Universitat de les Illes Balears

escola politècnica superior

Treball Final de Grau

GRAU DE MATEMÀTIQUES

El Mètode HyperLASSO i Aplicacions

MARIA DEL MAR BIBILONI FEMENIAS

Tutor
Jairo Enrique Rocha Cárdenas

Escola Politècnica Superior
Universitat de les Illes Balears
Palma, Setembre de 2017

Agraïments

Quiero agradecer a Jairo Rocha su tiempo y esfuerzo dedicado a orientarme en el desarrollo del Trabajo de Final de Grado, así como en despertar mi interés por la optimización y estadística aplicada al problema que aquí se describe.

També vull agrair a la meva família i amics per transmetre'm el seu incondicional suport i confiança. En especial als meus pares, als meus germans i a en David, per acompanyar-me en tot moment.

SUMARI

ÍNDEX DE FIGURES

Resum

En aquest treball de Final de Grau de Matemàtiques de la Universitat de les Illes Balears s'estudia el mètode d'HyperLASSO, descrit principalment en [3], que pretén resoldre el problema de selecció de variables quan el nombre de variables de la mostra excedeix en gran mesura al nombre de dades. Així, es descriuen les tècniques matemàtiques d'estadística i regressió penalitzada que donen lloc al mètode. A més de modificacions per fer l'algoritme més eficient.

El mètode d'HyperLASSO és principalment un mètode de regressió penalitzada, que sorgeix de suposar que els coeficients de les variables independents segueixen una distribució normal-exponencial-gamma de mitjana zero. Aquesta distribució s'obté de generalitzar la funció de densitat de Laplace, i la seva densitat presenta un cim més contret entorn al zero i coes més amples que la densitat de Laplace i d'una normal. Aquestes característiques són les que motiven l'ús d'aquesta distribució.

Amb l'objectiu d'observar l'efecte de HyperLASSO, s'ha afegit un capítol d'exemples amb dades reals de mutacions en els gens d'individus amb càncer de pulmó. D'aquesta manera, s'inclouen en el contest de GWAS. En aquest capítol es fan presents les poques variables seleccionades en el model, per certs casos.

CAPÍTOL 1

INTRODUCCIÓ

Establir un model matemàtic per explicar el comportament d'una variable en funció d'altres és una tècnica emprada sovint per intentar extreure informació d'un conjunt de dades. Així, suposem que tenim una variable Y i hi ha indicis de que el seu valor pot estar condicionat per algunes variables independents d'un conjunt $X = \{X_j\}_{j=1,\dots,m}$. El problema que es presenta és esbrinar quines d'aquestes covariables expliquen la variable Y, és a dir, seleccionar un subconjunt de variables del conjunt X [1]. Per resoldre aquest problema de selecció de variables existeixen diverses eines de regressió, que depenen del tipus de dades que es volen analitzar i la relació que es vol establir. En aquest treball ens centrarem en la regressió logística, que es basa en establir un model per explicar una variable binària Y que, per exemple, pot representar la presència o no d'una malaltia.

Un gran nombre de possibles variables explicatives dificulta l'obtenció del model, ja que les eines de regressió tradicionals seleccionen un subconjunt de X força gran, però l'objectiu és ser selectius, per tal d'explicar Y d'una manera senzilla, només amb aquelles variables que més li afectin [2]. L'obtenció de mètodes que identifiquin un conjunt reduït de covariables explicatives motiva l'estudi de noves eines de regressió i el tema del Treball de Final de Grau.

Concretament, en aquest treball s'estudia el mètode d'HyperLASSO proposat per Hoggart *et al.*, descrit principalment a [3], que pretén resoldre el problema d'associació de canvis en el genoma humà amb la presència de malalties. Així, s'elabora un mètode eficient per treballar amb un nombre elevat de variables independents. En el capítol 3 es descriu el model matemàtic, que es basa en regressió logística penalitzada amb una funció de penalització que presenta característiques útils per l'obtenció del model. També s'expliquen detalls de la implementació de HyperLASSO i com controlar l'error de Tipus I.

Els precursors principals d'HyperLASSO són els models de regressió logística i, en particular, LASSO. Per tant, per entendre i situar HyperLASSO s'ha redactat el capítol 2, una introducció a aquests models, on s'expliquen els canvis en la regressió logística tradicional que han conduït al tema que es tracta en aquest treball.

A més, en el capítol 4 es donen dos exemples on s'aplica HyperLASSO a dades reals d'individus amb càncer de pulmó. Així, veurem com funciona el mètode estudiat per un problema concret i com introdueix poques variables en el model. Finalment, es presenta un capítol amb les conclusions, les aportacions al treball i habilitats adquirides.

CAPÍTOL 2

REGRESSIÓ LOGÍSTICA PENALITZADA

En aquest capítol veurem com han anat canviat les eines de regressió logística en funció del nombre de variables que es vulguin seleccionar. Així, veurem quins són els precursors de HLASSO i com el model és cada vegada més fi.

2.1. Regressió Logística lineal

Donada una mostra $\{(X_i, Y_i)\}_{i=1,\dots,n}$, on X_i és un vector de variables i Y_i una variable binària que classifica cada X_i en la classe C_1 o C_2, volem resoldre el problema de selecció de variables, és a dir, determinar quines variables $X_{i1}, \dots, X_{im}$ afecten més al valor de Y_i, i per tant, en la classificació de X_i. En general, quan Y_i pren valors reals, es sol emprar regressió lineal suposant que $Y_i = \beta_0 + \beta^T X_i$, però aquest model no és útil quan Y_i és binària i cal emprar-ne un que s'adapti millor a les dades. A continuació, s'explica un model basat en regressió logística lineal.

Considerem un parell de variables aleatòries (X_i, Y_i) per a cada $i = 1, \dots, n$ de la mostra, amb $Y_i \in \{-1, 1\}$ i $X_i \in \mathbb{R}^m$. Sigui π_i la probabilitat de que Y_i prengui el valor 1 donada la nostra mostra, és a dir, $\pi_i = p(Y_i = 1|X)$, $X = (X_1, \dots, X_n)$, es defineix el paràmetre *logit* com $\lambda_i := \ln(\text{odds}(\pi_i))$ [4], on $\text{odds}(p) = \dfrac{p}{1-p}$.

Aquesta expressió permet aïllar π_i en funció de λ_i de la següent manera

$$\begin{aligned} -\lambda_i &= \ln\left(\frac{1-\pi_i}{\pi_i}\right), \\ \frac{1}{1+e^{-\lambda_i}} &= \pi_i. \end{aligned}$$

A més, el logaritme d'odds pren valors en $(-\infty, \infty)$, fet que facilita imposar linealitat de λ_i en X_i: $\lambda_i = \beta_0 + \beta^T X_i$, amb $\beta = (\beta_1, \dots, \beta_m)$ [5]. Així, el problema de selecció de variables serà determinar quines coordenades X_{ij} afecten més al valor de λ_i, és a dir,

quins coeficients β_j són significativament diferents de zero. Amb aquesta condició es defineix la *funció logística*,

$$\pi_i(X_i) = \frac{1}{1+e^{-(\beta_0+\beta^T X_i)}}.$$

Per unificar les expressions $p(Y_i = 1|X)$ y $p(Y_i = -1|X)$ en una única expressió $p(Y_i|X)$ [6], sigui $\bar{\beta} := (\beta_0, \beta_1, ..., \beta_m)$, considerem

$$p(Y_i|X,\bar{\beta}) = \frac{1}{1+e^{-Y_i(\beta_0+\beta^T X_i)}}, \tag{2.1}$$

que satisfà

$$\begin{aligned} p(Y_i = 1|X,\bar{\beta}) &= \frac{1}{1+e^{-(\beta_0+\beta^T X_i)}} = \pi_i, \\ p(Y_i = -1|X,\bar{\beta}) &= \frac{1}{1+e^{(\beta_0+\beta^T X_i)}} = \frac{1}{1+\frac{1}{e^{-(\beta_0+\beta^T X_i)}}} = \frac{e^{-(\beta_0+\beta^T X_i)}}{1+e^{-(\beta_0+\beta^T X_i)}} = 1-\pi_i. \end{aligned}$$

Notem que cada Y_i té una funció de probabilitat $p(Y_i|X,\bar{\beta})$ diferent, ja que la fórmula (2.1) depèn del valor en les components del vector X_i, i cada Y_i té un vector X_i diferent associat. Per tant, les variables aleatòries Y_i no són idènticament distribuïdes. Els paràmetres β_0 i β, en canvi, sí que són els mateixos per a cada Y_i, per definició del model.

Suposant que les variables Y_i són independents donades les covariables X_{ij}, podem aplicar el mètode de màxima versemblança per estimar els paràmetres desconeguts, imposant que la probabilitat d'obtenir la mostra donada sigui màxima, és a dir, maximitzant la funció de versemblança $\mathscr{L}(\bar{\beta}|Y,X)$, on $Y := (Y_1, ..., Y_n)$ i $X := (X_1, ..., X_n)$ representen la mostra. En aquest cas, tenim que

$$\mathscr{L}(\bar{\beta}|Y,X) = p(Y|X,\bar{\beta}) = \prod_{i=1}^{n} p(Y_i|X,\bar{\beta}) = \prod_{i=1}^{n} \frac{1}{1+e^{-Y_i(\beta_0+\beta^T X_i)}}. \tag{2.2}$$

Així, donat que maximitzar $\mathscr{L}(\bar{\beta}|Y,X)$ és equivalent a maximitzar $\ln \mathscr{L}(\bar{\beta}|Y,X)$, el problema consisteix a trobar els paràmetres β i β_0 tals que (2.3) sigui màxim,

$$\sum_{i=1}^{n} \ln\left(\frac{1}{1+e^{-Y_i(\beta_0+\beta^T X_i)}}\right). \tag{2.3}$$

La funció objectiu és diferenciable y còncava en $\bar{\beta}$. Per tant, no podem assegurar la unicitat de solucions.

La diferenciabilitat de la funció es deriva de que cada sumand és diferenciable, per ser-ho la funció logística que, a més, és estrictament positiva. Per tal de demostrar la concavitat, s'empra el següent Teorema, tal com s'indica a [7].

Teorema 2.1.1. *Sigui $S \subseteq \mathbb{R}^m$ un conjunt convex i $f : S \longrightarrow \mathbb{R}$ dues vegades diferenciable en S. Si la matriu hessiana de f en x, $H_f(x)$, es semidefinida positiva per a tot $x \in S$, aleshores f és convexa.*

Resultat 2.1.1. *Sigui $\ln \mathscr{L}(\bar{\beta}|Y,X)$ la funció definida en (2.3), $-\ln \mathscr{L}(\bar{\beta}|Y,X)$ és convexa en $\mathbb{R}^{m+1}$.*

Demostració. Sigui $h(\bar{\beta}) = -\ln p(Y_i|X,\bar{\beta})$, dues vegades diferenciable en tot $\mathbb{R}^{m+1}$, vegem que $H_h(\bar{\beta})$ és semidefinida positiva per a tot $\bar{\beta} \in \mathbb{R}^{m+1}$. És a dir, que es satisfà

$$a^T H_h(\bar{\beta}) a \geq 0, \quad \forall \bar{\beta} \in \mathbb{R}^{m+1}, \ \forall a \in \mathbb{R}^{m+1}.$$

D'aquesta manera, $-\ln \mathscr{L}(\bar{\beta}|Y,X)$ serà convexa per ser suma de funcions convexes.

Per simplificar la notació, considerem $X_i = (X_{i0}, X_{i1}, ..., X_{im})$ afegint $X_{i0} = 1$, per a tot i. Així, les derivades de primer ordre de h són,

$$\frac{\partial h}{\partial \beta_j} = \frac{-Y_i X_{ij}}{1 + e^{Y_i \bar{\beta}^T X_i}}, \quad j = 0, ..., m.$$

Per tant, les de segon ordre

$$\frac{\partial h^2}{\partial^2 \beta_j} = \frac{X_{ij}^2 e^{Y_i \bar{\beta}^T X_i}}{\left(1 + e^{Y_i \bar{\beta}^T X_i}\right)^2}, \quad \frac{\partial h^2}{\partial \beta_j \beta_k} = \frac{X_{ij} X_{ik} e^{Y_i \bar{\beta}^T X_i}}{\left(1 + e^{Y_i \bar{\beta}^T X_i}\right)^2}, \qquad j,k = 0, ..., m \text{ amb } j \neq k.$$

D'aquesta manera, la matriu hessiana ve determinada per

$$H_h(\bar{\beta}) = \frac{e^{Y_i \bar{\beta}^T X_i}}{\left(1 + e^{Y_i \bar{\beta}^T X_i}\right)^2} \begin{pmatrix} X_{i0}^2 & X_{i0}X_{i1} & X_{i0}X_{i2} & \cdots & X_{i0}X_{im} \\ X_{i1}X_{i0} & X_{i1}^2 & X_{i1}X_{i2} & \cdots & X_{i1}X_{im} \\ X_{i2}X_{i0} & X_{i2}X_{i1} & X_{i2}^2 & \cdots & X_{i2}X_{im} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{im}^2 & X_{im}X_{i1} & X_{im}X_{i2} & \cdots & X_{im}^2 \end{pmatrix}.$$

Vegem que és semidefinida positiva,

$$\begin{aligned} a^T H_h(\bar{\beta}) a &= \frac{e^{Y_i \bar{\beta}^T X_i}}{\left(1 + e^{Y_i \bar{\beta}^T X_i}\right)^2} \left(\sum_{k=0}^{m} a_k X_{ik} X_{i0}, \ldots, \sum_{k=0}^{m} a_k X_{ik} X_{im} \right) \cdot \begin{pmatrix} a_0 \\ \vdots \\ a_m \end{pmatrix} \\ &= \frac{e^{Y_i \bar{\beta}^T X_i}}{\left(1 + e^{Y_i \bar{\beta}^T X_i}\right)^2} \left(\sum_{k=0}^{m} a_k X_{ik} X_{i0} a_0 + \cdots + \sum_{k=0}^{m} a_k X_{ik} X_{im} a_m \right) \end{aligned}$$

$$
\begin{aligned}
&= \frac{e^{Y_i\bar{\beta}^T X_i}}{\left(1+e^{Y_i\bar{\beta}^T X_i}\right)^2}\left(\sum_{k=0}^{m}\sum_{l=0}^{n} a_k X_{ik} X_{il} a_l\right)\\
&= \frac{e^{Y_i\bar{\beta}^T X_i}}{\left(1+e^{Y_i\bar{\beta}^T X_i}\right)^2}\left(\sum_{k=0}^{m} a_k X_{ik}\left(\sum_{l=0}^{m} X_{il} a_l\right)\right)\\
&= \frac{e^{Y_i\bar{\beta}^T X_i}}{\left(1+e^{Y_i\bar{\beta}^T X_i}\right)^2}\left(\sum_{k=0}^{m} a_k X_{ik}\right)\left(\sum_{l=0}^{m} X_{il} a_l\right)\\
&= \frac{e^{Y_i\bar{\beta}^T X_i}}{\left(1+e^{Y_i\bar{\beta}^T X_i}\right)^2}\left(\sum_{k=0}^{m} a_k X_{ik}\right)^2 \geq 0.
\end{aligned}
$$

□

2.1.1. Sobreajustament dels coeficients

Suposem que tenim un conjunt de dades $\{X_i\}_{i=1,\dots,n}$ tals que, per a cada $i = 1,\dots,n$, $X_i \in \mathbb{R}^m$ i $m \geq n$. En aquest cas, el nombre de paràmetres β_j a determinar, que coincideix amb el nombre m de coordenades de les dades X_i, supera el nombre de dades del problema, per tant, els punts X_i són linealment separables, és a dir, podem trobar valors de β_0 i β tals que

$$\beta_0 + \beta^T X_i \geq 0 \ \ \forall i \ \text{ tq } \ Y_i = 1, \tag{2.4}$$

$$\beta_0 + \beta^T X_i < 0 \ \ \forall i \ \text{ tq } \ Y_i = -1. \tag{2.5}$$

Així, multiplicant per una constant positiva cada expressió anterior, podem escriure l'equació de l'hiperplà que separa les dades de manera que els coeficients preguin valors $|\beta_0|$ i $||\beta||_1$ tan grans com vulguem.

