
**LAERTE DIAS DE CARVALHO**

# CONTRIBUIÇÃO À GEOMETRIA ANALÍTICA DOS ESTIMADORES LASSO E ELASTIC NET

**LAVRAS - MG**

**2020**

**LAERTE DIAS DE CARVALHO**

# CONTRIBUIÇÃO À GEOMETRIA ANALÍTICA DOS ESTIMADORES LASSO E ELASTIC NET

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Prof. Dr. Lucas Monteiro Chaves
Orientador

Prof. Dr. Devanil Jaques de Souza
Coorientador

**LAVRAS - MG**
**2020**

**Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).**

Carvalho, Laerte Dias de.
Contribuição à geometria analítica dos estimadores Lasso e Elastic Net / Laerte Dias de Carvalho. - 2020.
82 p. : il.

Orientador: Lucas Monteiro Chaves.
Coorientador: Devanil Jaques de Souza.
Tese (doutorado) - Universidade Federal de Lavras, 2020.
Bibliografia.

1. Estimadores Ridge. 2. Geometria de Modelos Lineares. 3. Estimadores de encolhimento. I. Chaves, Lucas Monteiro. II. Souza, Devanil Jaques de. III. Título.

À minha mãe, Maria Alda Dias de Carvalho e ao meu pai, Expedito Valdemiro de Carvalho (in memoriam); os meus irmãos Alda Maria de Carvalho Rocha; Adriana Dias de Carvalho Vieira e Luciano Dias de Carvalho, os meus filhos, Leonardo Chagas Carvalho e Camila Chagas Carvalho e à minha esposa Cíntia Maria Chagas Carvalho pelo carinho, amizade, amor e força em todos os momentos.

**DEDICO**

“Por vezes sentimos que aquilo que fazemos não é senão uma gota de água no mar. Mas o mar seria menor se lhe faltasse uma gota.”

Santa Madre Teresa de Calcutá

## AGRADECIMENTOS

A Deus, pelo sol que ilumina os meus dias, pelos corpos celestes que clareiam as minhas noites mais escuras, pelas chuvas que lavam a minha alma, pelas primaveras que colorem a minha vida com as suas flores, pelos pássaros que alegram o meu espírito com os seus cantos, pelos alimentos que fortificam o meu corpo, pela verdade que aos poucos me liberta, pelo trabalho que me faz sentir útil, pela vida que me proporciona novas experiências e oportunidades de crescimento, e por tudo que ainda não tenho consciência de que ELE faz por mim.

Aos meus pais Alda e Valdemiro (in memoriam) que, pelo exemplo de humildade e dignidade, me mostram o sentido da vida.

À minha eterna companheira Cíntia, sinônimo de vida e esperança, por ser meu forte nas horas difíceis e meu encanto nas horas de paz.

À minha irmã Alda Maria de Carvalho Rocha e sua família, pelo apoio direto e indireto.

A todos os meus familiares e amigos, pelo apoio e carinho.

Ao meu orientador Lucas Monteiro Chaves, que pelos conhecimentos e esclarecimentos intelectuais e morais confiados a mim, me ensinou o verdadeiro sentido de se ensinar.

Ao Prof. Devanil Jaques de Souza que de forma síncrona com as ideias do professor Lucas, contribui para minha formação moral e intelectual. À Universidade Federal de Lavras (UFLA) e ao Departamento de Estatística (DES), pela oportunidade de cursar o doutorado.

Aos professores do programa de pós-graduação em estatística e experimentação agropecuária da UFLA, pelas contribuições na minha formação durante as suas disciplinas, em especial aos professores Renato Ribeiro de Lima, Júlio de Sousa Bueno Filho, Thelma Sáfadi e Daniel Furtado Ferreira.

Às funcionárias do DES: Nádia de Carvalho Ferreira, Vítor Fernando Terra, Josiane Cristina Pinto de Oliveira, Maria das Dores Correa Santos e Magali de Souza Arantes Pedroso, pela amizade e carinho, em especial a Nádia, pelo companheirismo e compreensão nas minhas várias ligações e solicitações burocráticas.

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela concessão da bolsa de estudos no início do doutorado, tornando financeiramente possível a realização do doutorado.

À Universidade Federal de Viçosa pela oportunidade concedida e aos meus colegas do Departamento de Matemática da UFV pelo apoio incondicional.

A todos os colegas de turma pela companhia nos estudos, nos momentos de lazer e nos momentos de tensão de véspera das provas.

A todos que de alguma forma contribuíram, mesmo inconsciente, para a realização deste trabalho.

**RESUMO**

Os métodos de estimação e seleção de variáveis em modelos lineares LASSO (Least Absolute Shrinkage and Selection Operator)(1996) e Elastic Net (2005) são atualmente amplamente utilizados. Usualmente apresentados como solução de problemas variacionais, nesse trabalho são obtidos utilizando uma abordagem geométrica. Tal abordagem possibilitou a obtenção de propriedades relativas à geometria analítica do método de sua construção o que permitiu uma relação entre o estimador Elastic Net e o estimador Ridge. Também é apresentado um algoritmo para obter o estimador LASSO.

Palavras-chave: Estimadores Ridge. Geometria de Modelos Lineares. Estimadores de encolhimento.

**ABSTRACT**

The estimation and variable selection methods in the linear models LASSO (Least Absolute Shrinkage and Selection Operator) (1996) and Elastic Net (2005) have been widely used. The estimators are usually presented as a solution to variational problems. In this study, a geometric approach was proposed. Such approach allowed obtaining properties related to the analytical geometry of the method of construction which showed a close relationship between the Elastic Net estimator and the Ridge estimator. An algorithm to obtain the LASSO estimator is also presented.

Keywords: Ridge Estimator. Geometry of Linear Models. Shrinkage estimators.

## LISTA DE FIGURAS

**SUMÁRIO**

# 1 INTRODUÇÃO

Otimização convexa é um ramo especial da área de otimização matemática em que a função objetivo e o conjunto factível são ambos convexos. Problemas de otimização convexa surgem em diversas aplicações e já possuem uma teoria rica e desenvolvida, sendo esses os principais fatores que possibilitaram a criação de métodos computacionais robustos e eficientes para resolvê-los.

Embora a teoria matemática em que se baseia essa área, seja relativamente antiga, foi a partir do final do século XX que houve um grande desenvolvimento no estudo de aplicações e na construção de métodos computacionais específicos para essa classe de problemas. A concepção dos métodos de pontos interiores para problemas de programação linear criou uma nova classe de algoritmos que foram generalizados e que hoje resolvem de forma rápida e confiável os problemas de otimização convexa.

As aplicações surgem na estatística e em diversos ramos da engenharia, como sistemas de controle, análise e processamento de sinais, análise de circuitos, projeto de sistemas de energia elétrica, entre outros. Estudaremos uma classe especial de estimadores e, para isso, teremos que estudar também o conceito de dualidade e sua relação com os multiplicadores de Lagrange, para podermos aplicar a teoria de otimização convexa.

A relevância da teoria de otimização na estatística reside, principalmente, no fato de que podemos modelar problemas de natureza essencialmente estatística como um problema de otimização. Um dos métodos de estimação mais utilizados em estatística é o método da máxima verossimilhança em que é necessário maximizar a função verossimilhança (ou função log-verossimilhança). Outra aplicação é o cálculo da moda que é o ponto de máximo da função densidade de probabilidade.

A maximização de funções de várias variáveis não se resume em resolver a equação $\nabla f = 0$, já que é comum que haja restrições nas variáveis como por exemplo, restrições intervalares ou simplesmente, positividade nas variáveis. Dessa forma, o processo de estimação requer teorias matemáticas mais sofisticadas, como a teoria de otimização com restrições. A teoria de otimização convexa está bem desenvolvida na literatura e o objetivo desse projeto é a utilização da otimização convexa na obtenção de estimadores. Pretendemos analisar alguns estimadores de uma classe específica, estudar as suas propriedades e aplicá-los em situações concretas de interesse prático.

Hoerl e Kennard (1970), a partir da constatação de que o estimador de mínimos quadrados não era aceitável quando da presença de quasimulticolinearidade, propuseram, então, o chamado estimador de encolhimento Ridge. Esse estimador deu origem a outros estimadores de encolhimento, tais como os estimadores Lasso e Elastic Net, dentre outros.

Neste trabalho, pretende-se estudar esses estimadores, utilizando-se uma abordagem geométrica e,

além disso, expressar o processo de otimização desses estimadores, empregando a teoria de otimização convexa. Como referência básica, utilizou-se o livro Convex Optimization, de autoria de Stephen Boyd e Lieven Vandenberghe (BOYD; VANDENBERGHE, 2008).

Vale destacar que, com o advento de um número cada vez maior de dados, é colocado o problema de se obterem novos e mais eficientes métodos estatísticos. Em particular, para a teoria da regressão linear, o número elevado de covariáveis demanda, além de um robusto método de estimação dos parâmetros, também um método de seleção de covariáveis. O método de estimação, utilizando quadrados mínimos, apresenta problemas quando se tem, por exemplo, quasi-colinearidade, isto é, presença de covariáveis altamente correlacionadas. Os dois métodos clássicos mais utilizados para se sanar tal deficiência são a estimação Ridge e o método denominado Subset Selection. No entanto, ambos os métodos apresentam problemas.

A regressão Ridge é muito mais estável, mas raramente gera estimativas nulas para os parâmetros. Em razão desses fatos, foi proposto por Tibshirani (1996) um novo método denominado LASSO (Least Absolute Shrinkage and Selection Operator). A principal característica desse método é que, no processo de encolhimento, as estimativas de muitas covariáveis se anulam, e, portanto, é também um método automático de seleção de covariáveis. Definida em termos de quadrados mínimos penalizados, apresenta um desenvolvimento analítico bastante complexo.

Um dos objetivos deste trabalho é apresentar a teoria do método LASSO, de forma mais geométrica possível. Além disso, apresentaremos resultados teóricos, juntamente, com suas demonstrações, algumas da quais não apresentadas na literatura, visando a tornar o texto uma referência para os artigos básicos da teoria. O método Lasso também apresenta problemas, por exemplo, no caso em que se tem mais covariáveis do que observações, o método seleciona, no máximo, o número de covariáveis igual ao número de observações. Tem-se também que, se duas covariáveis são altamente correlacionadas, o método Lasso tem tendência a selecionar apenas uma delas, o que, eventualmente, pode ser um problema. No sentido de se obter um método que possua as propriedades do Lasso, mas que consiga minimizar suas deficiências, um novo método denominado Elastic Net é proposto por (ZOU; HASTIE,2005). É também um método baseado na penalização da soma de quadrados.

Essa penalização é obtida como uma combinação ponderada, mais precisamente, uma combinação convexa entre as penalizações do Lasso e do método Ridge. Esse novo método, além das propriedades de seleção de covariáveis, tem a vantagem de fazer seleção por grupo, isto é, possui a tendência de que covariáveis altamente correlacionadas sejam simultaneamente selecionadas. Ambos os métodos Lasso e Elastic Net não possuem forma explícita para se determinar suas estimativas. Estas são obtidas, por meio de procedimentos algorítmicos, baseados no algoritmo LARS (Least Angle Regression) proposto por Efron et

al. (2004). Esse algoritmo de seleção de modelos é uma evolução dos algoritmos clássicos Forward Stepwise e Forward Stagewise. Possui importância em si mesmo e com pequenas alterações, calcula tanto a estimativa Lasso quanto a estimativa Elastic Net, sendo a base fundamental do pacote glmnet na linguagem R, que será o pacote por nós utilizado.

Com essa linguagem geométrica e com auxílio da teoria de otimização convexa o presente trabalho pretende apresentar a teoria dos estimadores Ridge, Lasso e Elastic Net. Exibimos uma análise completa no caso ortogonal e, no caso não ortogonal do estimador Lasso, apresentamos uma fórmula que, baseada em geometria, auxilia na obtenção desse estimador. Apresentamos uma ideia de implementação da fórmula computacionalmente, tarefa esta a ser desenvolvida na sequência do trabalho. Além disso, no caso do estimador Elastic Net, utilizamos uma translação no vetor de parâmetros obtendo um resultado interessante.

# 2 REFERENCIAL TEÓRICO

## 2.1 Problemas de otimização convexa

Introduziremos, nesta seção, algumas definições e a terminologia básica da teoria de otimização convexa. Inicialmente, vamos definir alguns conceitos básicos que aparecem com frequência, ao longo do texto.

**Definição 1** Um conjunto $S \subseteq \mathbb{R}^n$ é dito convexo se o segmento de reta entre quaisquer dois pontos em $S$ está contido em $S$, isto é, se para quaisquer $x$,$y \in S$ e $\alpha \in [0,1]$, verifica-se que

$$[(1-\alpha)x + \alpha y] \in S$$

Podemos generalizar esta definição, chamando um ponto da forma $\alpha_1 x_1 + \ldots + \alpha_k x_k$, onde $\alpha_1 + \ldots + \alpha_k = 1$ e $\alpha_i \geq 0$ ; $i = 1,\ldots,k$, uma combinação convexa dos pontos $x_1,\ldots,x_k$. Podemos considerar uma combinação convexa de pontos como a média aritmética ponderada desses pontos. Pode-se mostrar que um conjunto $S$ é convexo se, e somente se, todas as combinações convexas de seus pontos está em $S$.

**Definição 2** Seja $S$ um subconjunto convexo de $\mathbb{R}^n$. Uma função $f : S \to \mathbb{R}$ é dita convexa se

$$f((1-\alpha)x + \alpha y) \leq (1-\alpha)f(x) + \alpha f(y)\,;\, \forall x,y \in S, \forall \alpha \in [0,1].$$

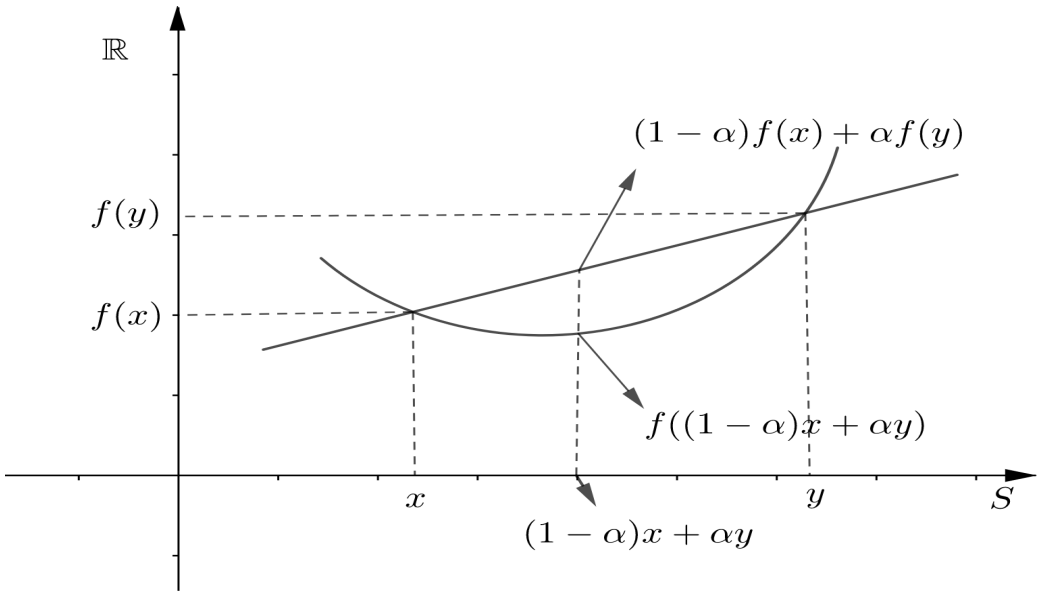
Na Figura 1, ilustra-se a definição de função convexa para $\alpha \in [0,1]$ fixado.

Nós usaremos a notação

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{restrito a} & f_i(x) \leq 0,\ i = 1,\ldots,m \\ & h_i(x) = 0,\ i = 1,\ldots,p \end{array} \tag{2.1}$$

para descrever o problema de encontrar $x$ que minimiza $f_0(x)$ dentre todos os $x$ que satisfaçam as condições $f_i(x) \leq 0,\ i = 1,\ldots\ldots,m$ e $h_i(x) = 0,\ i = 1,\ldots\ldots,p$. Nós chamaremos $x \in \mathbb{R}^n$ de variável otimizadora da função objetivo ou função custo $f_0 : \mathbb{R}^n \to \mathbb{R}$, as inequações $f_i(x) \leq 0$ são chamadas restrições de desigualdades, e as correspondentes funções $f_i : \mathbb{R}^n \to \mathbb{R}$ são chamadas de funções de restrições de desigualdades. As equações $h_i(x) = 0$ são chamadas de equações de restrição, e as funções $h_i : \mathbb{R}^n \to \mathbb{R}$ de funções de equações de restrição. Se não existem restrições (i.e., $m = p = 0$) nós dizemos que o problema (2.1) é sem restrições ou irrestrito. O conjunto de pontos em que a função objetivo e todas as restrições são definidas,

Figura 1 Representação geométrica de uma função convexa.




Fonte: Do autor (2019).

$$D = \bigcap_{i=0}^{m} dom\, f_i \cap \bigcap_{i=1}^{p} dom\, h_i$$

é chamado de domínio do problema de otimização (2.1). Um ponto $x \in D$ é factível se ele satisfaz as restrições $f_i(x) \leq 0,\ i = 1,\ldots,m$ e $h_i(x) = 0,\ i = 1,\ldots,p$. O problema (2.1) é dito ser factível se existem, pelo menos, um ponto factível e outro não factível. O conjunto de todos os pontos factíveis é chamado de conjunto factível ou conjunto de restrição.

O valor ótimo $p^*$ do problema (2.1) é definido como

$$p^* = \inf \{f_0(x)/f_i(x) \leq 0, i = 1,\ldots,m, h_i(x) = 0, i = 1,\ldots,p\}.$$

Nós permitiremos $p^*$ assumir os valores estendidos $\pm\infty$. Se o problema é não factível, nós temos $p^* = \infty$ (seguindo a convenção padrão de que o ínfimo do conjunto vazio é $\infty$). Se houver pontos factíveis $x_k$ com $f_0(x_k) \to -\infty$ quando $k \to \infty$, então $p^* = -\infty$ e dizemos, nesse caso, que o problema (2.1) é ilimitado inferiormente. Nós diremos que $x^*$ é um ponto ótimo, ou que é a solução do problema (2.1), se $x^*$ é factível e $f_0(x^*) = p^*$. O conjunto de todos os pontos ótimos é um conjunto ótimo, denotado por $X_{opt}$, ou seja

$$X_{opt} = \{x : f_i(x) \leq 0, i = 1,\ldots,m, h_i(x) = 0, i = 1,\ldots,p, f_0(x) = p^*\}.$$

Se existir um ponto ótimo para o problema (2.1), nós dizemos que o valor ótimo é alcançado ou

atingido, e o problema tem solução. Se $X_{opt}$ é vazio, nós dizemos que o valor ótimo não é alcançado ou não é atingido (isso sempre ocorre quando o problema é ilimitado inferiormente).

## 2.2 Alguns resultados importantes em otimização convexa

Nessa seção, serão apresentados algumas definições e resultados importantes em otimização convexa.

**Definição** Um problema de otimização convexa é da forma,

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{restrito a} & f_i(x) \leq 0,\ i = 1,\ldots,m \\ & h_i(x) = a'x - b_i = 0,\ \ i = 1,\ldots,p \end{array} \tag{2.2}$$

em que $f_0$,$f_1,\ldots,f_m$ são funções convexas

Observe que, nesse caso, o conjunto

$$D = \bigcap_{i=0}^{m} dom\, f_i \cap \bigcap_{i=1}^{p} dom\, h_i$$

é um conjunto convexo, e o conjunto dos pontos factíveis é a interseção de $D$ com os conjuntos $\{x\,,\, f_i(x) \leq 0\}$ e $\{x\,,\, h_i(x) = 0\}$ que são convexos e, portanto, é um conjunto convexo.

Neste texto, os problemas de otimização serão sempre considerados problemas de otimização convexa. Enunciaremos, a seguir, alguns resultados importantes em otimização convexa que vamos utilizar ao longo do texto. O primeiro resultado nos garante que a aproximação de Taylor de primeira ordem em qualquer ponto é uma cota inferior para a função em todo o seu domínio. Isso vai implicar que, se $\nabla f(x) = 0$, então, pela desigualdade $f(y) \geq f(x) + \nabla f(x)'(y - x)$, teremos que $f(y) \geq f(x), \forall y \in dom(f)$, isso é, $x$ é minimizador global de $f$. Passemos a enunciar o teorema e, na sequência, daremos uma demonstração desse resultado.

**Teorema 1 (Condições de Primeira Ordem)** Suponha $f$ uma função contínua com derivadas parciais até a primeira ordem também contínuas, isto é, $f \in C^1$. Então, $f$ é uma função convexa se, e somente se, seu domínio $dom(f)$ é um conjunto convexo e $f(y) \geq f(x) + \nabla f(x)'(y - x)$ é satisfeita $\forall x$,$y \in dom(f)$.

**Demonstração**

Vamos supor, inicialmente, que $f$ é convexa. Então para todo $\alpha \in [0,1]$,

$$f((1-\alpha)x+\alpha y) \leq (1-\alpha)f(x)+\alpha f(y).$$

Assim, para $\alpha \neq 0$,

$$\tfrac{f(x+\alpha(y-x))-f(x)}{\alpha} \leq f(y)-f(x).$$

Passando ao limite quando $\alpha \to 0$, obtemos

$$\lim_{\alpha\to 0} \tfrac{f(x+\alpha(y-x))-f(x)}{\alpha} = \partial_{y-x} f(x) = \nabla f(x)'(y-x),$$

que é a derivada direcional de $f$ na direção do vetor $y-x$. Portanto,

$$\nabla f(x)'(y-x) \leq f(y)-f(x).$$

Para demonstrar a recíproca, suponhamos agora que

$$f(y) \geq f(x)+\nabla f(x)'(y-x),\ \forall x,y \in dom(f).$$

Sejam $x_1$,$x_2 \in dom(f)$ e seja $x=\alpha x_1+(1-\alpha)x_2$, $\alpha \in [0,1]$. Note que, fazendo $y=x_1$ e $y=x_2$, obtemos, respectivamente,

$$f(x_1) \geq f(x)+\nabla f(x)'(x_1-x) \text{ e } f(x_2) \geq f(x)+\nabla f(x)'(x_2-x).$$

Multiplicando a primeira desigualdade por $\alpha$ e a segunda por $1-\alpha$, e somando as duas inequações, membro a membro, obtemos

$$\alpha f(x_1)+(1-\alpha)f(x_2) \geq f(x)+\nabla f(x)'[\alpha x_1+(1-\alpha)x_2-x].$$

Para concluir a demonstração, basta substituir $x=\alpha x_1+(1-\alpha)x_2$, donde resulta que

$$\alpha f(x_1)+(1-\alpha)f(x_2) \geq f(\alpha x_1+(1-\alpha)x_2).$$

■

Passemos, agora, a enunciar outro resultado importante que nos garante a convexidade de $f$, supondo que $f$ admita derivada de segunda ordem contínua. Ocorre que, em alguns casos, as condições, a seguir, são mais fáceis de serem verificadas na prática, apesar de serem mais restritivas do que a definição e as condições de primeira ordem. Vamos, então, enunciar e demonstrar esse resultado.

**Teorema 2 (Condições de Segunda Ordem)** Suponha $f$ uma função contínua com derivadas parciais até a segunda ordem também contínuas, isto é, $f \in C^2$, sendo $dom(f)$ um conjunto não vazio. Então $f$ é uma função convexa se, e somente, se sua matriz Hessiana é semidefinida positiva em todo o seu domínio, ou seja, $\nabla^2 f(x) \geq 0$, $\forall x \in dom(f)$.

**Demonstração**

Pelo Teorema de Taylor, temos

$$f(y) = f(x) + \nabla f(x)'(y-x) + \tfrac{1}{2}(y-x)'\nabla^2 f(x+\alpha(y-x))(y-x),$$

para algum $\alpha \in [0,1]$. Se a matriz Hessiana é semidefinida positiva, então

$$f(y) \geq f(x) + \nabla f(x)'(y-x),$$

de onde, observando-se as condições de primeira ordem, podemos concluir que $f$ é convexa.

Reciprocamente, por contrapositiva, suponhamos que a matriz Hessiana não seja positiva definida em algum $x \in dom(f)$. Então existe $y$ tal que

$$(y-x)'\nabla^2 f(x)(y-x) < 0.$$

Usando, novamente a continuidade da matriz Hessiana, podemos escolher $y$ tal que para todo $\alpha \in [0,1]$,

$$(y-x)'\nabla^2 f(x+\alpha(y-x))(y-x) < 0.$$

Desse fato e do Teorema de Taylor, tem-se que

$$f(y) \geq f(x) + \nabla f(x)'(y-x),$$

não é satisfeita. Portanto, $f$ não é convexa. E isso conclui a demonstração.

■

Vale ressaltar que esse teorema é relevante no sentido de nos fornecer a ideia geométrica de que uma função apenas é convexa se sua concavidade é voltada para cima em todo ponto.

Uma das propriedades mais importantes que os problemas de otimização convexa possuem é descrita, pelo teorema que se segue.

**Teorema 3** Se $f(x)$ é convexa então todo mínimo local do problema

$$\begin{aligned} &\text{Minimizar} \;\; f(x) \\ &\text{Sujeito a} \;\; g(x) \leq 0, \\ &\qquad\quad\;\; Ax = b, \end{aligned}$$

é um mínimo global.

