells de punts antipodals (és a dir, els punts que només es diferencien per un canvi de signe). Observem que les dades de l'exemple del paràgraf anterior no es troben en posició general.

Si les dades de X s'extreuen d'una distribució de probabilitat contínua, llavors amb probabilitat 1, les dades es troben en posició general i, per tant, les solucions lasso seran úniques. Com a resultat, la no singularitat de les solucions del lasso només es pot produir amb dades de valors discrets.

Observem que els algoritmes numèrics per calcular solucions del lasso normalment produeixen solucions vàlides en el cas no únic. No obstant això, la solució particular que ofereixen pot dependre de les particularitats de l'algorisme. Per exemple, amb el descens seguint les coordenades, l'elecció dels valors inicials pot afectar la solució final.

5.4 Càlcul de solucions: Descens seguint les coordenades

El problema del lasso és un programa convex, específicament un programa quadràtic amb una restricció convexa. Com a tal, hi ha molts mètodes sofisticats per resoldre'l. Tanmateix, hi ha un algorisme de càlcul especialment senzill i eficaç, que dona informació sobre el funcionament del lasso. Per comoditat, reescrivim el criteri en la forma Lagrangiana:

$$\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2N}\sum_{i=1}^{N}(y_i-\sum_{j=1}^{p}x_{ij}\beta_j)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\} \tag{5.7}$$

Com abans, assumirem que la resposta y_i està centrada i que els predictors estan estandartizats. En aquest cas, es pot ometre el terme β_0. La forma Lagrangiana és especialment convenient per al càlcul numèric de la solució mitjançant un procediment simple conegut com a descens seguint les coordenades (*coordinate descent*).

Cas amb només un predictor (Soft-thresholding)

Considerem la mostra $\{(z_i,y_i)\}_{i=1}^N$ amb un únic paràmetre predictor (per comoditat, canviem la notació de x_{ij} per z_i). El problema consisteix en resoldre

$$\min_{\beta}\left\{\frac{1}{2N}\sum_{i=1}^{N}(y_i-z_i\beta)^2+\lambda|\beta|\right\}. \tag{5.8}$$

L'enfocament estàndard d'aquest problema per minimitzar la fórmula seria calcular el gradient (primera derivada) respecte a β, i igualar-lo a zero. Hi ha una complicació, però, perquè la funció del valor absolut $|.|$ no té derivada a $\beta=0$. Tanmateix, podem procedir a la inspecció directa de la funció i trobar que

$$\hat{\beta}=\begin{cases}\frac{1}{N}\langle z,y\rangle-\lambda & \text{si} \quad \frac{1}{N}\langle z,y\rangle>\lambda\\ 0 & \text{si} \quad \frac{1}{N}\langle z,y\rangle\leq\lambda\\ \frac{1}{N}\langle z,y\rangle+\lambda & \text{si} \quad \frac{1}{N}\langle z,y\rangle<\lambda\end{cases} \tag{5.9}$$

que podem escriure com

$$\hat{\beta}=\mathcal{S}_\lambda(\frac{1}{N}\langle z,y\rangle). \tag{5.10}$$

On l'operador soft-thresholding és

$$\mathcal{S}_\lambda(x)=sign(x)(|x|-\lambda)_+ \tag{5.11}$$

que tradueix l'argument de x cap a zero per la quantitat λ, i el posa a zero si $|x|\leq\lambda$. Fixem-nos que per a les dades estandarditzades z_i és només una versió soft-thesholding de l'estimació habitual de mínims quadrats $\hat{\beta}=\frac{1}{N}\langle z,y\rangle$.

Cas amb diversos predictors (descens seguint les coordenades cíclic)

Utilitzant la idea del cas amb una sola variable, ara podem desenvolupar un mètode senzill per resoldre el problema complet del lasso. Més exactament, recorrem repetidament els predictors en un ordre fix (però arbitrari) (per exemple, $j = 1, 2, ..., p$), on en el pas j, actualitzem el coeficient β_j minimitzant la funció objectiu mantenint fixats tots els altres coeficients $\{\hat{\beta}_k, k \neq j\}$ als seus valors actuals.

La funció objectiu serà

$$\frac{1}{2N}\sum_{i=1}^{N}(y_i - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j)^2 + \lambda\sum_{k\neq j}|\beta_k| + \lambda|\beta_j|. \tag{5.12}$$

Veiem que cada β_j pot ser expressada en termes del residu parcial $r_i^{(j)} = y_i - \sum_{k\neq j} x_{ik}\hat{\beta}_k$ que elimina del resultat l'ajust actual de tots excepte el j-esim predictor.