A més, si escrivim la funció a maximitzar (2.2) segons els valors que pren Y_i, tenim que

$$\mathscr{L}(\bar{\beta}|Y,X) = \prod_{i:\ Y_i=1} \frac{1}{1+e^{-(\beta_0+\beta^T X_i)}} \prod_{i:\ Y_i=-1} \frac{1}{1+e^{\beta_0+\beta^T X_i}}.$$

Per tant, valors molt grans en valor absolut dels coeficients fan que $\mathscr{L}(\bar{\beta}|Y,X)$ sigui gairebé 1, que és el màxim que pren la funció quan $\beta_0 + \beta^T X_i$ és $+\infty$ o $-\infty$ en cada cas.

Aquest fet causa el problema de què el màxim s'assoleix quan els paràmetres són $+\infty$ o $-\infty$. Per tant, maximitzar la funció de màxima versemblança no resol el problema d'estimar el paràmetres del model descrit. A més, si el $|\beta_0|$ i la $||\beta||_1$ prenen valors molt grans, la funció logística s'ajusta força bé a les dades, reduint el nombre de punts X_i amb una $P(Y_i|X_i)$ entorn el 0.5 [6, 8]. Per aquest motiu, encara que l'objectiu d'estimar els paràmetres és maximitzar la funció (2.2), quan $m \geq n$ s'han sobre-ajustat les dades i el model resultant deixa de ser útil. Aquest fenomen es conegut per *overfitting* i es pot

evitar afegint un *terme de regularització* o *penalització*, per forçar que molts coeficients estiguin entorn al zero.

D'altra banda, si tots els coeficients β resultants són no nuls, es relacionaran totes les coordenades de les nostres dades amb la variable Y_i. Però el nostre objectiu és determinar algunes variables, les que més afectin a la variable Y_i. Per tant, obtenir una relació total no ens aporta informació rellevant.

Una idea per forçar que els coeficients estiguin entorn el zero és imposar que cada $\beta_j \sim \mathcal{N}(0,\sigma^2)$, $j = 1,....,m$. Donat que β_0 no és coeficient de cap variable, no imposam cap condició sobre ell, i suposam que $p(\bar{\beta}|X,Y) = p(\beta|X,Y)$.

Així, suposant que els β_i són independents, prenem la funció de densitat d'una normal de mitjana zero i mateixa variància com a funció a priori dels coeficients β,

$$p(\beta) = \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\beta_j^2}{2\sigma^2}}, \ \ \sigma > 0.$$

Per obtenir la densitat a posteriori dels paràmetres basta aplicar el Teorema de Bayes [9]

$$p(\bar{\beta}|X,Y) = \frac{p(Y|X,\bar{\beta})p(\bar{\beta})}{p(Y)}, \tag{2.6}$$

on el denominador es constant respecte dels β_j, ja que només depèn de la mostra. Per tant,

$$p(\bar{\beta}|X,Y) \propto p(Y|X,\bar{\beta})p(\bar{\beta}).$$

Donat que la funció de versemblança $\mathscr{L}(\bar{\beta}|X,Y)$ es $p(Y|X,\bar{\beta})$ i aplicant logaritmes obtenim

$$\begin{aligned}
\ln p(\bar{\beta}|X,Y) &= \ln\left(\mathscr{L}(\bar{\beta}|X,Y)\prod_{j=1}^{m}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\beta_j^2}{2\sigma^2}}\right)+c \\
&= \ln\mathscr{L}(\bar{\beta}|X,Y)+\sum_{j=1}^{m}\left(-\ln\sqrt{2\pi\sigma^2}-\frac{\beta_j^2}{2\sigma^2}\right)+c \qquad (2.7)\\
&= \ln\mathscr{L}(\bar{\beta}|X,Y)-m\ln\sqrt{2\pi\sigma^2}-\frac{1}{2\sigma^2}\sum_{j=1}^{m}\beta_j^2+c.
\end{aligned}$$

Ara, l'objectiu és maximitzar la funció $p(\bar{\beta}|X,Y)$ (equivalentment maximitzar el logaritme) per estimar els paràmetres β_j desconeguts; per tant, podem ignorar els termes constants de (2.7) i el problema d'optimització resulta

$$\text{maximitzar } \ln\mathscr{L}(\bar{\beta}|X,Y)-\frac{1}{2\sigma^2}||\beta||_2^2.$$

Fixem-nos que el problema és redueix a (2.3) amb un terme de regularització dels coeficients β que, com hem vist, prové de suposar que segueixen una $\mathcal{N}(0,\sigma^2)$. Tot junt, volem

$$\text{maximitzar } \sum_{i=1}^{n} \ln\left(\frac{1}{1+e^{Y_i(-\beta_0-\beta^T X_i)}}\right) - \frac{1}{2\sigma^2}||\beta||_2^2. \tag{2.8}$$

El model de regressió logística lineal, amb un terme quadràtic com a penalització, es coneix com *ridge regression* [10]. El logaritme de la funció de màxima versemblança de $\bar{\beta}$ és còncau, tal com s'ha demostrat amb anterioritat. A més, la resta de quadrats és una funció estrictament còncava. Així, la nova funció objectiu també ho és i, per tant, la solució és única.

Els coeficients resultants de la maximització, tal com s'indica a l'article [11], prenen valors propers a zero. Aquest fet dificulta la interpretació, ja que cercam valors significativament diferents de zero. D'altra banda, sovint interessa determinar un conjunt petit de covariables que afectin, o expliquin, la variable Y_i. Així, quan el nombre de variables és molt gran, es pot determinar un model més útil en aquest sentit, és a dir, que prengui un conjunt menor de coeficients no nuls. Amb l'objectiu de resoldre aquests problemes, obtenint un gran nombre de coeficients exactament zero, sorgeix LASSO.

2.2. LASSO per la regressió logística

LASSO (*least absolute shrinkage and selection operator*) és un mètode de regressió pensat per treballar amb un gran nombre de variables, fixant molts dels coeficients de les variables a zero. En particular, és útil quan el nombre d'elements de la mostra és menor al nombre de variables. El mètode consisteix en resoldre un determinat problema d'optimització, suposant a priori que els coeficients que determinen la regressió que es pretén realitzar segueixen una distribució de Laplace, enlloc d'una normal.

La funció de densitat d'una distribució de Laplace(μ,ψ) ve donada per [12]

$$p(\beta_j|\mu,\psi) = \frac{1}{2\psi} e^{-\frac{|\beta_j-\mu|}{\psi}} \quad \psi > 0,$$

on ψ denota el paràmetre d'escala i μ el de localització.

Si comparam una Laplace$(0,\psi)$ i una $\mathcal{N}(0,\sigma^2)$, amb la mateixa variància, la densitat de Laplace presenta un cim més elevat en el zero i coes més amples. Aquest fet, augmenta el nombre de paràmetres nuls, ja que la probabilitat d'obtenir un coeficient entron al zero és més elevada, i fa que els no nuls prenguin valors menys propers al zero [11].

De nou, per obtenir els paràmetres aplicam inferència Bayesiana i suposam independència dels paràmetres. D'aquesta manera, pel Teorema de Bayes (2.6)

$$p(\bar{\beta}|X,Y) \propto \mathcal{L}(\bar{\beta}|X,Y) \prod_{j=1}^{m} \frac{1}{2\psi} e^{-\frac{|\beta_j|}{\psi}}.$$

Prenent el logaritme tenim que

$$\ln p(\bar{\beta}|X,Y) = \ln \mathscr{L}(\bar{\beta}|X,Y) - m\ln 2\psi - \frac{1}{2\psi}||\bar{\beta}||_1 + c.$$

Així, el resultat de maximitzar $\ln p(\bar{\beta}|X,Y)$ per tal de d'estimar els paràmetres desconeguts β_j, sense tenir en compte els termes constants, és LASSO:

$$\text{maximitzar } \ln \mathscr{L}(\bar{\beta}|X,Y) - \eta||\bar{\beta}||_1,\ \eta > 0.$$

Segons la funció de versemblança $\mathscr{L}$, es defineix LASSO per un problema de regressió concret. En el nostre cas, ens centrarem amb LASSO per la regressió logística i a continuació es presenta el problema per obtenir els coeficients β que la determinen. Fixem-nos que tenim la mateixa funció objectiu (2.8), llevat del terme de regularització.

$$\text{maximitzar} \quad \sum_{i=1}^{n} \ln\left(\frac{1}{1+e^{Y_i(-\beta_0-\beta^T X_i)}}\right) - \eta||\bar{\beta}||_1,\ \eta > 0. \tag{2.9}$$

En afegir el terme de regularització amb la norma-1 es perd la diferenciabilitat de la funció objectiu i la seguretat d'unicitat de solucions [13]. Així, la funció objectiu esdevé còncava enlloc d'estricament còncava, però existeixen mètodes eficients per resoldre el problema.

En aquest capítol s'han presentat tres models: regressió logística lineal pura, ridge regression i, per acabar, LASSO. D'aquests, tenim un model sense penalització, un model amb penalització en norma-2 i un model amb penalització en norma-1. En el següent capítol s'explicarà un model amb penalització en que els coeficients entorn al zero s'estrenyen encara més cap al zero.

CAPÍTOL 3

HYPER LASSO

Al capítol anterior, hem imposat que els paràmetres β_j segueixin a priori una distribució de Laplace, ja que presenta característiques més útils que la normal a l'hora resoldre el problema de selecció de variables amb moltes més variables que dades. Així, la pregunta natural que sorgeix és, i si utilitzem una altra distribució? Podem millorar en alguns casos el model? La resposta és afirmativa i porta el nom d'HLASSO (Hyper LASSO).

3.1. Mescla de densitats

HLASSO és un model de regressió lineal penalitzada, on la nova penalització s'obté a partir d'una funció de densitat que es genera mitjançant la mescla de densitats. Així, per tal de conèixer la distribució, cal definir aquest concepte.

Definició 3.1.1. *La mescla o composició de densitats és la combinació convexa de funcions de densitat [14], és a dir, direm que una funció de densitat $p(x)$ és mescla de les densitats $p_1(x),...,p_k(x)$ si es pot escriure com*

$$p(x) = \sum_{i=1}^{k} \omega_i p_i(x),$$

on els pesos $\omega_1,...,\omega_k$ són estrictament positius i sumen 1 [14, 15].

En les aplicacions es sol emprar quan una població està dividida en k subgrups, i una variable aleatòria segueix una distribució diferent per a cada grup. Per exemple, es sap que l'alçada dels adults segueix una distribució normal, però la mitjana i variància pels homes difereix considerablement de les dones [16]. Així, la funció de densitat de l'altura s'explica millor com a composició de dues distribucions normals amb diferents

paràmetres.

Altres aplicacions, que són les que interessen en aquest treball, tracten la composició de densitats contínua. Quan es tenen un nombre infinit de densitats que composen la densitat d'una variable aleatòria, la suma esdevé una integral i la mescla resulta [15]

$$p(x) = \int_A p(x|a)\omega(a)da,$$

on $p(x|a)$ és una funció de densitat que depèn del paràmetre desconegut $a \in A$, que alhora segueix una distribució amb funció de densitat $\omega(a)$.

Aquesta composició resulta molt útil per generar funcions de densitat. Per exemple, donada una variable aleatòria que segueix una distribució normal amb variància desconeguda, si es sap que la variància segueix una distribució gamma inversa, aleshores la mescla de densitats origina una nova funció de densitat. Aquesta correspon amb la densitat d'una t de Student. Comparada amb la normal, la funció obtinguda està centrada en el mateix punt, però presenta coes més amples [17, 18].

En el nostre cas, la funció de densitat de Laplace, emprada en el model de LASSO, es pot generar com a mescla de densitats d'una normal, tal com segueix

$$p_L(\beta_j|0,\psi) = \int_0^\infty p_N(\beta_j|0,\sigma^2)p_{Ga}(\sigma^2|1,\frac{\psi^2}{2})d\sigma^2, \tag{3.1}$$

on p_N representa la funció de densitat d'una $\mathcal{N}(0,\sigma^2)$ en β_j i p_{Ga} d'una distribució Gamma$\left(1,\frac{\psi^2}{2}\right)$ en σ^2 [3, 19]. La funció de densitat d'una Gamma(a,b) es defineix a continuació [20]

$$p_{Ga}(x|a,b) = \begin{cases} \dfrac{b^a}{\Gamma(a)}x^{a-1}e^{-bx} & \text{si } x \geq 0, \\ 0 & \text{si } x < 0. \end{cases}$$

Aquesta expressió de la densitat de Laplace permet definir la nova densitat pel model d'HLASSO, tal com es calcula a la següent secció.

3.2. Model HLASSO per la regressió logística

Entendre LASSO com un model bastat en inferència Bayesiana permet millorar el terme de regularització de (2.8), prenent una distribució a priori millor que la Laplace, és a dir, amb una probabilitat major d'obtenir coeficients pròxims a zero. Així, el nou model pretén evitar millor l'*overfitting* quan hi ha més variables que dades del problema i relacionar menys covariables amb Y_i.

La funció de distribució a priori dels β_j proposada per Hoggart *et al.* és la distribució normal-exponencial-gamma. Aquesta, sorgeix de suposar que cada $\beta_j \sim$ Laplace$\left(0,\sqrt{2\psi}\right)$ i $\psi \sim$ Gamma$\left(\lambda,\gamma^2\right)$ [19]. Així, l'objectiu és generar una densitat amb un cim més alt

i coes més amples que la densitat de Laplace. D'aquesta manera, NEG resulta d'una doble mescla de densitats [21, 3]

$$\begin{aligned} p(\beta_j|\mu,\lambda,\gamma) &= \int_0^\infty p_L(\beta_j|\mu,\sqrt{2\psi})\,p_{Ga}\left(\psi,\gamma^2\right)d\psi \\ &\overset{(3.1)}{=} \int_0^\infty\int_0^\infty p_N(\beta_j|0,\sigma^2)\,p_{Ga}(\sigma^2|1,\psi)\,p_{Ga}(\psi|\lambda,\gamma^2)\,d\sigma^2\,d\psi \quad , \\ &= e^{\frac{(\beta_j-\mu)^2}{4\gamma^2}}\cdot D_{-(2\lambda+1)}\left(\frac{|\beta_j-\mu|}{\gamma}\right)\cdot\frac{2^\lambda\lambda}{\gamma\sqrt{\pi}}\Gamma\left(\lambda+\frac{1}{2}\right) \end{aligned} \tag{3.2}$$

on μ és la mitjana, λ el paràmetre de forma, γ el d'escala i $D_v(z)$ és una funció cilíndrica parabòlica. Direm que $\beta_j \sim NEG(\mu,\lambda,\gamma)$ quan segueixi aquesta distribució. En el nostre cas, donat que volem molts coeficients nuls, fixa'm $\mu = 0$ i definim la funció $NEG(\beta_j|\lambda,\gamma) := p(\beta_j|0,\lambda,\gamma)$.