**Demonstração**

De fato, seja $x^*$ um mínimo local do problema dado, ou seja, $x^*$ é factível e

$$f(x^*) = \inf\{f(z) \;\mid\; z \in D, \|z - x^*\| \leq \varepsilon\},$$

para algum $\varepsilon > 0$. Suponhamos, por contradição, que $x^*$ não é um minimizador global, isto é, existe $y$ factível tal que $f(y) < f(x^*)$. Evidentemente que $\|y - x^*\| > \varepsilon$, já que, em caso contrário, $f(x^*) \leq f(y)$. Considere o ponto $z$ dado por

$$z = (1-\theta)x^* + \theta y, \;\; \theta = \tfrac{\varepsilon}{2\|y - x^*\|}.$$

Então, temos que $\|z - x^*\| = \frac{\varepsilon}{2} < \varepsilon$, e pela convexidade do conjunto dos pontos factíveis, $z$ é factível. Agora, pela hipótese da convexidade de $f$, resulta que

$$f(z) \leq (1-\theta)f(x^*) + \theta f(y) < f(x^*),$$

e isso é uma contradição. Dessa forma, não existe $y$ factível com $f(y) < f(x^*)$, de onde temos que $x^*$ é otimizador global de $f$ em $D$.

■

Mais um resultado que vamos enunciar é um critério útil de otimalidade para função objetivo que seja diferenciável, fato este que geralmente ocorre na prática.

**Teorema 4 (Critério de Otimalidade para função objetivo diferenciável)** Para o problema de minimização com $f$ diferenciável

$$\begin{aligned} &\text{Minimizar} \;\; f(x) \\ &\text{Sujeito a} \;\; g(x) \leq 0, \\ &\qquad\quad\;\; Ax = b, \end{aligned}$$

seja $X$ seu conjunto dos pontos factíveis. Então $x^*$ é ótimo se, e somente se, $x^* \in X$ e

$$\nabla f(x^*)'(y - x^*) \geq 0, \; \forall y \in X.$$

**Demonstração**

Suponhamos, inicialmente, que $x^* \in X$ é tal que $x^*$ satisfaz a inequação

$$\nabla f(x^*)'(y - x^*) \geq 0, \, \forall y \in X.$$

Então, se $y \in X$ temos pelas condições de primeira ordem, vista no teorema 1, que $f(y) \geq f(x^*)$. Isso mostra que $x^*$ é minimizador global do problema dado.

Reciprocamente, suponhamos que $x^*$ é minimizador global de $f$ em $X$, tal que a condição

$$\nabla f(x^*)'(y - x^*) \geq 0, \, \forall y \in X,$$

não seja válida, isto é, para algum $y \in X$, temos

$$\nabla f(x^*)'(y - x^*) < 0.$$

Considere o ponto $z(t) = ty + (1-t)x^*$, sendo $t \in [0,1]$. Como $z(t)$ está no segmento de reta entre os pontos $x^*$ e $y$, e $X$ é convexo, $z(t)$ é factível. Agora, note que

$$\tfrac{d}{dt} f(z(t)) \,|_{t=0} = \nabla f(x^*)'(y - x^*) < 0,$$

o que mostra que para algum $t$ suficientemente pequeno, temos que $f(z(t)) < f(x^*)$, o que contradiz a hipótese. Logo, um minimizador $x^*$ deve satisfazer a inequação

$$\nabla f(x^*)(y - x^*) \geq 0, \, \forall y \in X.$$

Isso completa a demonstração do teorema.

■

O teorema acima apresenta a seguinte justificativa geométrica que nos diz que um ponto $x^*$ é minimizador local de $f$ em $X$ se, e somente se, não existem direções de decrescimento factíveis em $X$ a partir de $x^*$, ou seja, não existe $v \in \mathbb{R}^n$ tal que $f(x^* + v) < f(x^*)$, em que $x^* + v \in X$.

## 2.3 Dualidade

Nesta seção, estudaremos alguns tópicos que envolvem a teoria de dualidade, assunto central na área de otimização convexa, dentre eles os multiplicadores de Lagrange.

Considere o problema geral de otimização convexa em sua forma padrão,

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{restrito a} & f_i(x) \leq 0,\ i = 1,\ldots,m \\ & h_i(x) = 0,\ i = 1,\ldots,p \end{array} \tag{2.3}$$

com a variável $x \in \mathbb{R}^n$, sendo $f_0 : \mathbb{R}^n \to \mathbb{R}$, $f = (f_1,\ldots,f_m) : \mathbb{R}^n \to \mathbb{R}^m$ e $h = (h_1,\ldots,h_p) : \mathbb{R}^n \to \mathbb{R}^p$. De agora em diante, chamaremos esse problema de problema primal ou problema original. Nós assumiremos que o conjunto

$$D = \bigcap_{i=0}^{m} dom\, f_i \cap \bigcap_{i=1}^{p} dom\, h_i\,,$$

é não vazio, e denotaremos o valor ótimo de (2.3) por $p^*$, que é único. A ideia básica da dualidade lagrangeana é levar em conta as restrições, aumentando a função objetivo com a soma ponderada das restrições. Definimos a função lagrangeana $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ associada ao problema (2.3) como

$$L\,(x,\lambda,\nu) = f_0\,(x) + \sum_{i=1}^{m} \lambda_i f_i\,(x) + \sum_{i=1}^{p} \nu_i h_i\,(x), \tag{2.4}$$

com domínio $D \times \mathbb{R}^m \times \mathbb{R}^p$, em que $D \subseteq \mathbb{R}^n$ e $\lambda_i$ e $\nu_i$ são chamados multiplicadores de Lagrange associados com a *i*-ésima restrição de desigualdade $f_i\,(x) \leq 0$ e a *i*-ésima restrição de igualdade $h_i\,(x) = 0$, respectivamente. Os vetores $\lambda$ e $\nu$ são chamados de variáveis duais ou vetores multiplicadores de Lagrange associados ao problema (2.3).

## 2.4 Teoria geral da estimação linear

Considere o modelo de regressão linear $y = X\beta + \varepsilon$, em que $X$ é uma matriz, $n \times p$, definida pelos valores das covariáveis, com posto $r(X) = p = \min(n,p)$ e $\beta$ é um vetor $p$-dimensional de parâmetros, já considerando o modelo centralizado. A matriz $X$ pode ser vista como uma transformação linear do espaço de parâmetros $\mathbb{R}^p$ no espaço de dados $\mathbb{R}^n$, $X : \mathbb{R}^p \to \mathbb{R}^n$ (FIGURA 2). Como $\mu = E\,[y] = X\beta$ , o vetor de médias está contido no subespaço imagem de $X$ ($\mathrm{Im}\,(X)$). Portanto, o subespaço $\mathrm{Im}\,(X)$ é o nosso modelo estatístico.

De uma maneira mais geral, pode-se considerar qualquer subconjunto $K \subseteq \mathbb{R}^n$ como modelo estatístico, isto é, onde se supõe estar o vetor de médias $\mu = E\,[y]$. Uma boa opção para a escolha do subconjunto

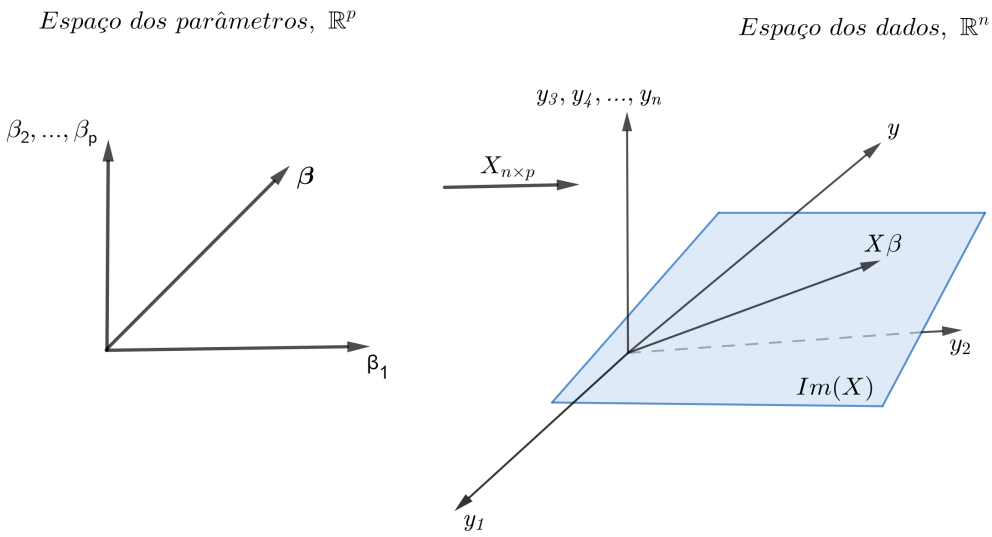
$K$ para que esse tenha propriedades matemáticas adequadas é que $K$ seja convexo e fechado. Uma vez observado o vetor de dados $y$ uma estimativa de $\mu$ é obtida simplesmente projetando-se o vetor $y$ em $K$. A projeção mais utilizada é a que minimiza a distância de $y$ a $K$, isto é, a de quadrados mínimos.

Seja $K$ um subconjunto de $\mathbb{R}^n$ onde, por hipótese, será suposto estar o vetor $\mu$. Portanto, nesse sentido, o subconjunto $K$ é denominado como sendo o modelo estatístico. Uma vez observado o vetor $y$ a estimativa de $\mu$ será obtida simplesmente projetando $y$ no subconjunto $K$.

Dessa forma, o modelo de regressão definido pelo convexo $K$ é justamente esse subespaço. Observe que este é o caso mais simples possível, uma vez que estaremos considerando $K$ fechado e convexo. Mais simples ainda, se tem que a projeção $\mu : \mathbb{R}^n \to \mathrm{Im}\,(X) \subseteq \mathbb{R}^n$ é uma projeção linear ortogonal. Dessa forma, a expressão de $\mu(y)$ pode ser obtida de forma explícita para o caso de X ser injetiva, como a fórmula matricial $\mu\,(y) = X(X'\,X)^{-1}X'\,y$ e $\hat{\beta}_{\mathrm{ols}} = (X'\,X)^{-1}X'\,y$ (ols, do inglês "ordinary least square").

Para expressar o problema de minimizar a distância do vetor de dados $y$ a esse subespaço (Figura 2), primeiramente, vamos expressá-lo como a solução de um sistema homogêneo de equações lineares, isto é, obter a caracterização da $\mathrm{Im}(X)$ como solução de $Ax = 0$. Utilizando o fato de que a $\mathrm{Im}(X)$ é perpendicular ao $\ker X'$, a matriz A mais adequada será a projeção ortogonal no espaço $\ker X'$. Como a projeção na $\mathrm{Im}X$ é dada por $X(XX')^{-1}X'$, tem-se a expressão para o projetor A como $A = I - X(XX')^{-1}X'$.

Figura 2 Representação geométrica do estimador de mínimos quadrados.




Fonte: Do autor (2019).

## 2.5 O estimador de mínimos quadrados

Nessa subseção, vamos caracterizar o estimador de mínimos quadrados sob a ótica da dualidade. O problema de mínimos quadrados no espaço de dados pode ser expresso como:

$$\begin{array}{ll} \min & f_0(x) = \|y - x\|^2 \\ \text{restrito} & Ax = 0 \end{array}. \tag{2.5}$$

A lagrangeana para o estimador de quadrados mínimos é dado por:

$$L\left(x,\nu\right) = \|y - x\|^2 + \nu' Ax = (y - x)' I\left(y - x\right) + \nu' Ax. \tag{2.6}$$

A função dual de Lagrange, neste caso específico, é a função $g : \mathbb{R}^p \to \mathbb{R}$ definida como o valor mínimo da função lagrangeana sobre $x$, para $\nu \in \mathbb{R}^p$, sendo dada por:

$$g\left(\nu\right) = \inf_{x \in \mathbb{R}^n} L\left(x,\nu\right) = \inf_{x \in \mathbb{R}^n} \left((y - x)' I\left(y - x\right) + \nu' Ax\right). \tag{2.7}$$

obtida por,

$$\nabla_x L\left(x,\nu\right) = -2\left(y - x\right) + A'\nu = 0. \tag{2.8}$$

Assim,

$$2\left(y - x\right) = A'\nu. \tag{2.9}$$

Resolvendo em relação a $x$ tem-se que:

$$x = y - \frac{1}{2} A'\nu. \tag{2.10}$$

Substituindo (2.10) em (2.7) tem-se que:

$$\begin{aligned} g\left(\nu\right) &= \left\| y - y + \frac{1}{2} A'\nu \right\|^2 + \nu' A \left( y - \frac{1}{2} A'\nu \right) \\ &= \frac{1}{4} \left\| A'\nu \right\|^2 + \nu' Ay - \frac{1}{2} \nu' AA'\nu \\ &= -\frac{1}{4} \nu' AA'\nu + \nu' Ay. \end{aligned} \tag{2.11}$$

A dualidade fraca estabelece que:

$$\max_{\nu} g\left(\nu\right) \leq f_0\left(x_{opt}\right). \tag{2.12}$$

para se obter o máximo de $g(v)$ tem-se,

$$\nabla_{\nu} g\left(\nu\right) = -\frac{1}{4} 2AA'\nu + Ay = 0. \tag{2.13}$$

Resolvendo em relação a $\nu$ tem-se que:

$$\nu = 2\big(AA'\big)^{-1} Ay. \tag{2.14}$$

Substituindo (2.14) em (2.11) tem-se que:

$$\begin{aligned} g\left(2\big(AA'\big)^{-1}Ay\right) &= -\frac{1}{4}\Big(2\big(AA'\big)^{-1}Ay\Big)' AA' \left(2\big(AA'\big)^{-1}Ay\right) + \Big(2\big(AA'\big)^{-1}Ay\Big)' Ay \\ &= -\frac{1}{4} 2y'A'\big(AA'\big)^{-1} AA' 2\big(AA'\big)^{-1} Ay + 2y'A'\big(AA'\big)^{-1} Ay \\ &= -y'A'\big(AA'\big)^{-1} Ay + 2y'A'\big(AA'\big)^{-1} Ay \\ &= y'A'\big(AA'\big)^{-1} Ay. \end{aligned} \tag{2.15}$$

Para se obter o valor ótimo de $x$, substituindo (2.14) em (2.10) tem-se que:

$$\begin{aligned} x_{opt} &= y - \frac{1}{2} A'\nu \\ &= y - \frac{1}{2} 2A'\big(AA'\big)^{-1} Ay \\ &= y - A'\big(AA'\big)^{-1} Ay. \end{aligned} \tag{2.16}$$

Portanto, vamos caracterizar o subespaço $\mathrm{Im}(X)$ como solução de um sistema linear homogêneo $Ax = 0$.

A transformação $X(X'X)^{-1}X'$ é a projeção ortogonal, em $\mathrm{Im}(X)$ e portanto $I - X(X'X)^{-1}X'$ é a projeção ortogonal no complemento ortogonal do subespaço $\mathrm{Im}(X)$. Os vetores de $\mathrm{Im}(X)$ são justamente os vetores que se projetam no ortogonal de $\mathrm{Im}(X)$ no vetor nulo. Portanto, $\mathrm{Im}(X)$ fica caracterizado como solução do sistema homogêneo $Ax = 0$ , em que $A = I - X(X'X)^{-1}X'$.

Portanto

$$x_{opt} = y - X\left(X'X\right)^{-1}X'y. \tag{2.17}$$

$$\text{Valor ótimo} = \left\| X\left(X'X\right)^{-1}X'y \right\|^2. \tag{2.18}$$

Portanto, o problema de quadrados mínimos no espaço de parâmetros é expresso como:

$$\min \ f_0(\beta) = \|X\beta - y\|^2, \ \beta \in \mathbb{R}^p. \tag{2.19}$$

Temos que:

$$\begin{aligned} \|X\beta - y\|^2 &= (X\beta - y)'(X\beta - y) \\ &= \beta'X'X\beta - \beta'X'y - y'X\beta + y'y \\ &= \beta'X'X\beta - 2\beta'X'y + y'y. \end{aligned} \tag{2.20}$$

Derivando em relação a $\beta$ e igualando a derivada a zero vem que:

$$2X'X\hat{\beta} - 2X'y = 0. \tag{2.21}$$

Resolvendo em relação $\hat{\beta}$, obtém-se o estimador de mínimos quadrados para $\beta$, que é dado por:

$$\hat{\beta}_{ols} = \left(X'X\right)^{-1}X'y, \tag{2.22}$$

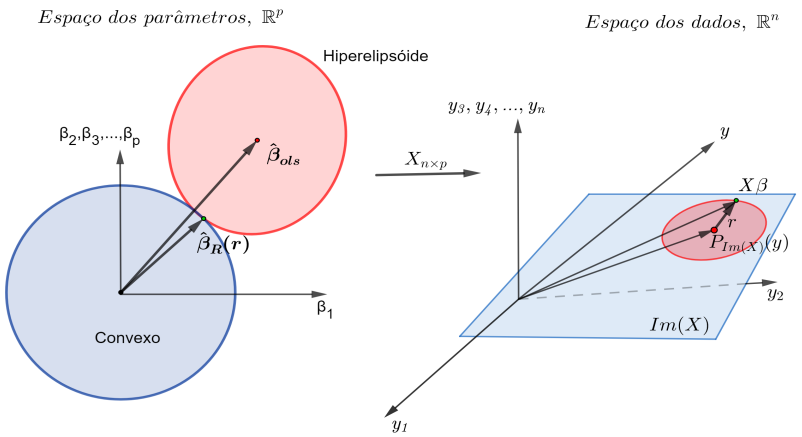
desde que a inversa de $X'X$ exista.

## 2.6 Regressão Ridge

Se o estimador de quadrados mínimos, $\hat{\beta}_{ols}$, em razão de presença de quasimulticolinearidade, gera estimativas com excessiva variância, uma construção para se contornar essa dificuldade, é a seguinte: vamos tomar uma esfera de raio $r$ em $\mathrm{Im}(X)$ centrada em $X\hat{\beta}_{ols}$. Todos os pontos dessa esfera estão à mesma distância do vetor de dados $y$. Pela filosofia dos quadrados mínimos todos eles geram estimadores igualmente plausíveis. Um procedimento usual na estatística é a abordagem conservadora de, entre duas estimativas igualmente plausíveis, se optar pela de menor tamanho ou de menor norma. Deve-se, então, escolher na

elipse, imagem inversa desse círculo, centrada em $\hat{\beta}_{ols}$, o vetor $\beta$ de menor norma, que será denominado $\hat{\beta}_R$ (FIGURA 3). Esse procedimento de estimação é denominado estimação Ridge (HOERL; KENNARD, 1970).

Figura 3 Representação geométrica referente à obtenção do estimador Ridge.




Fonte: Do autor (2019).

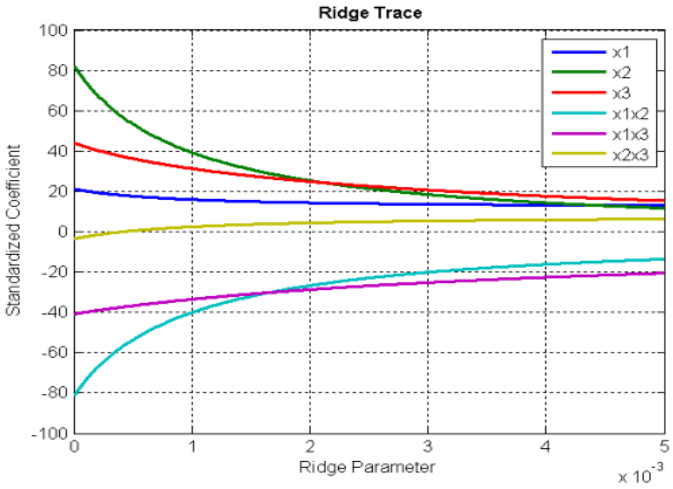
Dessa forma, o estimador Ridge $\hat{\beta}_R(r)$ é um estimador de encolhimento no sentido que $\left\|\hat{\beta}_R(r)\right\| < \left\|\hat{\beta}_{ols}(r)\right\|$ e também, evidentemente, um estimador viesado.

Uma das grandes vantagens do método de estimação Ridge é que, variando-se o parâmetro $r$ obtém-se uma curva $\hat{\beta}_R(r)$ no espaço de parâmetros. Tomando-se as coordenadas $\left(\hat{\beta}_R\right)_i(r)$ de $\hat{\beta}_R(r)$, é possível uma descrição bidimensional do comportamento de cada componente do estimador, traçando-se, simultaneamente, os gráficos de $\left(\hat{\beta}_R\right)_i(r)$ em função do parâmetro $r$. Tais gráficos são denominados *ridgetraces*, e permitem uma análise gráfica bastante útil do processo de estimação. Na Figura 4, apresenta-se um desses *ridgetraces*.

A questão de se determinar um valor ótimo para o parâmetro r, isto é, o quanto se deve ter de encolhimento para que seja mínimo o erro quadrático médio $E_\beta\left[\left\|\hat{\beta}_R-\beta\right\|^2\right]$, é uma das questões mais estudadas na teoria, envolvendo problemas analíticos complexos. Tal valor depende do vetor de parâmetro populacional $\beta$, que deve ser estimado obtendo-se estimadores bastante complexos do valor ótimo para $r$.

Uma forma gráfica de se estimar uma região para o valor ótimo de r é o intervalo em que as curvas de $ridgetrace$ se tornam, aproximadamente, paralelas ao eixo das abscissas.

Figura 4 Exemplo de $ridgetrace$ em $\mathbb{R}^6$.




Fonte: Tibishirani (2013).

Nessa subseção, será explicitada a obtenção do estimador $\hat{\beta}_R$. A ideia é a seguinte: os vetores $z$ que pertencem a essa esfera satisfazem a condição $\left\langle z - P_{\mathrm{Im}(X)}(y), z - P_{\mathrm{Im}(X)}(y) \right\rangle = r^2$. Considere, então, os vetores $\beta$ no espaço paramétrico, tais que $X\beta = z$. Então:

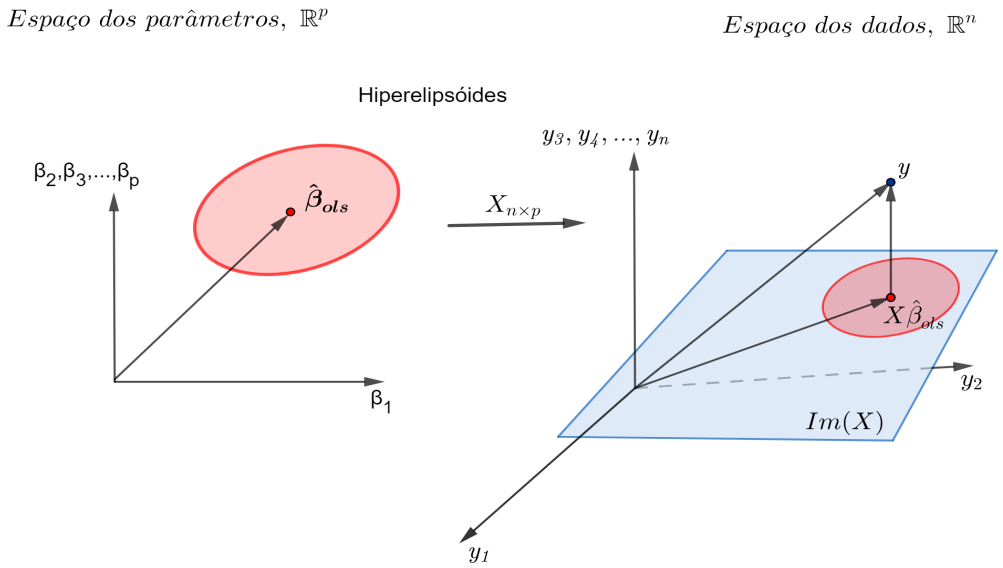
$$
\begin{aligned}
r^2 &= \left\langle z - P_{\mathrm{Im}(X)}(y), z - P_{\mathrm{Im}(X)}(y) \right\rangle \\
&= \left\langle X\beta - X\hat{\beta}_{ols}, X\beta - X\hat{\beta}_{ols} \right\rangle \\
&= \left\langle X\left(\beta - \hat{\beta}_{ols}\right), X\left(\beta - \hat{\beta}_{ols}\right) \right\rangle \\
&= \left[X\left(\beta - \hat{\beta}_{ols}\right)\right]' X\left(\beta - \hat{\beta}_{ols}\right) \\
&= \left(\beta - \hat{\beta}_{ols}\right)' X'X\left(\beta - \hat{\beta}_{ols}\right),
\end{aligned}
$$

de onde segue que a pré-imagem da esfera é um elipsoide centrado em $\hat{\beta}_{ols}$ no espaço paramétrico $\mathbb{R}^p$, conforme Figura 5.

Adota-se, então, uma atitude conservadora. Todos os $\beta$, nesse elipsóide, são estimativas viáveis. Opta-se, então, por aquela de menor norma, isto é, toma-se o $\beta$ obtido como tangente entre o elipsoide e uma esfera centrada na origem. A dedução analítica da expressão desse estimador é apresentada a seguir e é representada geometricamente pela Figura 6.