Per tant, el terme j-esim dels coeficients serà de la forma

$$\hat{\beta}_j = \mathcal{S}_\lambda(\frac{1}{N}\langle x_j, r^{(j)}\rangle), \tag{5.13}$$

o equivalentment

$$\hat{\beta}_j \leftarrow \mathcal{S}_\lambda(\hat{\beta}_j + \frac{1}{N}\langle x_j, r\rangle), \tag{5.14}$$

on $r_i = y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j$ són els residus totals.

L'algoritme funciona aplicant aquesta actualització del soft-thresholding (5.13) repetidament de manera cíclica, actualitzant les coordenades de $\hat{\beta}$ (i, per tant, els vectors dels residus) al llarg del procés. Aquest algoritme s'anomena descens seguint les coordenades cíclic (*cyclical coordinate descent*).

A la pràctica, sovint ens interessa trobar la solució del lasso no només per a un únic valor fix de λ, sinó tot el camí de solucions per a un interval de possibles valors de λ. Un mètode raonable per fer-ho és començar amb un valor prou gran per a que l'única solució òptima sigui el vector de zeros. Aquest valor és igual a $\lambda_{max} = \max_j |\frac{1}{N}\langle x_j, y\rangle|$. Després disminuïm una petita quantitat i executem el descens seguint les coordenades fins a la convergència. Reduint de nou i utilitzant la solució anterior com a nou inici, executem el descens seguint les coordenades fins a la convergència. I així successivament. D'aquesta manera podem calcular de manera eficient les solucions sobre un interval de valors de λ. Ens referim a aquest mètode com a descens seguint les coordenades per camins (*pathwise coordinate descent*).

5.5 Càlcul de solucions: LARS

La regressió LAR (Least Angle Regression), també coneguda com l'enfocament homotòpic, va ser introduïda originalment per Bradley Efron, Trevor Hastie, Iain Johnstone i Robert Tibshirani el 2004. [7]

És un procediment per resoldre el lasso que proporciona tot el camí de les solucions en funció del paràmetre de regularització λ. És un algorisme bastant eficient, però no arriba a funcionar correctament per a grans problemes, així com altres mètodes. Tanmateix, té una motivació estadística interessant i es pot veure com una mena de versió 'democràtica' de la Forward Regression o procediment 'pas a pas'.

El procés de selecció de variables anomenat Forward Regression o Forward Stepwise Regression, descrit per Weisberg (1980) crea un model seqüencialment, afegint una variable cada cop. A cada pas, identifica la millor variable que cal incloure al conjunt actiu i, a continuació, actualitza l'ajust de mínims quadrats per incloure totes les variables actives.

El LAR utilitza un procediment similar a la tècnica anterior, però incorpora els predictors de forma gradual, només entra 'tant' d'un predictor com es mereix, de manera que tots els predictors formen part del model final. Com en el mètode 'pas a pas' comencem amb tots els coeficients iguals a zero i trobem el predictor més correlacionat amb la resposta y, diguem x_j. En lloc d'ajustar completament aquesta variable, el mètode LAR mou el coeficient d'aquesta variable contínuament cap al seu valor de mínims quadrats (fent que la seva correlació amb el residu disminueixi en valor absolut). Quan una altra variable x_k té la mateixa correlació amb el residu actual com x_j, el procés s'atura i a continuació, la variable x_k s'uneix al conjunt actiu. Seguidament, es procedeix en direcció equiangular entre els dos predictors x_j, x_k, dit d'una altra manera, els seus coeficients es mouen junts de forma que es mantenen les seves correlacions lligades i decreixents, fins que una tercera variable x_l es guanya la seva posició en el conjunt de 'més correlacionats'. En altres paraules, el predictor x_l té tanta correlació amb el residu com $\beta_j x_j + \beta_k x_k$. Aquest procés continua fins que totes les variables es troben al model i acaba amb l'ajust per mínims quadrats complet.

Anem a veure com es calcularien aquests coeficients de manera algebraica. Primer estandaritzem els predictors i centrem les variables de la resposta. Comencem amb els residus inicials $r_o = y - \overline{y}$ i els coeficients iguals a zero $\beta^0 = (\beta_1, \beta_2, \cdots, \beta_p) = 0$.