Les funcions cilíndriques parabòliques són solucions de la següent equació diferencial [22]

$$\frac{d^2y}{dz^2} - \left(\frac{1}{4}z^4 + a\right)y = 0, \quad a\in\mathbb{R}. \tag{3.3}$$

Una solució de (3.3), que és la que empra NEG, es denota per $D_v(z) = U\left(-v-\frac{1}{2},z\right)$. L'expressió d'aquesta solució ve donada per

$$U(a,z) = \sqrt{\pi}2^{\frac{1}{2}a}\left(\frac{2^{-\frac{1}{4}}e^{\frac{1}{4}}z^2\;{}_1F_1\left(\frac{1}{2}z^2\middle|\frac{1}{2}a+\frac{1}{4},\frac{1}{2}\right)}{\Gamma\left(\frac{3}{4}+\frac{1}{2}a\right)} - \frac{2^{\frac{1}{4}}ze^{\frac{1}{4}z^2}\;{}_1F_1\left(-\frac{1}{2}z^2\middle|-\frac{1}{2}a+\frac{3}{4},\frac{3}{2}\right)}{\Gamma\left(\frac{1}{4}+\frac{1}{2}a\right)}\right),$$

on ${}_1F_1$ denota la funció hipergeomètrica confluent:

$${}_1F_1(z|a,c) = \sum_{k=0}^{\infty}\frac{\Gamma(a+k)}{\Gamma(c+k)}\frac{\Gamma(c)}{\Gamma(a)}\frac{z^k}{k!}\ [22].$$

El model d'HLASSO requereix calcular el valor de la funció cilíndrica parabòlica D_v. Per fer-ho, Hoggart *et al.* empren un algoritme en Fortran que es pot descarregar a `http://www.ebi.ac.uk/projects/BARGEN` i realitza el càlcul ràpidament. Aquest, es basa en l'algoritme descrit en [23].

Al capítol anterior, s'ha esmentat que el model amb la distribució de Laplace, enlloc de la normal, és més útil quan es treballa amb un gran nombre de variables en comparació al nombre de dades. Quant a NEG, considera encara més coeficients entorn al zero que la Laplace, i per tant, permet millorar el model en aquests casos. Amb l'objectiu de conèixer el comportament de la nova funció i comparar-ho amb les funcions de densitat que generen els models esmentats, s'ha obtingut la Figura 3.1. Per fer-ho, s'han considerat els paràmetres corresponents a cada distribució de manera que la variància

de β_j sigui la mateixa, concretament, s'ha pres $\text{Var}(\beta_j) = 2$. A la Taula 3.1 es recullen les fórmules pel càlcul de la variància segons cada distribució.

	Variance
$\mathcal{N}(\mu, \sigma^2)$	σ^2
Laplace(μ, ψ)	$2\psi^2$
$NEG(\mu, \lambda, \gamma)$	$\frac{\gamma^2}{\lambda - 1}$, $\lambda > 1$.

Taula 3.1: Variàncies d'una variable aleatòria que segueix una distribució normal, Laplace o normal-exponencial-gamma, en funció dels seus paràmetres. Les fórmules han estat extretes de [12, 21].

Fixem-nos en la Figura 3.1 (*a*), que el cim de la funció de densitat de NEG és el més elevat, fent que la probabilitat de que β_j prengui un valor entorn al zero sigui molt més alta que amb les altres distribucions. Així, es pretén aconseguir un major nombre de coeficients exactament zero. D'altra banda, en (*b*) s'observa que les *coes* són més amples que la densitat de Laplace i d'una normal. Com expliquen Hoggart *et al.*, les coes més amples provoquen que els valors no nuls no estiguin tan aprop del zero. Aquest fet, tal com s'ha comentat amb anterioritat, és important a l'hora de resoldre el problema de selecció de variables, ja que ajuda a identificar les covariables explicatives de Y_i. Així, l'efecte sobre les coes i el cim més elevat de la funció de densitat, fan que molts dels paràmetres estimats siguin nuls i que els no nuls es distingeixin significativament del zero.

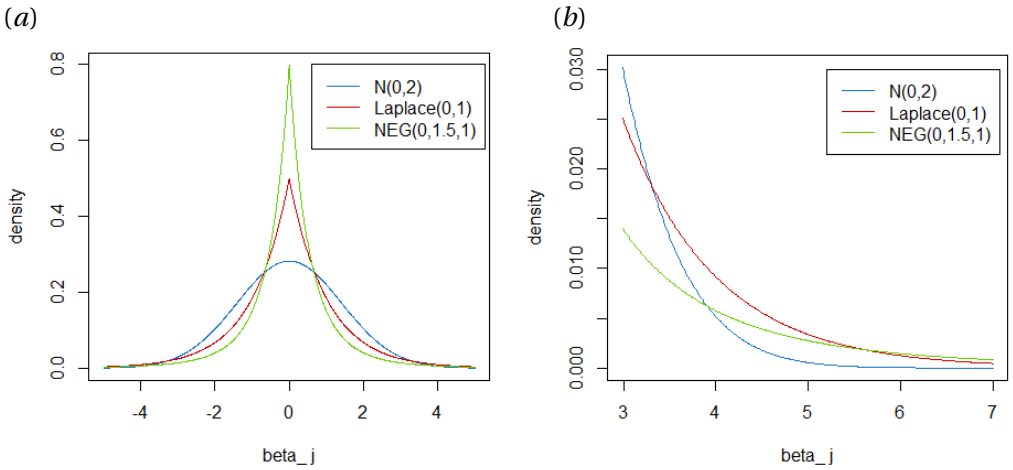

Figura 3.1: Densitats de les distribucions de β_j, segons si segueix una normal, Laplace o normal-exponencial-gamma, totes amb variància 2 i mitjana 0. En (*a*) podem observar com augmenta el cim de cada distribució en el zero, i per tant, la corba s'estreny. En (*b*) tenim les respectives coes; per $\beta_j > 6$ s'aprecia que la més ample és la de NEG. Les gràfiques s'han obtingut amb R.

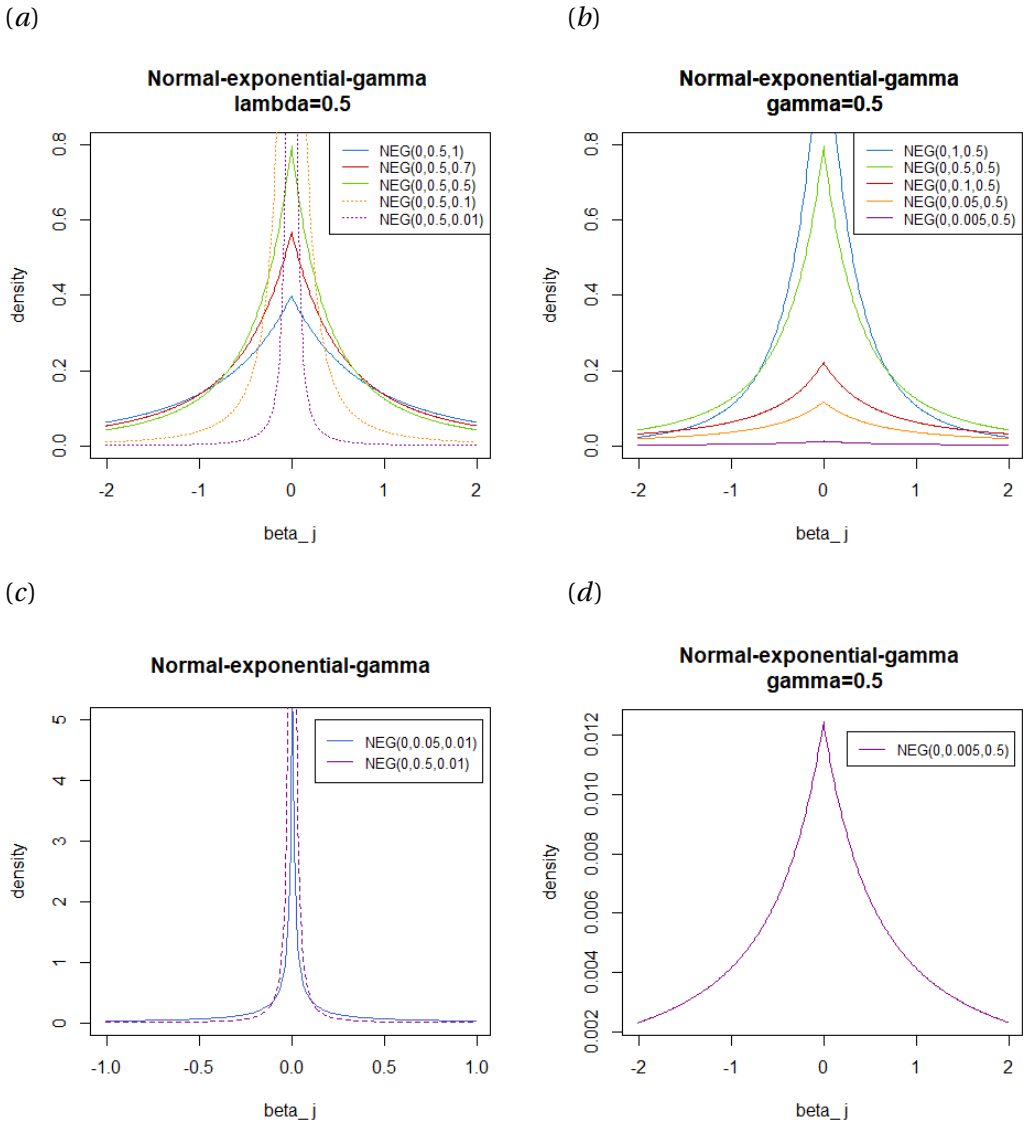

Figura 3.2: Gràfiques de la densitat de NEG per a diferents paràmetres de forma (λ) i escala (γ). En (a) totes les corbes tenen el mateix λ. En (b) comparteixen el mateix valor de γ; per observar-ne millor la densitat de $NEG(0, 0.005, 0.5)$ s'ha disminuït el rang a (d). A més, en (c) es comparen dues gràfiques estretes entorn al $\beta_j = 0$; una d'elles, la línia discontinua, ja dibuixada en (a). Les gràfiques s'han obtingut amb R.

El paràmetre de forma, λ, controla la forma de les coes de NEG, mentre que γ controla l'escala [19]. Així, ambdós defineixen la forma característica de NEG. Per tal de tenir una idea intuïtiva de com afecten els paràmetres, s'ha representat la densitat per a diferents valors de λ i γ. A la Figura 3.2 tenim el resultat. Totes les corbes en (*a*) tenen el mateix valor de $\lambda = 0.5$, i com més petit és γ més s'estreny la corba en $\beta_j = 0$, ja que el cim esdevé més elevat. D'altra banda, en (*b*) s'ha fitxat $\gamma = 0.05$ i es modifica l'amplada de les coes per diferents valors de λ. El resultat són corbes comprimides cap a l'eix d'abscisses per valors petits. Així i tot, tal com es veu en (*d*), NEG no es contreu cap

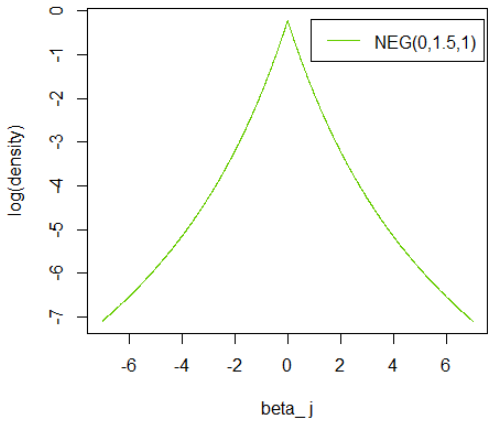

Figura 3.3: Logaritme de la densitat d'una normal-exponencial-gamma en β_j, amb paràmetres $\mu = 0$, $\lambda = 1.5$ i $\gamma = 1$. Correspon al logaritme de la densitat de NEG representada a la Figura 3.1. La gràfica s'ha obtingut amb R.

a $\beta_j = 0$, per tant, la probabilitat de que β_j estigui en un entorn reduït del zero no és suficientment elevada. És a dir, un λ petit no basta si volem potenciar les propietats de NEG en el problema de selecció de variables. Per exemple, ambdues gràfiques en *(c)*, per $\gamma = 0.01$, tenen un cim elevat que contreu la densitat cap al zero. Notem que per $\lambda = 0.05$, enlloc de 0.5, les coes són més amples.

Ara, vegem com estimam els coeficients emprant NEG. Aplicant el Teorema de Bayes i suposant independència dels paràmetres tenim que,

$$p(\bar{\beta}|X,Y) \propto \mathscr{L}(\bar{\beta}|X,Y) \prod_{j=1}^{m} NEG(\beta_j|\lambda,\gamma).$$

Així, pel mètode de màxima versemblança, per determinar els coeficients basta resoldre el següent problema d'optimització

$$\text{maximitzar } \ln \mathscr{L}(\bar{\beta}|X,Y) + \sum_{j=1}^{m} \ln NEG(\beta_j|\lambda,\gamma),$$

on $\sum_{j=1}^{m} \ln NEG(\beta_j|\lambda,\gamma)$ és el nou terme de regularització.

Per simplificar la notació, considerem $L(\bar{\beta}) = \ln \mathscr{L}(\bar{\beta}|X,Y)$ i $f(\beta) = -\sum_{j=1}^{m} \ln NEG(\beta_j|\lambda,\gamma)$, amb el signe menys per marcar que és la penalització [3] . D'aquesta manera, el problema resulta

$$\text{maximitzar } L(\bar{\beta}) - f(\beta). \tag{3.4}$$

El logaritme de la funció de versemblança és el mateix que pels models de LASSO o ridge regression, per tant, és còncau. Quant al logaritme de NEG, és no-còncau (i no-convex). Podem observar aquest fet a la Figura 3.3, per un cas particular. Fixem-nos, que hi ha punts pels quals la recta que els uneix es troba completament sota el logaritme,

i punts pels quals es troba sobre. D'aquesta manera, la suma dels logaritmes de NEG tampoc és còncava ni convexa. De fet, la funció objectiu resulta multi-modal [3]. Un altre problema, és que la funció no és diferenciable en el zero, com calia esperar d'una fórmula que empra el valor absolut.

3.3. Algoritme HLASSO

Per determinar el màxim de la densitat posterior $p(\bar{\beta}|X,Y)$, és a dir, resoldre el problema d'optimització (3.4), Hoggart *et al.* proposen emprar l'algoritme CCD (*cyclic coordinate descent*) amb múltiples execucions, per tal d'assolir varis cims de la funció objectiu [3].