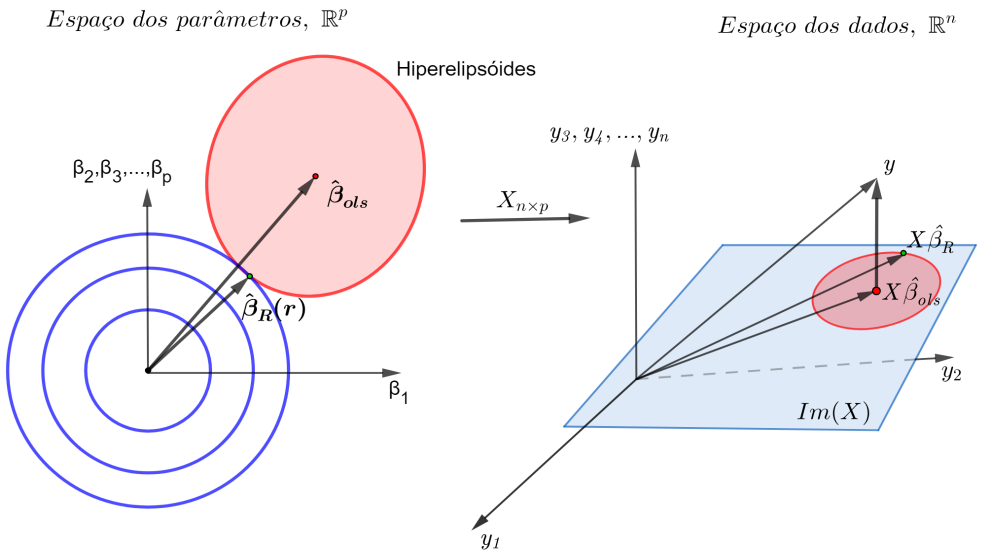
Para cada valor fixo de $r$, a estimativa $\hat{\beta}_R(r)$ é obtida como um problema de minimização. Ana-

Figura 5 Penalização na definição do estimador Ridge.




Fonte: Do autor (2019).

Figura 6 A geometria do estimador Ridge.




Fonte: Do autor (2019).

liticamente, esse problema pode ser descrito de duas formas equivalentes. Primeiramente no espaço de parâmetros.

Se quer minimizar a função $\min_{\beta} \|\beta\|^2$ sujeito à restrição

$$\left(\beta - \hat{\beta}_{ols}\right)' X'X \left(\beta - \hat{\beta}_{ols}\right) = r^2.$$

Utilizando o método dos multiplicadores de Lagrange, a Lagrangeana para esse problema é:

$$\begin{aligned} L(\beta,\lambda) &= \|\beta\|^2 + \lambda \left( \left(\beta - \hat{\beta}_{ols}\right)' X'X \left(\beta - \hat{\beta}_{ols}\right) - r^2 \right) \\ &= \beta'\beta + \lambda \left( \left(\beta - \hat{\beta}_{ols}\right)' X'X \left(\beta - \hat{\beta}_{ols}\right) \right) - \lambda r^2. \end{aligned}$$

Derivando em relação aos parâmetros e igualando à 0 tem-se

$$\begin{cases} \frac{\partial L}{\partial \beta} = 2\beta + \lambda \left[ 2X'X \left(\beta - \hat{\beta}_{ols}\right) \right] = 0 \\ \frac{\partial L}{\partial \lambda} = \left(\beta - \hat{\beta}_{ols}\right)' X'X \left(\beta - \hat{\beta}_{ols}\right) - r^2 = 0 \end{cases}.$$

Logo,

$$\beta + \lambda X'X\beta = \lambda X'X\hat{\beta}_{ols} \Rightarrow \left(I + \lambda X'X\right)\beta = \lambda X'X\hat{\beta}_{ols}.$$

Portanto, a solução é dada explicitamente por

$$\hat{\beta}_R(\lambda) = \left(I + \lambda X'X\right)^{-1} \lambda X'X\hat{\beta}_{ols} = \left(\frac{1}{\lambda} I + X'X\right)^{-1} X'X\hat{\beta}_{ols}.$$

Uma vez que $\hat{\beta}_{ols} = (X'X)^{-1}X'y$, então

$$\hat{\beta}_R(k) = \left(\frac{1}{\lambda} I + X'X\right)^{-1} X'y = \left(kI + X'X\right)^{-1} X'y,$$

em que $k = \frac{1}{\lambda}$.

O valor de $k$ em função de $r$ é obtido substituindo-se $\hat{\beta}_R(r)$ na restrição

$$\left(\hat{\beta}_R(r) - \hat{\beta}\right)' X'X \left(\hat{\beta}_R(r) - \hat{\beta}\right) = r^2.$$

$$\left(\left(kI + X'X\right)^{-1} X'y - \left(X'X\right)^{-1} X'y\right)' X'X \left(\left(kI + X'X\right)^{-1} X'y - \left(X'X\right)^{-1} X'y\right) = r^2.$$

Como é fácil atribuir um valor para $k$ e obter o valor correspondente de $r$, geralmente o estimador Ridge é expresso em função de $k$ no lugar de expressá-lo em função de $r$.

$$\hat{\beta}_R(k) = \left(kI + X'X\right)^{-1} X'y$$

Tal substituição é adequada, porém $k$ não tem um significado geométrico como o tem $r$.

O problema variacional que define o estimador Ridge admite outra forma equivalente, no espaço de dados. Considere $\beta$ sobre uma esfera de raio $r$, isto é, $\beta'\beta = r^2$. A imagem dessa esfera pela transformação $X$ é uma elipse.

A ideia agora é obter $\beta$ tal que $X\beta$ esteja o mais próximo possível de $y$, isto é,

$$\min_{\beta} \|y - X\beta\|^2$$

restrito à elipse, correspondente à imagem da esfera $\|\beta\|^2 = r^2$. A Lagrangeana desse problema é:

$$\begin{aligned} L(\beta,\lambda) &= \|y - X\beta\|^2 + \lambda\left(\beta'\beta - r^2\right) \\ &= (y - X\beta)'(y - X\beta) + \lambda\left(\beta'\beta - r^2\right) \\ &= y'y - y'X\beta - (X\beta)'y + (X\beta)'X\beta + \lambda\left(\beta'\beta - r^2\right) \\ &= y'y - 2\beta'X'y + \beta'X'X\beta + \lambda\left(\beta'\beta - r^2\right). \end{aligned}$$

Derivando-se $L(\beta,\lambda)$ em relação a $\beta$ e igualando-se a derivada a zero:

$$\frac{\partial L}{\partial \beta} = 2y'X + 2X'X\beta + \lambda(2\beta) = 0.$$

Assim,

$$-y'X + \left(X'X + \lambda I\right)'\beta = 0.$$

Resolvendo para $\beta$ tem-se a solução Ridge:

$$\hat{\beta}_R(r) = \left(\lambda I + X'X\right)^{-1} X'y.$$

No espaço de dados, tem-se a interpretação

$$\|y - X\beta\|^2 = \left\|y - X\hat{\beta}_{ols}\right\|^2 + \left\|X\beta - X\hat{\beta}_{ols}\right\|^2 = \text{cte} + \left\|X\hat{\beta}_{ols} - X\beta\right\|^2.$$

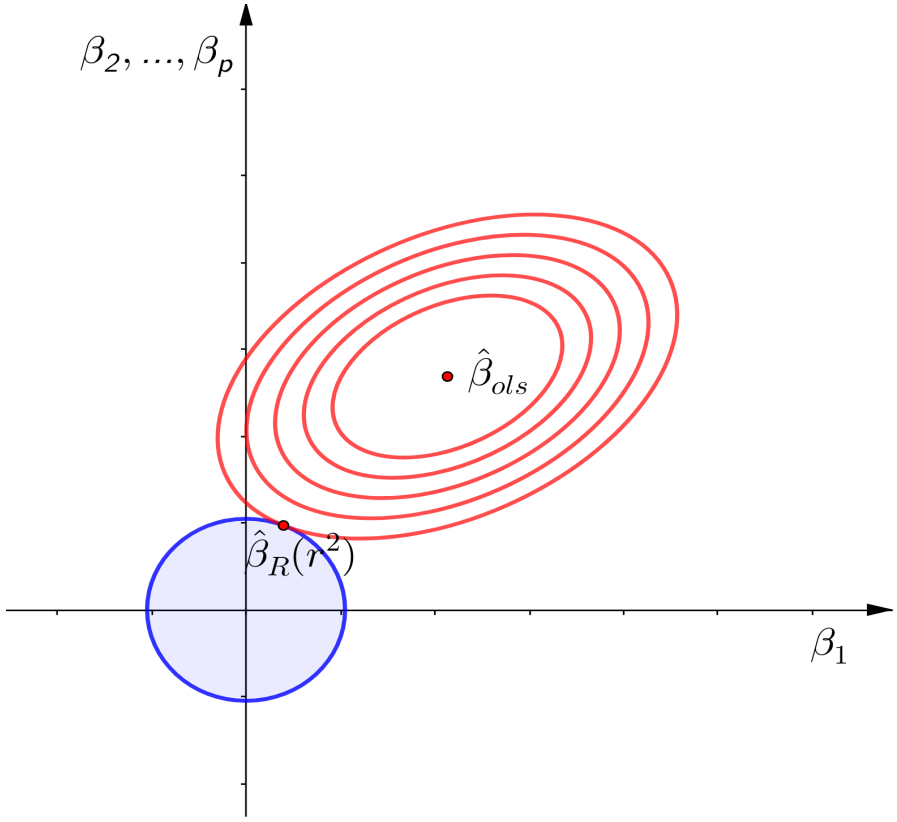
O que se quer, então, é a minimização

$$\min_{\beta} \left\|X\beta - X\hat{\beta}\right\|^2 = \min_{\beta} \left(\beta - \hat{\beta}\right)' X'X \left(\beta - \hat{\beta}\right),$$

sujeito à restrição $\beta'\beta = r^2$, isto é, para os vetores $\beta$ sobre a esfera obter aquele na elipse centrada em $\hat{\beta}_{ols}$ de menor tamanho pela métrica de Mahalanobis,$\langle u,v\rangle = u'X'Xv$, $\forall u,v \in \mathbb{R}^p$ (FIGURA 7). Tem-se, portanto, uma reformulação dual à primeiramente utilizada.

.

Figura 7 Projeção definida pela métrica de Mahalanobis.




Fonte: Do autor (2019).

## 2.7 Regressão Lasso

Em regressão linear, além do problema de estimação do vetor de parâmetros outro problema consiste na seleção de covariáveis. Na presença de muitas covariáveis, é razoável que o pesquisador queira selecionar apenas algumas delas que mais afetam a variável resposta, isto é, um modelo mais parcimonioso. Uma possibilidade é a de se obter $\hat{\beta}_{ols}$ e eliminar as covariáveis relativas às coordenadas de $\hat{\beta}_{ols}$ de valores relativamente muito menores que as outras. Se o estimador de quadrados mínimos apresentar alta variabilidade (quase multicolinearidade), esse procedimento evidentemente apresenta problemas e é conhecido que não é estável, ou seja, se um novo vetor de respostas $y$ é observado, com alta probabilidade, covariáveis diferentes serão selecionadas.

Uma alternativa seria a de se obter um processo de estimação que, além de apresentar pouca variabilidade, geralmente obtida por algum processo de encolhimento, gere, com alta probabilidade, estimativas $\hat{\beta}$ em que várias de suas componentes sejam nulas. Dessa forma, teria-se um processo automático de seleção de covariáveis. Essa é de fato a proposta de um processo de estimação denominado Lasso, acrônimo de *Least absolute shrinkage and selection operator*, proposto por Tibshirani (1996).

Consideremos, então, os dados $(x_i, y_i)$, com $i = 1, \ldots, n$ para $x_i = (x_{i1}, \ldots, x_{ip})$ sendo $x_i$ o vetor linha da matriz de delineamento, dado pelos valores das variáveis preditoras (covariáveis) e $y_i$ as respostas observadas. Se $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$, o estimador Lasso é definido por:

$$(\alpha,\beta) = min \left\{ \sum_{i=1}^{n} \left( y_i - \alpha - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\},$$

sujeito à restrição $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j| \leq t$.

Considerando o modelo em sua forma centralizada, conforme (RENCHER, A. C., SHAALJE, G. B.,2008) temos que $\hat{\alpha} = \bar{y}$. É adequado também, por uma transformação de dados, colocar o problema na forma normalizada $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0, \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$ e $\frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} = 0$. Dessa forma, o estimador pode ser escrito como

$$\hat{\beta}_{lasso} = \min \|y - X\beta\|^2 = \min_{\beta} \left( \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right),$$

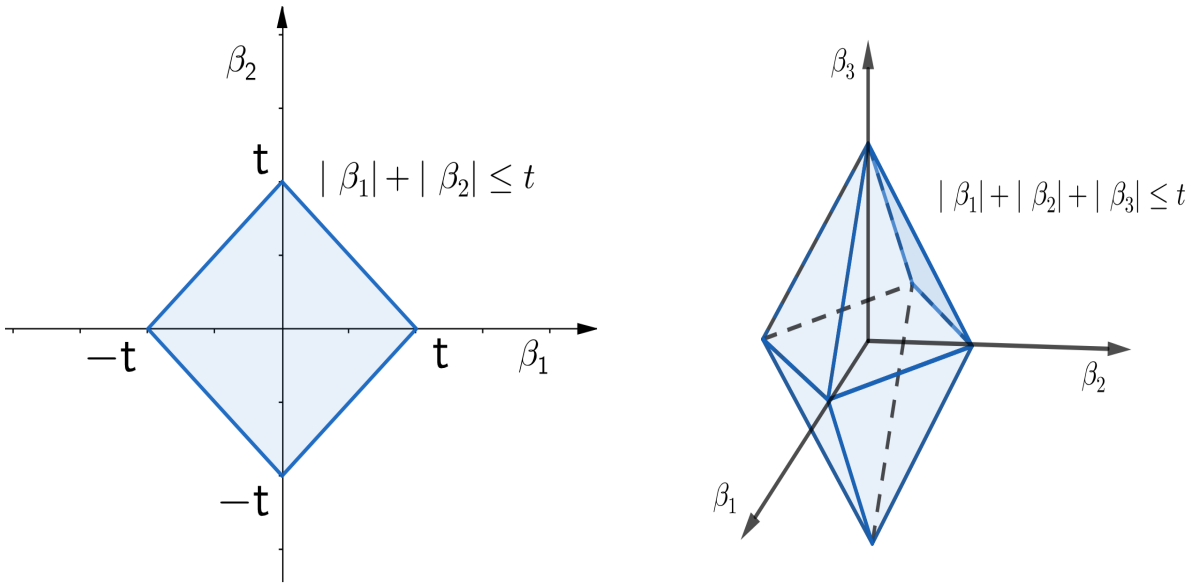
ou também pode ser interpretado como $\min \|y - X\beta\|^2$ sujeito a restrição $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j| \leq t$.

Essa última forma, fica representada na forma Lagrangeana por:

$$L(\beta,\lambda) = \|y - X\beta\|^2 + \lambda(\|\beta\|_1 - t).$$

Em termos geométricos, o subconjunto fechado e convexo $K_p$, definido por $\sum_{j=1}^{p} |\beta_j| \leq t$, é um hiperpirâmide, cujas diagonais estão sobre os eixos coordenados e o centro sobre a origem. O valor $t \geq 0$ pode ser considerado como um parâmetro de ajuste. Para o caso bidimensional e tridimensional, $K_p$ é como se segue na Figura 8.

Figura 8 Restrição $K_p$ Lasso.




(a) Caso bidimensional. (b) Caso tridimensional.

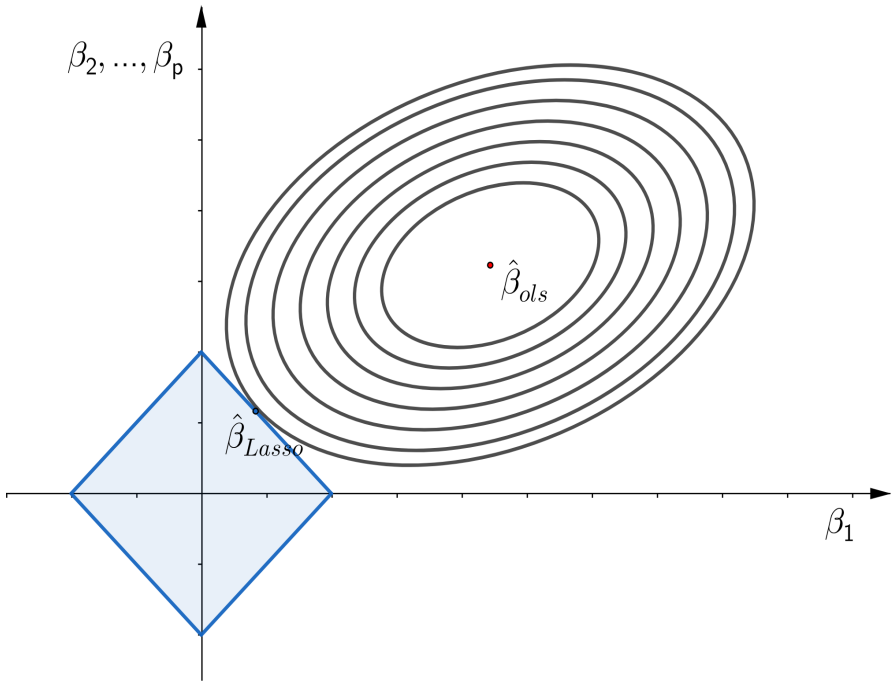
Fonte: Do autor (2019).

No espaço de dados $\mathbb{R}^n$ , o convexo $K = X\,(K_p)$ é semelhante a $K_p$. O estimador Lasso, $\hat{\beta}_{lasso}$, é ,então, obtido da forma anteriormente descrita. Obtém-se, sobre o convexo $K$, o ponto mais próximo do vetor de dados $y$, o qual será denominado $P_K\,(y)$. Dessa forma, $X\left(\hat{\beta}_{lasso}\right) = P_K\,(y)$. O ponto $P_K\,(y)$ também pode ser obtido de maneira geométrica fazendo uso do triângulo fundamental, ilustrado mais adiante, na Figura 12, como o ponto de $K$ mais próximo de $X\left(\hat{\beta}_{ols}\right) = X(X'X)^{-1}X'y$.

A obtenção de $\hat{\beta}_{lasso}$ também pode ser descrita no espaço paramétrico, uma vez que minimizar a distância do vetor $y$ ao convexo $K$ é, como também descrito anteriormente, equivalente a se minimizar a distância da estimativa de mínimos quadrados $\hat{\beta}_{ols}$ ao convexo $K_p$. Entretanto, como essa distância é definida pela métrica $\langle v_1,v_2\rangle_p = v'_1X'Xv_2,\ v_1,v_2 \in \mathbb{R}^p$, a obtenção do ponto de $K_p$ mais próximo de $\hat{\beta}_{ols}$ pode ser descrita como o primeiro ponto de $K_p$ que tangencia um elipsoide centrado em $\hat{\beta}_{ols}$, conforme

Figura 9.

Figura 9 Obtenção do estimador Lasso.




Fonte: Do autor (2019).

O conjunto convexo $K_p$ impõe duas restrições que possuem justificativas estatísticas. A primeira delas é a de manter os valores das estimativas $\left(\hat{\beta}_{lasso}\right)_i$ limitadas. Essa é uma posição conservadora, para se evitar estimativas excessivas. A segunda justificativa é bem mais sofisticada. O convexo $K_p$ tem como borda uma hipersuperfície, formada de planos tais que essas se intersectam em superfícies formadas de planos de dimensões menores, que podemos pensar como "arestas". Como as dimensões das arestas variam, ocorre que os pontos nas arestas possuem algumas coordenadas nulas. É intuitivo pensar que essas arestas têm uma maior probabilidade de conterem o ponto mais próximo de $\hat{\beta}_{ols}$. Portanto, o estimador $\hat{\beta}_{lasso}$ tem uma maior probabilidade de possuir algumas coordenadas nulas. Nesse sentido, o estimador funciona também como um processo de seleção de covariáveis e é essa propriedade que mais impulsionou seu estudo e a sua utilização.

De forma análoga ao estimador Ridge, para o qual se pode traçar o *ridgetrace* para as componentes do vetor de estimativas, considerando como parâmetro de ajuste o valor $t$ da definição do estimador Lasso, é possível traçar curvas para as componentes da estimativa Lasso. Um aspecto interessante para essas curvas é que se obtém a sequência ordenada em que as covariáveis se anulam. Dessa forma, tem-se uma descrição gráfica para a seleção das covariáveis (TIBSHIRANI, 1996).

## 2.8 Regressão Elastic Net

Esse método de regressão, pretende obter o melhor entre os dois métodos de regressão, Ridge e Lasso, porém com novas propriedades de que esses métodos carecem. A ideia, nessa metodologia, é simples, consistindo em minimizar a soma de quadrados do resíduo restrito a uma combinação linear das restrições dos métodos Ridge e Lasso.

Para uma regressão múltipla $y = X\beta + \epsilon$, considerando o modelo centrado e as covariáveis na forma normalizada, isto é, $\sum_{i=1}^{n} y_i = 0$, $\sum_{i=1}^{n} x_{ij} = 0$, $\sum_{i=1}^{n} x_{ij}^2 = 1$ com $j = 1, \ldots, p$, a penalização da soma de quadrados é dada por uma combinação convexa da penalização Ridge com a penalização Lasso:

$$\eta_2 \|\beta\|^2 + \eta_1 \|\beta\|_1 \leq t,$$

em que $\eta_1, \eta_2 \geq 0$ e $\|\beta\|^2 = \sum_{j=1}^{p} \beta_j^2$ , $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$. Portanto, a função a ser minimizada é dada por $L(\beta,\lambda,\eta_1,\eta_2) = \|y - X\beta\|^2 + \lambda \left[\eta_2 \|\beta\|^2 + \eta_1 \|\beta\|_1 - t\right]$. O estimador obtido será denominado estimador *Elastic Net Ingênuo*, representado por

$$\hat{\beta}_{eni} = \arg\min_{\beta} L(\beta,\lambda).$$

Em termos de soma de quadrados restrita (penalizada), minimizar a função $L(\beta,\lambda,\eta_1,\eta_2)$ é equivalente a minimizar $\|y - X\beta\|$ sujeito à restrição

$$\begin{aligned}
\eta_1 \|\beta\|_1 + \eta_2 \|\beta\|^2 \leq t &\Rightarrow \frac{\eta_1}{\eta_1 + \eta_2} \|\beta\|_1 + \frac{\eta_2}{\eta_1 + \eta_2} \|\beta\|^2 \leq \frac{t}{\eta_1 + \eta_2} \\
&\Rightarrow \left(1 - \frac{\eta_1}{\eta_1 + \eta_2}\right) \|\beta\|^2 + \frac{\eta_1}{\eta_1 + \eta_2} \|\beta\|_1 \leq \frac{t}{\eta_1 + \eta_2} \\
&\Rightarrow (1 - \alpha) \|\beta\|^2 + \alpha \|\beta\|_1 \leq t',
\end{aligned}$$

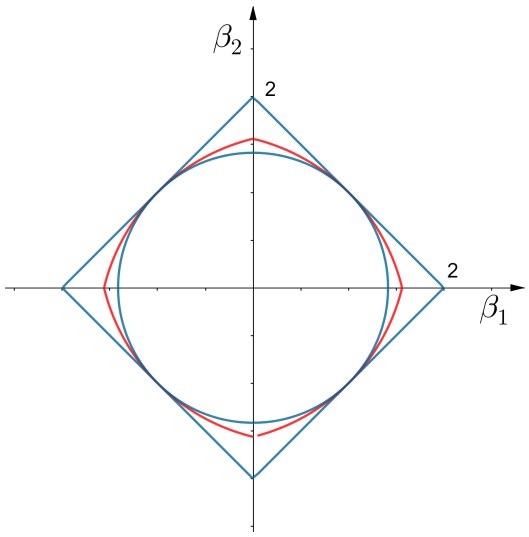
em que, $0 \leq \alpha \leq 1$. A penalidade $(1 - \alpha) \|\beta\|^2 + \alpha \|\beta\|_1$ é denominada de penalidade Elastic Net e é uma combinação convexa entre as penalidades que definem a estimação Ridge e Lasso. Pode-se observar que para $\alpha = 0$ se tem a estimação Ridge e se $\alpha = 1$ a estimação resultante é a do Lasso.

O conjunto $\|\beta\|_1 \leq t$ é convexo, mas não estritamente convexo. Já o conjunto $\|\beta\|^2 \leq t$ é estritamente convexo, e portanto, para $\alpha \neq 1$, o conjunto definido por $(1 - \alpha) \|\beta\|^2 + \alpha \|\beta\|_1$ é estritamente convexo. Logo, o modelo $K_p$ é o convexo

$$K_p = \left\{ \beta \in \mathbb{R}^p, (1-\alpha) \left\| \beta \right\|^2 + \alpha \| \beta \|_1 \leq t \right\}.$$

Para analisar a forma desse conjunto convexo, basta analisar o seu bordo, isto é, analisar a igualdade $(1-\alpha) \left\| \beta \right\|^2 + \alpha \| \beta \|_1 = t$. Para o caso $\beta \in \mathbb{R}^2$, tome, por exemplo, $t = 2$, $\alpha = 0{,}5$ e $p = 2$ onde se tem $(0{,}5) \left\| \beta \right\|^2 + (0{,}5) \left\| \beta \right\|_1 = 2$. Assim, se $\left\| \beta \right\|^2 = 2$ e $\left\| \beta \right\|_1 = 2$, tem-se a representação, conforme Figura 10.

Figura 10 $K_p$ na estimação Elastic Net, com as restrições Ridge e Lasso (azul) e Elastic Net (vermelho).




Fonte: Do autor (2019).