A continuació trobem el prodictor x_j que estigui més correlacionat amb els residus r_0, és a dir, el que tingui el valor més gran de

$$\frac{1}{N}|\langle x_j, r_0\rangle| = \lambda_0.$$

Aleshores el nostre conjunt actiu serà $\mathcal{A} = \{j\}$, i la matriu $X_\mathcal{A}$ estarà formada només pel predictor x_j.

Després, per a $k = 2, 3, \cdots, K = \min(N-1, p)$, primer calcularem la direcció de mínims quadrats

$$\delta = \frac{1}{N\lambda_{k-1}}(X'_\mathcal{A} X_\mathcal{A})^{-1} X'_\mathcal{A} r_{k-1}$$

i definirem el p-vector Δ com $\Delta_\mathcal{A} = \delta$ i els altres elements iguals a zero.

Ara hem de moure els coeficients β de β^{k-1} en direcció Δ cap a les seves solucions de MQO en $X_{\mathcal{A}}$:

$$\beta(\lambda) = \beta^{k-1} + (\lambda_{k-1} - \lambda)\Delta,$$

per a $0 < \lambda \leq \lambda_{k-1}$, calculant al manteix temps els residus

$$r(\lambda) = y - X\beta = r_{k-1} - (\lambda_{k-1} - \lambda)X_{\mathcal{A}}\delta.$$

Tanmateix anirem calculant les correlacions dels altres predictors,

$$\frac{1}{N}|\langle x_l, r(\lambda)\rangle|,$$

tal que $l \notin \mathcal{A}$, per seleccionar la més elevada que entrarà al conjunt actiu, és a dir, quan es compleixi que

$$\frac{1}{N}|\langle x_l, r(\lambda)\rangle| = \lambda,$$

definint així el nou λ_k. Finalment actualitzarem el conjunt actiu $\mathcal{A} = \mathcal{A} \cup \{l\}$, els nous coeficients $\beta^k = \beta(\lambda_k) = \beta^{k-1} + (\lambda_{k-1} - \lambda_k)\Delta_{\mathcal{A}}$ i els residus $r^k = y - X\beta^k$.

El nom de 'Least Angle Regression' deriva del fet que quan augmentem els coeficients de β^{k-1}, el vector ajustat evoluciona en la direcció $X\Delta = X_{\mathcal{A}}\delta$, i el seu producte intern amb cada vector actiu ve donat per $X'_{\mathcal{A}}X_{\mathcal{A}}\delta = \frac{1}{N\lambda_{k-1}}X'_{\mathcal{A}}r_{k-1}$. Com que totes les columnes de X tenen norma 1, tenim que els angles entre cada vector actiu i el vector ajustat són iguals i, per tant, mínims.

Modificació del algoritme LAR

Es pot demostrar que, amb una petita modificació, aquest procediment dona lloc a tot el camí de les solucions del lasso.

La modificació necessària és: si un coeficient diferent de zero arriba a zero, aquest s'elimina del conjunt actiu $\mathcal{A}$ de predictors i es torna a calcular la direcció de mínims quadrats conjunta.

Podem donar un argument heurístic de per què el LAR i el Lasso són dos procediments tan similars. Tal com hem observat, a qualsevol pas de l'algoritme tenim

$$x'_j(y - X\beta(\lambda)) = \lambda s_j, \forall j \in \mathcal{A}, \tag{5.15}$$

on $s_j \in \{-1, 1\}$ indica el signe del producte intern i λ el seu valor absolut. També tenim que per definició del conjunt actiu del LAR, $|x'_k(y - X\beta(\lambda))| \leq \lambda, \forall k \notin \mathcal{A}$.

Ara considerem el criteri del Lasso:

$$R(\beta) = \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

Sigui $\mathcal{B}$ el conjunt actiu de variables de la solució per a un determinat λ. Per a aquestes variables, $R(\beta)$ és diferenciable i

$$x'_j(y - X\beta) = \lambda sign(\beta_j), \forall j \in \mathcal{B}. \tag{5.16}$$

Si comparem (5.15) i (5.16), veiem que són identiques només quan el signe de β_j coincideix amb el signe del producte intern. Es per això que l'algoritme del LAR i el del lasso són diferents quan un coeficient del conjunt actiu passa per zero; la condició (5.16) no es compleix per aquest tipus de variables, i s'elimina del conjunt actiu $\mathcal{B}$ quan s'aplica la modificació del LAR per al lasso.