L'algoritme CCD és un algoritme per resoldre problemes d'optimització en diverses variables. El mètode consisteix en resoldre el problema respecte cada variable, és a dir, es fixa una variable i maximitza (o minimitza) la funció objectiu suposant les altres variables constants [10]. Per fer-ho, siguin $\alpha_1,...,\alpha_m$ les variables, es prenen els valors inicials α_i^0 per $i=2,...,m$ i s'obté el valor α_1^{new} que determina el màxim. El segon pas, és maximitzar respecte de α_2 amb els termes constants inicials α_1^{new} i α_i^0, $i=3,...,m$. En general, al pas j resolem el problema respecte α_j amb les constants α_k^{new}, per $k=1,...,j-1$ i α_i^0, $i=j+1,...,m$. Així, un cop determinat el valor de totes les variables, repetim el procés fins satisfer el criteri de convergència.

En el nostre cas, s'empra el criteri descrit a l'article [10]. Segons aquest, l'algoritme acaba quan

$$\frac{\sum_{i=1}^{n}|\Delta\eta_i|}{1+\sum_{i=1}^{n}|\eta_i|}\leq 0.005, \tag{3.5}$$

on $\eta_i = Y_i(\beta_0+\beta^T X_i)$. La idea és calcular els canvis del valor de la funció lineal en cada iteració i aturar quan aquest valor sigui prou petit. Així, tal com s'explica a l'article, una altra opció seria acabar quan el numerador fos petit, encara que HLASSO empra la fracció (3.5) proposada.

Un cop definit l'algoritme CCD, s'ha d'escollir un mètode, dels diversos que es poden emprar, per resoldre el problema d'optimització en una variable. En en nostre cas, HLASSO aplica el mètode de Newton en una variable per trobar un zero de $\frac{\partial}{\partial\beta_j}\ln p(\bar{\beta}|X,Y)$ [3] . Així, l'actualització de pesos a cada pas ve donada per

$$\beta_j^{new} \longleftarrow \beta_j - \frac{\frac{\partial}{\partial\beta_j}\ln p(\bar{\beta}|X,Y)}{\frac{\partial^2}{\partial\beta_j^2}\ln p(\bar{\beta}|X,Y)}, \quad j=0,...,m. \tag{3.6}$$

El mètode de Newton requereix que el denominador no s'anul·li. Aquest fet es pot donar, ja que la derivada parcial de segon odre de p es pot escriure com a suma de les derivades segones de L i $-f$, i com es veurà més endavant tenen signes oposats. Per tal d'evitar el conflicte, si el denominador és nul no s'actualitza el valor del coeficient, i per tant, $\beta_j^{new} \longleftarrow \beta_j$.
D'altra banda, la derivada de NEG no està ben definida per $\beta_j = 0$, per tant, tampoc podem aplicar el mètode de Newton en aquest cas. Així, la idea per actualitzar els valors dels coeficients, exposada a l'article [3], és calcular el límits laterals. Fent el límit per la dreta

$$\beta_j^{new} \longleftarrow 0 - \left. \frac{\frac{\partial}{\partial \beta_j} \ln p(\bar{\beta}|X,Y)}{\frac{\partial^2}{\partial \beta_j^2} \ln p(\bar{\beta}|X,Y)} \right|_{\beta_j = 0^+},$$

es mira el signe de β_j^{new}. Si és positiu, s'empra aquesta actualització. En cas contrari, és fa el límit per l'esquerra i si β_j^{new} resulta negatiu, es pren aquest nou valor. Altrament, es manté a zero. Més endavant, quedarà definit el càlcul d'aquests límits.

Donat que la densitat $p(\bar{\beta}|X,Y)$ es multi-modal, l'algoritme pot acabar en un màxim local. D'aquesta manera, Hoggart *et al.* executen l'algoritme varies vegades, amb els mateixos valors inicials y una permutació aleatòria en l'ordre de les dades (X_i, Y_i). Així, l'objectiu és trobar varis cims de la funció i prendre els paràmetres corresponents al major valor de p. De fet, l'algoritme pren sempre els mateixos valors inicials, independentment de les dades del problema [3]. Tots els coeficients s'inicialitzen a zero menys β_0, que pren el valor $\ln\left(\frac{n_1}{n_0}\right)$, on n_1 i n_0 representen el nombre de casos $(Y_i = -1)$ i controls $(Y_i = 1)$, respectivament. Aquest valor correspon al màxim de la funció de versemblança si tots els altres coeficients són nuls, tal com exposa el següent resultat.

Resultat 3.3.1. *Prenent $\beta = \mathbf{0}$ constant, el màxim de la funció definida a (2.3) s'assoleix en $\beta_0 = \ln\left(\frac{n_1}{n_0}\right)$.*

Demostració. Considerant $\beta = \mathbf{0}$, la funció de versemblança es redueix a

$$\sum_{i=1}^{n} \ln\left(\frac{1}{1+e^{-Y_i\beta_0}}\right).$$

Anem a resoldre el problema en una variable, β_0, cercant un zero de la derivada,

$$\frac{\partial L(\beta_0|\beta = \mathbf{0})}{\partial \beta_0} = \sum_{i=1}^{n} \frac{-Y_i}{1+e^{Y_i\beta_0}} = 0.$$

En la següent subsecció s'explica en detall l'expressió de la parcial.

Descomponent la suma segons el valor que pren Y_i i resolent,

$$\sum_{i:Y_i=1}^{n_1} \frac{-1}{1+e^{\beta_0}} + \sum_{i:Y_i=-1}^{n_0} \frac{1}{1+e^{-\beta_0}} = 0,$$

$$\frac{-n_1 + n_0 e^{\beta_0}}{1+e^{\beta_0}} = 0,$$

$$\beta_0 = \ln\left(\frac{n_1}{n_0}\right).$$

Donada la convexitat de L, tenim el màxim en β_0. □

3.3.1. Càlcul de les derivades

Per tal d'implementar HLASSO, s'han de calcular les derivades parcials del logaritme de la densitat posterior, de primer i segon ordre

$$\frac{\partial \ln p(\bar{\beta}|X,Y)}{\partial \beta_0} = \frac{\partial L(\bar{\beta})}{\partial \beta_0}, \qquad \frac{\partial^2 \ln p(\bar{\beta}|X,Y)}{\partial \beta_0^2} = \frac{\partial^2 L(\bar{\beta})}{\partial \beta_0^2},$$

$$\frac{\partial \ln p(\bar{\beta}|X,Y)}{\partial \beta_j} = \frac{\partial L(\bar{\beta})}{\partial \beta_j} - \frac{\partial f(\beta)}{\partial \beta_j}, \qquad \frac{\partial^2 \ln p(\bar{\beta}|X,Y)}{\partial \beta_j^2} = \frac{\partial^2 L(\bar{\beta})}{\partial \beta_j^2} - \frac{\partial^2 f(\beta)}{\partial \beta_j^2}, \quad j = 1,\ldots,m.$$

Cal tenir en compte, que el càlcul de les parcials de $f(\beta)$ es reduirà a calcular la derivada de $NEG(\beta_j)$, tal com es veu a continuació.

$$\begin{aligned}\frac{\partial f(\beta)}{\partial \beta_j} &= \frac{\partial}{\partial \beta_j}\left(-\sum_{i=1}^{n} \ln NEG(\beta_i|\lambda,\gamma)\right) = -\frac{d}{d\beta_j}\left(\ln NEG(\beta_j|\lambda,\gamma)\right) \\ &= -\frac{1}{NEG(\beta_j|\lambda,\gamma)}\frac{d}{d\beta_j} NEG(\beta_j|\lambda,\gamma).\end{aligned}$$

Notem que la funció NEG no és diferenciable en $\beta_j = 0$ (veure Figura 3.1), per tant, no podrem emprar el mètode de Newton (3.6) si s'inicialitza algun coeficient β_j a zero o $\beta_j^{new} = 0$. A la subsecció 3.3.2 es defineix l'algoritme en aquest cas.

Per fer el càlcul, es considera $\kappa = \frac{2^\lambda \lambda}{\gamma\sqrt{\pi}}\Gamma\left(\lambda + \frac{1}{2}\right)$ i aleshores

$$\frac{d}{d\beta_j} NEG(\beta_j|\lambda,\gamma) = \kappa \frac{d}{d\beta_j}\left(e^{\frac{\beta_j^2}{4\gamma^2}} \cdot D_{-(2\lambda+1)}\left(\frac{|\beta_j|}{\gamma}\right)\right). \tag{3.7}$$

Amb la nova expressió de la derivada, utilitzarem les següents integrals conegudes per obtenir-ne el resultat [3, supplem.],

$$\int_0^\infty x^{\nu-1}(x+\alpha)^{-\nu+\frac{1}{2}} e^{-\mu x} dx = 2^{\nu-\frac{1}{2}}\Gamma(\nu)\mu^{-\frac{1}{2}} e^{\frac{\alpha\mu}{2}} D_{1-2\nu}\left(\sqrt{2\alpha\mu}\right), \tag{3.8}$$

$$\int_0^\infty x^{\nu-1}(x+\alpha)^{-\nu-\frac{1}{2}} e^{-\mu x} dx = 2^{\nu}\Gamma(\nu)\alpha^{-\frac{1}{2}} e^{\frac{\alpha\mu}{2}} D_{-2\nu}\left(\sqrt{2\alpha\mu}\right). \tag{3.9}$$

Ara, prenent $\upsilon = \lambda + \frac{1}{2}$, $\mu = \frac{\beta_j^2}{2}$, $\alpha = \frac{1}{\gamma^2}$ en (3.9) i aïllant l'exponencial pel cilindre parabòlic, s'aconsegueix una expressió igual a la que es vol derivar.

$$\begin{aligned}\frac{1}{2^{\lambda+\frac{1}{2}}\Gamma\left(\lambda+\frac{1}{2}\right)\gamma}\int_0^\infty x^{\lambda-\frac{1}{2}}\left(x+\frac{1}{\gamma^2}\right)^{-(\lambda+1)}e^{-\frac{\beta_j^2}{2}x}dx &= e^{\frac{\beta_j^2}{4\gamma^2}}D_{-2\left(\lambda+\frac{1}{2}\right)}\left(\sqrt{\frac{\beta^2}{\gamma^2}}\right)\\ &= e^{\frac{\beta_j^2}{4\gamma^2}}D_{-(2\lambda+1)}\left(\frac{|\beta_j|}{\gamma}\right).\end{aligned}$$

Per tant, multiplicant per κ i derivant respecte de β_j cada membre de la igualtat

$$\kappa\frac{d}{d\beta_j}\left(\frac{1}{2^{\lambda+\frac{1}{2}}\Gamma\left(\lambda+\frac{1}{2}\right)\gamma}\int_0^\infty x^{\lambda-\frac{1}{2}}\left(x+\frac{1}{\gamma^2}\right)^{-(\lambda+1)}e^{-\frac{\beta_j^2}{2}x}dx\right)=\kappa\frac{d}{d\beta_j}\left(e^{\frac{\beta_j^2}{4\gamma^2}}D_{-(2\lambda+1)}\left(\frac{|\beta_j|}{\gamma}\right)\right).$$

D'aquesta manera, derivant baix el signe integral s'arriba a la següent expressió

$$\begin{aligned}\frac{d}{d\beta_j}NEG(\beta_j|\lambda,\gamma) &= \frac{\kappa}{2^{\lambda+\frac{1}{2}}\Gamma\left(\lambda+\frac{1}{2}\right)\gamma}\int_0^\infty\frac{\partial}{\partial\beta_j}\left(x^{\lambda-\frac{1}{2}}\left(x+\frac{1}{\gamma^2}\right)^{-(\lambda+1)}e^{\frac{-1}{2}\beta_j^2x}dx\right)\\ &= \frac{\kappa}{2^{\lambda+\frac{1}{2}}\Gamma\left(\lambda+\frac{1}{2}\right)\gamma}\int_0^\infty x^{\lambda-\frac{1}{2}}\left(x+\frac{1}{\gamma^2}\right)^{-(\lambda+1)}e^{-\frac{1}{2}\beta_j^2x}(-\beta_jx)dx\\ &= -\frac{\beta_j\kappa}{2^{\lambda+\frac{1}{2}}\Gamma\left(\lambda+\frac{1}{2}\right)\gamma}\int_0^\infty x^{\lambda+\frac{1}{2}}\left(x+\frac{1}{\gamma^2}\right)^{-(\lambda+1)}e^{-\frac{1}{2}\beta_j^2x}dx.\end{aligned} \tag{3.10}$$

D'altra banda, si substituïm $\upsilon = \lambda + \frac{3}{2}, \mu = \frac{\beta_j^2}{2}, \alpha = \frac{1}{\gamma^2}$ en (3.9) s'obté

$$\begin{aligned}\int_0^\infty x^{\lambda+\frac{1}{2}}\left(x+\frac{1}{\gamma^2}\right)^{-(\lambda+1)}e^{-\frac{\beta_j^2}{2}x}dx &= 2^{\lambda+1}\Gamma\left(\lambda+\frac{3}{2}\right)\left(\frac{\beta_j^2}{2}\right)^{-\frac{1}{2}}e^{\frac{\beta_j^2}{4\gamma^2}}D_{1-2\left(\lambda+\frac{3}{2}\right)}\left(\sqrt{\frac{\beta_j^2}{\gamma^2}}\right)\\ &= 2^{\lambda+1}\Gamma\left(\lambda+\frac{3}{2}\right)\frac{\sqrt{2}}{|\beta_j|}e^{\frac{\beta_j^2}{4\gamma^2}}D_{-(2\lambda+2)}\left(\frac{|\beta_j|}{\gamma}\right).\end{aligned}$$

Així, substituint el resultat de la integral en (3.10), l'expressió queda determinada de la següent manera

$$\frac{d}{d\beta_j}NEG(\beta_j|\lambda,\gamma) = -\frac{\beta_j\kappa}{2^{\lambda+\frac{1}{2}}\Gamma\left(\lambda+\frac{1}{2}\right)\gamma}2^{\lambda+1}\Gamma\left(\lambda+\frac{3}{2}\right)\frac{\sqrt{2}}{|\beta_j|}e^{\frac{\beta_j^2}{4\gamma^2}}D_{-(2\lambda+2)}\left(\frac{|\beta_j|}{\gamma}\right)$$

$$
\begin{aligned}
&= -\frac{2\kappa \text{sign}(\beta_j)}{\Gamma\left(\lambda+\frac{1}{2}\right)\gamma}\Gamma\left(\lambda+\frac{3}{2}\right)e^{\frac{\beta_j^2}{4\gamma^2}}D_{-(2\lambda+2)}\left(\frac{|\beta_j|}{\gamma}\right) \\
&= -\frac{2\kappa \text{sign}(\beta_j)\left(\lambda+\frac{1}{2}\right)}{\gamma}e^{\frac{\beta_j^2}{4\gamma^2}}D_{-(2\lambda+2)}\left(\frac{|\beta_j|}{\gamma}\right),
\end{aligned}
\tag{3.11}
$$

on la darrera igualtat es satisfà per propietats de la funció Γ [24] :

$$
\Gamma\left(\lambda+\frac{3}{2}\right)=\Gamma\left(\left(\lambda+\frac{1}{2}\right)+1\right)=\left(\lambda+\frac{1}{2}\right)\Gamma\left(\lambda+\frac{1}{2}\right).
$$