Novamente, o processo de estimação Elastic Net consiste em projetar com distância mínima o estimador de quadrados mínimos $\hat{\beta}_{ols}$ em $K_p$. Geometricamente a projeção é obtida como o ponto de tangência entre a família de elipses centrada em $\hat{\beta}_{ols}$ e a curva $(1-\alpha) \left\| \beta \right\|^2 + \alpha \| \beta \|_1 = t$, conforme representação na Figura 11.

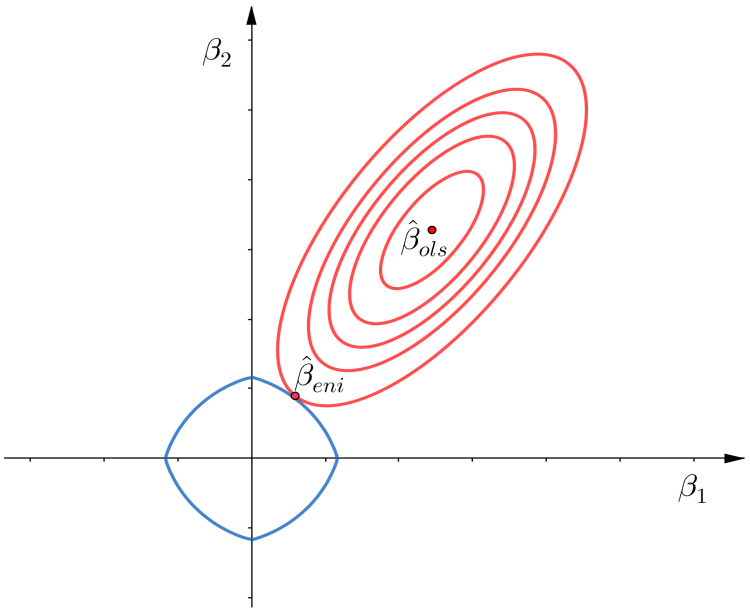
Observando-se, novamente, a Figura 11 tem-se que nos pontos $\left( \mp \frac{\alpha}{2(1-\alpha)} \pm \sqrt{\frac{t}{1-\alpha} + \left( \frac{\alpha}{2(1-\alpha)} \right)^2} \ , \ 0 \right)$ e $\left( 0 \ , \ \mp \frac{\alpha}{2(1-\alpha)} \pm \sqrt{\frac{t}{1-\alpha} + \left( \frac{\alpha}{2(1-\alpha)} \right)^2} \right)$ a curva não é diferenciável, isto é, se têm "pontas". Para o caso em que $\beta \in \mathbb{R}^p$, sempre que uma das componentes de $\beta$ é nula, a função $\left\| \beta \right\|_1$ não é diferenciável.

Neste sentido o convexo

$$K_p = \left\{ \beta \in \mathbb{R}^p, (1-\alpha) \left\| \beta \right\|^2 + \alpha \| \beta \|_1 \leq t \right\}.$$

possui "faces" para as quais não se têm derivadas, no sentido de que as derivadas laterais são 1 e -1. Essas "faces" então são como pontas no convexo $K_p$.

Figura 11 Geometria do estimador Elastic Net.




Fonte: Do autor (2019).

Intuitivamente, essas "pontas" tem maior probabilidade de, primeiramente, interceptarem os hiperelipsóides e, portanto, a propriedade de seleção de covariáveis também ocorre para a estimação Elastic Net.

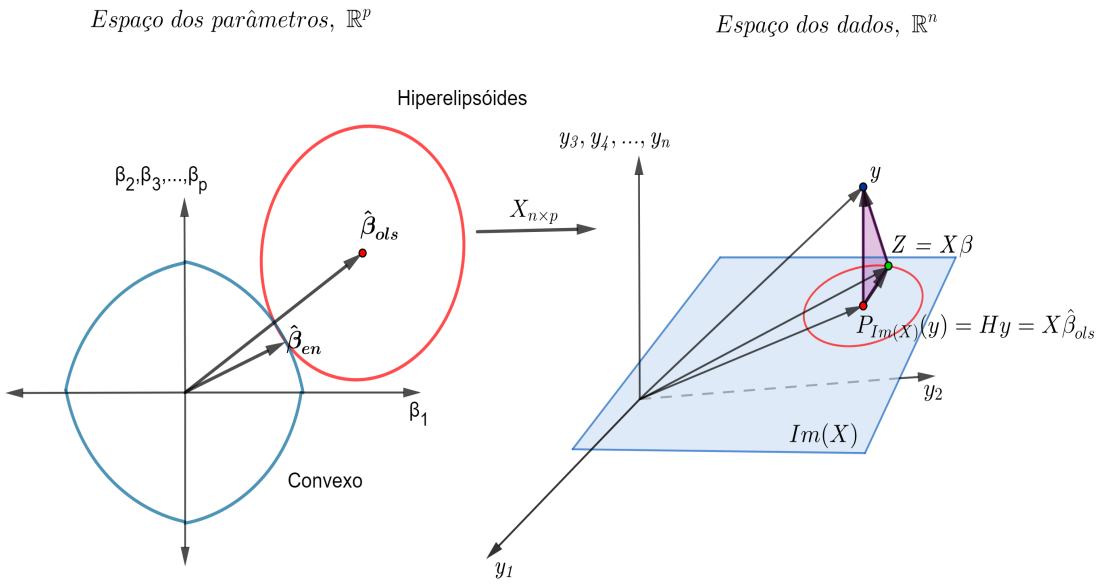
No espaço de dados, o método pode ser visualizado conforme descrito a seguir. A aplicação $X$ leva o conjunto convexo $K_p$, $(1-\alpha)\left\|\beta\right\|^2 + \alpha\|\beta\|_1 = t$ em um conjunto convexo $K \subset \mathrm{Im}\,(X)$. O processo é, simplesmente, o de se obter o vetor $X\beta$ em $K$ que esteja o mais próximo de $y$ (FIGURA 12).

O estimador, assim obtido, será denominado estimador Elastic Net Ingênuo e denotado por $\hat{\beta}_{eni}$.

O processo do Elastic Net Ingênuo pode ser visto como um processo de duas etapas: uma regressão Ridge, seguida de uma regressão Lasso. Esse fato será fundamental na construção de algoritmos eficientes para o cálculo de estimativas Elastic Net. A ideia é que a regressão Ridge pode ser obtida com o emprego de estimadores mistos (COSTA, 2015; GRUBER, 1998), isto é, utilizando-se um modelo linear aumentado a partir do modelo original. Esse procedimento está descrito a seguir.

Primeiramente, faz-se uma reparametrização, em que os novos parâmetros são da forma $\beta^* = \sqrt{1+\gamma}\ \beta$. A partir do conjunto de dados $(y_{n\times 1}, X_{n\times p})$ define-se um novo conjunto de dados (dados aumentados) que podem, por exemplo, ser provenientes de um outro experimento, anteriormente, realizado. Observe que, como estamos com o modelo normalizado, os valores aumentados são nulos, isto é, provenientes de novos dados com média zero, obtidos com apenas uma covariável de cada vez.

Figura 12 O método de regressão Elastic Net.




Fonte: Do autor (2019).

Com esses dados aumentados, a regressão fica da forma $\left(y^*_{(n+p)\times 1}, X^*_{(n+p)\times p}\right)$, em que $y^*_{(n+p)\times 1} = \begin{pmatrix} y \\ 0 \end{pmatrix}$ e $X^*_{(n+p)\times p} = \frac{1}{\sqrt{1+\gamma}} \begin{pmatrix} X_{n\times p} \\ \sqrt{\gamma} I_{p\times p} \end{pmatrix}$. Assim, tem-se, então, uma nova regressão, em relação aos novos parâmetros $\beta^*$ dada por $y^* = X^*\beta^* + \epsilon$. O estimador de quadrados mínimos desta regressão é

$$\begin{aligned}
\hat{\beta}^*_{ols} &= \left(X^{*\prime}X^*\right)^{-1} X^{*\prime} y^* \\
&= \frac{1+\gamma}{\sqrt{1+\gamma}} \left( \left(X', \sqrt{\gamma} I\right) \begin{pmatrix} X \\ \sqrt{\gamma} I \end{pmatrix} \right)^{-1} \left(X', \sqrt{\gamma} I\right) \begin{pmatrix} y \\ 0 \end{pmatrix} \\
&= \sqrt{1+\gamma} \left(X'X + \gamma I\right)^{-1} X' y.
\end{aligned}$$

Logo, $\hat{\beta}^*_{ols} = \sqrt{1+\gamma}\hat{\beta}$, em que $\hat{\beta} = (X'X + \gamma I)^{-1} X'y = \hat{\beta}_R(\gamma)$. Portanto, o estimador de quadrados mínimos satisfaz $\hat{\beta}^*_{ols} = \sqrt{1+\gamma}\hat{\beta}_R(\gamma)$, isto é, exatamente o valor obtido pela reparametrização da estimativa Ridge da regressão original $y = X\beta + \epsilon$.

A estimativa Lasso para a regressão $y^* = X^*\beta^* + \epsilon$ é obtida minimizando a Lagrangeana

$$L\left(\lambda,\beta^*\right) = \|y - X^*\beta^*\|^2 + \lambda\|\beta^*\|_1,$$

isto é, $\beta_L^* = \underset{\beta^*}{\arg\min}\, L\left(\lambda,\beta^*\right)$, em que $\lambda = \frac{\eta_1}{\sqrt{1+\eta_2}}$.

Como

$$\begin{aligned}
\|y^* - X^*\beta^*\|^2 + \lambda\|\beta^*\|_1 &= \left\| \begin{pmatrix} y \\ 0 \end{pmatrix} - \frac{1}{\sqrt{1+\eta_2}} \begin{pmatrix} X \\ \sqrt{\eta_2} I \end{pmatrix} \sqrt{1+\eta_2}\beta \right\|^2 + \lambda \left\| \sqrt{1+\eta_2}\beta \right\|_1 \\
&= \left\| \begin{pmatrix} y \\ 0 \end{pmatrix} - \frac{1}{1} \begin{pmatrix} X\beta \\ \sqrt{\eta_2}\beta \end{pmatrix} \right\|^2 + \sqrt{1+\eta_2}\lambda\|\beta\|_1 \\
&= \|y - X\beta\|^2 + \eta_2\|\beta\|^2 + \sqrt{1+\eta_2}\lambda\|\beta\|_1 \\
&= \|y - X\beta\|^2 + \eta_2\|\beta\|^2 + \eta_1\|\beta\|_1,
\end{aligned}$$

tem-se que o estimador Lasso $\hat{\beta}_{lasso}^*$ é exatamente o estimador $\hat{\beta}_{eni}$ pois são definidos pela minimização da mesma função. Em termos dos parâmetros originais, $\hat{\beta}_{eni} = \frac{1}{\sqrt{1+\eta_2}}\hat{\beta}_{lasso}^*$. Dessa forma, o cálculo do estimador Elastic Net Ingênuo pode ser obtido como o estimador Lasso de um sistema aumentado.

Para exemplificar, vamos calcular a estimativa Elastic Net Ingênuo para o caso ortogonal $X'X = I$. O estimador Lasso para uma regressão ortogonal é dado por

$$\left(\hat{\beta}_{lasso}\right)_j = \text{sinal}\left(\hat{\beta}_{ols}\right)_j \left( \left| \left(\hat{\beta}_{ols}\right)_j \right| - \gamma \right)^+.$$

A regressão com os dados aumentados satisfaz

$$X^* = \frac{1}{\sqrt{1+\eta_2}} \begin{pmatrix} X \\ \sqrt{\eta_2} I \end{pmatrix}.$$

Assim,

$$\begin{aligned}(X^*)'(X^*) &= \frac{1}{1+\eta_2}\left(X', \sqrt{\eta_2}I\right)\begin{pmatrix} X \\ \sqrt{\eta_2}I \end{pmatrix} \\ &= \frac{1}{1+\eta_2}\left(X'X + \eta_2 I\right) \\ &= \frac{1}{1+\eta_2}\left(I + \eta_2 I\right) \\ &= I.\end{aligned}$$

Portanto o sistema aumentado também é ortogonal. O Lasso do sistema aumentado é

$$\left(\hat{\beta}^*_{lasso}\right)_j = \left(\left|\left(\hat{\beta}^*_{ols}\right)_j\right| - \gamma\right)^+ \text{sinal}\left(\hat{\beta}^*_{ols}\right)_j.$$

Como demonstrado,

$$\begin{aligned}\hat{\beta}^*_{ols} &= \sqrt{1+\eta_2}\hat{\beta}_R(\eta_2) \\ &= \sqrt{1+\eta_2}\left(X'X + \eta_2 I\right)^{-1}X'y \\ &= \sqrt{1+\eta_2}\left(\left(1+\eta_2\right)I\right)^{-1}X'y \\ &= \frac{1}{\sqrt{1+\eta_2}}X'y.\end{aligned}$$

Mas $X'y$ é o estimador de quadrados mínimos da regressão original, e portanto $\hat{\beta}^*_{ols} = \frac{1}{\sqrt{1+\eta_2}}\hat{\beta}_{ols}$. Como o estimador Elastic Net Ingênuo satisfaz $\hat{\beta}_{eni} = \frac{1}{\sqrt{1+\eta_2}}\hat{\beta}^*_{ols}$, então

$$
\begin{aligned}
(\hat{\beta}_{eni})_j &= \frac{1}{\sqrt{1+\eta_2}}\left(\hat{\beta}^*_{lasso}\right)_j \\
&= \frac{1}{\sqrt{1+\eta_2}}\left(\left|\left(\hat{\beta}^*_{ols}\right)_j\right| - \gamma\right)^+ \operatorname{sinal}\left(\hat{\beta}^*_{ols}\right)_j \\
&= \frac{1}{\sqrt{1+\eta_2}}\left(\left|\frac{1}{\sqrt{1+\eta_2}}\left(\hat{\beta}_{ols}\right)_j\right| - \gamma\right)^+ \operatorname{sinal}\left(\hat{\beta}_{ols}\right)_j \\
&= \frac{\left(\left|\left(\hat{\beta}_{ols}\right)_j\right| - \sqrt{1+\eta_2}\gamma\right)^+}{1+\eta_2} \operatorname{sinal}\left(\hat{\beta}_{ols}\right)_j \\
&= \frac{\left(\left|\left(\hat{\beta}_{ols}\right)_j\right| - \frac{\eta_1}{2}\right)^+}{1+\eta_2} \operatorname{sinal}\left(\hat{\beta}_{ols}\right)_j,
\end{aligned}
$$

conforme descrito em (ZOU, HASTIE,2005).

A grande vantagem do método Elastic Net, em relação ao método Lasso, é que se tem a propriedade de efeito de grupo, isto é, variáveis altamente correlacionadas têm alta probabilidade de gerar estimativas relacionadas a essas covariáveis com valores próximos. Essa propriedade decorre do fato de que a penalidade empregada no método Elastic Net define um modelo $K_p$, estritamente convexo, ao contrário do método Lasso que define um modelo $K_p$ apenas convexo.

A diferença entre uma penalização convexa e uma estritamente convexa é explicitada na Proposição 1. Seja $J(\cdot)$ uma função positiva e o problema de minimização

$$
\hat{\beta} = \arg\min_{\beta} \left\{ \|y - X\beta\|^2 + \lambda J(\beta) \right\}.
$$

Tem-se

**Proposição 1:** Se $x_i = x_j$, com $i,j \in \{1,\ldots,p\}$,

1. Se $J(\cdot)$ é estritamente convexa, então $\hat{\beta}_i = \hat{\beta}_j, \forall \lambda > 0$.
2. Se $J(\beta) = \|\beta\|_1$, portanto, convexa, mas não estritamente convexa, então $\hat{\beta}_i\hat{\beta}_j \geq 0$ e

$$
\hat{\beta}^*_k = \begin{cases} \hat{\beta}_k, & k \neq i,j \\ \left(\hat{\beta}_i + \hat{\beta}_j\right) s, & k = i \\ \left(\hat{\beta}_i + \hat{\beta}_j\right)(1-s), & k = j \end{cases},
$$

é uma outra solução do problema de minimização para $s \in [0,1]$.

Um resultado de maior aplicabilidade vai garantir uma cota para a diferença entre as estimativas $\left(\hat{\beta}_{eni}\right)_i$ e $\left(\hat{\beta}_{eni}\right)_j$ em termos da correlação amostral $\rho = x'_i x_j$.

**Teorema 5 (Efeito de grupo no Elastic Net Ingênuo)** Considere o conjunto de dados $(y,X)$ e os parâmetros $(\lambda_1,\lambda_2)$, em que a resposta $y$ está centrada e os preditores $X$ estão padronizados. Seja $\hat{\beta}(\lambda_1,\lambda_2)$ o estimador Elastic Net Ingênuo. Suponha que $\hat{\beta}_i(\lambda_1,\lambda_2)\hat{\beta}_j(\lambda_1,\lambda_2) > 0$. Defina,

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

Então, $D_{\lambda_1,\lambda_2}(i,j) \leq \frac{1}{\lambda_2}\sqrt{2(1-\rho)}$, em que $\rho = x'_i x_j$ é a correlação amostral.

**Demonstração**

Se $\hat{\beta}_i(\lambda_1,\lambda_2)\hat{\beta}_j(\lambda_1,\lambda_2) > 0$, então ambos $\hat{\beta}_i(\lambda_1,\lambda_2)$ e $\hat{\beta}_j(\lambda_1,\lambda_2)$ são não nulos, e têm o mesmo sinal, isto é, $\text{sgn}\{\hat{\beta}_i(\lambda_1,\lambda_2)\} = \text{sgn}\{\hat{\beta}_j(\lambda_1,\lambda_2)\}$. Por hipótese,

$$\hat{\beta} = \arg\min_{\beta} \{L(\lambda_1,\lambda_2,\beta)\}.$$

em que,

$$L(\lambda_1,\lambda_2,\beta) = \|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1.$$

Então, $\hat{\beta}(\lambda_1,\lambda_2)$ deve satisfazer a relação,

$$\frac{\partial L(\lambda_1,\lambda_2,\beta)}{\partial \beta_k}|_{\beta=\hat{\beta}(\lambda_1,\lambda_2)} = 0, \text{ se } \hat{\beta}_k(\lambda_1,\lambda_2) \neq 0.$$

Assim,

$$\begin{aligned} L(\lambda_1,\lambda_2,\beta) &= \|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1 \\ &= \|y\|^2 - 2\beta' X'y + \beta' X'X\beta + \beta'\lambda_2 I\beta + \lambda_1\beta' v, \end{aligned}$$

em que $v$ denota o vetor de sinais de $\hat{\beta} = \hat{\beta}_{ols}$. Dessa forma, diferenciando em relação a $\beta_k$ , obtemos:

$$\frac{\partial L}{\partial \beta_k} = -2x'_k\left[y - X\hat{\beta}(\lambda_1,\lambda_2)\right] + \lambda_1 \text{sgn}\left\{\hat{\beta}_k(\lambda_1,\lambda_2)\right\} + 2\lambda_2\hat{\beta}_k(\lambda_1,\lambda_2). \tag{2.23}$$

Aplicando em $\hat{\beta}(\lambda_1,\lambda_2)$ e fazendo $k = i$ e $k = j$, obtemos:

$$-2x'_i\left[y - X\hat{\beta}(\lambda_1,\lambda_2)\right] + \lambda_1 \text{sgn}\left\{\hat{\beta}_i(\lambda_1,\lambda_2)\right\} + 2\lambda_2\hat{\beta}_i(\lambda_1,\lambda_2) = 0. \tag{2.24}$$

$$-2x'_j \left[y - X\hat{\beta}(\lambda_1,\lambda_2)\right] + \lambda_1 \text{sgn}\left\{\hat{\beta}_j(\lambda_1,\lambda_2)\right\} + 2\lambda_2\hat{\beta}_j(\lambda_1,\lambda_2) = 0. \tag{2.25}$$

Subtraindo (2.24) de (2.25) , obtemos, após usar que $\text{sgn}\left\{\hat{\beta}_i(\lambda_1,\lambda_2)\right\} = \text{sgn}\left\{\hat{\beta}_j(\lambda_1,\lambda_2)\right\}$,

$$(x'_j - x'_i)\left\{y - X\hat{\beta}(\lambda_1,\lambda_2)\right\} + \lambda_2\left\{\hat{\beta}_i(\lambda_1,\lambda_2) - \hat{\beta}_j(\lambda_1,\lambda_2)\right\} = 0,$$

que é equivalente a

$$\hat{\beta}_i(\lambda_1,\lambda_2) - \hat{\beta}_j(\lambda_1,\lambda_2) = \tfrac{1}{\lambda_2}(x'_i - x'_j)'\ \hat{r}(\lambda_1,\lambda_2),$$

em que $\hat{r}(\lambda_1,\lambda_2) = y - X\hat{\beta}(\lambda_1,\lambda_2)$ é o vetor residual.

Visto que $X$ está padronizada, podemos escrever

$$\begin{aligned}\|x_i - x_j\|^2 &= (x_i - x_j)'(x_i - x_j)\\ &= x'_i x_i - 2x'_i x_j + x'_j x_j\\ &= \|x_i\|^2 + \|x_j\|^2 - 2\rho\\ &= 2(1-\rho),\end{aligned}$$

em que, $\rho = x'_i x_j$.

Agora, como $\hat{\beta} = \arg\min_{\beta}\{L(\lambda_1,\lambda_2,\beta)\}$, em particular, temos

$$L(\lambda_1,\lambda_2,\hat{\beta}(\lambda_1,\lambda_2)) \leq L(\lambda_1,\lambda_2,\beta = 0),$$

isto é,

$$\|\hat{r}(\lambda_1,\lambda_2)\|^2 + \lambda_2\left\|\hat{\beta}(\lambda_1,\lambda_2)\right\|^2 + \lambda_1\left\|\hat{\beta}(\lambda_1,\lambda_2)\right\|_1 \leq \|y\|^2,$$

uma vez que $\|\hat{r}(\lambda_1,\lambda_2)\| \leq \|y\|$. Além disso, como

$$\hat{\beta}_i(\lambda_1,\lambda_2) - \hat{\beta}_j(\lambda_1,\lambda_2) = \tfrac{1}{\lambda_2}(x'_i - x'_j)'\ \hat{r}(\lambda_1,\lambda_2),$$

resulta que,

$$
\begin{aligned}
D_{\lambda_1,\lambda_2}(i,j) &= \frac{1}{\|y\|_1}\left|\hat{\beta}_i(\lambda_1,\lambda_2) - \hat{\beta}_j(\lambda_1,\lambda_2)\right| \\
&\leq \frac{1}{\lambda_2}\frac{\|\hat{r}(\lambda_1,\lambda_2)\|}{\|y\|_1}\,\|x_i - x_j\|^2 \\
&= \frac{1}{\lambda_2}\frac{\|\hat{r}(\lambda_1,\lambda_2)\|}{\|y\|_1}\,\sqrt{2(1-\rho)} \\
&\leq \frac{1}{\lambda_2}\,\sqrt{2(1-\rho)}.
\end{aligned}
$$

Assim, $D_{\lambda_1,\lambda_2}(i,j) \leq \frac{1}{\lambda_2}\,\sqrt{2(1-\rho)}$, e isso completa a demonstração do teorema.

■

O teorema que acabamos de demonstrar nos diz que a quantidade adimensional $D_{\lambda_1,\lambda_2}(i,j)$ descreve a diferença entre os valores dos coeficientes dos preditores $i$ e $j$. Se $x_i$ e $x_j$ são altamente correlacionados, isto é, $\rho \cong 1$ ( Se $\rho \cong -1$ , então considere $-x_j$), no teorema, afirma-se que a diferença entre os valores dos coeficientes dos preditores $i$ e $j$ é quase nula. O limite superior da desigualdade acima, fornece uma descrição quantitativa para o efeito de agrupamento do Elastic Net Ingênuo. O Lasso não tem esse efeito de grupo, conforme demonstrado em (EFRON et al., 2004).

Segue do teorema 5, visto, anteriormente, que se $x_i$ e $x_j$ são altamente correlacionados, então a diferença das estimativas dos parâmetros relativos a estas covariáveis é próxima de zero. Essa propriedade é que dá ao estimador sua propriedade de seleção por grupo, isto é, grupos de variáveis altamente correlacionadas tendem a ter estimativas correspondentes próximas. Essa propriedade é muito interessante em aplicações. O Lasso, em geral, não possui tal propriedade.

Como observado, anteriormente, o estimador Elastic Net Ingênuo é obtido por um procedimento em dois estágios: uma estimativa Ridge e, em seguida, um encolhimento tipo Lasso. Portanto, se tem um duplo encolhimento e estudos de simulação comprovam que tal fato afeta a performance do estimador. Para que se possa corrigir tal situação, o estimador Elastic Net é definido como um reescalonamento do estimador Elastic Net Ingênuo $\hat{\beta}_{en} = (1+\lambda_2)\,\hat{\beta}_{eni}$.