6 La regressió Bridge

La regressió Ridge i la regressió Lasso formen part d'una gran família de regressions penalitzades anomenada regressió Bridge. Aquesta regressió va ser introduïda per primera vegada el 1993 per Frank i Friedman a la revista *Technometrics* [2].

L'estimació del Bridge es pot obenir minimitzant la següent expressió per a $\gamma \geq 0$,

$$\min_{\beta_0,\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\} \quad \text{tal que} \quad \sum_{j=1}^{p} |\beta_j|^\gamma \leq t. \tag{6.1}$$

per a un paràmetre de regularització $t \geq 0$, o de manera equivalent,

$$\min_{\beta_0,\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^\gamma \right\}. \tag{6.2}$$

per a un paràmetre de penalització $\lambda \geq 0$.

Els problemes (6.1) i (6.2) són equivalents. Per a una $0 \leq \lambda \leq +\infty$ donada, existeix una $t \geq 0$, tal que els dos problemes tenen la mateixa solució, i viceversa.

El problema del Bridge utilitza la norma l_γ ($\gamma > 0$) com a penalització i, per tant, inclou el ridge ($\gamma = 2$) i el lasso ($\gamma = 1$) com a casos especials. Se sap que si $0 < \gamma \leq 1$, els estimadors mitjançant el bridge produeixen models econòmics en predictors, i quan $\gamma > 1$ redueix els coeficients cap a 0.

És desitjable optimitzar el paràmetre γ, juntament amb el paràmetre de regularització o penalització, mitjançant una validació encreuada generalitzada (Fu, 1998 [4]).

7 Validació encreuada

El paràmetre t, lligat a la restricció del Lasso (5.1), controla la complexitat del model; valors més grans de t alliberen més paràmetres i permeten que el model s'adapti més a les dades d'entrenament. Per contra, valors més petits de t restringeixen més els parmètres, fet que condueix a models amb pocs predictors i més interpretables que s'ajusten a les dades de forma menys acurada. Oblidant-nos de la interpretabilitat, podem demanar el valor de t que proporciona el model més precís per predir les dades de prova, independents de la mateixa població. Aquesta precisió s'anomena capacitat de generalització del model. Un valor de t massa petit pot evitar que el Lasso capti el senyal principal de les dades, mentre que un valor massa gran pot provocar un ajust excessiu. En aquest darrer cas, el model s'adapta al soroll, és a dir, als predictors amb menys capacitat de predicció que hi ha a les dades d'entrenament. En ambdós casos, l'error de predicció del conjunt de proves s'inflarà. Normalment hi ha un valor intermedi de t que assoleix un bon equilibri entre aquests dos extrems i, en el procés, produeix un model amb alguns coeficients iguals a zero.

Per tal d'estimar el millor valor de t, podem crear conjunts d'entrenament i de prova dividint a l'atzar el conjunt total de les dades i estimant el rendiment amb les dades de prova, mitjançant un procediment conegut com a validació encreuada.

Amb més detall, primer dividim aleatòriament el conjunt de dades complet en K grups amb $K > 1$. Les opcions típiques de K poden ser 5 o 10 i, de vegades, N (*leave one out*). Fixem un grup com a conjunt de prova i designem els $K - 1$ grups restants com el conjunt d'entrenament. A continuació, apliquem el Lasso a les dades d'entrenament per a un rang de valors de t diferents i fem servir cada model ajustat per predir les respostes del conjunt de dades de prova, calculant la mitjana dels errors de predicció al quadrat per a cada valor de t. Aquest procés es repeteix un total de K vegades, on cadascun dels grups K jugarà el paper de dades de prova, amb els grups $K - 1$ restants utilitzats com a dades d'entrenament. D'aquesta manera, obtenim K diferents estimacions de l'error de predicció en un interval de valors de t. Per aquestes K estimacions de l'error de predicció es calcula la mitjana per a cada valor de t, produint així una corba d'error de la validació encreuada.

Per tant, el nostre paràmetre de regularització òptim serà aquell que tingui l'error mitjà de la validació encreuada més petit.

Observem que el procés de validació encreuada anterior està centrat en el paràmetre t. També es pot dur a terme una validació creuada utilitzant la forma lagrangiana (5.5), centrant-se en el paràmetre λ. Els dos mètodes donaran resultats similars però no idèntics, ja que el mapatge entre t i λ depèn de les dades.