Per tant, la derivada parcial de f ve donada per

$$
\begin{aligned}
\frac{\partial f(\bar{\beta}_j)}{\partial \beta_j} &= -\frac{1}{\kappa e^{\frac{\beta_j^2}{4\gamma^2}}D_{-(2\lambda+1)}\left(\frac{|\beta_j|}{\gamma}\right)}\cdot\frac{-2\kappa \text{sign}(\beta_j)\left(\lambda+\frac{1}{2}\right)}{\gamma}e^{\frac{\beta_j^2}{4\gamma^2}}D_{-(2\lambda+2)}\left(\frac{|\beta_j|}{\gamma}\right) \\
&= \frac{2\text{sign}(\beta_j)\left(\lambda+\frac{1}{2}\right)D_{-(2\lambda+2)}\left(\frac{|\beta_j|}{\gamma}\right)}{\gamma D_{-(2\lambda+1)}\left(\frac{|\beta_j|}{\gamma}\right)}.
\end{aligned}
\tag{3.12}
$$

Pel càlcul de la derivada parcial de segon ordre s'obté

$$
\frac{\partial^2 f(\beta)}{\partial \beta_j^2}=\frac{NEG(\beta_j|\lambda,\gamma)\frac{d^2}{d\beta_j^2}NEG(\beta_j|\lambda,\gamma)-\left(\frac{d}{d\beta_j}NEG(\beta_j|\lambda,\gamma)\right)^2}{NEG^2(\beta_j|\lambda,\gamma)}.
$$

Per tant, cal calcular

$$
\frac{\partial^2}{\partial \beta_j^2}NEG(\beta_j|\lambda,\gamma)=-\frac{2\kappa \text{sign}(\beta_j)\left(\lambda+\frac{1}{2}\right)}{\gamma}\frac{\partial}{\partial \beta_j}e^{\frac{\beta_j^2}{4\gamma^2}}D_{-(2\lambda+2)}\left(\frac{|\beta_j|}{\gamma}\right).
$$

Prenent $\upsilon=\lambda+1$, $\mu=\dfrac{\beta_j^2}{2}$ i $\alpha=\dfrac{1}{\gamma^2}$ en (3.9), tal com es mostra a continuació,

$$
\int_0^\infty x^\lambda\left(x+\frac{1}{\gamma^2}\right)^{-\lambda-\frac{3}{2}}e^{-\frac{\beta_j^2}{2}x}dx = 2^{\lambda+1}\Gamma(\lambda+1)\gamma e^{\frac{\beta_j^2}{4\gamma^2}}D_{-(2\lambda+2)}\left(\frac{|\beta_j|}{\gamma}\right),
$$

s'obté

$$
\begin{aligned}
\frac{d^2}{d\beta_j^2}NEG(\beta_j|\lambda,\gamma) &= -\frac{\kappa \text{sign}(\beta_j)\left(\lambda+\frac{1}{2}\right)}{2^\lambda\Gamma(\lambda+1)\gamma^2}\frac{\partial}{\partial \beta_j}\int_0^\infty x^\lambda\left(x+\frac{1}{\gamma^2}\right)^{-\lambda-\frac{3}{2}}e^{-\frac{\beta_j^2}{2}x}dx \\
&= \frac{\kappa|\beta_j|\left(\lambda+\frac{1}{2}\right)}{2^\lambda\Gamma(\lambda+1)\gamma^2}\int_0^\infty x^{\lambda+1}\left(x+\frac{1}{\gamma^2}\right)^{-\lambda-\frac{3}{2}}e^{-\frac{\beta_j^2}{2}x}dx.
\end{aligned}
$$

D'altra banda, prenent $\upsilon = \lambda + 2$, $\mu = \dfrac{\beta_j^2}{2}$ i $\alpha = \dfrac{1}{\gamma^2}$ en (3.8) s'obté la següent expressió de la integral

$$\int_0^\infty x^{\lambda+1}\left(x+\frac{1}{\gamma^2}\right)^{-\lambda-\frac{3}{2}} e^{-\frac{\beta_j^2}{2}x}dx = 2^{\lambda+\frac{3}{2}}\Gamma(\lambda+2)\frac{\sqrt{2}}{|\beta_j|}e^{\frac{\beta_j^2}{4\gamma^2}}D_{-(2\lambda+3)}\left(\frac{|\beta_j|}{\gamma}\right).$$

Aleshores,

$$\begin{aligned}\frac{d^2}{d\beta_j^2}NEG(\beta_j|\lambda,\gamma) &= \frac{4\kappa\left(\lambda+\frac{1}{2}\right)\Gamma(\lambda+2)}{\Gamma(\lambda+1)\gamma^2}e^{\frac{\beta_j^2}{4\gamma^2}}D_{-(2\lambda+3)}\left(\frac{|\beta_j|}{\gamma}\right)\\ &= \frac{4\kappa\left(\lambda+\frac{1}{2}\right)(\lambda+1)}{\gamma^2}e^{\frac{\beta_j^2}{4\gamma^2}}D_{-(2\lambda+3)}\left(\frac{|\beta_j|}{\gamma}\right),\end{aligned}$$

ja que $\Gamma(\lambda+2) = (\lambda+1)\Gamma(\lambda+1)$.

D'aquesta manera,

$$\begin{aligned}\frac{\partial^2 f(\beta)}{\partial\beta_j^2} &= \frac{\kappa\left(e^{\frac{\beta_j^2}{4\gamma^2}}D_{-(2\lambda+1)}\left(\frac{|\beta_j|}{\gamma}\right)\right)\cdot\frac{2\kappa\left(\lambda+\frac{1}{2}\right)(\lambda+1)}{\gamma^2}e^{\frac{\beta_j^2}{4\gamma^2}}D_{-(2\lambda+3)}\left(\frac{|\beta_j|}{\gamma}\right)-\left(-\frac{2\kappa\,\mathrm{sign}(\beta_j)\left(\lambda+\frac{1}{2}\right)}{\gamma}e^{\frac{\beta_j^2}{4\gamma^2}}D_{-(2\lambda+2)}\left(\frac{|\beta_j|}{\gamma}\right)\right)^2}{\kappa^2\left(e^{\frac{\beta_j^2}{4\gamma^2}}D_{-(2\lambda+1)}\left(\frac{|\beta_j|}{\gamma}\right)\right)^2}\\ &= \frac{4\left(\lambda+\frac{1}{2}\right)(\lambda+1)D_{-(2\lambda+1)}\left(\frac{|\beta_j|}{\gamma}\right)D_{(-(2\lambda+3)}\left(\frac{|\beta_j|}{\gamma}\right)-4\left(\lambda+\frac{1}{2}\right)^2\left(D_{(-(2\lambda+2)}\left(\frac{|\beta_j|}{\gamma}\right)\right)^2}{\left(D_{-(2\lambda+1)}\left(\frac{|\beta_j|}{\gamma}\right)\right)^2}\\ &= \frac{4}{\gamma^2}\left(\left(\lambda+\frac{1}{2}\right)(\lambda+1)\frac{D_{-(2\lambda+3)}\left(\frac{|\beta_j|}{\gamma}\right)}{D_{-(2\lambda+1)}\left(\frac{|\beta_j|}{\gamma}\right)}-\left(\lambda+\frac{1}{2}\right)^2\left(\frac{D_{-(2\lambda+2)}\left(\frac{|\beta_j|}{\gamma}\right)}{D_{-(2\lambda+1)}\left(\frac{|\beta_j|}{\gamma}\right)}\right)^2\right).\end{aligned}$$

Per calcular les parcials $\dfrac{\partial L(\bar\beta)}{\partial\beta_j}$ i $\dfrac{\partial^2 L(\bar\beta)}{\partial\beta_j^2}$, cal tenir en compte que estem aplicant HLASSO per la regressió logística. D'aquesta manera, es té que

$$L(\bar\beta) = \ln\mathscr{L}(\bar\beta|X,Y) = \sum_{i=1}^n \ln\left(\frac{1}{1+e^{-Y_i(\beta_0+\beta^T X_i)}}\right) = -\sum_{i=1}^n \ln\left(1+e^{-Y_i(\beta_0+\beta^T X_i)}\right).$$

Així, derivant respecte de β_j s'obté

$$\frac{\partial L(\bar\beta)}{\partial\beta_j} = \sum_{i=1}^n \frac{e^{-Y_i(\beta_0+\beta^T X_i)}(Y_i X_{ij})}{1+e^{-Y_i(\beta_0+\beta^T X_i)}} = \sum_{i=1}^n \frac{Y_i X_{ij}}{\left(1+e^{-Y_i(\beta_0+\beta^T X_i)}\right)e^{Y_i(\beta_0+\beta^T X_i)}}$$

$$= \sum_{i=1}^{n} \frac{Y_i X_{ij}}{1+e^{Y_i(\beta_0+\beta^T X_i)}}, \tag{3.13}$$

$$\frac{\partial L(\bar{\beta})}{\partial \beta_0} = \sum_{i=1}^{n} \frac{Y_i e^{-Y_i(\beta_0+\beta^T X_i)}}{1+e^{-Y_i(\beta_0+\beta^T X_i)}} = \sum_{i=1}^{n} \frac{Y_i}{1+e^{Y_i(\beta_0+\beta^T X_i)}}, \tag{3.14}$$

$$\frac{\partial^2 L(\bar{\beta})}{\partial \beta_j^2} = \sum_{i=1}^{n} \frac{-Y_i X_{ij} e^{Y_i(\beta_0+\beta^T X_i)} Y_i X_{ij}}{\left(1+e^{Y_i(\beta_0+\beta^T X_i)}\right)^2} = {}^{1} \sum_{i=1}^{n} -X_{ij}^2 \frac{e^{Y_i(\beta_0+\beta^T X_i)}}{\left(1+e^{Y_i(\beta_0+\beta^T X_i)}\right)^2}, \tag{3.15}$$

(3.16)

$$\frac{\partial^2 L(\bar{\beta})}{\partial \beta_0^2} = \sum_{i=1}^{n} \frac{-Y_i e^{Y_i(\beta_0+\beta^T X_i)} Y_i}{\left(1+e^{Y_i(\beta_0+\beta^T X_i)}\right)^2} = \sum_{i=1}^{n} -\frac{e^{Y_i(\beta_0+\beta^T X_i)}}{\left(1+e^{Y_i(\beta_0+\beta^T X_i)}\right)^2}, \tag{3.17}$$

on X_{ij} denota la component j-èssima del vector X_i.

3.3.2. Variacions de l'algoritme

El mètode de Newton té un desavantatge, i és que β_j^{new} pot oscil·lar entorn al màxim sense arribar a convergir. Aquest fet es pot donar si es produeixen grans salts, és a dir, si l'actualització de pesos modifica notablement el valor de β_j a cada pas, fent que aquest passi de positiu a negatiu. A més, la densitat posterior és multi-modal, per tant, també es poden generar problemes amb la convergència de l'algoritme si no es controlen els salts, encara que el signe del coeficient es mantengui. Per resoldre aquest problema, Hoggart *et al.* realitzen varies modificacions al mètode de Newton.

La primera modificació és imposar que, si $\beta_j \cdot \beta_j^{new} < 0$, aleshores $\beta_j^{new} \leftarrow 0$. Així, s'eviten els canvis de signe en l'actualització dels paràmetres. En cas que un β_j sigui zero es segueix el mateix criteri explicat amb anterioritat, per tant, s'actualitza el valor de β_j si el mètode de Newton en 0^+ és positiu, o en 0^- és negatiu. Per no haver de calcular els límits laterals, llevat que sigui necessari, s'empra el següent teorema que ens dóna una condició necessària i suficient per a què no es doni el canvi de signe quan $\beta_j = 0$ [25, 3].

Teorema 3.3.1. *Si $\beta_j = 0$, no es dóna un canvi de signe si, i només sí,*

$$\left|\frac{\partial L(\bar{\beta})}{\partial \beta_j}\right|_{\beta_j=0} > \left.\frac{\partial f(\beta)}{\partial \beta_j}\right|_{\beta_j=0^+}. \tag{3.18}$$

Demostració. Si $\beta_j = 0$, tenim que la parcial de segon ordre del logaritme de la densitat posterior és negativa [3]. Aquest fet es deriva de que la segona parcial de f en $\beta_j = 0$ és positiva. Vegem-ho.

[1] Donat que $Y_i \in \{-1, 1\}$, $Y_i^2 = 1$.

$$\left.\frac{\partial^2 f(\beta)}{\partial \beta_j^2}\right|_{\beta_j=0^{\pm}} = \frac{4\left(\lambda+\frac{1}{2}\right)(\lambda+1)D_{-(2\lambda+1)}(0)\,D_{(-(2\lambda+3)}(0) - 4\left(\lambda+\frac{1}{2}\right)^2\left(D_{-(2\lambda+2)}(0)\right)^2}{\left(D_{-(2\lambda+1)}(0)\right)^2}. \tag{3.19}$$

Les funcions cilíndriques parabòliques avaluades a l'origen prenen el següent valor, que depèn del parámetre $a = -\frac{1}{2} - \nu$ [26],

$$D_\nu(0) = \frac{\sqrt{\pi}}{2^{\frac{1}{2}a+\frac{1}{4}}\Gamma\left(\frac{3}{4}+\frac{1}{2}a\right)}. \tag{3.20}$$

Així, donat que $\lambda > 0$, perquè (3.19) sigui positiva basta que ho sigui la següent expressió,

$$\frac{(\lambda+1)\pi}{2^{2\lambda+2}\Gamma(\lambda+2)\Gamma(\lambda+1)} - \left(\lambda+\frac{1}{2}\right)\frac{\pi}{2^{2\lambda+2}\left(\Gamma\left(\lambda+\frac{3}{2}\right)\right)^2}.$$

Per la propietat de recurrència de la funció Γ, $\Gamma(\lambda+2) = (\lambda+1)\Gamma(\lambda+1)$. Per tant, la expressió ens queda

$$\frac{\pi}{2^{2\lambda+2}}\left(\frac{1}{(\Gamma(\lambda+1))^2} - \frac{\lambda+\frac{1}{2}}{\left(\Gamma\left(\lambda+\frac{3}{2}\right)\right)^2}\right) = \frac{\pi}{2^{2\lambda+2}}\left(\frac{\left(\Gamma\left(\lambda+\frac{3}{2}\right)\right)^2 - \left(\lambda+\frac{1}{2}\right)(\Gamma(\lambda+1))^2}{\left(\Gamma(\lambda+1)\Gamma\left(\lambda+\frac{3}{2}\right)\right)^2}\right). \tag{3.21}$$

De nou, basta veure que el numerador és positiu, i ho és segur [3].