# 3 RESULTADOS

## 3.1 Análise geométrica da restrição Elastic Net.

Nesta seção, vamos fazer uma análise geométrica da restrição Elastic Net, visando a conhecer propriedades geométricas dessa restrição. Pela convexidade da restrição, é suficiente fazer uma análise na fronteira da restrição que é dada por,

$$(1-\alpha)\|\beta\|^2 + \alpha\|\beta\|_1 = t, \text{ em que } \alpha \in [0,1] \text{ e } t \geq 0.$$

Primeiramente, devemos observar que, pela natureza da restrição em que aparecem termos quadráticos ou modulares, existe uma simetria da restrição, em relação aos hiperplanos coordenados. Para analisarmos essa restrição, visando à generalização, vamos definir o vetor de sinais do vetor $\beta' = (\beta_1,\beta_2,\ldots,\beta_p) \in \mathbb{R}^p$, denotado por

$$v' = (v_1,v_2,\ldots,v_p) \in \mathbb{R}^p$$

como,

$$v_i = \begin{cases} +1 & ,\text{se } \beta_i \geq 0 \\ -1 & ,\text{se } \beta_i < 0. \end{cases}$$

Observe que $v$ é um vetor coluna $p$-dimensional de $1's$ e $-1's$, que é fixo em cada hiperoctante. Dessa forma, iremos analisar a restrição em um determinado hiperoctante, cujo vetor de sinais seja $v$, e o restante da restrição obter por simetria. Para $\alpha = 1$, temos a restrição do lasso que será discutida em detalhes adiante. Para $\alpha \neq 1$, podemos escrever

$$(1-\alpha)\|\beta\|^2 + \alpha\|\beta\|_1 = t. \tag{3.1}$$

Dividindo ambos os membros da Equação (3.1) por $1-\alpha$ resulta que

$$\|\beta\|^2 + \frac{\alpha}{1-\alpha}\|\beta\|_1 = \frac{t}{1-\alpha}.$$

Como $\|\beta\|^2 = \beta'\beta$ e $\|\beta\|_1 = \beta'v = v'\beta$, em que $v$ é o vetor de sinais $p$-dimensional que é fixo no hiperoctante que estejamos trabalhando, segue que

$$\beta'\beta + \frac{\alpha}{1-\alpha}\beta'v = \frac{t}{1-\alpha}.$$

Assim, adicionando em ambos os membros o termo $\frac{\alpha^2}{4(1-\alpha)^2}v'v$, tem-se

$$\beta'\beta + 2\frac{\alpha}{2(1-\alpha)}\beta'v + \frac{\alpha^2}{4(1-\alpha)^2}v'v = \frac{t}{1-\alpha} + \frac{\alpha^2}{4(1-\alpha)^2}v'v.$$

A equação acima, pode ser escrita mais concisamente na forma,

$$\left(\beta + \frac{\alpha}{2(1-\alpha)}v\right)'\left(\beta + \frac{\alpha}{2(1-\alpha)}v\right) = \frac{t}{1-\alpha} + \frac{\alpha^2}{4(1-\alpha)^2}v'v.$$

ou, equivalentemente, observando que $v'v = p$.

$$\left\|\beta + \tfrac{\alpha}{2(1-\alpha)}v\right\|^2 = \tfrac{t}{1-\alpha} + \tfrac{\alpha^2 p}{4(1-\alpha)^2} \Rightarrow \|\beta - C\|^2 = R^2 \Rightarrow \|\beta - C\| = R$$

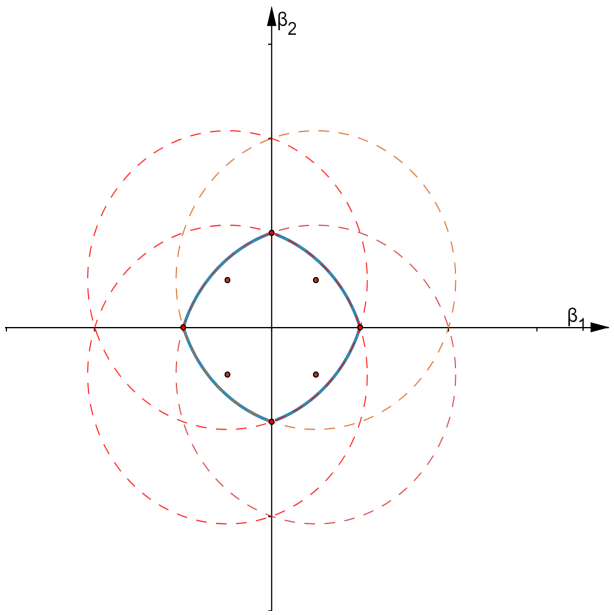
Essa equação representa uma esfera $p$-dimensional de centro $C$ e raio $R$ dados por,

$$C = -\tfrac{\alpha}{2(1-\alpha)}v \quad \text{e raio } R = \sqrt{\tfrac{t}{1-\alpha} + \tfrac{\alpha^2 p}{4(1-\alpha)^2}}.$$

É importante observar que o vetor $v$ é ortogonal ao hiperplano $\beta'v = v'\beta = t$. Cabe esclarecer que o vetor $v$ descrito acima modifica de acordo com o hiperoctante em que se encontra a curva. Dessa forma, por simetria, conhecemos toda a curva da restrição.

A seguir, na Figura 13, ilustramos a forma geométrica da restrição Elastic Net no plano, para $\alpha = 0{,}5$ e $t = 1$. As linhas tracejadas mostram as circunferências que contribuem na composição da restrição.

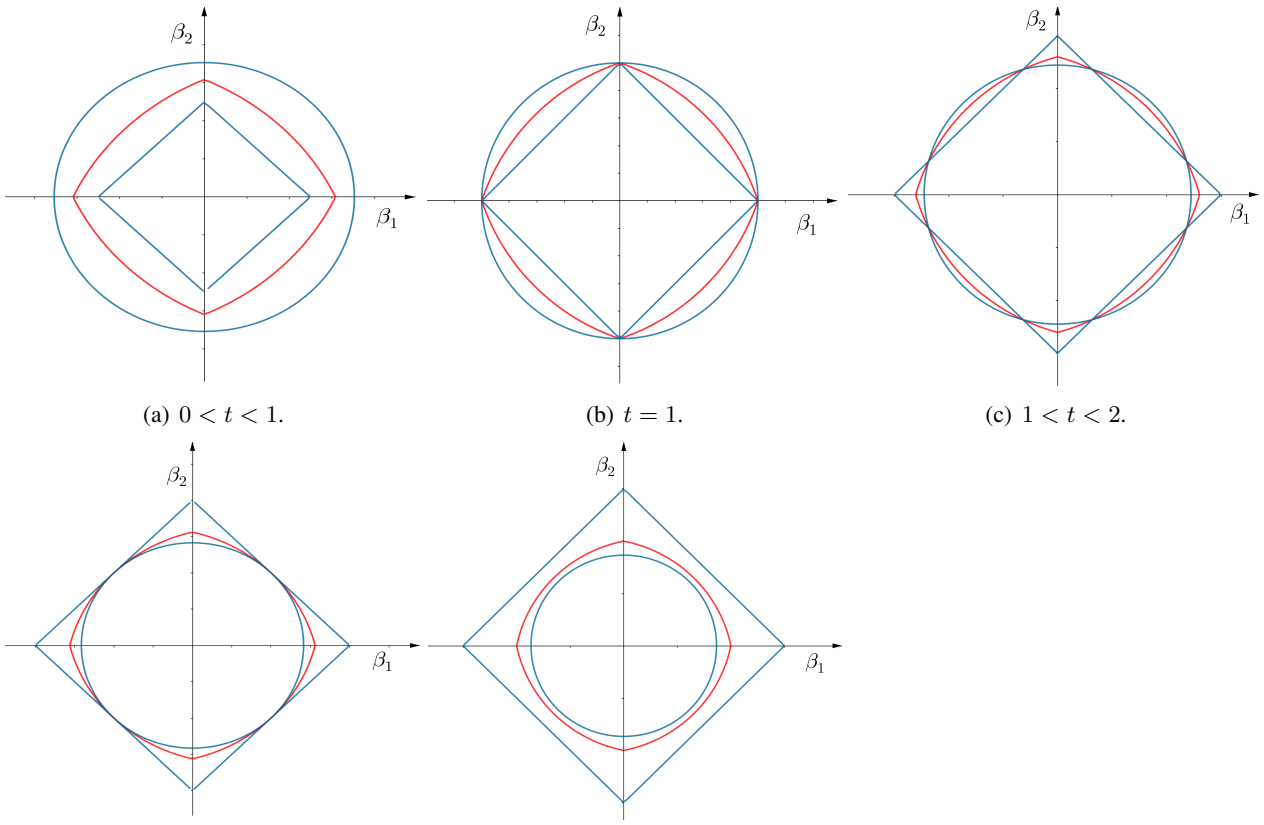
Figura 13 As esferas no plano da restrição Elastic Net, para $\alpha = 0{,}5$ e $t = 1$ .




Fonte: Do autor (2019).

Um fato relevante é que podemos classificar em cinco configurações distintas a natureza das restrições, Ridge, Lasso e ElasticNet, de acordo com a variação do parâmetro $t$. De fato, a seguir, na Figura 14, ilustramos a forma geométrica da fronteira das restrições, no plano em que consideramos $\alpha = 0{,}5$ na restrição Elastic Net.

Figura 14 Restrições Ridge, Lasso e Elastic Net com $t$ variando, para $\alpha = 0{,}5$.




(a) $0 < t < 1$. (b) $t = 1$. (c) $1 < t < 2$.

(d) $t = 2$. (e) $t > 2$.

Fonte: Do autor (2019).

Podemos extrair, dessa análise, algumas conclusões relativas ao mínimo da função $f(\beta) = \|y - X\beta\|^2$ sujeita às restrições do tipo Ridge, Lasso e Elastic Net. Para fazermos essa análise, basta observar que, se uma função contínua $f$ tem um mínimo em $X \subseteq \mathbb{R}^p$ e $Y \subseteq X$, então, o mínimo em $X$ é menor ou igual ao mínimo em $Y$. Por exemplo, podemos afirmar que, no caso (a) e (b), em que $0 < t \leq 1$, teremos a relação $f(\hat{\beta}_{Ridge}) \leq f(\hat{\beta}_{en}) \leq f(\hat{\beta}_{Lasso})$, independentemente do valor de $\alpha \in [0{,}1]$. Os casos (d) e (e) são análogos, com as devidas modificações. O caso (c) não permite concluir de imediato alguma relação mais geral.

### 3.2 Análise das curvas de nível da função erro quadrático.

Nesta seção, vamos analisar as curvas de nível de $f(\beta) = \|y - X\beta\|^2$ que será útil no desenvolvimento que se seguirá. Com efeito,

$$\begin{aligned}
f(\beta) = \|y - X\beta\|^2 &= \left\|(y - X\hat{\beta}_{ols}) - (X\beta - X\hat{\beta}_{ols})\right\|^2 \\
&= \left[(y - X\hat{\beta}_{ols}) - (X\beta - X\hat{\beta}_{ols})\right]' \left[(y - X\hat{\beta}_{ols}) - (X\beta - X\hat{\beta}_{ols})\right] \\
&= \left[(y - X\hat{\beta}_{ols})' - (X\beta - X\hat{\beta}_{ols})'\right] \left[(y - X\hat{\beta}_{ols}) - (X\beta - X\hat{\beta}_{ols})\right] \\
&= (y - X\hat{\beta}_{ols})'(y - X\hat{\beta}_{ols}) - 2(y - X\hat{\beta}_{ols})'(X\beta - X\hat{\beta}_{ols}) + (X\beta - X\hat{\beta}_{ols})'(X\beta - X\hat{\beta}_{ols}) \\
&= \left\|y - X\hat{\beta}_{ols}\right\|^2 + \left\|X\beta - X\hat{\beta}_{ols}\right\|^2.
\end{aligned}$$

Note que o termo $(y - X\hat{\beta}_{ols})'(X\beta - X\hat{\beta}_{ols}) = 0$, conforme vimos anteriormente. Assim, como estamos estudando as curvas de nível de $f(\beta)$ e já que $\left\|y - X\hat{\beta}_{ols}\right\|^2$ é constante, resulta que

$$\begin{aligned}
\left\|X\beta - X\hat{\beta}_{ols}\right\|^2 &= f(\beta) - \left\|y - X\hat{\beta}_{ols}\right\|^2 \\
&= \left\|X(\beta - \hat{\beta}_{ols})\right\|^2 \\
&= \left[X(\beta - \hat{\beta}_{ols})\right]' \left[X(\beta - \hat{\beta}_{ols})\right] \\
&= (\beta - \hat{\beta}_{ols})'X'X(\beta - \hat{\beta}_{ols}) = C,
\end{aligned}$$

em que $C$ é uma constante.

Dessa forma, as curvas de nível de $f$ são hiperelipsóides centrados em $\hat{\beta}_{ols}$, isto é,

$$(\beta - \hat{\beta}_{ols})'X'X(\beta - \hat{\beta}_{ols}) = C.$$

Além disso, se nos restringirmos ao caso ortogonal, $X'X = I$, temos que as curvas de nível são hiperesferas centradas em $\hat{\beta}_{ols}$, a saber, $\beta$ satisfaz a equação

$$(\beta - \hat{\beta}_{ols})'(\beta - \hat{\beta}_{ols}) = C.$$

## 3.3 Análise dos estimadores de encolhimento Ridge, Lasso e Elastic Net para o caso ortogonal

### 3.3.1 Introdução

Hoerl e Kennard (1970), propuseram, a partir da constatação de que o estimador de mínimos quadrados não era admissível para dimensões maiores que dois, o então chamado estimador de encolhimento Ridge. Esse estimador deu origem a outros estimadores de encolhimento, tais como Lasso (Least Absolute Shrinkage and Selection Operator) e Elastic Net, dentre outros. Nesta seção, os estimadores Ridge e Lasso são derivados como casos particulares do problema de minimização que define o estimador Elastic Net. Tal abordagem, confere unidade à teoria, que na literatura, é, geralmente, tratada de forma compartimentada. Tal unificação é possível com o uso do vetor de sinais $v$, definido pelo hiperoctante em que se encontra a curva de restrição.

Considere o modelo de regressão linear $y = X\beta + \varepsilon$, em que $X$ é uma matriz, $n \times p$, definida pelos valores das covariáveis, com posto $r(X) = p = \min(n,p)$ e $\beta$ é um vetor $p$-dimensional de parâmetros, já considerando o modelo centralizado.

Primeiramente, com o enfoque de otimização convexa, precisamos minimizar uma função convexa $f(\beta) = \|y - X\beta\|^2$ restrita ao conjunto convexo $(1-\alpha)\|\beta\|^2 + \alpha\|\beta\|_1 = t$, em que $\alpha \in [0,1]$ , o que caracteriza um problema de otimização convexa, conforme definimos anteriormente. Para isso, consideremos a Lagrangeana associada ao problema,

$$L(\beta,\lambda) = \|y - X\beta\|^2 + \lambda\left[(1-\alpha)\|\beta\|^2 + \alpha\|\beta\|_1 - t\right].$$

Vamos admitir que a solução que buscamos esteja no hiperoctante que tem $v$ como seu vetor de sinais. Dessa forma, podemos escrever que $\|\beta\|_1 = \beta' v = v'\beta$. Visando à diferenciação da função matricial, vamos reescrever a função na forma.

$$\begin{aligned}
L(\beta,\lambda) &= \|y - X\beta\|^2 + \lambda\left[(1-\alpha)\|\beta\|^2 + \alpha\|\beta\|_1 - t\right] \\
&= (y - X\beta)'(y - X\beta) + \lambda(1-\alpha)\|\beta\|^2 + \lambda\alpha\|\beta\|_1 - \lambda t \\
&= \left(y' - \beta' X'\right)(y - X\beta) + \lambda(1-\alpha)\beta'\beta + \lambda\alpha\beta' v - \lambda t \\
&= y'y - 2\beta' X' y + \beta' X' X\beta + \beta'\lambda(1-\alpha)I\beta + 2\beta'\frac{\lambda\alpha}{2}v - \lambda t \\
&= y'y - 2\beta'\left(X'y - \frac{\lambda\alpha}{2}v\right) + \beta'\left(X'X + \lambda(1-\alpha)I\right)\beta - \lambda t.
\end{aligned}$$

Assim, obtemos por diferenciação, que

$$\begin{cases} \frac{\partial L(\beta,\lambda)}{\partial \beta} = -2\left(X'y - \frac{\lambda\alpha}{2}v\right) + 2\left(X'X + \lambda(1-\alpha)I\right)\beta \\ \frac{\partial L(\beta,\lambda)}{\partial \lambda} = (1-\alpha)\|\beta\|^2 + \alpha\|\beta\|_1 - t \end{cases}.$$

Igualando a zero essas derivadas parciais, obtemos

$$\begin{cases} -2\left(X'y - \frac{\lambda\alpha}{2}v\right) + 2\left(X'X + \lambda(1-\alpha)I\right)\beta = 0 \\ (1-\alpha)\|\beta\|^2 + \alpha\|\beta\|_1 - t = 0 \end{cases}.$$

O que resulta em,

$$\begin{cases} \beta = \left(X'X + \lambda(1-\alpha)I\right)^{-1}\left(X'y - \frac{\lambda\alpha}{2}v\right) \\ (1-\alpha)\|\beta\|^2 + \alpha\|\beta\|_1 = t \end{cases}. \tag{3.2}$$

Note que a fórmula (3.2) é uma generalização das fórmulas correspondentes para os estimadores de mínimos quadrados, Ridge e Lasso. De fato, para $\lambda = 0$, não há restrição e o resultado independe do vetor de sinais e é geral. Nesse caso, temos a solução de mínimos quadrados, $\hat{\beta}_{ols} = (X'X)^{-1}X'y$.

Para $\lambda \neq 0$, temos dois casos particulares.

Para $\alpha = 0$, temos o estimador Ridge,

$$\hat{\beta}_{ridge} = (X'X + \lambda I)^{-1}X'y.$$

Para $\alpha = 1$, temos o estimador Lasso,

$$\hat{\beta}_{lasso} = (X'X)^{-1}\left(X'y - \tfrac{\lambda}{2}v\right).$$

Note que, para exibir o estimador Ridge, Lasso ou Elastic Net há necessidade ainda de explicitar $\lambda$ em função de $t$. Mostraremos adiante que, em algumas situações, isso será possível e em outras não. Devido a essa impossibilidade, faz-se necessária a utilização de recursos computacionais, visando à solução do sistema.

### 3.3.2 O estimador Ridge no caso ortogonal

Nesta seção, vamos apresentar o estimador Ridge, no caso ortogonal, cujo resultado é conhecido, todavia apresentaremos, utilizando a fórmula (3.2) obtida com o auxílio do vetor de sinais. Nesse caso, temos que minimizar $f(\beta) = \|y - X\beta\|^2$, sujeito à restrição $\|\beta\|^2 = t$. Vamos proceder por dois caminhos distintos, visando a explicitar o estimador nesse caso ortogonal.

Primeiramente, com o enfoque de otimização convexa, precisamos minimizar uma função convexa $f$ restrita ao conjunto convexo $\|\beta\|^2 \leq t$. Sabemos que o mínimo ocorrerá na fronteira, devido às curvas de nível de $f$. Assim, de acordo com a equação (3.2), fazendo $\alpha = 0$ e $X'X = I$, obtemos

$$\begin{cases} \beta = \frac{1}{1+\lambda} X'y \\ \|\beta\|^2 = t \end{cases}.$$

Além disso, podemos melhorar um pouco mais a fórmula ao observarmos que, no caso ortogonal,

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

Portanto, podemos escrever a expressão explícita para o estimador Ridge, no caso ortogonal como,

$$\beta = \frac{1}{1+\lambda}\hat{\beta}_{ols}. \tag{3.3}$$

Para finalizar a expressão do estimador Ridge, devemos exibir o parâmetro $\lambda$, em função do parâmetro $t$. Com efeito,

$$\begin{aligned} \|\beta\|^2 = t \Rightarrow \left\| \tfrac{1}{1+\lambda}\hat{\beta}_{ols} \right\|^2 = t \Rightarrow \tfrac{1}{(1+\lambda)^2}\left\|\hat{\beta}_{ols}\right\|^2 = t \\ \Rightarrow (1+\lambda)^2 = \frac{\left\|\hat{\beta}_{ols}\right\|^2}{t} \Rightarrow 1+\lambda = \frac{\left\|\hat{\beta}_{ols}\right\|}{\sqrt{t}} \\ \Rightarrow \lambda = \frac{\left\|\hat{\beta}_{ols}\right\|}{\sqrt{t}} - 1. \end{aligned} \tag{3.4}$$

Daí resulta que, substituindo (3.4) em (3.3) temos:

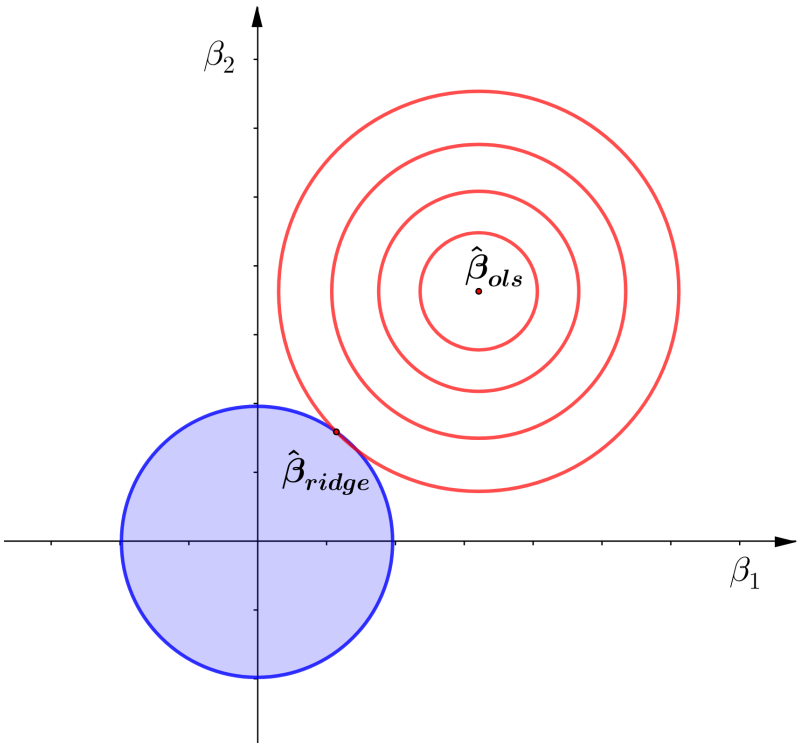
$$\hat{\beta}_{ridge} = \sqrt{t}\frac{\hat{\beta}_{ols}}{\left\|\hat{\beta}_{ols}\right\|}.$$

Passemos agora a analisar o mesmo problema do ponto de vista geométrico. Vimos que as curvas de nível da função $f(\beta) = \|y - X\beta\|^2$ são, de fato, hiperelipsóides. Além disso, como estamos restritos ao caso ortogonal, $X'X = I$, temos que as curvas de nível são hipercírculos centrados em $\hat{\beta}_{ols}$, a saber, $\beta$ satisfaz a equação

$$(\beta - \hat{\beta}_{ols})'(\beta - \hat{\beta}_{ols}) = C.$$

Não é difícil concluir, que no caso do estimador Ridge, os hipercírculos centrados em $\hat{\beta}_{ols}$ ao crescerem, irão tocar a restrição Ridge ortogonalmente, em dois pontos, um será o máximo e o outro, que nos interessa, será o mínimo. O ponto de mínimo será onde os hipercírculos tocam primeiro a restrição, conforme ilustra-se na Figura 15.

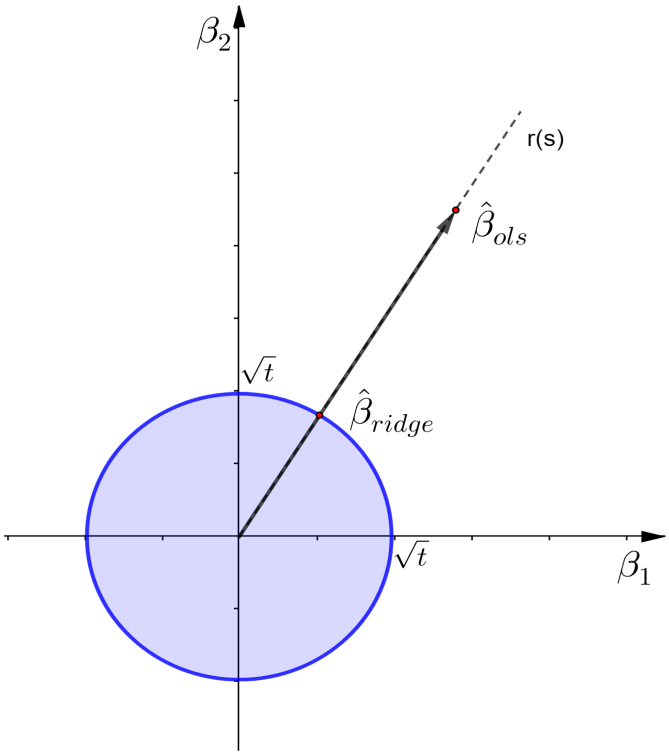
Figura 15 Curvas de nível no caso Ridge.




Fonte: Do autor (2019).

O procedimento para se determinar o estimador, neste caso ortogonal, é muito simples. Consideremos a reta $r(s)$, que passa pela origem na direção do vetor $\hat{\beta}_{ols}$, a saber $r(s) = s\hat{\beta}_{ols}$, em que o parâmetro, não estatístico $s$ é não negativo. Dessa forma, devemos determinar o parâmetro $s$ para que $r(s)$ pertença à restrição $\|\beta\|^2 = t$ o que ocorre se, e somente se, $\|\beta\| = \sqrt{t}$. Conforme ilustra a Figura 16.

Figura 16 Estimador Ridge no caso ortogonal.




Fonte: Do autor (2019).