8 Elastic net

Limitacions del Lasso

Encara que el Lasso funcioni correctament en moltes situacions, té algunes limitacions. Considerem els diferents casos:

a. Quan $p > N$, el lasso selecciona almenys N variables abans que es saturi, degut a la naturalesa del problema convex d'optimització. Això sembla un problema bastant limitant en un mètode de selecció de variables. A més, el lasso no està ben definit tret que el límit de la norma l_1 dels coeficients sigui inferior a un valor determinat.

b. Si hi ha un grup de variables entre les quals les correlacions dos a dos són molt elevades, el lasso tendeix a seleccionar només una variable del grup sense importar quina és la seleccionada.

c. Per a situacions on $N > p$, si hi ha correlacions elevades entre predictors, s'ha observat empíricament que el rendiment de predicció del lasso és inferior que el de la regressió ridge.

D'aquesta manera sorgeix l'elastic net, una nova tècnica de regularització introduïda per Zou i Hastie el 2005 [6].

Naïve Elastic net

De manera similar al lasso, la regularització elastic net fa simultàniament la selecció i contracció contínua de les variables, i pot seleccionar grups de variables correlacionades.

Suposem les variables resposta centrades i els predictors estandaritzats.

Per a qualsevol valor fixat de $\lambda_1, \lambda_2 \geq 0$, definim el criteri del naïve elastic net, com l'estimador $\hat{\beta}$ que minimitza la funció següent:

$$\min_{\beta} \left\{ \frac{1}{2N} \|y - X\beta\|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \right\} \tag{8.1}$$

Aquest procediment també es pot escriure mitjançant una penalització. Sigui $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$; aleshores el problema (8.1) és equivalent a:

$$\min_{\beta} \left\{ \frac{1}{2N} \|y - X\beta\|_2^2 \right\} \quad \text{tal que} \quad (1-\alpha)\|\beta\|_1 + \alpha\|\beta\|_2^2 \leqslant t, \tag{8.2}$$

on $P_\alpha = (1-\alpha)\|\beta\|_1 + \alpha\|\beta\|_2^2$ és la penalització elastic net, una combinació convexa de les penalitzacions de la regressió ridge ($\alpha = 0$) i la penalització del lasso ($\alpha = 1$). A continuació considerarem només el cas $\alpha < 1$.

L'elastic net amb $\alpha = 1 - \epsilon$, per a un $\epsilon > 0$ petit, té un rendiment similar al lasso, però elimina qulsevol comportament arbitrari causat per correlacions extremes. Més generalment, tota la família P_α crea un compromís entre el rigde i el lasso. A mesura que α augmenta de 0 a 1, per a una λ donada, el número coeficients de la regressió iguals a zero augmenta monotònicament de 0 a el número coeficients de la regressió iguals a zero de la solució del lasso.

Càlcul de solucions

Resulta que resoldre l'equació (8.1) és equivalent a un problema d'optimització com el lasso. Aquest fet implica que el naïve elastic net gaudeix de les mateixes avantatges computacionals que el lasso.

Lema 8.1. *Donat (X, y) un conjunt de dades i (λ_1, λ_2), es defineix el conjunt de dades artificials (X^*, y^*) com*

$$X^*_{(N+p)\times p} = (1+\lambda_2)^{\frac{-1}{2}} \begin{pmatrix} X \\ \sqrt{\lambda_2} I \end{pmatrix}, \qquad y^*_{(N+p)} = \begin{pmatrix} y \\ 0 \end{pmatrix}.$$

Sigui $\gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}$ i $\beta^* = \sqrt{1+\lambda_2}\beta$. Aleshores el criteri de naïve elastic net es pot reescriure com

$$\min_{\beta^*} \left\{ \frac{1}{2N} \|y^* - X^*\beta^*\|_2^2 + \gamma \|\beta^*\|_1 \right\}.$$

Finalment tenim

$$\hat{\beta} = \frac{1}{\sqrt{1+\lambda_2}} \hat{\beta}^*$$

El lema (8.1) ens diu que podem transformar el problema de naïve elastic net en un problema de tipus lasso equivalent amb dades augmentades. La mida de la mostra en les dades augmentades és $N + p$ i la matriu X^* té rang p, el que comporta que aquest procediment pot seleccionar els p predictors en qualsevol situació. Això resol la primera de les limitacions del lasso explicada en l'escenari (a).

Aquest mateix lema també ens mostra que el naïve elastic net pot efectuar una selecció de variables de manera similar al lasso, però aquest mètode té la propietat d'efectuar selecció de variables agrupades, una qualitat que el lasso no comparteix.

Selecció de variables agrupades

En el cas $p \gg N$, la propietat seleccionadora de variables agrupades té una gran importància.