Així, com el denominador del mètode de Newton és una funció negativa menys una positiva, tenim que és negatiu. A més, com NEG és sempre positiva i decreixent per valors de $\beta_j > 0$ es té que

$$\left.\frac{\partial f(\beta)}{\partial \beta_j}\right|_{\beta_j=0^+} = -\frac{1}{NEG(\beta_j|\lambda,\gamma)} \cdot \left.\frac{\partial NEG(\beta_j|\lambda,\gamma)}{\partial \beta_j}\right|_{\beta_j=0^+} > 0. \tag{3.22}$$

($\Leftarrow$)

Ara, estudiem β_j^{new} si es dóna (3.18). Si la parcial de $L(\bar{\beta})$ és positiva, tenim que

$$\left.\frac{\partial L(\bar{\beta})}{\partial \beta_j}\right|_{\beta_j=0^+} - \left.\frac{\partial f(\beta)}{\partial \beta_j}\right|_{\beta_j=0^+} > 0. \tag{3.23}$$

Per tant, $\beta_j^{new} > 0$ i no hi ha canvi de signe:

$$\beta_j^{new} \leftarrow 0^+ - \left.\frac{\frac{\partial \ln p(\bar{\beta}|X,Y)}{\partial \beta_j}(>0)}{\frac{\partial^2 \ln p(\bar{\beta}|X,Y)}{\partial \beta_j^2}(<0)}\right|_{\beta_j=0^+}.$$

D'altra banda, si la parcial de $L(\bar{\beta})$ és negativa (3.18) es redueix a

$$\left.\frac{\partial L(\bar{\beta})}{\partial \beta_j}\right|_{\beta_j=0} + \left.\frac{\partial f(\beta)}{\partial \beta_j}\right|_{\beta_j=0^+} < 0. \tag{3.24}$$

Així, com NEG és una funció parell, les pendents de β_j oposats tenen signes oposats, i per tant, la funció derivada és senar. D'aquesta manera,

$$\left.\frac{\partial f(\beta)}{\partial \beta_j}\right|_{\beta_j=0^+} = -\left.\frac{\partial f(\beta)}{\partial \beta_j}\right|_{\beta_j=0^-}.$$

Per tant, podem concloure que en aquest cas, tampoc hi ha canvi de signe:

$$\beta_j^{new} \leftarrow 0^- - \left.\frac{\frac{\partial \ln p(\bar{\beta}|X,Y)}{\partial \beta_j}(<0)}{\frac{\partial^2 \ln p(\bar{\beta}|X,Y)}{\partial \beta_j^2}(<0)}\right|_{\beta_j=0^-}.$$

(⇒)

Per provar l'altre implicació, basta veure que si no hi ha canvi de signe al voltant de l'origen és perquè es dóna (3.23) o (3.24).

□

Calcular la derivada parcial de L és costós, per tant, Hoggart *et al.* empren el Teorema 3.3.1 per establir un criteri que el eviti fer el càlcul quan no sigui necessari. Així, afiten superior i inferiorment la parcial de L. Donat un $\beta_j = 0$, si el valor absolut d'ambdues fites és menor que la parcial de f en $\beta_j = 0^+$, també ho serà el valor absolut de la derivada de L i el teorema ens assegura que hi haurà canvi de signe, per tant, el coeficient es manté a zero i no cal calcular la parcial. D'aquesta manera, si algun dels valors absoluts de les fites és major que la derivada de f, no podem assegurar que hi hagi un canvi de signe. En aquest cas, es calcula la parcial de L i es comproven els límits laterals.

Les fites proposades són les definides a continuació [3, supplem.],

$$\begin{aligned}
\frac{\partial L(\bar{\beta})}{\partial \beta_j} &> -\frac{\sum_{i=1}^{n} I(Y_i X_{ij} < 0)|X_{ij}|}{1+e^{\eta_{min}}} + \frac{\sum_{i=1}^{n} I(Y_i X_{ij} > 0)|X_{ij}|}{1+e^{\eta_{max}}} \\
\frac{\partial L(\bar{\beta})}{\partial \beta_j} &< -\frac{\sum_{i=1}^{n} I(Y_i X_{ij} < 0)|X_{ij}|}{1+e^{\eta_{max}}} + \frac{\sum_{i=1}^{n} I(Y_i X_{ij} > 0)|X_{ij}|}{1+e^{\eta_{min}}},
\end{aligned} \tag{3.25}$$

on

$$I(E) = \begin{cases} 0 & \text{Si E és fals,} \\ 1 & \text{altrament,} \end{cases}$$

i η_{max}, η_{min} són fita superior i inferior de $\{Y_i(\beta_0 + \beta^T X_i)\}_{i=1,\dots,n}$, respectivament. Donat que aquestes fites depenen dels paràmetres β_j, s'actualitzen a cada iteració. Així, el valor inicial és

$$\eta_{max} = \max_i \left(Y_i \left(\ln \left(\frac{n_1}{n_0} \right) \right) \right), \quad \eta_{min} = \min_i \left(Y_i \left(\ln \left(\frac{n_1}{n_0} \right) \right) \right).$$

Quan es calcula cada nou $Y_i(\beta_0 + \beta^T X_i)$ es compara amb les fites. Si el seu valor és menor que η_{min}, passa a ser fita inferior. Si supera η_{max}, esdevé fita superior. D'aquesta manera, es mantindran les fites per molt que variïn els β_j.

Teorema 3.3.2. *La derivada parcial de L respecte de β_j està afitada superior i inferiorment, tal com es defineix en (3.25).*

Demostració. La derivada parcial de L respecte de β_j es pot escriure segons els valors que pren Y_i, emprant la funció $I(E)$:

$$\frac{\partial L(\tilde{\beta})}{\partial \beta_0} = \sum_{i=1}^{n} \frac{I(Y_i > 0)|X_{ij}|}{1 + e^{Y_i(\beta_0 + \beta^T X_i)}} - \sum_{i=1}^{n} \frac{I(Y_i < 0)|X_{ij}|}{1 + e^{Y_i(\beta_0 + \beta^T X_i)}}.$$

D'altra banda, emprant les fites tenim que

$$\begin{aligned}
-\frac{\sum_{i=1}^{n} I(Y_i X_{ij} < 0)|X_{ij}|}{1 + e^{\eta_{min}}} &< -\sum_{i=1}^{n} \frac{I(Y_i < 0)|X_{ij}|}{1 + e^{Y_i(\beta_0 + \beta^T X_i)}} < -\frac{\sum_{i=1}^{n} I(Y_i X_{ij} < 0)|X_{ij}|}{1 + e^{\eta_{max}}}, \\
\frac{\sum_{i=1}^{n} I(Y_i X_{ij} > 0)|X_{ij}|}{1 + e^{\eta_{max}}} &< \sum_{i=1}^{n} \frac{I(Y_i > 0)|X_{ij}|}{1 + e^{Y_i(\beta_0 + \beta^T X_i)}} < \frac{\sum_{i=1}^{n} I(Y_i X_{ij} > 0)|X_{ij}|}{1 + e^{\eta_{min}}}.
\end{aligned}$$

D'aquesta manera, es segueix (3.25).

□

La següent idea per controlar el pas, és fer el denominador de Newton (3.6) més gran en valor absolut, emprant una fita inferior de la parcial de segon ordre de L [3]. Per fer-ho, Hoggart *et al.* es basen en les modificacions del mètode de Newton, aplicat per resoldre el problema de ridge regression, exposades a l'article [10]. Així, s'empra la funció $F(r, \delta)$ amb $\delta \geq 0$ definida a continuació,

$$F(\eta, \delta) = \begin{cases} 0.25 & \text{si } |\eta| \leq \delta, \\ \dfrac{1}{2 + e^{|\eta| - \delta} + e^{\delta - |\eta|}} & \text{altrament.} \end{cases} \tag{3.26}$$

Tal com prova el següent teorema, F és fita superior de certa funció. Aquesta propietat, és la que permet determinar una fita inferior de la parcial.

Teorema 3.3.3. *Sigui $F(\eta, \delta)$ la funció definida a (3.26), per a tot $\eta \in \mathbb{R}$ i $\delta \in \mathbb{R}^+ \cup \{0\}$ es satisfà*

$$F(\eta, \delta) \geq \frac{e^{\eta}}{(1 + e^{\eta})^2}.$$

Demostració. Sigui $\eta \in \mathbb{R}$ i $\delta \in \mathbb{R}^+ \cup \{0\}$, vegem que 0.25 sempre es fita superior de $\frac{e^\eta}{(1+e^\eta)^2}$. Així, en particular també ho serà quan $|\eta| \leq \delta$.

Tenim que,

$$0 \leq \left(1-e^\eta\right)^2 = \left(1-2e^\eta+e^{2\eta}\right).$$

Aleshores, multiplicant per $e^{-\eta} > 0$,

$$\begin{array}{rcl} 0 & \leq & e^{-\eta}-2+e^\eta, \\ 4 & \leq & e^{-\eta}+2+e^\eta. \end{array}$$

Per tant,

$$\frac{1}{4} \geq \frac{1}{e^{-\eta}+2+e^\eta} = \frac{e^\eta}{(1+e^\eta)^2}\,. \tag{3.27}$$

D'altra banda, cal provar que si $|\eta| > \delta$ tenim la cota superior

$$\frac{1}{2+e^{|\eta|-\delta}+e^{\delta-|\eta|}}.$$

La funció $h(\theta) = e^\theta + e^{-\theta}$ és creixent per $\theta > 0$. Així, com estam en el cas $|\eta|-\delta > 0$ i, a més, $|\eta|-\delta < |\eta|$, tenim que

$$\begin{array}{rcll} e^{|\eta|-\delta}+e^{\delta-|\eta|} & \leq & e^{|\eta|}+e^{-|\eta|} & = \quad e^\eta+e^{-\eta}, \\ 2+e^{|\eta|-\delta}+e^{\delta-|\eta|} & \leq & 2+e^\eta+e^{-\eta}, & \end{array}$$

Aleshores,

$$\frac{1}{2+e^{|\eta|-\delta}+e^{\delta-|\eta|}} \geq \frac{e^\eta}{(1+e^\eta)^2}, \quad |\eta| \geq \delta. \tag{3.28}$$

Així, per (3.27) i (3.28) es segueix que $F(\eta,\delta)$ amb $\delta \geq 0$ és una fita superior de $\frac{e^\eta}{(1+e^\eta)^2}$. □

Aplicant el Teorema 3.3.3 i prenent $F\left(Y_i(\beta_0+\beta^T X_i),\delta\right)$ es té

$$F\left(Y_i(\beta_0+\beta^T X_i),\delta\right) \geq \frac{e^{Y_i(\beta_0+\beta^T X_i)}}{\left(1+e^{Y_i(\beta_0+\beta^T X_i)}\right)^2} > 0, \quad \forall i = 1,\ldots,n.$$

Per tant, una fita inferior de la derivada és

$$-\sum_{i=1}^n X_{ij}^2 F\left(Y_i(\beta_0+\beta^T X_i),\delta\right) \leq \frac{\partial^2 L(\bar{\beta})}{\partial \beta_j^2}.$$

Notem que per $\delta = 0$ es dóna la igualtat.

Aquesta fita resulta útil a l'article al que Hoggart *et al.* fan referència, ja que es basa en ridge regression [10]. Com la derivada de segon ordre de la penalització que prové d'una normal és positiva, diguem-li h, basta substituir la derivada parcial de segon ordre de L per la fita inferior. Així, el denominador de Newton serà més gran en valor absolut:

$$-\sum_{i=1}^{n} X_{ij}^2 F\left(Y_i(\beta_0+\beta^T X_i),\delta\right)-\frac{\partial^2 h(\beta)}{\partial\beta_j^2}\leq\frac{\partial^2 L(\bar{\beta})}{\partial\beta_j^2}-\frac{\partial^2 h(\beta)}{\partial\beta_j^2}<0.$$

En el nostre cas, $f(\beta)$ com a funció d'un β_j és còncava a $(-\infty,0)$ i $(0,+\infty)$ (no a la unió), malgrat no sigui una funció còncava. D'aquesta manera, la derivada parcial de segon ordre de f es negativa en els punts on es pot calcular. Aquest fet, fa que no basti substituir la derivada del logaritme de la funció de versemblança per la fita inferior obtinguda, és a dir, el denominador de Newton només reduirà el pas quan es satisfacin les hipòtesis del següent resultat. Altrament, augmentarà. Així hi tot, Hoggart *et al.* opten per emprar la fita inferior enlloc de la parcial, menys quan es tracta d'actualitzar β_0.

Resultat 3.3.2. *Si* $\dfrac{\partial^2 L(\bar{\beta})}{\partial\beta_j^2}\leq\dfrac{\partial^2 f(\beta)}{\partial\beta_j^2}$, *aleshores*

$$\left|\frac{\partial^2 L(\bar{\beta})}{\partial\beta_j^2}-\frac{\partial^2 f(\beta)}{\partial\beta_j^2}\right|\leq\left|-\sum_{i=1}^{n} X_{ij}^2 F\left(Y_i(\beta_0+\beta^T X_i),\delta\right)-\frac{\partial^2 f(\beta)}{\partial\beta_j^2}\right|.$$

Demostració. Donada la fita inferior de la derivada de segon ordre de L, baix les hipòtesis del resultat tenim que

$$-\sum_{i=1}^{n} X_{ij}^2 F\left(Y_i(\beta_0+\beta^T X_i),\delta\right)\leq\frac{\partial^2 L(\bar{\beta})}{\partial\beta_j^2}\leq\frac{\partial^2 f(\beta)}{\partial\beta_j^2}.$$

Per tant, s'obté el que volíem demostrar com segueix a continuació.

$$\begin{array}{ccc}
-\dfrac{\partial^2 L(\bar{\beta})}{\partial\beta_j^2} & \leq & \displaystyle\sum_{i=1}^{n} X_{ij}^2 F\left(Y_i(\beta_0+\beta^T X_i),\delta\right), \\
-\dfrac{\partial^2 L(\bar{\beta})}{\partial\beta_j^2}+\dfrac{\partial^2 f(\beta)}{\partial\beta_j^2} & \leq & \displaystyle\sum_{i=1}^{n} X_{ij}^2 F\left(Y_i(\beta_0+\beta^T X_i),\delta\right)+\dfrac{\partial^2 f(\beta)}{\partial\beta_j^2} \quad . \\
\| & & \| \\
\left|\dfrac{\partial^2 L(\bar{\beta})}{\partial\beta_j^2}-\dfrac{\partial^2 f(\beta)}{\partial\beta_j^2}\right| & & \left|-\displaystyle\sum_{i=1}^{n} X_{ij}^2 F\left(Y_i(\beta_0+\beta^T X_i),\delta\right)-\dfrac{\partial^2 f(\beta)}{\partial\beta_j^2}\right|
\end{array}$$

□

Notem que tenim una fita inferior per qualsevol $\delta>0$. Així hi tot, cal anar en compte a l'hora d'escollir el paràmetre, ja que la fita emprada es pot entendre com la derivada de segon ordre de L aplicada a un $\beta_j+\epsilon$, per algun ϵ. Si aquest valor s'allunya de β_j, el

valor de β_j^{new} pot ser molt diferent de l'esperat al emprar la derivada. Donat que no ens interesa que aquest fet ocorri, cal emprar un δ que fasi que $\beta_j + \epsilon$ es mantengui entorn β_j, és a dir, dins la regió de confiança [10].