Dessa forma, devemos ter

$$\|r(s)\| = \sqrt{t} \Leftrightarrow \left\|s\hat{\beta}_{ols}\right\| = \sqrt{t} \Leftrightarrow s\left\|\hat{\beta}_{ols}\right\| = \sqrt{t} \Leftrightarrow s = \frac{\sqrt{t}}{\|\hat{\beta}_{ols}\|}.$$

Portanto, voltando à equação paramétrica da reta, obtemos

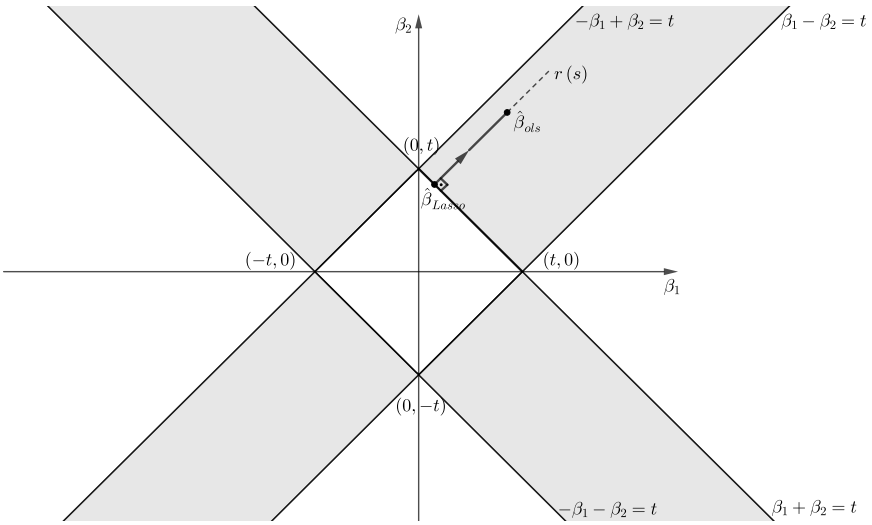
$$\hat{\beta}_{ridge} = \sqrt{t}\frac{\hat{\beta}_{ols}}{\|\hat{\beta}_{ols}\|}.$$

Obtemos a mesma expressão para o estimador, como era esperado. Isso finaliza nossa discussão para o caso ortogonal do estimador Ridge.

### 3.3.3 O estimador Lasso no caso ortogonal

O estimador Lasso é obtido pela projeção ortogonal do estimador de quadrados mínimos no convexo $\|\beta\|_1 \leq t$. Devemos salientar que, no exterior do convexo, fixado o parâmetro $t$, existem duas regiões distintas, aquela que denominaremos de "região principal", em que o vetor $\hat{\beta}_{ols}$ se projeta, ortogonalmente, em um hiperplano do convexo, e a outra região, que denominaremos de "região de atração", em que não há projeção ortogonal no convexo. Nesse último caso, o estimador Lasso estará em uma face singular do convexo, isto é, o estimador $\hat{\beta}_{lasso}$ terá uma ou mais coordenadas nulas. Essa é uma propriedade das mais importantes do método, pois, além de uma estimação, o método também seleciona covariáveis, uma vez que as coordenadas que se anulam em $\hat{\beta}_{lasso}$ implicam que as covariáveis correspondentes não terão efeito na equação de predição. Essa face singular é a que se encontra mais próxima do estimador de quadrados mínimos. Para esclarecer a ideia, o caso bidimensional está ilustrado, na Figura 17, onde podemos observar a região hachurada que corresponde à região principal, ilustrada no primeiro quadrante, observe que a região depende do parâmetro $t$.

Figura 17 Estimador Lasso no caso ortogonal.




Fonte:Do autor (2019).

Passemos a exibir o estimador $\hat{\beta}_{lasso}$ no caso em que $\hat{\beta}_{ols}$ se projeta, ortogonalmente, num hiperplano do convexo. Vamos considerar o vetor de sinais $v = (v_1, v_2, \ldots, v_p)' \in \mathbb{R}^p$ do estimador de mínimos quadrados como,

$$v_i = \begin{cases} +1 & , \text{se} \left(\hat{\beta}_{ols}\right)_i \geq 0 \\ -1 & , \text{se} \left(\hat{\beta}_{ols}\right)_i < 0 \end{cases} , \forall i = 1, \ldots, p$$

Inicialmente, faremos um tratamento, usando otimização convexa. Nesse caso, observe que $\hat{\beta}_{lasso}$ tem os mesmos sinais do estimador de mínimos quadrados $\hat{\beta}_{ols}$, além disso, podemos escrever $\|\beta\|_1 = v'\beta = \beta' v$.

De fato, precisamos minimizar a função convexa $f(\beta) = \|y - X\beta\|^2$, restrita ao conjunto convexo $\|\beta\|_1 \leq t$, o que caracteriza um problema de otimização convexa, conforme definimos anteriormente. Assim, de acordo com a equação (3.2), fazendo $\alpha = 1$ e $X'X = I$, obtemos

$$\begin{cases} \beta = X'y - \frac{\lambda}{2}v \\ v'\beta = t \end{cases}.$$

Como estamos no caso ortogonal, podemos reescrever na forma,

$$\beta = X'y - \tfrac{\lambda}{2}v = \hat{\beta}_{ols} - \tfrac{\lambda}{2}v.$$

Assim, precisamos agora, visando a obter o estimador Lasso, expressar $\lambda$ em função de $t$. Com efeito

$$v'\beta = t \Rightarrow v'\left[\hat{\beta}_{ols} - \tfrac{\lambda}{2}v\right] = v'\hat{\beta}_{ols} - \tfrac{\lambda}{2}v'v = v'\hat{\beta}_{ols} - \tfrac{\lambda}{2}\|v\|^2 = t.$$

Assim, $-\frac{\lambda}{2} = \frac{t - v'\hat{\beta}_{ols}}{\|v\|^2} \Leftrightarrow \lambda = -2\left(\frac{t - v'\hat{\beta}_{ols}}{\|v\|^2}\right)$, e, portanto, podemos obter a expressão final para o estimador Lasso, como

$$\hat{\beta}_{lasso}(t,v) = \hat{\beta}_{ols} + \left(\tfrac{t - v'\hat{\beta}_{ols}}{p}\right) v.$$

Cabe aqui uma observação interessante. Com a variação do parâmetro $t$, o estimador $\hat{\beta}_{lasso}$ descreve uma reta que passa por $\hat{\beta}_{ols}$ na direção do vetor de sinais $v$. Observe que, no processo, iniciamos com o parâmetro valendo $t_0 = \left\|\hat{\beta}_{ols}\right\|_1$ e vamos diminuindo o valor do parâmetro até que uma das coordenadas do estimador Lasso se anule, após o que não haverá projeção ortogonal no convexo.

Vamos discutir, brevemente o caso ortogonal para a situação de duas covariáveis. Usaremos, para isso, a fórmula do Lasso para o caso ortogonal

$$\hat{\beta}_{lasso}(t,v) = \hat{\beta}_{ols} + \left(\tfrac{t - v'\hat{\beta}_{ols}}{p}\right) v.$$

Note que para $p = 2$, assumindo que $\hat{\beta}_{ols} = \left(\hat{\beta}_1^0, \hat{\beta}_2^0\right)$ pertença ao primeiro quadrante, podemos escrever

$$\hat{\beta}_{lasso}(t,v) = \hat{\beta}_{ols} + \left(\tfrac{t - v'\hat{\beta}_{ols}}{p}\right) v = \begin{pmatrix} \hat{\beta}_1^0 \\ \hat{\beta}_2^0 \end{pmatrix} + \left(\tfrac{t - (\hat{\beta}_1^0 + \hat{\beta}_2^0)}{2}\right) \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} \hat{\beta}_1^0 + \frac{t - (\hat{\beta}_1^0 + \hat{\beta}_2^0)}{2} \\ \hat{\beta}_2^0 + \frac{t - (\hat{\beta}_1^0 + \hat{\beta}_2^0)}{2} \end{pmatrix} = \begin{pmatrix} \frac{t}{2} + \frac{\hat{\beta}_1^0 - \hat{\beta}_2^0}{2} \\ \frac{t}{2} - \frac{\hat{\beta}_1^0 - \hat{\beta}_2^0}{2} \end{pmatrix}$$

Dessa forma, "o estimador Lasso pode ser dado por uma única relação" (TIBISHIRANI, 1996, p.272). Nas situações em que $\hat{\beta}_{ols}$ se encontra em outros quadrantes, a análise segue de forma análoga bastando alterar o vetor de sinais adequadamente. Passemos a analisar, geometricamente, o problema. Para isso, considere a equação paramétrica da reta $r(s)$ que passa por $\hat{\beta}_{ols}$ na direção do vetor $v$, a saber

$$r(s) = \hat{\beta}_{ols} + sv, \text{ em que } s < 0,$$

Desejamos determinar $s$, de modo que $r(s)$ pertença ao hiperplano $v'\beta = t$. Assim, podemos escrever

$$\begin{aligned} v'r(s) = t \Leftrightarrow v'\left(\hat{\beta}_{ols} + sv\right) = t \Leftrightarrow v'\hat{\beta}_{ols} + sv'v = t \\ \Leftrightarrow v'\hat{\beta}_{ols} + s\|v\|^2 = t \Leftrightarrow s = \frac{t - v'\hat{\beta}_{ols}}{\|v\|^2} \Leftrightarrow s = \frac{t - v'\hat{\beta}_{ols}}{p}. \end{aligned}$$

Substituindo o valor de $s$ encontrado na equação paramétrica da reta, obtemos a mesma expressão para o estimador Lasso,

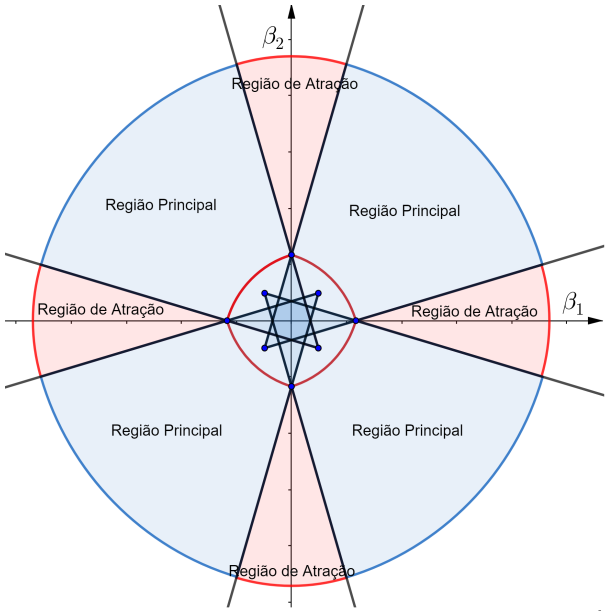
$$\hat{\beta}_{lasso}(t,v) = \hat{\beta}_{ols} + \left(\tfrac{t - v'\hat{\beta}_{ols}}{p}\right) v.$$

Note que, de maneira análoga à seção anterior do caso ortogonal do Ridge, independente do procedimento, obtemos a mesma expressão para o estimador Lasso, como era esperado.

### 3.3.4 O estimador Elastic Net no caso ortogonal

Nesse caso, iremos considerar o vetor de sinais $v = (v_1, v_2, \ldots, v_p)' \in \mathbb{R}^p$ do estimador de mínimos quadrados, o qual tem o mesmo sinal dos três estimadores. O estimador Elastic Net é obtido, pela projeção ortogonal do estimador de quadrados mínimos no convexo $(1-\alpha)\|\beta\|^2 + \alpha\|\beta\|_1 \leq t$. Devemos salientar que, fixado o valor do parâmetro $t$, existem no exterior do convexo duas regiões distintas, aquela que denominaremos de "região principal", em que o vetor $\hat{\beta}_{ols}$ se projeta, ortogonalmente, numa hiperface do convexo, ou na região complementar, que denominaremos de "região de atração", em que não há projeção ortogonal no convexo, analogamente ao procedimento utilizado para o Lasso no caso ortogonal. Nesse último caso, o estimador $\hat{\beta}_{en}$ ocorrerá em uma face singular (uma ou mais coordenadas nulas) mais próxima do estimador de quadrados mínimos. Para esclarecer a ideia, o caso bidimensional está ilustrado na Figura 18. Note que podemos observar a região hachurada (em azul) que corresponde à região principal. Aqui, vale ressaltar que as regiões se modificam em função do parâmetro $t$, todavia mantém a forma semelhante.

Figura 18 Região Principal no caso ortogonal (Elastic Net).




Fonte: Do autor (2019).

Passemos a exibir o estimador $\hat{\beta}_{en}$ no caso em que $\hat{\beta}_{ols}$ se projeta ortogonalmente na hiperface do convexo.

Primeiramente, com o enfoque de otimização convexa, precisamos minimizar uma função convexa $f(\beta) = \|y - X\beta\|^2$ restrita ao conjunto convexo $(1-\alpha)\|\beta\|^2 + \alpha\|\beta\|_1 = t$, em que $\alpha \in [0,1]$ , o que caracteriza um problema de otimização convexa conforme definimos anteriormente.

Assim, de acordo com a equação (3.2), fazendo $X'X = I$, obtemos

$$\begin{cases} \beta = [1+\lambda(1-\alpha)]^{-1}\left(\hat{\beta}_{ols} - \frac{\lambda\alpha}{2}v\right) \\ (1-\alpha)\|\beta\|^2 + \alpha\|\beta\|_1 = t \end{cases}$$

Assim, precisamos agora, visando a obter o estimador Lasso, expressar $\lambda$ em função de $t$. Observe que, nesse caso, não parece trivial essa tarefa. Contudo, de acordo com a seção (3.1), temos para $\alpha \neq 1$, que

$$(1-\alpha)\|\beta\|^2 + \alpha\|\beta\|_1 = t \Leftrightarrow \|\beta - C\| = R,$$

em que, o centro $C$ e o raio $R$, são dados por

$$C = -\tfrac{\alpha}{2(1-\alpha)}v \ \text{ e } \ \mathrm{R} = \sqrt{\tfrac{t}{1-\alpha} + \left(\tfrac{\alpha}{2(1-\alpha)}\right)^2 p}.$$

Assim podemos escrever mais concisamente o sistema a ser resolvido na forma,

$$\begin{cases} \beta = \frac{1}{1+\lambda(1-\alpha)}\left(\hat{\beta}_{ols} - \frac{\lambda}{2}\alpha v\right) \\ \|\beta - C\| = R \end{cases}$$

Note que a nova equação da restrição nos indica o caminho de como escrever $\lambda$ em função do parâmetro $t$. Manipulando, algebricamente, de maneira adequada, podemos escrever:

$$\begin{aligned} \beta - C &= \tfrac{1}{1+\lambda(1-\alpha)}\left(\hat{\beta}_{ols} - \tfrac{\lambda}{2}\alpha v\right) - C \\ &= \tfrac{1}{1+\lambda(1-\alpha)}\hat{\beta}_{ols} - \tfrac{\lambda\alpha}{2[1+\lambda(1-\alpha)]}v + \tfrac{\alpha}{2(1-\alpha)}v \\ &= \tfrac{1}{1+\lambda(1-\alpha)}\hat{\beta}_{ols} + \left[\tfrac{1}{2(1-\alpha)} - \tfrac{\lambda}{2[1+\lambda(1-\alpha)]}\right]\alpha v \\ &= \tfrac{1}{1+\lambda(1-\alpha)}\hat{\beta}_{ols} + \left[\tfrac{1+\lambda(1-\alpha)-\lambda(1-\alpha)}{2(1-\alpha)[1+\lambda(1-\alpha)]}\right]\alpha v \\ &= \tfrac{1}{1+\lambda(1-\alpha)}\hat{\beta}_{ols} + \left[\tfrac{1}{2(1-\alpha)[1+\lambda(1-\alpha)]}\right]\alpha v \\ &= \tfrac{1}{1+\lambda(1-\alpha)}\left[\hat{\beta}_{ols} + \tfrac{\alpha}{2(1-\alpha)}v\right] \\ &= \tfrac{1}{1+\lambda(1-\alpha)}\left[\hat{\beta}_{ols} - C\right] \end{aligned}$$

Dessa forma, tem-se que:

$$\beta - C = \tfrac{1}{1+\lambda(1-\alpha)}\left[\hat{\beta}_{ols} - C\right] \ .$$

Submetendo essa última relação à restrição, visando a isolar $\lambda$ em função do parâmetro $t$, obtemos:

$$\|\beta - C\| = \left\|\tfrac{1}{1+\lambda(1-\alpha)}\left[\hat{\beta}_{ols} - C\right]\right\| = \tfrac{1}{1+\lambda(1-\alpha)}\left\|\hat{\beta}_{ols} - C\right\|.$$

Daí resulta que,

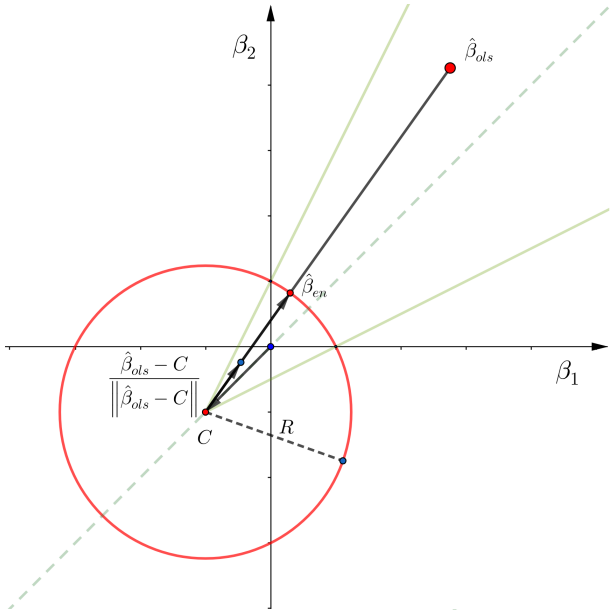
$$1 + \lambda(1 - \alpha) = \frac{\|\hat{\beta}_{ols} - C\|}{R}$$

Dessa forma, de posse da expressão obtida acima, podemos escrever a expressão explícita do estimador Elastic Net obtida por otimização convexa no caso ortogonal, para $\alpha \neq 1$, que é dado por:

$$\hat{\beta}_{en} = C + R\,\frac{\hat{\beta}_{ols} - C}{\|\hat{\beta}_{ols} - C\|}$$

Todo esse processo está descrito, geometricamente, na figura 19.

Figura 19 Estimador Elastic Net no caso ortogonal.




Fonte: Do autor (2019).

## 3.4 Análise dos estimadores de encolhimento Ridge, Lasso e Elastic Net para o caso não ortogonal.

Nesta seção, vamos analisar os estimadores Ridge, Lasso e Elastic Net no caso geral. Para isso, considere o modelo de regressão linear $y = X\beta + \varepsilon$, em que $X$ é uma matriz, $n \times p$, definida pelos valores das covariáveis, com posto $r(X) = p = \min(n,p)$ e $\beta$ é um vetor $p$-dimensional de parâmetros, já considerando o modelo centralizado.

### 3.4.1 O estimador Ridge

O estimador Ridge é obtido pela projeção ortogonal, na norma de Mahalanobis, $\langle u,v\rangle = u'X'Xv$, $\forall u,v \in \mathbb{R}^p$, do estimador de quadrados mínimos no convexo $\|\beta\|^2 \leq t$.

Nesse caso, temos que minimizar $f(\beta) = \|y - X\beta\|^2$, sujeito à restrição $\|\beta\|^2 \leq t$. Aqui podemos ser mais precisos, como vimos no caso ortogonal, as curvas de nível de $f$ são hiperelipsóides e devido a este fato, o mínimo ocorrerá no conjunto mais restritivo $\|\beta\|^2 = t$, isto é, na fronteira do convexo. Esse problema caracteriza um problema de otimização convexa, conforme definimos anteriormente. Assim, de acordo com a equação (3.2), fazendo $\alpha = 0$, obtemos o sistema a ser resolvido dado por,

$$\begin{cases} \beta = (X'X + \lambda I)^{-1}X'y \\ \|\beta\|^2 = t \end{cases}.$$

Para finalizar a expressão do estimador Ridge, devemos exibir o parâmetro $\lambda$ em função do parâmetro $t$. Todavia, no caso não ortogonal, essa tarefa não é simples. No entanto, podemos fazer uso do teorema da decomposição espectral para escrever a matriz simétrica $X'X$, na forma

$$X'X = P\Lambda P^{-1} = P\Lambda P',$$

em que $P$ é uma matriz de mudança de coordenadas ortonormal, cujas colunas são os autovetores (ortonormais) de $X'X$, e portanto $P^{-1} = P'$ , em que $\Lambda$ é uma matriz diagonal, de ordem $p$, cujos elementos da diagonal são os autovalores de $X'X$, digamos que $\Lambda = diag(\lambda_1,\lambda_2,\ldots,\lambda_p)$. Dessa forma, podemos escrever

$$X'X + \lambda I = P\Lambda P' + P\lambda I P' = P(\Lambda + \lambda I)P',$$

Assim, podemos escrever que

$$\begin{aligned} \beta &= [X'X + \lambda I]^{-1}X'y = [X'X + \lambda I]^{-1}X'X(X'X)^{-1}X'y \\ &= [X'X + \lambda I]^{-1}X'X\hat{\beta}_{ols} = [P(\Lambda + \lambda I)P']^{-1}P\Lambda P'\hat{\beta}_{ols} \\ &= P(\Lambda + \lambda I)^{-1}P'P\Lambda P'\hat{\beta}_{ols} = P(\Lambda + \lambda I)^{-1}\Lambda P'\hat{\beta}_{ols} \\ &= PD_0P'\hat{\beta}_{ols} \end{aligned}$$

em que,

$$D_0 = \begin{pmatrix} \frac{\lambda_1}{\lambda_1+\lambda} & 0 & 0 & 0 \\ 0 & \frac{\lambda_2}{\lambda_2+\lambda} & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{\lambda_p}{\lambda_p+\lambda} \end{pmatrix}$$

Note que submetendo à restrição, tem-se

$$\begin{aligned} \|\beta\|^2 = t \Leftrightarrow \beta'\beta = t \Leftrightarrow \left(PD_0P'\hat{\beta}_{ols}\right)' \left(PD_0P'\hat{\beta}_{ols}\right) = t \\ \Leftrightarrow \left(P'\hat{\beta}_{ols}\right)' D_0^2 \left(P'\hat{\beta}_{ols}\right) = t \end{aligned}$$

No entanto, nesse caso, não sendo possível escrever $\lambda$ em função de $t$, recursos computacionais são necessários para resolver o problema, conforme observado em (HOERL; KENNARD, 1970). Isso conclui nossa análise no caso Ridge.

### 3.4.2 O estimador Lasso

Nesta seção, vamos analisar o estimador Lasso (Least Absolute Shrinkage and Selection Operator) no caso mais geral. Para isso, considere o modelo de regressão linear $y = X\beta + \varepsilon$, em que $X$ é uma matriz, $n \times p$, definida pelos valores das covariáveis, com posto $r(X) = p = \min(n,p)$ e $\beta$ é um vetor $p$-dimensional de parâmetros, já considerando o modelo centralizado. Pretendemos exibir a equação de uma reta que à medida que o parâmetro t varia, ela fornece o estimador Lasso, em um intervalo apropriado que veremos mais adiante. Esse procedimento se diferencia do procedimento adotado por Tibshirani (1996).

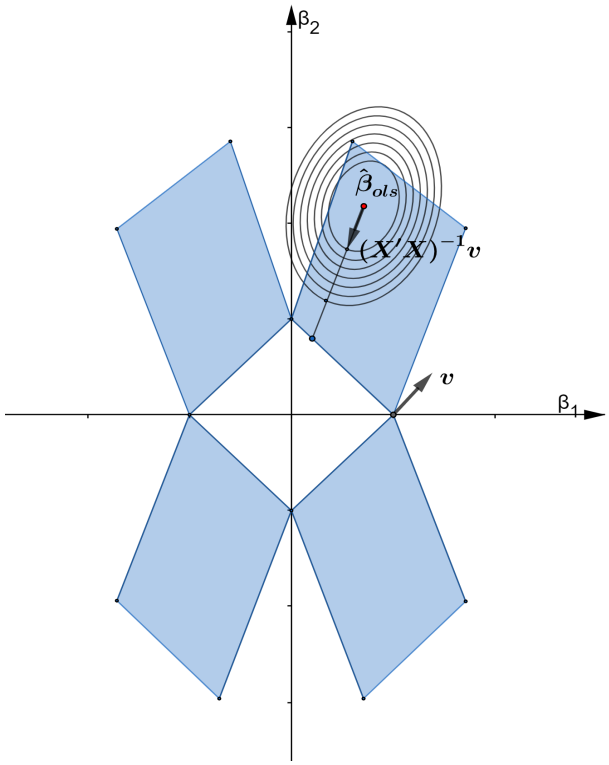
O estimador Lasso é obtido pela projeção ortogonal, na métrica de Mahalanobis, $\langle u,v\rangle = u'X'Xv$, para todo $u,v \in \mathbb{R}^p$, do estimador de quadrados mínimos no convexo $\|\beta\|_1 \leq t$, mais precisamente, na fronteira desse convexo, $\|\beta\|_1 = t$. Devemos salientar que, no exterior do convexo, fixado o parâmetro $t$, existem duas regiões distintas, aquela que denominaremos de "região principal", em que o vetor $\hat{\beta}_{ols}$ se projeta, ortogonalmente, na métrica de Mahalanobis, num hiperplano do convexo, ou na região complementar, que denominaremos de "região de atração", em que não há projeção ortogonal no convexo, analogamente ao procedimento utilizado no caso Lasso, no caso ortogonal. Nesse último caso, o estimador $\hat{\beta}_{lasso}$ estará em uma face singular mais próximo do estimador de quadrados mínimos. Para esclarecer a ideia, o caso bidimensional está ilustrado, na Figura 20, onde podemos observar a região hachurada que corresponde à região principal, que varia em função do parâmetro $t$.

Passemos a exibir o estimador $\hat{\beta}_{lasso}$ no caso em que $\hat{\beta}_{ols}$ se projeta, ortogonalmente, num hiperplano do convexo. Para analisarmos essa restrição, visando à generalização, vamos considerar o vetor de sinais $v$. Assim, de acordo com a equação (3.2), fazendo $\alpha = 1$, obtemos

$$\begin{cases} \beta = (X'X)^{-1}\left(X'y - \frac{\lambda v}{2}\right) = \hat{\beta}_{ols} - \frac{\lambda}{2}(X'X)^{-1}v \\ v'\beta = \beta' v = t \end{cases}.$$

Assim, precisamos agora, visando a obter o estimador Lasso, expressar $\lambda$ em função de $t$, com efeito

Figura 20 Estimador Lasso no caso não ortogonal.




Fonte: Do autor (2019).

$$
\begin{aligned}
v'\beta = t &\Leftrightarrow v'\left[\hat{\beta}_{ols} - \frac{\lambda}{2}(X'X)^{-1}v\right] = t \\
&\Leftrightarrow v'\hat{\beta}_{ols} - \frac{\lambda}{2}v'(X'X)^{-1}v = t \\
&\Leftrightarrow -\frac{\lambda}{2}v'(X'X)^{-1}v = t - v'\hat{\beta}_{ols} \\
&\Leftrightarrow -\frac{\lambda}{2} = \frac{t - v'\hat{\beta}_{ols}}{v'(X'X)^{-1}v} \\
&\Leftrightarrow \lambda = -2\frac{t - v'\hat{\beta}_{ols}}{v'(X'X)^{-1}v}.
\end{aligned}
$$

Assim, obtemos a expressão final para o estimador Lasso como

$$
\hat{\beta}_{lasso}(t,v) = \hat{\beta}_{ols} + \left(\frac{t - v'\hat{\beta}_{ols}}{v'(X'X)^{-1}v}\right)(X'X)^{-1}v. \tag{3.5}
$$

Observe que com a variação do parâmetro $t$, o estimador $\hat{\beta}_{lasso}$ descreve uma reta que passa por $\hat{\beta}_{ols}$

na direção do vetor $(X'X)^{-1}v$.

Passemos a analisar, geometricamente, o problema. Para isso, considere as funções

$$g(\beta) = (\beta - \hat{\beta}_{ols})'X'X(\beta - \hat{\beta}_{ols}) \quad \text{e} \quad h(\beta) = \beta'v - t.$$

Portanto, por diferenciação, obtemos os gradientes.

$$\begin{cases} \nabla g(\beta) = 2X'X(\beta - \hat{\beta}_{ols}) \\ \nabla h(\beta) = v \end{cases}.$$

Assim, no ponto de tangência, esses vetores devem ser paralelos, ou equivalentemente, existe $\gamma \in \mathbb{R}$ tal que

$$\nabla g(\beta) = \gamma \, \nabla h(\beta).$$

Isso implica que devemos ter,

$$\begin{aligned} 2X'X(\beta - \hat{\beta}_{ols}) = \gamma v \Leftrightarrow \beta - \hat{\beta}_{ols} &= \frac{\gamma}{2}(X'X)^{-1}v \\ \Leftrightarrow \beta &= \hat{\beta}_{ols} + \frac{\gamma}{2}(X'X)^{-1}v. \end{aligned}$$

Devemos agora determinar $\gamma$ para que $\beta$ pertence ao convexo, isto é, devemos ter que $\beta$ pertença à hiperface do convexo, $v'\beta = t$. Com efeito,

$$\begin{aligned} v'\beta = t \Leftrightarrow v' \left[\hat{\beta}_{ols} + \frac{\gamma}{2}(X'X)^{-1}v\right] &= t \\ \Leftrightarrow v'\hat{\beta}_{ols} + \frac{\gamma}{2}v'(X'X)^{-1}v &= t \\ \Leftrightarrow \frac{\gamma}{2}v'(X'X)^{-1}v &= t - v'\hat{\beta}_{ols} \\ \Leftrightarrow \frac{\gamma}{2} &= \frac{t - v'\hat{\beta}_{ols}}{v'(X'X)^{-1}v}. \end{aligned}$$

Portanto, temos, finalmente, a expressão do estimador Lasso, a saber,

$$\hat{\beta}_{lasso}(t,v) = \hat{\beta}_{ols} + \left(\frac{t - v'\hat{\beta}_{ols}}{v'(X'X)^{-1}v}\right)(X'X)^{-1}v. \tag{3.6}$$

Na Equação (3.6), pode-se notar que, após a observação dos dados $(X,y)$, a única quantidade desconhecida é o parâmetro não estatístico $t$ $(t \geq 0)$. Esse parâmetro é conhecido como parâmetro de *Tunning*, ou seja, um parâmetro de ajuste do modelo.

Primeiramente, devemos observar que a fórmula (3.6) descreve a equação paramétrica de uma reta que passa por $\hat{\beta}_{ols}$ na direção do vetor $(X'X)^{-1}v$.

Note também que, na realidade, a fórmula (3.6) deve ser analisada apenas no hiperoctante que tem $v$ como seu vetor de sinais, e portanto, trata-se de um segmento de reta. É importante observar que em cada hiperoctante que tem $v$ como seu vetor de sinais, existe uma e apenas uma face do convexo. Essa face, por construção, terá $v$ como seu vetor normal, cuja equação é dada por $\|\beta\|_1 = \beta'v = v'\beta = t$.

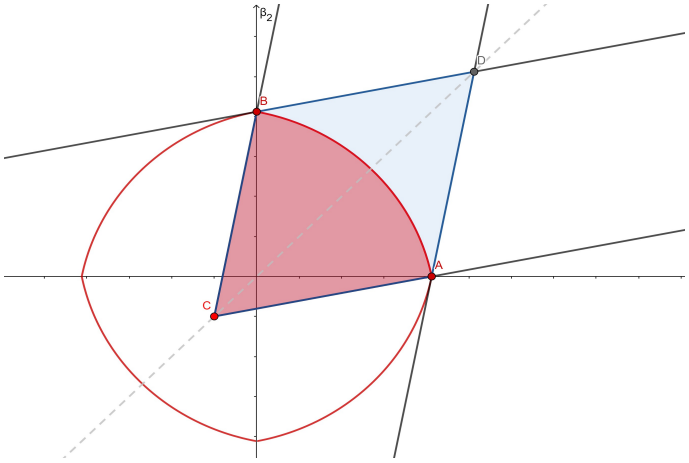
Vamos definir como "face estendida" o hiperplano que contém a face, exceto a própria face. Para decidirmos se um ponto pertencente ao hiperplano pertence à face ou à face estendida, é suficiente para estar na face que o ponto tenha os mesmos sinais de $v$. Outra forma de se testar se o ponto de fato pertence à sua face é verificar se satisfaz à restrição, ou seja, se $\hat{\beta}_{lasso}(t,v) \leq t$.

### 3.4.3 O estimador Elastic Net

Nesta seção, vamos analisar o estimador Elastic Net no caso não ortogonal. Para isso, considere o modelo de regressão linear $y = X\beta+\varepsilon$, em que $X$ é uma matriz $n\times p$, definida pelos valores das covariáveis, com posto $r(X) = p = \min(n,p)$, e $\beta$ é um vetor $p$-dimensional de parâmetros, já considerando o modelo centralizado.

O estimador Elastic Net é obtido pela projeção ortogonal, na métrica de Mahalanobis, $\langle u,v\rangle = u'X'Xv$, $\forall u,v \in \mathbb{R}^p$, do estimador de quadrados mínimos no convexo $(1-\alpha)\|\beta\|^2 + \alpha\|\beta\|_1 = t$. Devemos salientar que, no exterior do convexo, fixado o parâmetro $t$, existem duas regiões distintas, aquela que denominaremos de "região principal", em que o vetor $\hat{\beta}_{ols}$ se projeta, ortogonalmente, nesta norma, numa hiperface do convexo, e uma região complementar que denominaremos de "região de atração", em que não há projeção ortogonal no convexo. Para esclarecer a ideia, o caso bidimensional está ilustrado, na Figura 21, onde podemos observar a região hachurada (em azul) que corresponde à região principal ilustrada no complementar que denominaremos primeiro quadrante que, apesar de variar em função do parâmetro $t$, é uma região bem definida.

Figura 21 Região Principal no caso não ortogonal (Elastic Net).




Fonte: Do autor (2019).

Passemos a exibir o estimador $\hat{\beta}_{en}$ no caso em que $\hat{\beta}_{ols}$ se projeta, ortogonalmente, num hiperplano do convexo. Para analisarmos essa restrição, visando à generalização vamos considerar o vetor de sinais $v$. Assim, de acordo com a equação (3.2), obtemos o sistema

$$\begin{cases} \beta = (X'X + \lambda(1-\alpha)I)^{-1}\left(X'y - \frac{\lambda\alpha v}{2}\right) \\ (1-\alpha)\|\beta\|^2 + \alpha\|\beta\|_1 = t \end{cases}.$$

Observando que

$$\begin{aligned}(X'X+\lambda(1-\alpha)I)^{-1} &= \left(X'X+\lambda(1-\alpha)(X'X)(X'X)^{-1}\right)^{-1}\\ &= \left(X'X\left[I+\lambda(1-\alpha)(X'X)^{-1}\right]\right)^{-1}\\ &= \left[I+\lambda(1-\alpha)(X'X)^{-1}\right]^{-1}(X'X)^{-1},\end{aligned}$$

podemos escrever o sistema na forma mais apropriada

$$\begin{cases} \beta = \left(I+\lambda(1-\alpha)(X'X)^{-1}\right)^{-1}\left(\hat{\beta}_{ols}-\frac{\lambda\alpha}{2}(X'X)^{-1}v\right)\\ (1-\alpha)\|\beta\|^2+\alpha\|\beta\|_1 = t \end{cases}.$$

Note que, para exibir o estimador Elastic Net, há necessidade ainda de explicitar $\lambda$ em função de $t$. Nesse caso, essa tarefa não é possível, todavia dispomos do procedimento a seguir. De acordo com a seção (3.1), temos para $\alpha \neq 1$, que

$$(1-\alpha)\|\beta\|^2+\alpha\|\beta\|_1 = t \Leftrightarrow \|\beta - C\| = R,$$

em que, o centro $C$ e o raio $R$, são dados por

$$C = -\tfrac{\alpha}{2(1-\alpha)}v \ \text{ e } \ \mathrm{R} = \sqrt{\tfrac{t}{1-\alpha}+\left(\tfrac{\alpha}{2(1-\alpha)}\right)^2 p}.$$

Assim, podemos escrever mais concisamente o sistema a ser resolvido na forma,

$$\begin{cases} \beta = \left(I+\lambda(1-\alpha)(X'X)^{-1}\right)^{-1}\left(\hat{\beta}_{ols}-\frac{\lambda\alpha}{2}(X'X)^{-1}v\right)\\ \|\beta - C\| = R \end{cases}$$

Denotemos por $M = I+\lambda(1-\alpha)(X'X)^{-1}$. Assim, manipulando algebricamente de maneira adequada, podemos escrever:

$$\begin{aligned}\beta - C &= M^{-1}\left(\hat{\beta}_{ols}-\tfrac{\lambda\alpha}{2}(X'X)^{-1}v\right)-C\\ &= M^{-1}\hat{\beta}_{ols}-\tfrac{\lambda\alpha}{2}M^{-1}(X'X)^{-1}v+\tfrac{\alpha}{2(1-\alpha)}v\\ &= M^{-1}\hat{\beta}_{ols}-M^{-1}\tfrac{\lambda\alpha}{2}(X'X)^{-1}v+M^{-1}M\tfrac{\alpha}{2(1-\alpha)}v\\ &= M^{-1}\hat{\beta}_{ols}+M^{-1}\left[-\tfrac{\lambda}{2}(X'X)^{-1}+(I+\lambda(1-\alpha)(X'X)^{-1})\tfrac{1}{2(1-\alpha)}\right]\alpha v\\ &= M^{-1}\hat{\beta}_{ols}+M^{-1}\left[-\tfrac{\lambda}{2}(X'X)^{-1}+\tfrac{1}{2(1-\alpha)}I+\tfrac{\lambda}{2}(X'X)^{-1}\right]\alpha v\\ &= M^{-1}\hat{\beta}_{ols}+M^{-1}\tfrac{1}{2(1-\alpha)}\alpha v\\ &= M^{-1}\left[\hat{\beta}_{ols}+\tfrac{1}{2(1-\alpha)}\alpha v\right]\\ &= M^{-1}\left[\hat{\beta}_{ols}-C\right]\end{aligned}$$

Dessa forma, tem-se que:

$$\beta - C = \left[I + \lambda(1-\alpha)(X'X)^{-1}\right]^{-1}\left[\hat{\beta}_{ols} - C\right].$$

Note que com esse procedimento eliminamos um dos $\lambda$ na primeira equação, mantendo um único $\lambda$ na expressão de $M$.

Podemos resumir o que até agora obtivemos de forma simplificada no sistema,

$$\begin{cases} \beta - C = M^{-1}\left[\hat{\beta}_{ols} - C\right] \\ \|\beta - C\| = R \end{cases}$$

em que,

$$C = -\tfrac{\alpha}{2(1-\alpha)}v\,, \mathrm{R} = \sqrt{\tfrac{t}{1-\alpha} + \left(\tfrac{\alpha}{2(1-\alpha)}\right)^2 p} \quad \text{e} \quad M = I + \lambda(1-\alpha)(X'X)^{-1}.$$

Podemos fazer uso do teorema da decomposição espectral para escrever a matriz simétrica $X'X$, na forma

$$X'X = P\Lambda P^{-1} = P\Lambda P',$$

em que $P$ é uma matriz de mudança de coordenadas ortonormal, cujas colunas são os autovetores (ortonormais) de $X'X$, e $\Lambda$ é uma matriz diagonal, de ordem $p$, cujos elementos da diagonal são os autovalores de $X'X$, $\Lambda = diag(\lambda_1,\lambda_2,\ldots,\lambda_p)$. Dessa forma, podemos escrever

$$\begin{aligned} M &= I + \lambda(1-\alpha)(X'X)^{-1} = PIP' + \lambda(1-\alpha)(P\Lambda P')^{-1} \\ &= PIP' + P\lambda(1-\alpha)\Lambda^{-1}P' = P(I + \lambda(1-\alpha)\Lambda^{-1})P' \end{aligned}$$

Assim, podemos escrever que

$$\begin{aligned} \beta - C &= \left[I + \lambda(1-\alpha)(X'X)^{-1}\right]^{-1}(\hat{\beta}_{ols} - C) \\ &= \left[P(I + \lambda(1-\alpha)\Lambda^{-1})P'\right]^{-1}(\hat{\beta}_{ols} - C) \\ &= P\left[I + \lambda(1-\alpha)\Lambda^{-1}\right]^{-1}P'(\hat{\beta}_{ols} - C) \\ &= PD_\alpha P'(\hat{\beta}_{ols} - C) \end{aligned}$$

em que,

$$D_\alpha = \begin{pmatrix} \frac{\lambda_1}{\lambda_1+(1-\alpha)\lambda} & 0 & 0 & 0 \\ 0 & \frac{\lambda_2}{\lambda_2+(1-\alpha)\lambda} & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{\lambda_p}{\lambda_p+(1-\alpha)\lambda} \end{pmatrix}$$

Note que submetendo à restrição, tem-se

$$\|\beta - C\|^2 = R^2 \Leftrightarrow (\beta - C)'(\beta - C) = R^2 \Leftrightarrow \left[PD_\alpha P'(\hat{\beta}_{ols} - C)\right]' \left[PD_\alpha P'(\hat{\beta}_{ols} - C)\right] = R^2$$
$$\Leftrightarrow \left[P'(\hat{\beta}_{ols} - C)\right]' D_\alpha^2 \left[P'(\hat{\beta}_{ols} - C)\right] = R^2$$

No entanto, nesse caso, não sendo possível escrever $\lambda$ em função de $t$, recursos computacionais são necessários para resolver o problema. Note que, fazendo $\alpha = 0$, obtemos a mesma expressão obtida, na seção 3.4.1, observando que $C = -\frac{\alpha}{2(1-\alpha)}v = 0$. Isso conclui a nossa análise do estimador Elastic Net no caso mais geral.

### 3.4.4 Implementação da reta do estimador Lasso

Vamos iniciar esta seção, observando que pretendemos explicitar o estimador Lasso para todo $t \geq 0$. Note que para $t \geq t_0 = \left\|\hat{\beta}_{ols}\right\|_1$, teremos $\hat{\beta}_{lasso}(t,v) = \hat{\beta}_{ols}$. Assim, resta-nos explicitar o estimador no intervalo $0 \leq t < t_0 = \left\|\hat{\beta}_{ols}\right\|_1$.

O convexo $\|\beta\|_1 \leq t$ tem um número de faces igual ao número de hiperoctantes. Cada face define um hiperplano afim. Os vetores normais a esses hiperplanos são dados pelos vetores de sinal $v = (\pm 1, \pm 1, \ldots, \pm 1)'$ associados aos correspondentes hiperoctantes . Temos então $2^p$ direções possíveis para o vetor $v$. Observe que esses hiperplanos são 2 a 2 paralelos, e portanto, possuem os mesmos vetores normais. Assim, na realidade temos $2^{p-1}$ direções que, eventualmente, poderão participar da construção do estimador que será uma curva linear por partes. A ideia do algorítmo é a partir de $\hat{\beta}_{ols}$ construir uma reta para cada uma dessas direções $v$ utilizando a Fórmula do Lasso (3.6), obtendo a intercessão dessas retas com os hiperplanos. As soluções que nos interessam são aquelas que satisfazem à restrição $\|\beta\|_1 \leq t$.

A fórmula do Lasso (3.6) deve ser analisada, inicialmente, na direção dada pelo vetor de sinais de $\hat{\beta}_{ols}$. Após obtermos a expressão do estimador em função de $t$, em que todas as coordenadas são funções lineares de $t$, devemos resolver o sistema de equações $\hat{\beta}_{lasso}(t,v) = 0$ obtendo $p$ valores de $t$. Tomemos o maior valor de $t$, restrito ao intervalo $0 \leq t < t_0 = \left\|\hat{\beta}_{ols}\right\|_1$, que denotaremos por $\tilde{t}_1$. Assim, obteremos o estimador no intervalo $\tilde{t}_1 \leq t < t_0$.

Devemos agora analisar as $2^{p-1} - 1$ direções restantes restritas aos intervalo $0 \leq t < \tilde{t}_1$. Quando escolhermos uma direção distinta do vetor de sinais de $\hat{\beta}_{ols}$, devemos resolver o sistema de equações $\hat{\beta}_{lasso}(t,v) = 0$. Nessa etapa, obteremos $p$ valores de $t$, alguns negativos e outros positivos. Existe a possibilidade de obtermos algum valor nulo, mas isso ocorrerá com probabilidade zero. Devemos descartar os valores negativos e considerar apenas os valores positivos de $t$ tais que $\hat{\beta}_{lasso}(t,v)$ tenha o mesmo sinal de $v$, ou equivalentemente, $\|\beta\|_1 \leq t$. Isso nos garantirá que estamos na face do convexo.

Devido à continuidade do estimador, dentre as $2^{p-1} - 1$ direções restantes, teremos apenas duas possibilidades, ou teremos o conjunto vazio, caso em que a direção não participa da construção do estimador, ou um intervalo de valores de $t's$, caso em que essas direções participam da construção do estimador. Esses intervalos serão necessariamente disjuntos dois a dois, uma vez que os hiperelipsóides só podem tangenciar o convexo em uma face de cada vez.

Observe que só algumas direções deverão participar da construção do estimador. Uma questão que emerge é quantas e quais direções participam da construção do estimador. Após percorrermos todas as direções que geraram intervalos disjuntos devemos ordená-los pelo seu limite superior, por exemplo.

O próximo passo será unir a imagem desses segmentos, que serão segmentos de reta em $\mathbb{R}^p$, por

meio de segmentos de reta de $A$ a $B$ utilizando a fórmula,

$$L(t) = \left(1 - \frac{b-t}{b-a}\right) B + \frac{b-t}{b-a} A, \text{ no intervalo } a \leq t \leq b.$$

Assim, obteremos o estimador $\hat{\beta}_{lasso}(t,v)$ para todo $t \geq 0$. Esse processo será ilustrado em detalhes, por meio de um exemplo mais adiante.

Vale ressaltar que as direções que participam da construção do estimador são, de fato, soluções do problema variacional que define o estimador Lasso, essas soluções são descritas pela Fórmula do Lasso (3.6). Quando não tivermos tangência, teremos que o estimador toca o convexo numa aresta singular do mesmo.

### 3.4.5 Exemplo do estimador Lasso

Nesta seção, vamos apresentar um exemplo do processo de obtenção do estimador Lasso, usando a implementação da Fórmula do Lasso, para um exemplo extraído do Livro de Rencher e Shaalje (2008). Considere um conjunto de dados $(X,y)$ dados pelas matrizes.

$$y = \begin{pmatrix} 2 \\ 3 \\ 2 \\ 7 \\ 6 \\ 8 \\ 10 \\ 7 \\ 8 \\ 12 \\ 11 \\ 14 \end{pmatrix}, X = \begin{pmatrix} 1 & 0 & 2 \\ 1 & 2 & 6 \\ 1 & 2 & 7 \\ 1 & 2 & 5 \\ 1 & 4 & 9 \\ 1 & 4 & 8 \\ 1 & 4 & 7 \\ 1 & 6 & 10 \\ 1 & 6 & 11 \\ 1 & 6 & 9 \\ 1 & 8 & 15 \\ 1 & 8 & 13 \end{pmatrix}.$$

Inicialmente, vamos calcular o estimador de mínimos quadrados $\hat{\beta}_{ols} = (X'X)^{-1}X'y$. Para isso, devemos determinar as matrizes,

$$X'X = \begin{pmatrix} 12 & 52 & 102 \\ 52 & 296 & 536 \\ 102 & 536 & 1004 \end{pmatrix}, X'y = \begin{pmatrix} 90 \\ 482 \\ 872 \end{pmatrix},$$

$$(X'X)^{-1} = \begin{pmatrix} 0{,}97476 & 0{,}24290 & -0{,}22871 \\ 0{,}24290 & 0{,}16207 & -0{,}11120 \\ -0{,}22871 & -0{,}11120 & 0{,}08360 \end{pmatrix}.$$

Assim, obtemos

$$\hat{\beta}_{ols} = (X'X)^{-1}X'y = \begin{pmatrix} 5{,}3754 \\ 3{,}0118 \\ -1{,}2855 \end{pmatrix}.$$

O próximo passo é aplicar a fórmula

$$\hat{\beta}_{lasso}(t,v) = \hat{\beta}_{ols} + \left( \frac{t - v'\hat{\beta}_{ols}}{v'(X'X)^{-1}v} \right) (X'X)^{-1}v. \tag{3.7}$$

Note que a fórmula (3.7) é a equação paramétrica de uma reta em $\mathbb{R}^p$ que passa pelo ponto $\hat{\beta}_{ols}$ na direção do vetor $(X'X)^{-1}v$. Essa reta deve ser analisada apenas no hiperoctante que tem $v$ como seu vetor de sinais, observe que esse vetor $v$ é ortogonal à face do convexo contida no hiperoctante que tem $v$ como seu vetor de sinais. Note também que apenas uma face do convexo pertence a esse hiperoctante. Vamos determinar a expressão da fórmula (3.7) para cada $v$. A primeira direção a ser escolhida é sempre a do vetor de sinais de $\hat{\beta}_{ols}$. Vejamos:

- Análise na direção $v_5 = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$.

$$\rightarrow (X'X)^{-1}v_5 = \begin{pmatrix} 1{,}44637 \\ 0{,}51617 \\ -0{,}42351 \end{pmatrix}$$

$$\rightarrow {v'}_5(X'X)^{-1}v_5 = 2{,}38605$$

$$\rightarrow \left\|\hat{\beta}_{ols}\right\|_1 = {v'}_5\hat{\beta}_{ols} = 9{,}67270$$

$$\rightarrow \hat{\beta}_{lasso}(t,v_5) = \hat{\beta}_{ols} + \left(\frac{t-{v'}_5\hat{\beta}_{ols}}{{v'}_5(X'X)^{-1}v_5}\right)(X'X)^{-1}v_5 = \begin{pmatrix} 0{,}60618t - 0{,}48797 \\ 0{,}21633t + 0{,}91932 \\ -0{,}17749t + 0{,}43135 \end{pmatrix}$$

Fazendo $\hat{\beta}_{lasso}(t,v_5) = 0$, obtemos os respectivos valores de $t$, a saber

$$t_1 = 0{,}80499 \rightarrow \hat{\beta}_{lasso}(t_1,v_5) = \begin{pmatrix} 0 \\ 1{,}09346 \\ 0{,}28847 \end{pmatrix}$$

$$t_2 = -4{,}24962 \rightarrow \hat{\beta}_{lasso}(t_2,v_5) = \begin{pmatrix} -3{,}06400 \\ 0 \\ 1{,}18562 \end{pmatrix}$$

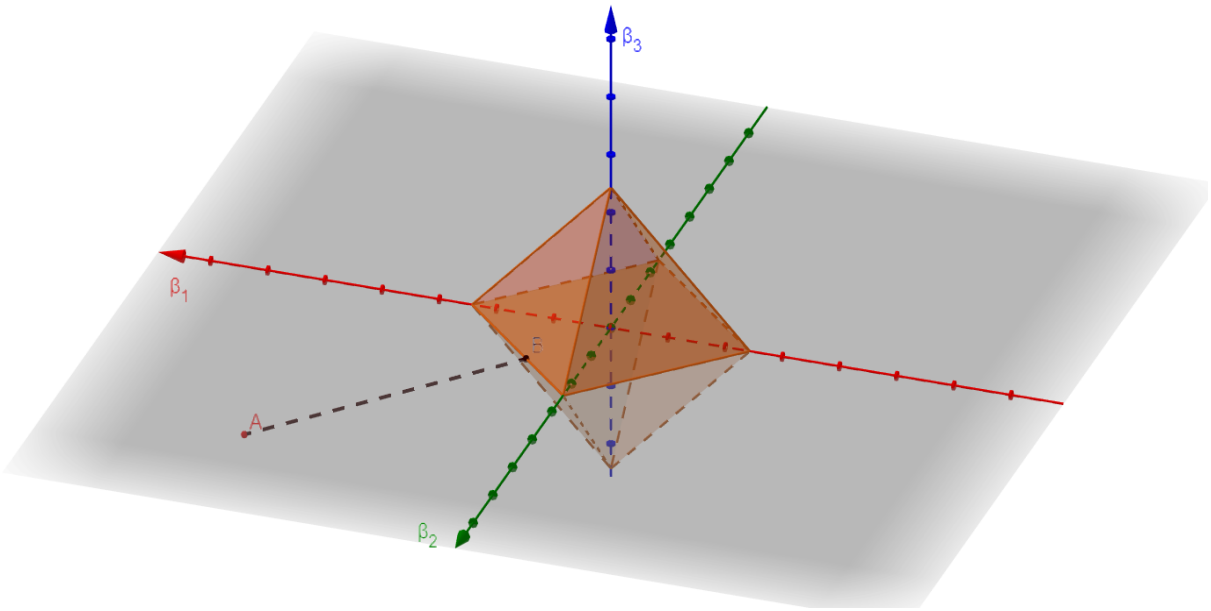
$$t_3 = 2{,}43028 \rightarrow \hat{\beta}_{lasso}(t_3,v_5) = \begin{pmatrix} 0{,}98522 \\ 1{,}44506 \\ 0 \end{pmatrix}$$

Devemos considerar apenas os valores do parâmetro $\mathrm{t} \geq 0$, tais que $\hat{\beta}_{lasso}(t,v_5) \leq t$. O que implicará que o ponto pertença à face correspondente do convexo e não à sua face estendida. Desse modo, apenas $\hat{\beta}_{lasso}(t_3,v_5)$ pertence à face correspondente.

Observe também que $v_5$ é o vetor de sinais de $\hat{\beta}_{ols}$, então nesse caso, devemos tomar o maior valor de $t$, tal que $0 < t < t_0 = \left\|\hat{\beta}_{ols}\right\|_1$, que no caso em questão é o valor $t_3 = 2{,}43028$.

Na Figura 22, ilustra-se a primeira pirâmide que, no caso, corresponde à primeira restrição, onde a trajetória do estimador muda de direção.

Figura 22 Trajetória inicial do estimador Lasso.




Fonte: Do autor (2019).

• Análise na direção $v_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$.

$$\rightarrow (X'X)^{-1}v_1 = \begin{pmatrix} 0{,}98895 \\ 0{,}29377 \\ -0{,}25631 \end{pmatrix}$$

$$\rightarrow v'_1(X'X)^{-1}v_1 = 1{,}02641$$

$$\rightarrow v'_1\hat{\beta}_{ols} = 7{,}10170$$

$$\rightarrow \hat{\beta}_{lasso}(t,v_1) = \hat{\beta}_{ols} + \left(\frac{t - v'_1\hat{\beta}_{ols}}{v'_1(X'X)^{-1}v_1}\right)(X'X)^{-1}v_1 = \begin{pmatrix} 0{,}96350t - 1{,}46712 \\ 0{,}28621t + 0{,}97921 \\ -0{,}24972t + 0{,}48790 \end{pmatrix}$$

Fazendo $\hat{\beta}_{lasso}(t,v_1) = 0$, obtemos os respectivos valores de $t$, a saber

$$t_1 = 1{,}52270 \rightarrow \hat{\beta}_{lasso}(t_1, v_1) = \begin{pmatrix} 0 \\ 1{,}41502 \\ 0{,}10765 \end{pmatrix}$$

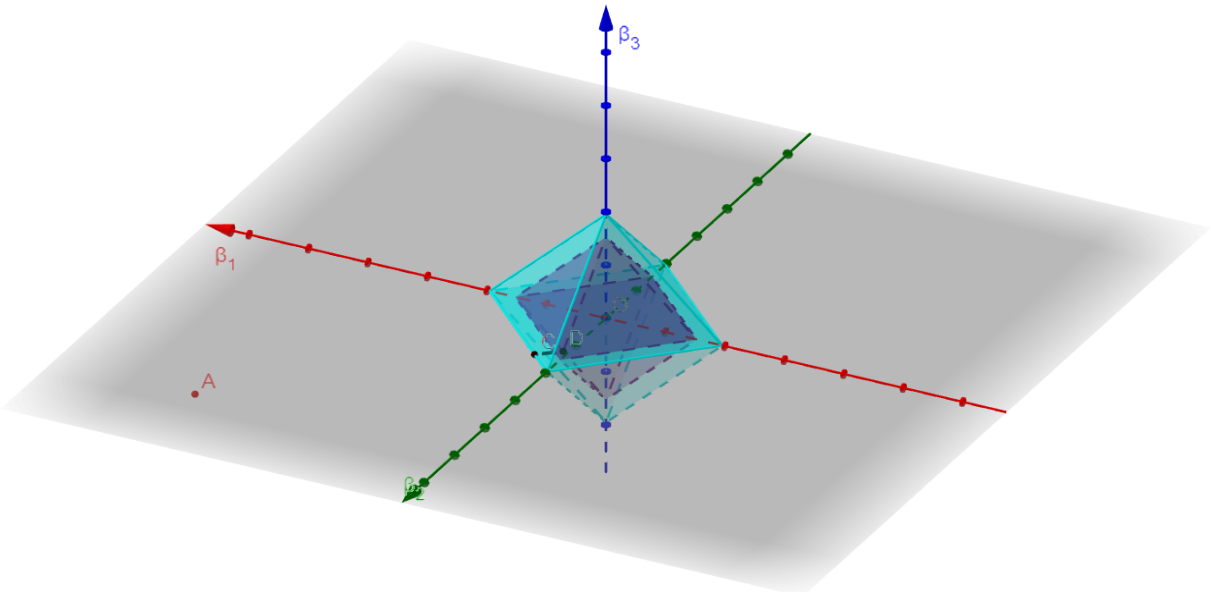
$$t_2 = -3{,}42130 \rightarrow \hat{\beta}_{lasso}(t_2, v_1) = \begin{pmatrix} -4{,}76354 \\ 0 \\ 1{,}34227 \end{pmatrix}$$

$$t_3 = 1{,}95379 \rightarrow \hat{\beta}_{lasso}(t_3, v_1) = \begin{pmatrix} 0{,}41536 \\ 1{,}53840 \\ 0 \end{pmatrix}$$

Novamente, vamos destacar que devemos considerar apenas os valores do parâmetro $0 \leq t \leq t_0 = \left\|\hat{\beta}_{ols}\right\|_1$, tais que $\hat{\beta}_{lasso}(t, v_1)$ tenha os mesmos sinais de $v_1$, o que implicará que o ponto pertença à face correspondente do convexo e não à sua face estendida. Desse modo, os pontos $\hat{\beta}_{lasso}(t_1, v_1)$ e $\hat{\beta}_{lasso}(t_3, v_1)$ pertencem à face correspondente.

Na Figura 23, ilustra-se a segunda e a terceira pirâmides onde a trajetória do estimador muda de direção.

Figura 23 Terceira etapa da trajetória do estimador Lasso.




Fonte: Do autor (2019).

• Análise na direção $v_2 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$.

$\rightarrow (X'X)^{-1}v_2 = \begin{pmatrix} -0{,}96057 \\ -0{,}19203 \\ 0{,}20111 \end{pmatrix}$

$\rightarrow v'_2(X'X)^{-1}v_2 = 0{,}96965$

$\rightarrow v'_2\hat{\beta}_{ols} = -3{,}64910$

$\rightarrow \hat{\beta}_{lasso}(t,v_2) = \hat{\beta}_{ols} + \left(\frac{t - v'_2\hat{\beta}_{ols}}{v'_2(X'X)^{-1}v_2}\right)(X'X)^{-1}v_2 = \begin{pmatrix} -0{,}99064t + 1{,}76047 \\ -0{,}19804t + 2{,}28913 \\ 0{,}20740t - 0{,}52866 \end{pmatrix}$

Fazendo $\hat{\beta}_{lasso}(t,v_2) = 0$, obtemos os respectivos valores de $t$, a saber

$$t_1 = 1{,}77710 \rightarrow \hat{\beta}_{lasso}(t_1,v_2) = \begin{pmatrix} 0 \\ 1{,}93719 \\ -0{,}16009 \end{pmatrix}$$

$$t_2 = 11{,}55893 \rightarrow \hat{\beta}_{lasso}(t_2,v_2) = \begin{pmatrix} -9{,}76047 \\ 0 \\ 1{,}86866 \end{pmatrix}$$

$$t_3 = 2{,}54899 \rightarrow \hat{\beta}_{lasso}(t_3,v_2) = \begin{pmatrix} -0{,}76047 \\ 1{,}78433 \\ 0 \end{pmatrix}$$

Nesse caso, nenhum ponto pertence à face.

• Análise na direção $v_3 = \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}$.

$\rightarrow (X'X)^{-1}v_3 = \begin{pmatrix} -1{,}44637 \\ -0{,}51617 \\ 0{,}42351 \end{pmatrix}$

$\rightarrow v'_3(X'X)^{-1}v_3 = 2{,}38605$

$\rightarrow v'_3\hat{\beta}_{ols} = -9{,}67270$

$\rightarrow \hat{\beta}_{lasso}(t,v_3) = \hat{\beta}_{ols} + \left(\frac{t - v'_3\hat{\beta}_{ols}}{v'_3(X'X)^{-1}v_3}\right)(X'X)^{-1}v_3 = \begin{pmatrix} -0{,}60618t - 0{,}48797 \\ -0{,}21633t + 0{,}91932 \\ 0{,}17749t + 0{,}43135 \end{pmatrix}$

Fazendo $\hat{\beta}_{lasso}(t,v_3) = 0$, obtemos os respectivos valores de $t$, a saber

$$t_1 = -0{,}80499 \rightarrow \hat{\beta}_{lasso}(t_1,v_3) = \begin{pmatrix} 0 \\ 1{,}09346 \\ 0{,}28847 \end{pmatrix}$$

$$t_2 = 4{,}24962 \rightarrow \hat{\beta}_{lasso}(t_2,v_3) = \begin{pmatrix} -3{,}06400 \\ 0 \\ 1{,}18562 \end{pmatrix}$$

$$t_3 = -2{,}43028 \rightarrow \hat{\beta}_{lasso}(t_3,v_3) = \begin{pmatrix} 0{,}98522 \\ 1{,}44506 \\ 0 \end{pmatrix}$$

Nesse caso, nenhum ponto pertence à face.

• Análise na direção $v_4 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$.

$$\rightarrow (X'X)^{-1}v_4 = \begin{pmatrix} 0{,}50315 \\ -0{,}03037 \\ -0{,}03391 \end{pmatrix}$$

$$\rightarrow v'_4(X'X)^{-1}v_4 = 0{,}49961$$

$$\rightarrow v'_4\hat{\beta}_{ols} = 1{,}07810$$

$$\rightarrow \hat{\beta}_{lasso}(t,v_4) = \hat{\beta}_{ols} + \left(\frac{t - v'_4\hat{\beta}_{ols}}{v'_4(X'X)^{-1}v_4}\right)(X'X)^{-1}v_4 = \begin{pmatrix} 1{,}00709t + 4{,}28966 \\ -0{,}06079t + 3{,}07733 \\ -0{,}06787t - 1{,}21233 \end{pmatrix}$$

Fazendo $\hat{\beta}_{lasso}(t,v_4) = 0$, obtemos os respectivos valores de $t$, a saber

$$t_1 = -4{,}25946 \rightarrow \hat{\beta}_{lasso}(t_1,v_4) = \begin{pmatrix} 0 \\ 3{,}33626 \\ -0{,}92324 \end{pmatrix}$$

$$t_2 = 50{,}62231 \rightarrow \hat{\beta}_{lasso}(t_2,v_4) = \begin{pmatrix} 55{,}27088 \\ 0 \\ -4{,}64807 \end{pmatrix}$$

$$t_3 = -17{,}86253 \rightarrow \hat{\beta}_{lasso}(t_3,v_4) = \begin{pmatrix} -13{,}69952 \\ 4{,}16319 \\ 0 \end{pmatrix}$$

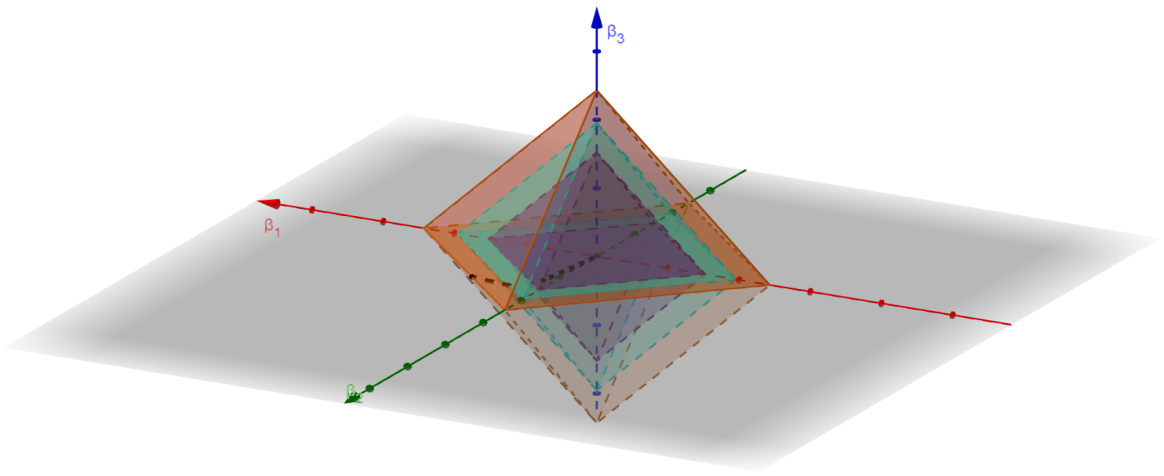
Nesse caso, nenhum ponto pertence à face.

Dessa forma, analisamos todas as faces que eventualmente tem ponto de tangência com os elipsóides e constatamos que apenas as faces correspondentes aos vetores $v_5$ e $v_1$ participam da construção do estimador. Devemos agora unir tais segmentos por segmentos de reta, segmentos esses que constituirão a segunda e quarta etapa do estimador em que não haverá pontos de tangência, utilizando a expressão

$$L(t) = [A,B] = \left(1 - \tfrac{b-t}{b-a}\right)B + \tfrac{b-t}{b-a}A,\ a \leq t \leq b.$$

Na Figura 24, ilustram-se os segmentos que unem os pontos de tangência obtidos nos passos anteriores.

Figura 24 Segmentos de reta unindo os pontos de tangência.



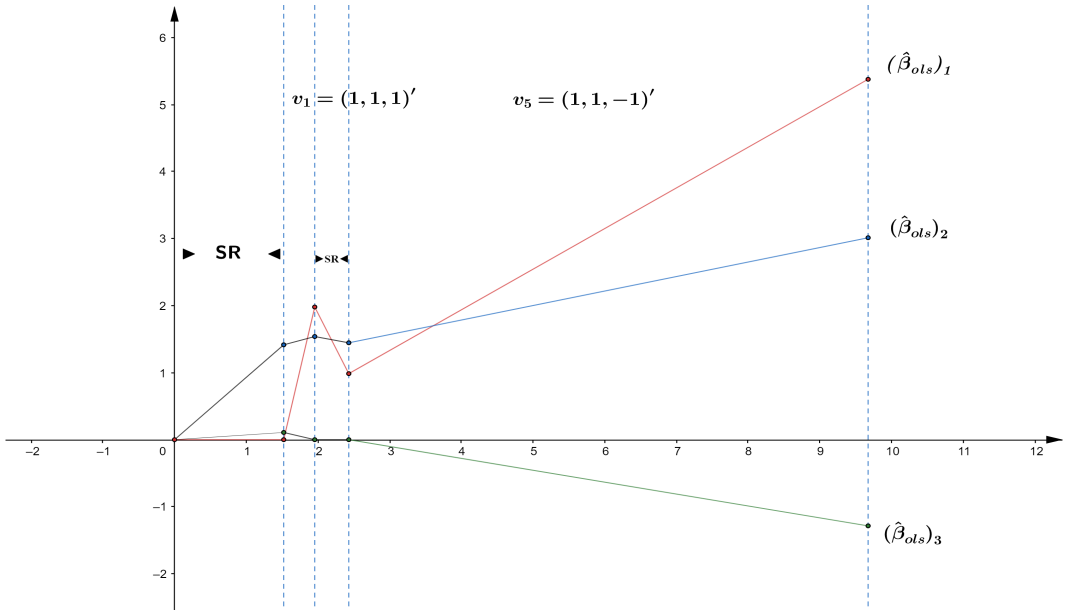

Fonte: Do autor (2019).

Dessa forma podemos finalmente exibir para cada valor do parâmetro não estatístico $t$, a expressão fechada do estimador Lasso nesse exemplo em particular.

$$\hat{\beta}_{lasso}(t) = \begin{cases} \begin{pmatrix} 5{,}3754 \\ 3{,}0118 \\ -1{,}2855 \end{pmatrix} = \hat{\beta}_{ols} & ; t \geq t_0 = \left\| \hat{\beta}_{ols} \right\|_1 = 9{,}67270 \\ \begin{pmatrix} 0{,}60618t - 0{,}48797 \\ 0{,}21633t + 0{,}91932 \\ -0{,}17749t + 0{,}43135 \end{pmatrix} & ; \tilde{t}_1 = 2{,}43028 \leq t < t_0 \\ \begin{pmatrix} 1{,}19996t - 1{,}93102 \\ -0{,}19655t + 1{,}92273 \\ 0 \end{pmatrix} & ; 1{,}955379 \leq t < 2{,}43028 \\ \begin{pmatrix} 0{,}96350t - 1{,}46712 \\ 0{,}28621t + 0{,}97921 \\ -0{,}24972t + 0{,}48790 \end{pmatrix} & ; 1{,}52270 \leq t \leq 1{,}955379 \\ \begin{pmatrix} 0 \\ 0{,}92928t \\ 0{,}07070t \end{pmatrix} & ; 0 \leq t < 1{,}52270. \end{cases}$$

A trajetória dos Lassotraces do estimador Lasso, $\hat{\beta}_{lasso}(t)$, está ilustrada na Figura 25 onde podemos observar o processo de encolhimento dos parâmetros. Usamos a notação $SR$ para descrever a parte da curva que unimos, por segmentos de reta. Observe que exibimos os LassoTraces em função do parâmetro $t$.

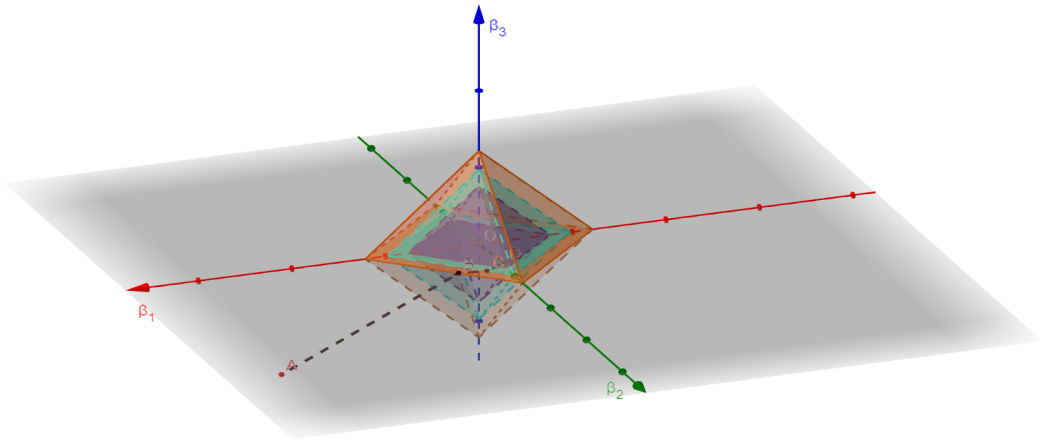
Figura 25 Trajetória dos Lasso Traces do estimador Lasso.




Fonte: Do autor (2019).

A trajetória do estimador $\hat{\beta}_{lasso}(t)$ está ilustrado, na Figura 26, onde podemos observar a presença de três pirâmides (restrições) onde a trajetória muda de direção.

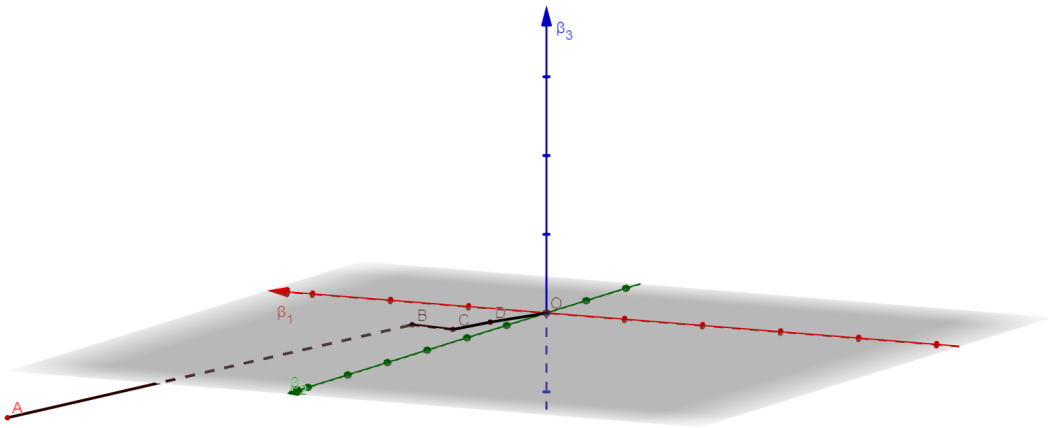
Figura 26 Trajetória do estimador Lasso e das pirâmides onde a direção muda.




Fonte: Do autor (2019).

Na Figura 27, ilustra-se a trajetória do estimador, no espaço, sem a presença das restrições (Pirâmides) em que o estimador muda de direção. No segmento de reta de $A$ a $B$ o estimador é tangente à face, assim como no segmento de $C$ a $D$. Já nos segmentos de reta de $B$ a $C$ e de $D$ a $O$ não há tangência.

Figura 27 Trajetória do estimador Lasso.




Fonte: Do autor (2019).

# 4 CONCLUSÃO

A geometria analítica do estimador Lasso não é, excessivamente, complexa e sendo explicitada facilita a sua compreensão. O estimador Elastic Net revelou relações com o estimador Ridge que é de construção mais simples.

Os estimadores Lasso e Elastic Net são apresentados, na literatura, como solução de um problema variacional. A abordagem geométrica a essa teoria é possível e, sem dúvida, contribui para a compreensão das propriedades inerentes a esses estimadores.

Pretendemos, na sequência do trabalho, fazer um estudo computacional do algoritmo proposto e obter mais propriedades do estimador, a partir da explicitação da geometria analítica.

# REFERÊNCIAS

BOYD, S.; VANDENBERGHE, L. **Convex Optimization**. New York: Cambridge University Press, 2008. 716 p.

COSTA, L. A. **Novo estimador de cumeeira de Rao com aplicação em seleção genômica.** Tese (Doutorado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras-MG, 2015, p. 126.

EFRON, B. The estimation of prediction error: covariance penalties and cross-validation. **Journal of the American Statistical Association**, v. 99, n. 467. p. 619-632, 2004.

GRUBER, M. H. J. **Improving efficiency by shrinkage: the James-Stein and ridge regression estimator**. 2nd ed. New York: M. Dekker, 1998. 632 p.

HOERL, A.E; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. **Thecnometrics**, v.12, n. 1, p. 51-67, 1970.

JAMES G.; WITTEN D.; HASTIE T. ; TIBISHIRANI, R. **An Introduction to Statistical Learning with Applications in R**, Springer Science Business Media New York, 2013.

PEREIRA, L. S. **Geometria dos métodos métodos de regressão LARS, LASSO e Elastic Net com uma aplicação em seleção genômica.** 2017. 167 p. Tese (Doutorado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras-MG, 2017.

RENCHER, A. C; SCHAALJE, G.B. **Linear Models in statistics.** New Jersey: John Wiley, 2008. 672 p.

TIBSHIRANI, R. Regression Shrinkage and Selection via the LASSO. **Journal of the Royal Statistical Society**, v. 58, n. 1, p. 267-288, 1996.

ZOU, H; HASTIE, Y. Regularization and variable selection via the elastic net. **Journal of the Royal Statistical Society**, v. 67, n. 2, p. 301-320, 2005.