Considerem un mètode de penalitzaicó genèric:

$$\hat{\beta} = \min_{\beta} \|y - X\beta\|_2^2 + \lambda J(\beta), \tag{8.3}$$

on $J(\cdot)$ és positiu per a $\beta \neq 0$.

Qualitativament parlant, un mètode de regressió posseeix l'efecte d'agrupament si els coeficients de regressió d'un grup de variables altament correlacionades solen ser iguals (fins i tot un canvi de signe si es correlaciona negativament). En particular, en la situació extrema en què algunes variables són exactament iguals, el mètode de regressió hauria d'assignar coeficients idèntics a les variables idèntiques.

Lema 8.2. *Suposem que $x_i = x_j$ per a $i, j \in \{1, \cdots, p\}$.*

a. *Si $J(\cdot)$ és estrictament convex, aleshores $\hat{\beta}_i = \hat{\beta}_j$, $\forall \lambda > 0$.*

b. *Si $J(\beta) = |\beta|_1$, aleshores $\hat{\beta}_i \hat{\beta}_j \geq 0$ i $\hat{\beta}^*$ també és un mínim de l'equació (8.3) on*

$$\hat{\beta}_k^* = \begin{cases} \hat{\beta}_k & si \quad k \neq i \quad i \quad k \neq j \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot (s) & si \qquad k = i \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot (1 - s) & si \qquad k = j \end{cases}$$

per qualsevol $s \in [0, 1]$.

El lema (8.2) mostra una clara distinció entre les funcions penalitzades estictament convexes i la penalització del lasso. La convexitat estricta garanteix l'efecte d'agrupament en la situació extrema amb predictors idèntics. En canvi, el lasso ni tan sols té una solució única. La penalització de l'elastic net amb $\lambda_2 > 0$ és estrictament convexa i, per tant, posseeix aquesta propietat.

Teorema 8.3. *Donat (X, y) un conjunt de dades i (λ_1, λ_2), amb la resposta centrada i els predictors estandaritzats. Sigui $\hat{\beta}(\lambda_1, \lambda_2)$ la solució del naïve elastic net. Suposem que $\hat{\beta}_i(\lambda_1, \lambda_2)\hat{\beta}_j(\lambda_1, \lambda_2) > 0$. Definim*

$$D_{\lambda_1, \lambda_2}(i, j) = \frac{1}{|y|_1} |\hat{\beta}_i(\lambda_1, \lambda_2) - \hat{\beta}_j(\lambda_1, \lambda_2)|;$$

aleshores

$$D_{\lambda_1, \lambda_2}(i, j) \leq \frac{1}{\lambda_2} \sqrt{2(1 - \rho)},$$

on $\rho = x_i' x_j$ el coeficient de correlació entre x_i i x_j.

$D_{\lambda_1, \lambda_2}(i, j)$ descriu la diferència entre els camins dels predictors i i j. Si x_i i x_j estan molt correlacionats, és a dir, $\rho \approx 1$ (si $\rho \approx -1$ considerem $-x_j$), el teorema ens diu que la diferència entre els coeficients associats als predictors i i j és gairabé 0. El límit superior de l'anterior desigualtat proporciona una descripció quantitativa de l'efecte d'agrupació del naïve elastic net.

Com a mètode de selecció de variables agrupades, la naïve elastic net supera les limitacions del lasso en els escenaris (a) i (b). Tot i això, evidències empíriques mostren que aquest procediment no té un rendiment satisfactori a menys que estigui molt a prop de la regressió ridge o del lasso. És per aquest motiu pel qual en diem 'naïve' elastic net.

El mètode de naïve elastic net és un procediment en dues etapes: per a cada λ_2 fix, primer troba els coeficients de regressió ridge i després s'aplica la contracció del lasso. És a dir, fem una doble contracció dels coeficients, que no col·labora a reduir les variàncies i introdueix un biaix extra innecessari, comparat amb les regressions ridge o lasso per separat. Però hi ha una manera per tal de corregir aquesta doble contracció.