Basant-se en les modificacions de l'algoritme explicades en [10], Hoggart *et al.* empren un paràmetre diferent per a cada β_j i, per tant, per a cada iteració. Considerem el paràmetre Δ_j amb la següent actualització a cada pas,

$$\Delta_j^{new} \longleftarrow \max\left(2|\Delta\beta_j|, \frac{\Delta_j}{2}\right), \quad j = 1, ..., m, \tag{3.29}$$

on $\Delta\beta_j = \beta_j^{new} - \beta_j$, i Δ_j s'inicialitza en 1. Amb aquest nou valor, que serà el δ del següent pas, la funció emprada és $F\left(Y_i(\beta_0 + \beta^T X_i), \left|\Delta_j \cdot X_{ij}\right|\right)$.

Finalment, de nou amb l'objectiu de controlar el pas, Hoggart *et al.* realitzen una modificació més a l'actualització de pesos de Newton, basant-se amb l'article [10]. Considerem la següent variació a cada pas, tal com ha quedat definida amb F,

$$\Delta v_j = \frac{\dfrac{\partial}{\partial\beta_j} L(\bar{\beta}) - \dfrac{\partial}{\partial\beta_j} f(\bar{\beta})}{-\displaystyle\sum_{i=1}^{n} X_{ij}^2 F\left(Y_i(\beta_0 + \beta^T X_i), \delta\right) - \dfrac{\partial^2}{\partial\beta_j^2} f(\beta)}, \quad j = 1, ..., m.$$

Recordem que per β_0, no s'utilitza F, per tant, Δv_0 manté la forma original del mètode de Newton.

D'aquesta manera, només es restarà Δv_j quan no excedeixi el valor Δ_j, que s'ha definit a 3.29 per controlar que cada β_j es mantengui dins una regió de confiança. Així, considerem

$$\Delta\beta_j = \begin{cases} -\Delta_j & \text{Si } \Delta v_j < -\Delta_j, \\ \Delta v_j & \text{Si } |\Delta v_j| \leq \Delta_j, \\ \Delta_j & \text{Si } \Delta j < -\Delta v_j. \end{cases}$$

En resum, l'actualització a cada pas resulta

$$\beta_j^{new} \longleftarrow \beta_j - \Delta\beta_j.$$

3.4. Error de Tipus I

Un cop establit el model, interessa conèixer la probabilitat de l'error de Tipus I que s'origina al estimar els paràmetres pel mètode d'Hlasso. Per fer-ho, s'obté una fita superior de la probabilitat de l'error que depèn únicament del nombre de casos, de controls i dels paràmetres λ i γ, és a dir, no depèn del valor dels coeficients β_j obtinguts.

Les Hipòtesis nul·les que volem contrastar són les següents,

$$H_{0j} : \beta_j = 0, \quad j = 1, ..., m.$$

És a dir, que la covariable X_{ij} no influeix en el valor de la variable Y_i. D'aquesta manera, rebutjarem H_{0j} si obtenim el coeficient diferent de zero [3].

Suposant H_{0j} certa, es vol obtenir la probabilitat de que β_j es calculi com a diferent de zero en la maximització, en altres paraules, la probabilitat de l'error de Tipus I pel nostre contrast. Sigui α la probabilitat d'error per β_j , amb els càlculs de [3, supplem.], Hoggart *et al.* calculen la següent fita de la probabilitat de l'error, quan les dades estan estandarditzades i hi ha el mateix nombre de casos i controls,

$$\alpha_j \leq 2\left(1-\Phi\left(\sqrt{\frac{n_0+n_1}{n_0 n_1}}\cdot\left.\frac{\partial f(\beta)}{\partial \beta_j}\right|_{\beta_j=0^+}\right)\right), \tag{3.30}$$

on Φ és la funció de distribució d'una normal estàndard.

A més, coneixem el valor de la parcial en $\beta_j = 0^+$,

$$\left.\frac{\partial f(\beta)}{\partial \beta_j}\right|_{\beta_j=0^+} = \frac{2\left(\lambda+\frac{1}{2}\right)D_{-(2\lambda+2)}(0)}{\gamma D_{-(2\lambda+1)}(0)},$$

on $D_v(0)$ es pot calcular a partir de (3.20).

D'aquesta manera,

$$\begin{aligned}\alpha_j &\leq 2\left(1-\Phi\left(\sqrt{\frac{n_0+n_1}{n_0 n_1}}\cdot\frac{2\left(\lambda+\frac{1}{2}\right)}{\gamma}\cdot\frac{2^{\lambda+\frac{1}{2}}\Gamma(\lambda+1)}{2^{\lambda+1}\Gamma\left(\lambda+\frac{3}{2}\right)}\right)\right)\\ &= 2\left(1-\Phi\left(\sqrt{\frac{n_0+n_1}{n_0 n_1}}\cdot\frac{\sqrt{2}}{\gamma}\cdot\frac{\Gamma(\lambda+1)}{\Gamma\left(\lambda+\frac{1}{2}\right)}\right)\right).\end{aligned} \tag{3.31}$$

Com podem observar, els paràmetres λ i γ de la funció NEG permeten determinar una fita petita de l'error de Tipus I. Així, Hoggart *et al.* solen fixar λ, normalment en 0.05, i calculen γ de manera que l'error sigui tan petit com es vulgui, per exemple $\alpha = 10^{-5}$. A més, per certs casos la desigualtat anterior és estricta [25].

CAPÍTOL 4

EXEMPLES

Al capítol anterior s'ha explicat el model matemàtic que recolza HyperLasso, així com els detalls en la seva implementació. L'objectiu d'aquest capítol és explicar com emprar el programa i com interpretar-ne la sortida. Per fer-ho, es donen dos exemples aplicats a dades reals del genoma d'individus sans i d'individus que presenten una malaltia. Així, es pretén situar els exemples en el context de GWAS (*Genome-wide association study*).

4.1. Respresentació de les dades en GWAS

Moltes malalties comuns en els humans tenen un fort component genètic, a més dels factors ambientals [27]. Aquest fet impulsa l'estudi d'associació en tot el genoma, GWAS, que és l'estudi del genoma humà per tal d'identificar quines variacions genètiques influeixen en major grau en la presència o no d'una malaltia específica [28]. Conèixer aquestes variacions suposa una gran ajuda en l'estudi de la malaltia, ja que pot aportar informació rellevant a l'hora de millorar els mètodes de diagnòstic i prevenció [29]. Les dades dels exemples que veurem en aquest capítol es basen, precisament, en variacions genètiques simples. Per entendre que són i com es codifiquen, cal conèixer com es representa el genoma.

Un gen és una llista ordenada de bases que es representen per lletres en el conjunt $\{G, C, T, A\}$. Cada humà té aproximadament 20000 parells de gens, un de la mare i un del pare, que composen els 23 parells de cromosomes. A més, tots els humans presenten diferències en bases en la mateixa posició de diversos gens, algunes d'elles bastant comunes. Per poder fer comparacions, s'empra una llista de gens de referència que es coneix per HRG (*Human reference genome)*. En aquest *humà de referència* es representa cada cromosoma individualment, és a dir, la llista no esta formada per parells cromosomes [30].

Una variació genètica simple (SNV, *Single Nucleotide Variant*) és un canvi de base en una posició concreta d'un gen, respecte l'HRG [31]. En GWAS s'estableix la hipòtesis de que la manifestació d'una malaltia està associada a certs SNVs. Així, l'estudi es basa en determinar quines covariables de SNVs afecten a la variable dependent de la presència o no de la malaltia. Per fer-ho, es recullen dades del parells d'SNVS de milers de persones sanes (controls) i milers de persones de característiques similars amb la malaltia(casos) [28].

Amb l'objectiu de comparar els SNVs de la mostra, s'assigna un valor a cada parell d'SNVs segons el tipus de variació genètica que representa. Si l'individu presenta un canvi de base a l'SNV del gen del pare o de la mare, respecte l'humà de referència, s'indica amb un 1, si ho presenta en ambdós gens, amb un 2, i si no hi ha canvi, amb un 0. D'aquesta manera, per a cada individu es coneix un llista de valors en $\{0,1,2\}$ que representen els canvis de cada parell d'SNVs respecte l'SNV de l'humà de referència.

4.2. El programa HyperLasso

La implementació de l'algoritme explicat a la secció 3.3, capítol 3, ha estat elaborada per Hoggart i es pot descarregar en `http://www.ebi.ac.uk/projects/BARGEN`. Al *readme.txt* del programa s'explica com instal·lar HLASSO en Linux, cal esmentar que no s'ha presentat cap problema en la instal·lació. A més, conté exemples i explicacions sobre el model d'HLASSO que complementa la informació de l'article principal [3].

Per a cada individu de la mostra, es necessita saber el valor de la variable dependent $Y_i \in \{0,1\}$ i el valor de les covariables $X_1,...,X_n$ de SNVs que es vulguin analitzar, que representen el tipus de mutació observada. Això és la llista de valors en $\{0,1,2\}$. A diferència del model explicat, els controls es denoten per 0 i 1, i no per -1 i 1, aquest fet es deu a que el programa canvia internament els 0 per -1 abans de realitzar la maximització.

La cridada al programa, amb els paràmetres de entrada que s'empraran, es la següent.

```
./runHLasso -genotypes dvFile -target ivFile -shape value -scale value
 -std -o outputFile -iter 10
```

El programa HLASSO reb la informació dels elements de la mostra en dues taules: una amb el valor de les variables dependents Y_i, que s'introdueix en `-target` i l'altre amb les variables independents $X_1,...,X_n$ de SNVs, que s'especifica en `-genotypes`. Quant al fitxer on volem guardar els resultats, que serà un `.R`, s'indica en `-o`.

De les altres variables d'entrada ens interessen les que afecten a les propietats del model d'HLASSO. Així, s'introdueix el paràmetre de forma, λ, i el d'escala, γ, en `-shape` i `-scale`, respectivament. La implementació d'HLASSO [3] suggereix emprar un λ prou petit, perquè el cim i les coes de NEG presentin les propietats desitjades. En els seus

exemples Hoggart *et al.* empren $\lambda = 0.05$, i expliquen que és el menor valor que no els genera problemes computacionals.

A més, el programa permet emprar un nou paràmetre, que anomena *penalty*, enlloc de γ. Si s'especifica el seu valor en `-penalty`, la rutina calcula el paràmetre γ a través de la següent expressió, que relaciona ambdues variables

$$\text{penalty} = \frac{\sqrt{2}}{\gamma} \cdot \frac{\Gamma(\lambda+1)}{\Gamma\left(\lambda+\frac{1}{2}\right)}.$$

Fixem-nos que penalty és, de fet, és el limit lateral per la dreta de la parcial de f.

Per acabar, donat el caràcter multi-modal del model d'HLASSO, són necessàries varies execucions del programa. Per fer-ho, només cal indicar en `-iter` en nombre desitjat. Recordem que cada execució es fa sobre una permutació aleatòria en l'ordre dels SNVs, i per tant, en l'actualització dels β_j. En els exemples també s'emprarà l'opció `-std`, que estandaritza les dades dels SNVs de la mostra. Així, podrem calcular la fita de la probabilitat de l'error de Tipus emprant (3.31).

4.3. Exemples en GWAS

En els exemples d'aquest capítol s'empra una mostra de 625 persones amb càncer de pulmó, que ha estat obtinguda a partir de *The Cancer Genome Atlas Consortium*. Per a cada individu, es coneixen dues llistes de 5266 covariables de SNVs, amb els respectius valors en $\{0,1,2\}$, una del teixit tumoral i l'altre de teixit normal, és a dir, una mostra de teixit no proper al tumor. Això suposa un total de 1250 variables independents, amb el mateix nombre de casos i controls. El problema d'aquestes dades és que les files de les matrius d'entrada no són independents, ja que a cada persona li pertanyen dues files. Tot i així, aplicarem HLASSO.

Els resultats complets dels posteriors exemples, així com les dades de SNVs, es poden trobar en `http://bass.uib.es/~mar/`. Per llegir la sortida de HLASSO només cal emprar la instrucció `dget()` de `R`, per exemple, `dget('example1_models.R')`. D'altra banda, per llegir la matriu de covariables de SNVs, basta fer

```
SNV<-read.table('SnvMatrix',header=TRUE)
```

D'aquesta manera, es pot emprar `R` per accedir a la informació d'ambdues matrius, i així calcular la informació necessària per interpretar els resultats.

4.3.1. Exemple 1

Primer, executarem el programa demanant que estandaritzi les dades. Per escollir els paràmetres s'ha fixa't $\lambda = 0.05$ i s'ha calculat el valor de γ per a que la fita superior de l'error sigui 10^{-5}, segons (3.31). Així, s'ha pres $\gamma = 0.0109096$, que equival a penalty= 78.0853.

Imposant 10 iteracions, la cridada resulta:

```
./runHLasso -genotypes ~/hyper/data/SnvMatrix  -target ~/hyper/data/
phenotype -shape 0.05 -scale 0.0109096 -std -o data/example1
 -iter 10
```

S'han trobat 10 modes, és a dir, 10 solucions diferents al problema de maximització. De tots els cims, prenem el màxim, que es troba en la quarta iteració i el seu valor és 863.3188. A continuació, tenim part de la sortida d'aquesta iteració, llegida amb R.

```
     [,1]         [,2]                    [,3]
[1,] "-183.8015" "133"                   "309"
[2,] "863.3188"  "9:33798574:G:A:PRSS3" "17:45249335:T:G:CDC27"
[3,] "18"        "-1.494693"             "-0.5645863"
[4,] "0"         "0"                     "0"
                           ...

     [,22]                [,23]                       [,24]
[1,] "5193"               "5241"                      "0"
[2,] "9:172167:C:T:CBWD1" "1:144871782:A:G:PDE4DIP" "Intercept"
[3,] "-1.548222"          "-2.16474"                  "2.130448"
[4,] "0"                  "0"                         "0"
```

La primera columna de la matriu tornada en cada iteració aporta la informació sobre la funció objectiu. Així, en la posició $(1,1)$ tenim el valor que pren la funció de versemblança L i en la posició $(2,1)$ el valor total, afegint-li la penalització: $L - f$. Comparant aquest valor es pot determinar quin dels màxims locals obtinguts és el major. Les altres columnes indiquen la informació dels SNVs seleccionats pel model d'HLASSO: la primera fila conté la posició dels SNVs en la permutació, la segona els noms i la tercera els coeficients β_j otinguts.

En aquest cas, de 5266 variables s'han seleccionat només 22 covariables d'SNVs. Cal destacar, que els coeficients són significativament diferents de zero, de fet, tots superen el 0.5 en valor absolut. En total, hi ha 9 coeficients positius, a més de β_0. Quant a la probabilitat de l'error de Tipus I és menor que 10^{-5}, ja que s'ha escollit el paràmetre γ amb aquest objectiu.