Elastic Net

Donat un conjunt de dades (y, X), els paràmetres de penalització (λ_1, λ_2) i les dades augmentades (y^*, X^*), el naïve elastic net soluciona el problema de tipus lasso:

$$\hat{\beta}^* = \min_{\beta^*} \left\{ \frac{1}{2N} \|y^* - X^*\beta^*\|_2^2 \frac{\lambda_1}{\sqrt{1+\lambda_2}} \|\beta^*\|_1 \right\}.$$

Lestimació de l'elastic net (corregit) $\hat{\beta}$ es defineix com:

$$\hat{\beta}(\text{elastic net}) = \sqrt{1+\lambda_2}\hat{\beta}^*$$

Recordem que $\hat{\beta} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\beta}^*$,

aleshores

$$\hat{\beta}(\text{elastic net}) = (1+\lambda_2)\hat{\beta}(\text{naïve elastic net}).$$

Per tant, l'estimador de l'elastic net és l'estimador del naïve elastic net redimensionat. Aquesta transformació d'escala fa que es conservi la propietat seleccionadora del naïve elastic net i, a més, és la manera més senzilla de desfer la contracció. Empíricament hem descobert que l'elastic net té un bon rendiment en comparació amb les regressions ridge i lasso.

El següent teorema ens dona una altre presentació del Elastic Net, en el que l'argument de descorrelació és més explícit.

Teorema 8.4. *Donat (X, y) un conjunt de dades i (λ_1, λ_2), aleshores l'estimador de l'elastic net $\hat{\beta}$ ve donat per la següent expressió:*

$$\hat{\beta} = \min_{\beta} \beta' \left(\frac{X'X + \lambda_2 I}{1+\lambda_2} \right) \beta - 2y'X\beta + \lambda_1|\beta|_1 \tag{8.4}$$

El Teorema (8.4) interpreta el elastic net com una versió estabilitzada del lasso, ja que

l'estimador ve donat per

$$\hat{\beta}(lasso) = \min_{\beta} \beta'(X'X)\beta - 2y'X\beta + \lambda_1|\beta|_1. \tag{8.5}$$

Veiem que $\hat{\Sigma} = X'X$ és una versó reduïda de la matriu de correlacions Σ, i

$$\frac{X'X + \lambda_2 I}{1 + \lambda_2} = (1 - \gamma)\hat{\Sigma} + \gamma I,$$

amb $\gamma = \frac{\lambda_2}{1+\lambda_2}$ que contrau la matriu $\hat{\Sigma}$ cap a la identitat.

Les equacions (8.4) i (8.5) mostren que reescalar després de la penalització de l'elastic net és matemàticament equivalent a reemplaçar $\hat{\Sigma}$ per la seva versió reduïda del lasso.

9 Implementació en R

En els annexos es poden trobar quatre notebooks, escrits en R, que contenen exemples explicats en detall d'alguns dels procediments que hem vist durant tot el treball i on s'utilitzen els paquets que s'expliquen a continuació. Els notebooks són els següents:

I. Prostate

II. Hitters

III. lu2004

IV. LARS Diabetes

9.1 Paquet glmnet

El paquet glmnet de R realitza procediments extremadament ràpids i eficients per ajustar la regularització del Ridge, Lasso o Elastic net per a models de regressió lineal, logística i multinomials, la regressió de Poisson, el model del Cox, el model de resposta múltiple de Gauss i la regressió multinomial agrupada. Calcula tots els camins de les solucions mitjançant el descens seguint les coordenades i es poden fer diverses prediccions a partir dels models ajustats.

A continuació, s'explicaran algunes de les funcions que es poden utilitzar a l'hora de realitzar l'anàlisi de regressió d'un model de regressió lineal.

glmnet

Ajusta el model lineal generalitzat mitjançant la máxima versemblança penalitzada. El camí de regularització es calcula per a una quadrícula de valors del paràmetre de regularització λ.

glmnet(x, y, alpha, lambda ...)

- **x**: matriu d'entrada formada pels predictors de dimensió $N \times p$.
- **y**: variable resposta.
- **alpha**: paràmetre amb $0 \leq \alpha \leq 1$. La penalització es defineix com

$$\frac{1-\alpha}{2}\|\beta\|_2^2 + \alpha|\beta|_1$$

  Així per a $alpha = 0$, la penalització coincideix amb la de la regressió Ridge, i per a $alpha = 1$ coincideix amb la de la regressió Lasso.

- **lambda**: seqüència de valors de lambda.

plot

Produeix un gràfic de la trajectòria dels coeficients estimats mitjançant la funció glmnet.

plot(x, xvar=c('norm', 'lambda', 'dev') ...)

- **x**: model creat anteriorment amb la funció glmnet.
- **xvar**: eix x: 'norm' norma L1 dels coeficients, 'lambda' log(lambda) i 'dev' el percentatge de desviació explicada.

cv.glmnet

Efectua la validació encreuada en K passos, produeix un gràfic, i retorna el valor de lambda que fa mínim l'error de la validació encreuada.

cv.glmnet(x, y, lambda, alpha ...)

- **x**: matriu de predictors com en la funció glmnet.
- **y**:variable resposta com en la funció glmnet.
- **alpha**: paràmetre que correspon a la penalització de la funció glmnet.
- **lambda**: seqüència de valors de lambda.

predict

De manera similar a altres mètodes de predicció, aquesta funció prediu valors ajustats, logit, coeficients i més objectes a partir d'un model 'glmnet' ajustat.

predict(object, s, type=c('link', 'response', 'coefficients', 'nonzero', 'class') ...)

- **object**: model creat anteriorment amb la funció glmnet en el que volem predir.
- **s**: valor del paràmetre de penalització lambda en què es requereixen les prediccions.
- **type**: tipus de predicció (només utilitzarem el de predir coeficients).

9.2 Paquet lars

El paquet lars de R realitza procediments eficients per calcular tota una seqüència de solucions del lasso amb el cost d'un ajust per mínims quadrats. La regressió LAR i la regressió Forward Stagewise estan relacionades amb el lasso, tal com es descriu al document on s'introdueïx el mètode LAR.

A continuació, s'explicaran algunes de les funcions que es poden utilitzar a l'hora de realitzar l'anàlisi de regressió d'un model de regressió lineal.

lars

Calcula totes les variants del Lasso i proporciona tota la seqüència de coeficients i ajustos, començant des de zero, fins a l'ajust de mínims quadrats. Amb l'opció 'lasso', calcula la solució completa de lasso simultàniament per a tots els valors del paràmetre de contracció amb en el mateix cost computacional que l'ajust per mínims quadrats.

lars(x, y, type = c('lasso', 'lar', 'forward.stagewise', 'stepwise')...)

- **x**: matriu de predictors.
- **y**: variable resposta.
- **type**: calcula les següents regressions: lasso, lar, forward.stagewise or stepwise.

cv.lars

Calcula l'error quadràtic mitjà de predicció mitjançant la validació encreuada de K passos per al lars, el lasso o la Forward Stagewise.

cv.lars(x, y, type = c('lasso', 'lar', 'forward.stagewise', 'stepwise'), mode=c('fraction', 'step'),...)

- **x**: matriu de predictors.
- **y**: variable resposta.
- **type**: calcula les següents regressions: lasso, lar, forward.stagewise or stepwise.
- **mode**: es refereix a l'índex que s'utilitza per a la validació encreuada.

plot.lars

Realitza una representació gràfica d'un ajust realitzat amb la funció lars.

plot(x, xvar= c('norm', 'df', 'arc.length', 'step'),...)

predict.lars

Mentre que la funció lars proporciona la trajectòria de la solució completa, la funció predict-lars, permet extreure una predicció en un punt en particular al llarg del camí de solucions.

predict(object, s, type = c('fit', 'coefficients'), mode = c('step', 'fraction', 'norm', 'lambda'), ...)

Referències

[1] Hoerl, A.E.; Kennard, R.W. (1970). Ridge regression: biased estimation for non-orthogonal problems, *Technometrics*. Vol 12, No. 1, pp. 55-67.

[2] Frank, I. E.: Friedman, J. H. (1993). A statistical view of some chemometrics regression tools, *Technometrics*, Vol. 35, No. 2, pp. 109-135.

[3] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 58, No. 1, pp. 267-288.

[4] Fu, W. J. (1998). Penalized regression: the Bridge versus the Lasso, *Journal of Computational and Graphical Statistics*. Vol 7, No. 3, pp.397-416.

[5] `https://en.wikipedia.org/wiki/Tikhonov_regularization`.

[6] Zou, H.: Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*. 67, Part 2, pp. 301-320.

[7] Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. (2004). Least Angle Regression. *Annals of Statistics*. Vol 32, No.2, pp. 407-499

[8] Efron, B.; Hastie, T. (2016). *Computer Age Statistical Inference*, Cambridge University Press, pp. 298-321.

[9] James, G.; Witten, D.; Hastie, T.; Tibshirani, R; Witten, D. (2013). *An Introduction to Statistical Learning: with Applications in R*, pp. 203-259.

[10] Hastie, T.; Tibshirani, R; Wainwright, M. (2015). *Statistical learning with sparsity*. Monographs on Statistics and Applied Probability 143.

[11] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.