4.3.2. Exemple 2

El programa d'HLASSO s'ha executat de nou amb les dades estandaritzades, canviant els paràmetres de forma i escala. Així prendrem $\lambda = 0.5$, un valor prou més gran que l'anterior, i penalty= 12, que equival a $\gamma = 0.1044428$. Així, la cridada és

```
./runHLasso -genotypes ~/hyper/data/SnvMatrix  -target ~/hyper/data/
phenotype -shape 0.5 -penalty 12 -std -o data/example2
 -iter 10
```

Aquesta vegada el màxim valor de la funció objectiu en les 10 iteracions és -218.7049, i s'assoleix en la vuitena iteració. Notem que per aquests paràmetres, prou més grans que els anteriors, s'han seleccionat 50 SNVs dels 5266. D'altra banda, no tots els coeficients són significativament diferents de zero, per exemple, el menor coeficient en valor absolut és $3.104209 \cdot 10^{-13}$. Aquest fet dificulta la relació d'alguns SNVs amb la variable binària Y_i, encara que trobam valors com 4.577333.

A continuació tenim part de la sortida en la vuitena iteració.

```
     [,1]        [,2]                  [,3]
[1,] "-83.36026" "133"                 "205"
[2,] "-218.7049" "9:33798574:G:A:PRSS3" "17:21319087:G:A:KCNJ18"
[3,] "66"        "-1.963986"           "-1.374144"
[4,] "0"         "0"                   "0"
                          ...

     [,50]               [,51]                      [,52]
[1,] "5193"              "5241"                     "0"
[2,] "9:172167:C:T:CBWD1" "1:144871782:A:G:PDE4DIP" "Intercept"
[3,] "-1.538583"         "-3.120206"                "0.03509385"
[4,] "0"                 "0"                        "0"
```

En aquest exemple, la fita de la probabilitat de l'error de Tipus I és molt gran, concretament 0.4972503. Per tant, no tenim raons estadístiques per suggerir que aquests SNVs puguin estar relacionats amb el càncer de pulmó.

Comparant els SNVs seleccionats en ambdós models, s'ha vist que hi ha 21 SNVs comuns en la selecció. D'aquesta manera, dels 22 SNPs associats a l'exemple 1, 21 apareixen a l'exemple 2. A més, els coeficients per a cada SNP tenen el mateix signe en cada model i valors propers. Aquests resultats es recullen a la Taula 4.1.

Aplicar el programa HLASSO a dades de SNVs, amb els paràmetres adequats, pot seleccionar un subconjunt reduït SNVs amb una probabilitat de l'error de Tipus I per coeficient força petit. Aquesta informació, juntament amb altres anàlisis estadístics i més informació biològica pot ajudar a descobrir la associació de diversos SNVs amb una malaltia específica.

Coeficient Ex1	Coeficient Ex2	SNV
-1.494693	-1.963986	9:33798574:G:A:PRSS3
-0.5645863	-0.4030229	17:45249335:T:G:CDC27
0.9239384	0.97442	11:1017084:G:A:MUC6
-1.45358	-1.486691	2:112615888:C:G:ANAPC1
-1.328894	-1.837389	2:97845632:T:C:ANKRD36
-1.996361	-2.684205	17:21319786:G:A:KCNJ18
-1.084921	-1.756122	1:143767643:T:C:PPIAL4G
-1.345099	-0.9970422	7:72413593:T:C:POM121
1.393971	1.634494	11:1017789:A:C:MUC6
2.687177	2.749099	1:146398387:G:C:NBPF12
1.549591	1.410283	21:11058226:G:C:BAGE3
-1.408707	-1.644662	3:113524266:G:C:ATP6V1A
1.217281	1.643215	11:1017069:G:A:MUC6
1.549459	2.035154	1:144220807:A:C:NBPF20
-0.9714698	-1.067016	9:33385863:G:T:AQP7
0.6417706	0.3852579	2:130832292:T:A:POTEF
1.632943	1.691106	11:1017325:A:C:MUC6
-0.7213303	-1.219892	5:115249078:C:T:AP3S1
-1.092169	-0.4957887	12:52865925:C:T:KRT6C
-1.548222	-1.538583	9:172167:C:T:CBWD1
-2.16474	-3.120206	1:144871782:A:G:PDE4DIP

Taula 4.1: Valor dels coeficients dels SNVs comuns seleccionats en els Exemples 1 y 2. Per l'exemple 1 s'ha pres $\lambda = 0.05$ i $\gamma = 0.0109096$; els resultats mostrats són els obtinguts en la quarta iteració d'HLASSO. Per l'exemple 2 els paràmetres són $\lambda = 0.5$ i penalty= 12 (equivalent a $\gamma = 0.1044428$); els resultats s'han obtingut en la vuitena iteració.

CAPÍTOL 5

CONCLUSIONS

El Treball de Final de Gau, que s'exposa en aquest document, sorgeix del meu l'interès per aprofundir en els coneixements sobre optimització adquirits en el Grau de Matemàtiques a la UIB. El tutor del treball, Jairo Rocha, em va proposar aquest tema d'optimització aplicada, que es relaciona amb probabilitat i estadística.

El problema d'optimització que s'ha estudiat, el mètode d'HyperLASSO, s'aplica per resoldre el problema de selecció de variables quan la mostra conté un gran nombre de possibles variables explicatives. Aquest mètode es basa en regressió logística, amb una penalització obtinguda a partir de la funció de densitat de NEG. Tal com s'ha vist en el capítol 4 d'exemples, els paràmetres λ i γ d'aquesta distribució afecten al nombre de variables seleccionades i en l'estimació de coeficients significativament diferent de zero. Tot i les observacions sobre aquests coeficients, l'estudi analític de com afecten als resultats s'escapa de l'objectiu de la memòria. Quant a les observacions, indiquen que, per les nostres dades, convé prendre un valor de γ prop del 0.01, mentre que per λ es segueixen les indicacions de Hoggart *et al.* de prendre un valor petit, que no sigui menor que 0.05. D'altra banda, en el exemple 1 s'ha vist el potencial de HyperLASSO per obtenir un conjunt molt petit de variables independents associades a Y_i. Una altre avantatge del mètode és la rapidesa en realitzar les iteracions, fet que es deu a les modificacions per millorar l'eficiència del mètode de Newton.

Durant la realització del treball he aprofundit els meus coneixements de regressió penalitzada, estudiant els mètodes a partir de l'objectiu de resoldre el problema de selecció de variables, enlloc de per intentar fer prediccions. A més, he après com s'obté el problema d'optimització a partir d'aplicar inferència Bayesiana, veient així una altra aplicació del Teorema de Bayes. D'altra banda, he conegut que són les funcions cilíndriques parabòliques i algunes propietats. Així mateix, abans de realitzar el treball no coneixia la distribució normal-exponential-gamma, ni el concepte de mescla de densitats.

En el capítol 3, secció 3.3, he après noves tècniques per millorar l'eficiència d'un algoritme i com es poden emprar mètodes per resoldre problemes d'optimització convexa, com el mètode de Newton, per resoldre problemes d'optimització no convexa i no diferenciable. Així, només cal modificar els algoritmes adequadament, sempre anant en compte de que els coeficients estimats en cada iteració no s'allunyin del valor que haurien d'obtenir sense les modificacions, sinó la solució del mètode por ser molt diferent de l'esperada.

Per poder realitzar la memòria i contrastar la informació rebuda principalment pels articles [3] i [10], he hagut d'aprendre un poc de Linux: fitxers i execució de programes. Sense fer cerques sobre el programa `CLG.cc` de HyperLASSO hagués resultat impossible la comprensió i aprofundiment sobre certs aspectes de l'algoritme. També he conegut com es representen les mutacions del genoma humà a les dades de SNVs, juntament amb conceptes de biologia bàsica que no coneixia o tenia oblidats.

Quant a les contribucions al treball, s'ha recolzat la definició de la distribució normal-exponential-gamma amb gràfiques que no apareixien als articles. Aquestes ajuden a tenir una idea intuïtiva de perquè el mètode de HyperLASSO resulta més útil en el problema de selecció de variables, amb un gran nombre de variables independents, que el mètode de LASSO i ridge regression.

A més de les gràfiques, les aportacions principals del treball resideixen en els detalls de l'algoritme d'HyperLASSO. Com s'ha comentat, molts d'aquests s'han extret de [10], contrastant la informació descrita amb el codi de HyperLasso. A més, totes les demostracions dels resultats i teoremes del capítol 3 són pròpies, llevat del Teorema 3.3.1, que dóna la condició de canvi de signe. D'aquest es tenia la indicació de que el denominador era negatiu a l'origen, sense demostra-ho, i que la derivada parcial de f és simètrica respecte el zero.

Finalment, aquest treball deixa obert el camí per obtenir una millor estimació de l'error de Tipus I, sense que sigui necessari estandaritzar les dades i tenir el mateix nombre de casos i controls. A més d'aplicar el mètode en GWAS per tal d'associar mutacions amb malalties específiques, continuant amb un estudi estadístic i biològic per poder obtenir resultats concloents. D'altra banda, també seria interessant analitzar les diferències amb altres mètodes de regressió penalitzada o extendre HyperLASSO en el tema de models lineals generalitzats, així com estudiar com afecta que les variables independents siguin discretes i ordenades als resultats. Quant al problema de independència de les variables dependents, Y_i, es pot estudiar com aplicar *random effects* a HyperLASSO.

A part dels coneixements adquirits en aquest treball, a nivell personal m'ha aportat la possibilitat d'aplicar l'optimització a dades reals en el context de GWAS, motivant el meu interès per HLASSO i el problema de selecció de variables, a més d'aplicar els meus coneixement d'estadística relacionats amb el mètode.

BIBLIOGRAFIA

[1] E. I. George, "The Variable Selection Problem," *Journal of the American Statistical Association*, vol. 95, no. 452, pp. 1304–1308, Dec. 2000. [Online]. Available: http://www-stat.wharton.upenn.edu/~edgeorge/Research_papers/George00JASA.pdf

[2] I. Ruczinski, "Variable Selection," Department of Biostatistics, Johns Hopkins University course. [Online]. Available: http://www.biostat.jhsph.edu/~iruczins/teaching/jf/ch10.pdf

[3] C. J. Hoggart, J. C. Whittaker, M. De Iorio, and D. J. Balding, "Simultaneous Analysis of All SNPs in Genome-Wide and Re-Sequencing Association Studies," *PLoSGenet*, vol. 4, no. 7, 2008.

[4] B. Efron and T. Hastie, *Computer Age Statistical Inference*. Cambridge, 2016.

[5] S. Omayma, "Interpreting Odd Ratios in Logistic Regression." [Online]. Available: http://rstudio-pubs-static.s3.amazonaws.com/182726_aef0a3092d4240f3830c2a7a9546916a.html

[6] X. Zhu, "Logistic Regression," Department of Computer Sciences, University of Winconsin-Madison. [Online]. Available: http://pages.cs.wisc.edu/~jerryzhu/cs769/lr.pdf

[7] J. D. M. Rennie, "Regularized Logistic Regression is Strictly Convex," 2005. [Online]. Available: http://qwone.com/~jason/writing/convexLR.pdf

[8] A. Doucet, "Uniqueness of MLE estimates in logistic regression," Department of Statistics, Oxford University. [Online]. Available: https://www.cs.ubc.ca/~arnaud/cs340/HW5_q2.pdf

[9] M. Eichler, "Bayes' Theorem," Department of Statistics, University of Chicago. [Online]. Available: http://galton.uchicago.edu/~eichler/stat24600/Handouts/l06.pdf

[10] A. Genkin, D. D.Lewis, and D. Madigan, "Large-Scale Bayesian Logistic Regression for Text Categorization," Section 4.2: The CLG Algorithm for Ridge Logistic Regression. [Online]. Available: http://yaroslavvb.com/papers/genkin-large.pdf

[11] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society.*, vol. 58, pp. 267–288, 1996.

[12] R. Stockute and P. Johnson, “Laplace Distribution,” 2013. [Online]. Available: http://pj.freefaculty.org/guides/stat/Distributions/DistributionWriteups/Laplace/Laplace-03.pdf

[13] R. J. Tibshirani, “The Lasso Problem and Uniqueness,” 2012, Carnegie Mellon University. [Online]. Available: https://arxiv.org/pdf/1206.0313.pdf

[14] E. Slud, “Handout on mixtures of densities and distributions,” Department of Mathematics, University of Maryland. [Online]. Available: http://www.math.umd.edu/~slud/s700/Mixtures.pdf

[15] Wikipedia, “Mixture distribution.” [Online]. Available: https://en.wikipedia.org/wiki/Mixture_distribution

[16] J. D. Cook, “Adult heights and mixture distributions.” [Online]. Available: https://www.johndcook.com/blog/mixture_distribution/

[17] Wikipedia, “Compound probability distribution.” [Online]. Available: https://en.wikipedia.org/wiki/Compound_probability_distribution

[18] ——, “Student’s t-distribution.” [Online]. Available: https://en.wikipedia.org/wiki/Student%27s_t-distribution

[19] J. E. Griffin and P. J. Brown, “Bayesian Hyper-LASSOS with non-convex penalization,” *Australian and New Zealand Journal of Statistics*, vol. 53, pp. 423–442, 2011.

[20] D. Panchenko, “Gamma, Chi-squared, Student T and fisher F Distributions,” 2006, Massachusetts Institute of Technology: MIT OpenCourseWare. License: Creative Commons BY-NC-SA. [Online]. Available: https://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/lecture-notes/lecture6.pdf

[21] Wikipedia, “Normal-exponential-gamma distribution.” [Online]. Available: https://en.wikipedia.org/wiki/Normal-exponential-gamma_distribution

[22] Digital Library of Mathematical Functions, “Numerical and asymptotic aspects of parabolic cylinder functions,” *Journal of Computational and Applied Mathematics*, vol. 121, pp. 221–246, 2000. [Online]. Available: http://ac.els-cdn.com/S0377042700003472/1-s2.0-S0377042700003472-main.pdf?_tid=6ccb2e4e-968a-11e7-8ae1-00000aab0f01&acdnat=1505090873_363ae8fe6f13f40a90f56b140e6e0c09

[23] S Zhang and J Jin, “Computation of Special Functions.” 1996, New York: Wiley.

[24] S. K. Hyde, “Properties of the Gamma function,” Department of Mathematics, Brigham Young University course. [Online]. Available: http://www.jekyll.math.byuh.edu/courses/m321/handouts/gammaproperties.pdf

[25] F. Frommlet, “Model Selection Procedures for Genome Wide Association Studies Slides,” pp. 75–88, May, 2013, Montefiore Institute. Department of Electrical Engineering and Computer Science. [Online]. Available: http://www.montefiore.ulg.ac.be/~mishra/systmod_presentations/2013_may_17_frommlet_presentation.pdf

[26] Digital Library of Mathematical Functions, “Parabolic Cylinder Functions. Properties.” [Online]. Available: http://dlmf.nist.gov/12.2

[27] F. Frommlet, M. Bogdan, and D. Ramsey, *Phenotypes and Genotypes*. Springer, 2014.

[28] Wikipedia, “Genome-wide association study.” [Online]. Available: https://en.wikipedia.org/wiki/Genome-wide_association_study

[29] “Estudios de asociación en todo el genoma,” NIH (National Human Genome Research Institute). [Online]. Available: https://www.genome.gov/27562846/estudios-de-asociacin-en-todo-el-genoma/

[30] Wikipedia, “Human genome.” [Online]. Available: https://en.wikipedia.org/wiki/Human_genome#SNP_frequency_across_the_human_genome

[31] ——, “Single-nucleotide polymorphism.” [Online]. Available: https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism