LASSO Clássico e Bayesiano

Kévin Allan Sales Rodrigues*

Instituto de Matemática e Estatística
Universidade de São Paulo

11 de novembro de 2018

Resumo

Abordaremos o método de seleção de variáveis LASSO (Least Absolute Shrinkage and Selection Operator). Além da abordagem clássica do LASSO também abordaremos uma abordagem bayesiana desta metodologia de seleção de variáveis. Focaremos na justificativa da utilização deste método e destacaremos suas vantagens e desvantagens em relação aos outros métodos de seleção de variáveis.

Palavras-chave: LASSO; Ridge regression; Stepwise.

1 Introdução

Existem casos em que o número de covariáveis é muito grande, ou mesmo, maior que o tamanho da amostra ($p > n$). Neste caso é necessário selecionar quais covariáveis serão utilizadas no modelo e quais não serão selecionadas. Os principais métodos para a seleção de variáveis são: seleção do melhor subconjunto de covariáveis, *backward*, *forward*, *stepwise* e LASSO (Least Absolute Shrinkage and Selection Operator).

O método da seleção do melhor subconjunto de covariáveis se baseia em ajustar todos os modelos com k covariáveis (o número total de modelos é $\binom{p}{k}$, em que p é o número total de covariáveis disponíveis) e posteriormente escolher o melhor dentre eles com base em algum critério. Para aplicar este método de seleção de covariáveis é necessário definir o número de covariáveis do modelo, k, e também definir o critério de comparação a ser utilizado a priori. Essa metodologia assegura que o melhor modelo será selecionado, contudo, quando o número de variáveis é realmente muito grande este método é inviável pois o número total de modelos, cresce absurdamente e consequentemente ajustar todos os modelos requer muito esforço computacional.

Note que $2^{10} = 1024$, $2^{20} = 1048576$ e $2^{30} = 1073741824$, ou seja, com apenas trinta covariáveis nós teremos que ajustar mais de um bilhão de modelos. Portanto essa abordagem é inviável computacionalmente em um contexto onde o número de covariáveis é realmente grande, como no caso em que existem cem covariáveis. Além disso é um método sensível porque é um processo discreto, isto é, as covariáveis são mantidas ou descartadas, o coeficiente de regressão associado a covariável (β_i) é zero ou um, sendo assim, pequenas mudanças nos dados podem gerar modelos muito diferentes além de reduzir a acurácia da predição (Tibshirani, 1996).

Os métodos *backward*, *forward* e *stepwise* não garantem que o melhor modelo, considerando algum critério, será encontrado.

*kevin.asr@outlook.com

O LASSO é um método de seleção e redução de variáveis apresentado por Tibshirani (1996) que possui a propriedade da regressão de cristas (*ridge regression*) de reduzir o valor das estimativas dos parâmetros mas é capaz de produzir estimativas iguais a zero para os parâmetros do modelo como no método da seleção do melhor subconjunto de covariáveis gerando assim modelos interpretáveis. Este método de seleção e redução de covariáveis foi adaptado para vários modelos como séries temporais (Audrino e Camponovo, 2013), processos autoregressivos com caudas pesadas (Sang e Sun, 2013) e também para regressão L_1 (Wang $et\ al.$, 2007).

O artigo de Wang $et\ al.$ (2007) é uma contribuição singular para os modelos de regressão pois combina a robustez da regressão L_1 com a seleção e redução de variáveis do LASSO.

Inicialmente abordaremos o LASSO no contexto do modelo linear geral e posteriormente estenderemos o LASSO para contextos mais gerais como modelos lineares generalizados.

2 Modelo Linear

O modelo de regressão linear é dado pela Equação (1),

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \epsilon_i,\ i = 1, 2, ..., n \tag{1}$$

em que Y_i é a i-ésima variável resposta, β_0 é o intercepto do modelo, $\beta_j,\ j = 1, \ldots, p$, é o parâmetro associado à j-ésima covariável, $X_{ji},\ j = 1, \ldots, p,\ i = 1, \ldots, n$ é a j-ésima covariável da i-ésima observação e ϵ_i é o erro associado à i-ésima observação. Podemos expressar o modelo de forma mais compacta, como:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

em que, $\boldsymbol{Y}_{n\times 1} = (Y_1, Y_2, \ldots, Y_n)^\top$, $\boldsymbol{X}_{n\times(p+1)} = (\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \ldots, \boldsymbol{X}_n^\top)^\top$, $\boldsymbol{X}_k = (1, X_{1k}, \ldots, X_{pk})$, $\boldsymbol{\beta}_{(p+1)\times 1} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ e $\boldsymbol{\epsilon}_{n\times 1} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^\top$. Uma suposição comumente utilizada é que $\boldsymbol{\epsilon} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_n\sigma^2)$.

O ajuste do modelo linear via mínimos quadrados é dado pela solução do problema

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \boldsymbol{X_i}\boldsymbol{\beta})^2.$$

Se $p \leq n$ então a matriz $\boldsymbol{X}$ terá posto culuna completo (na maioria das situações práticas), consequentemente $\boldsymbol{X}^\top\boldsymbol{X}$ tem posto completo e sua inversa existe. Então, obtemos a conhecida solução de mínimos quadrados

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{Y},$$

mas perceba que a solução só é única se $p \leq n$, caso contrário a solução de mínimos quadrados não será única. Isto é, quando temos dados em que $p > n$ não temos como ajustar todos os modelos (no sentido de obter ajuste único para cada modelo) pois nem sempre existirá a inversa de $\boldsymbol{X}^\top\boldsymbol{X}$.

3 Conceitos Básicos Sobre o LASSO

O estimador via LASSO no contexto da regressão linear é dado pela resolução do problema

$$\hat{\boldsymbol{\beta}}_L = \underset{\boldsymbol{\beta}}{\arg\min} \left(\sum_{i=1}^{n} (Y_i - \boldsymbol{X_i}\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right), \lambda \geq 0, \tag{2}$$

em que λ é um valor que deve ser escolhido previamente. Comentaremos como escolher λ na próxima seção. Note que, se $\lambda = 0$, então o estimador via LASSO será igual ao estimador via mínimos quadrados, isto é, o estimador de mínimos quadrados de $\boldsymbol{\beta}$ pode ser visto como um caso particular do estimador via LASSO.

O estimador via LASSO é basicamente o problema de minimização usual dos erros ao quadrado acrescido de um penalização L_1 em relação aos parâmetros de posição do modelo de regressão. Um detalhe importante acerca da penalidade é que o parâmetro β_0 não é restringido pela penalização. Note que quanto maior for o valor de λ maior será a penalização, ou seja, quanto maior for λ mais o vetor $\hat{\boldsymbol{\beta}}_L$ se aproximará do vetor $(\hat{\beta}_0, \mathbf{0}^\top)^\top$, pois, se $\lambda \to \infty$, a penalização tenderá ao infinito, isto é,

$$\lim_{\lambda \to \infty} \lambda \sum_{j=1}^{p} |\beta_j| = \infty$$

e, consequentemente, todas as estimativas dos parâmetros associados às covariáveis serão excluídas do modelo; assim teremos um modelo somente com intercepto.

De forma geral, este problema não possui solução analítica como o problema de mínimos quadrados. Portanto, é necessário utilizar algoritmos para obter a estimativa via LASSO em problemas reais. Um caso bem conhecido em que o estimador assume forma fechada é o caso em que a matriz de especificação do modelo, $\boldsymbol{X}$, é ortonormal. Como este caso é bem raro na prática (exceto em casos onde o experimento é previamente delineado para que a matriz $\boldsymbol{X}$ seja ortonormal) vamos ignorá-lo.

Outra forma de escrever o problema de minimização dado em (2) é

$$\hat{\boldsymbol{\beta}}_L = \underset{\boldsymbol{\beta}}{\arg\min} \left(\sum_{i=1}^{n} (Y_i - \boldsymbol{X_i}\boldsymbol{\beta})^2 \right) \text{ restrito a } \sum_{j=1}^{p} |\beta_j| \leq s, s \geq 0. \tag{3}$$

Note que λ não está presente em (3) e surgiu um s. A relação entre λ e s é de grandezas inversamente proporcionais, ou seja, quando o valor de λ aumenta o valor de s diminui e vice-versa. De forma mais clara, considere duas aplicações do LASSO, uma com $\lambda = \lambda_1$, cujo problema pode ser escrito na forma de (3) com $s = s_1$, e outra com $\lambda = \lambda_2$ e $s = s_2$. Então, vale

$$\lambda_1 > \lambda_2 \Leftrightarrow s_1 < s_2.$$

Uma observação importante é que, em (3), se $s \geq \sum_{j=1}^{p} |\hat{\beta}_j|$, em que $\hat{\beta}_j,\ j = 1, \ldots, p$, são as estimativas de mínimos quadrados dos parâmetros de posição do modelo, então o estimador via LASSO de $\boldsymbol{\beta}$ coincidirá com o estimador de mínimos quadrados de $\boldsymbol{\beta}$, porque, neste caso, não há restrição para as estimativas dos parâmetros. Por outro lado, se $s = 0$, então somente a estimativa do intercepto será não nula.

Antes de discutirmos uma das mais importantes características do estimador LASSO apresentaremos o estimador via *ridge regression* que é dado por

$$\hat{\boldsymbol{\beta}}_R = \underset{\boldsymbol{\beta}}{\arg\min} \left(\sum_{i=1}^{n} (Y_i - \boldsymbol{X_i}\boldsymbol{\beta})^2 + \eta \sum_{j=1}^{p} \beta_j^2 \right), \eta \geq 0. \tag{4}$$

De modo análogo ao estimador via LASSO, podemos reescrever (4) na forma

$$\hat{\boldsymbol{\beta}}_R = \arg\min_{\boldsymbol{\beta}} \left(\sum_{i=1}^{n} (Y_i - \boldsymbol{X_i}\boldsymbol{\beta})^2 \right) \text{ restrito a } \sum_{j=1}^{p} \beta_j^2 \leq r, r \geq 0. \tag{5}$$

Considerações similares às apresentadas para o parâmetro λ do LASSO podem ser feitas sobre o parâmetro η da *ridge regression*.

O estimador via *ridge regression* foi proposto antes do LASSO surgir. Perceba que a única diferença entre as equações (2) e (4) é a forma da penalização; enquanto a penalização L_1 é utilizada no LASSO, a penalização L_2 é usada na *ridge regression*. Essa pequena mudança interfere na capacidade das estimativas dos parâmetros do modelo efetivamente poderem ser nulas.

A Figura 1 apresenta a representação gráfica das equações (2) (no lado esquerdo da figura) e (4) (no lado direito da figura) no caso em que temos um total de duas covariáveis.

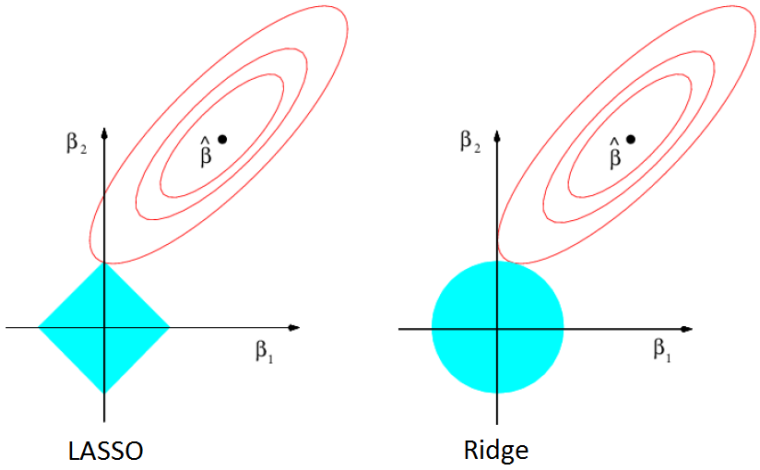

Figura 1: *Esquema gráfico do LASSO e Ridge regression no caso em que existem apenas duas variáveis explicativas.*

Fonte: Hastie, Tibshirani e Friedman (2008) (com adaptações).

Na Figura 1 as elipses de ambos os lados representam regiões de confiança para o vetor de parâmetros $(\beta_1, \beta_2)^\top$; quanto maior o nível de confiança maior a região. Note que a penalização L_1 do LASSO gera uma restrição com vértices em forma de losango enquanto a penalização L_2 da *ridge regression* gera uma região sem vértices em formato de círculo. É natural que o estimador via LASSO do vetor $(\beta_1, \beta_2)^\top$ seja o primeiro ponto da elipse que tocar a região de restrição, podendo ser inclusive em um dos vértices do losango. No caso da Figura 1 o estimador LASSO do vetor $(\beta_1, \beta_2)^\top$ é dado por $(0, s)^\top$ (devido ao fato de $\hat{\boldsymbol{\beta}}$ pertencer ao primeiro quadrante e a elipse tocar o losango no vértice onde $\beta_1 = 0$ e $|\beta_2| = s$). Verificando o lado direito da Figura 1 vemos que raramente a elipse tocará o círculo em um ponto em que $\beta_1 = 0$ ou $\beta_2 = 0$.

Situações similares ocorrem em casos em que o número de covariáveis é superior a dois. Em vez de termos elipses como regiões de confiança para os parâmetros, polígono como região de restrição do LASSO e círculo como região de restrição do estimador via *ridge regression* teremos, respectivamente, elipsóide (elipse multidimensional), polítopo

(polígono multidimensional) e hiperesfera (esfera multidimensional). Neste caso a ideia é a mesma, o polítopo terá vértices enquanto a hiperesfera não.

Outro fato importante sobre o LASSO é que o estimador oriundo deste procedimento é viciado para o parâmetro, porém o estimador proveniente do LASSO possui erro quadrático médio inferior ao erro quadrático médio do melhor estimador linear não viciado de $\boldsymbol{\beta}$ (Tibshirani, 1996) . No contexto da equação (1), o melhor estimador linear não viciado de $\boldsymbol{\beta}$ é simplesmente o estimador de mínimos quadrados, como assegurado pelo teorema de Gauss-Markov. Isto é, embora o estimador via LASSO seja viesado o seu viés é compensado pela diminuição da sua variância (em relação à variância do estimador via mínimos quadrados).

Um último detalhe importante é que o estimador via LASSO não é invariante por escala. Então, é necessário que as variáveis explicativas sejam padronizadas antes do processo de estimação.

4 Como Escolher λ

Escolher um valor de λ é fundamental para que o método funcione adequadamente. Lembre que, se $\lambda = 0$, então o método de estimação é simplesmente o método de mínimos quadrados (o que não nos ajuda em nada pois o método de mínimos quadrados não seleciona as covariáveis que entrarão no modelo) e se escolhemos um valor extremamente grande para λ então provavelmente acabaremos com um modelo sem covariáveis, isto é, só possui intercepto.

Um modo de escolher o valor de λ é a validação cruzada, que é um método bastante fácil de entender e de implementar. Basicamente dividiremos, aleatoriamente, a amostra em k partes iguais ou pelo menos aproximadamente iguais. Escolheremos a primeira parte para ser os dados de "validação" e as demais para ser dados de "treinamento", ajustaremos o LASSO com $\lambda = \lambda_0$ aos dados de "treinamento" e usaremos esse modelo para tentar prever os dados de "validação", então calcularemos o erro de predição. Repetiremos esse mesmo procedimento mais $k - 1$ vezes para as outras partes restantes. Após terminar as k iterações teremos k erros de predição e calcularemos a média dos k erros de predição. Esse procedimento será feito para vários valores λ_0 distintos. Finalmente, escolheremos o valor λ_0 que minimize o erro de predição médio.

Tibshirani (1996) recomenda $k = 5$ ou $k = 10$, mas naturalmente quaisquer valores inteiros entre cinco e dez podem ser usados. Não há uma regra para escolha de k. Claro que o número de observações deve ser "suficientemente grande" para ser dividido entre as k partes.

É importante salientar que este método de escolha do valor de λ não depende do conhecimento do número de parâmetros ("graus de liberdade") do modelo e também não depende de uma estimativa do parâmetro de escala do modelo. Embora na situação tradicional em que se usa o estimador de mínimos quadrados, o número de parâmetros do modelo ajustado esteja bem definido e seja fácil achar uma estimativa para o parâmetro de escala, no caso do LASSO isso não é trivial, pois λ também é uma quantidade que tem impacto no ajuste do modelo (embora λ não apareça de forma explícita no modelo ajustado). Logo, esse método é proposto por Tibshirani por contornar esses problemas.

O erro de predição relacionado a λ_0 é dado por

$$EP_{\lambda_0} = \frac{1}{n}\sum_{i=1}^{k} n_i EQM_i,$$

em que n_i é o número de observações da i-ésima parte dos dados e

$$EQM_i = \frac{1}{n_i}\sum_{j=1}^{n_i}(Y_j - \hat{Y}_j)^2,$$

ou seja, EQM_i é o erro quadrático médio quando usamos a i-ésima parte dos dados como "validação" e o erro de predição é simplesmente uma média ponderada dos EQM's. Essa é a quantidade que nos guiará na escolha do valor de λ; quanto menor o erro de predição, mais adequado é o valor de λ.

Em geral, não se utiliza o valor de λ que minimiza o erro de predição. Em vez disso se utiliza o maior valor de λ cujo erro de predição esteja a um desvio padrão do erro de predição associado ao valor de λ que minimiza o erro de predição. Tibshirani (1996) argumenta que como não conhecemos a real forma da curva do erro de predição em termos do valor de λ (afinal temos apenas uma curva aleatória, pois a validação cruzada seleciona aleatoriamente quais observações pertencerão a dada parte dos dados) devemos escolher o valor de λ como definido anteriormente, favorecendo um valor de λ maior e consequentemente um modelo mais simples (menos parâmetros).

5 Inferência no LASSO

A inferência no LASSO ainda é um tópico de pesquisa aberto. O tópico de pesquisa que engloba esse tipo de problema se chama "inferência após seleção de variáveis". A razão para o surgimento desta área é que ao selecionarmos as covariáveis do modelo estamos influenciando a signifiância dos testes de hipóteses e o valor-P.

Por exemplo, suponha que temos uma variável resposta e cinco mil variáveis explicativas. Vamos ajustar todos os modelos lineares simples possíveis. Então, encontramos uma variável explicativa cuja estimativa do parâmetro associado a ela tenha valor-P igual a 0,01. Será que, de fato, essa variável explicativa é realmente "importante" para o modelo? Ou há apenas uma associação espúria entre a variável explicativa e a variável resposta, já que é natural esperar que pelo menos uma das cinco mil variáveis terá forte correlação com a variável explicativa?

Existem propostas para a inferência após a aplicação do LASSO. Para mais detalhes veja Taylor e Tibshirani (2015) e Lee $et\ al.$ (2016). Um pacote também foi desenvolvido para este fim, *selectiveInference*, veja Tibshirani $et\ al.$ (2017).

6 LASSO no R

Para usar o LASSO no R (R Core Team, 2018) usaremos o pacote *glmnet* (Friedman $et\ al.$, 2010), que permite o ajuste de modelos lineares, modelos lineares generalizados, modelos de Cox (modelo semi-paramétrico) via LASSO, *ridge regression* e outras penalizações. Nesta seção aplicaremos o LASSO no contexto de modelo linear.

Utilizaremos o conjunto de dados chamado Hitters do pacote *ISLR* do R. Esse conjunto de dados contém 263 observações sobre jogadores de baseball e 20 variáveis. Ajustaremos um modelo de regressão linear via LASSO aos dados, mas antes excluiremos observações com dados faltantes. O nosso objetivo é modelar a variável resposta, salário anual dos jogadores, por meio do LASSO. Neste caso, temos um total de 19 variáveis explicativas.

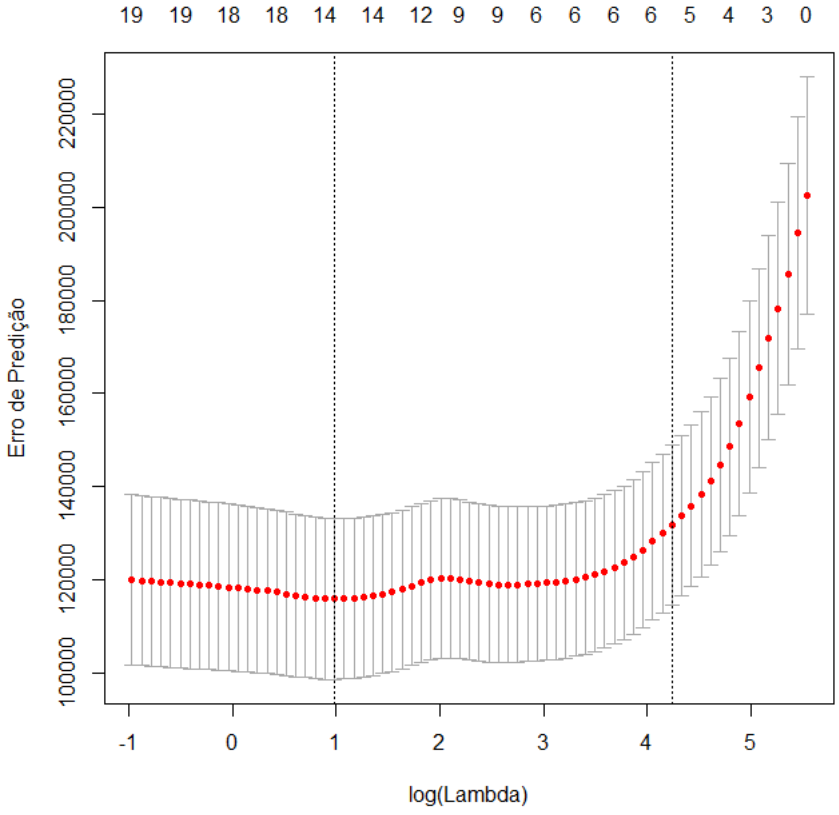

Figura 2: *Gráfico da validação cruzada com $k = 10$ para o conjunto de dado Hitters na escala logarítmica. Com retas verticais no ponto $\lambda = 2,674375$ que minimiza o erro de predição e o ponto a um desvio padrão a mais de distância $\lambda = 69,40069$.*

A Figura 2 mostra o gráfico de validação cruzada com $k = 10$ para o conjunto de dado Hitters na escala logarítmica. Observe que no ponto $\lambda = 2,674375$ ($\log \lambda = 0,983$), que minimiza o erro de predição, ainda existem 14 variáveis explicativas no modelo e no ponto $\lambda = 69,40069$ ($\log \lambda = 4,23$) existem 6 covariáveis no modelo. Portanto, avaliaremos o modelo ajustado com $\lambda = 69,40069$, seguindo o critério proposto por Tibshirani.

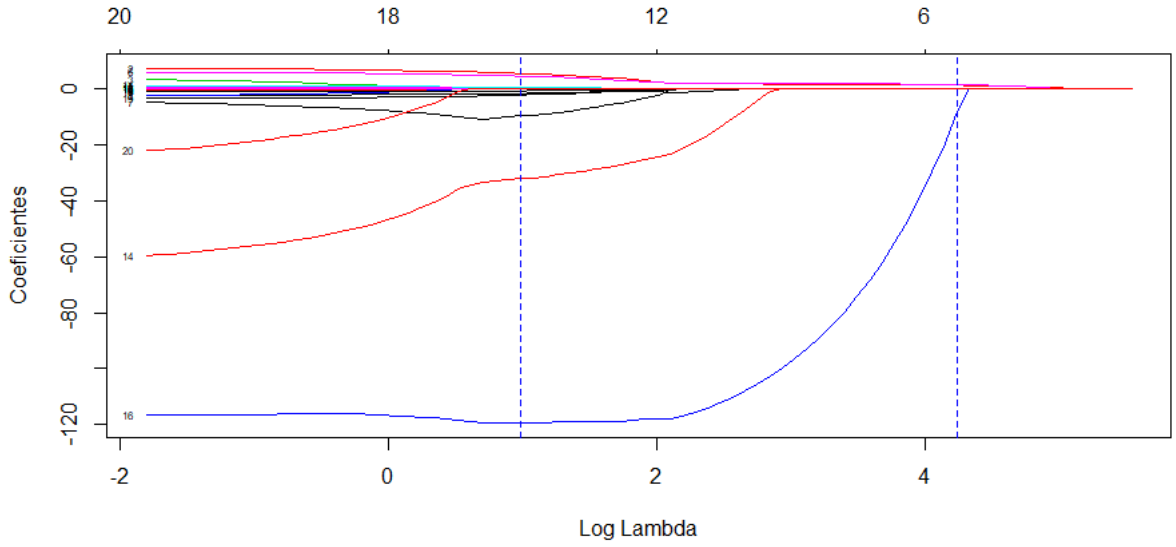

Figura 3: *Gráfico da estimativa dos parâmetros via LASSO à medida que λ varia na escala logarítmica para o conjunto de dados Hitters, com retas verticais no ponto $\lambda = 2,674375$, que minimiza o erro de predição, e o ponto a um desvio padrão a mais de distância $\lambda = 69,40069$.*

A Figura 3 mostra o gráfico da estimativa dos parâmetros via LASSO à medida que λ varia na escala logarítmica para o conjunto de dados Hitters. Observe que no ponto $\lambda = 69,40069$ ainda existem 6 covariáveis no modelo.

A Tabela 1 contém as estimativas dos parâmetros do modelo ajustado via LASSO com $\lambda = 69,40069$. Perceba que das 19 variáveis explicativas o LASSO selecionaou apenas 6 para o modelo final.

Tabela 1: *Estimativas dos parâmetros do modelo ajustado via LASSO com* $\lambda = 69,40069$.

Parâmetros	Estimativas
Intercepto	127,96
AtBat	0,00
Hits	1,42
HmRun	0,00
Runs	0,00
RBI	0,00
Walks	1,58
Years	0,00
CAtBat	0,00
CHits	0,00
CHmRun	0,00
CRuns	0,16
CRBI	0,34
CWalks	0,00
LeagueA	0,00
LeagueN	0,00
DivisionW	-8,06
PutOuts	0,08
Assists	0,00
Errors	0,00
NewLeagueN	0,00

7 LASSO Bayesiano

O LASSO bayesiano, também denominado BLASSO, foi proposto ainda no artigo seminal de Tibshirani (1996) mas só foi estudado com mais profundidade em Park e Casella (2008). Cometeremos uma breve digressão na notação apresentada até agora para acompanhar a notação de Park e Casella (2008).

Tibshirani (1996) notou que a estimativa do LASSO pode ser vista como a moda a posteriori quando $\beta_j|\sigma^2$ têm distribuição Laplace$(0, \sqrt{\sigma^2}/\lambda)$ independentes, isto é,

$$\pi\left(\boldsymbol{\beta}|\sigma^2\right) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}}\right) = \frac{\lambda^p}{2^p\sqrt{\sigma^{2^p}}} \exp\left(-\frac{\lambda}{\sqrt{\sigma^2}}\sum_{j=1}^{p}|\beta_j|\right).$$

Observe-se que condicionar a σ^2 é importante, pois garante que a posteriori $\pi\left(\boldsymbol{\beta}, \sigma^2|\tilde{\boldsymbol{y}}\right)$ será unimodal. Afinal, se a posteriori não fosse unimodal qual seria a moda a posteriori?

Antes de partirmos para a estrutura hierárquica do modelo destacaremos um resultado fundamental para o BLASSO. A distribuição Laplace pode ser expressa como uma mistura na escala da normal (com densidade exponencial), ou seja,

$$\frac{a}{2}e^{-a|z|} = \int_0^{+\infty} \frac{1}{\sqrt{2\pi s}} e^{-z^2/(2s)} \frac{a^2}{2} e^{-a^2 s/2} ds,\ a > 0. \tag{6}$$

Perceba que $Z|S = s \sim N(0, s)$ e $S \sim$ Exp$(a^2/2)$.

A estrutura hierárquica do modelo linear bayesiano é dada por

$$\boldsymbol{y}|\mu, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \sim N_p\left(\mu \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\right),$$
$$\boldsymbol{\beta}|\sigma^2, \tau_1^2, \ldots, \tau_p^2 \sim N_p\left(\mathbf{0}_p, \sigma^2 \boldsymbol{D}_\tau\right), \boldsymbol{D}_\tau = \text{diag}\left(\tau_1^2, \ldots, \tau_p^2\right),$$
$$\sigma^2 \sim \pi\left(\sigma^2\right) \perp \tau_j^2 \overset{iid}{\sim} \text{Exp}\left(\frac{\lambda^2}{2}\right).$$

Note que para obter a distribuição de $\boldsymbol{\beta}|\sigma^2$ basta integrar o produto de $N_p\left(\mathbf{0}_p, \sigma^2 \boldsymbol{D}_\tau\right)$ por $\prod_{j=1}^p \text{Exp}\left(\frac{\lambda^2}{2}\right)$ em relação a $\tau_1^2, \ldots, \tau_p^2$. Temos

$$\pi\left(\boldsymbol{\beta}|\sigma^2\right) = \int \prod_{j=1}^p \frac{1}{\sqrt{2\pi\sigma^2\tau_j^2}} \exp\left(-\frac{\beta_j^2}{2\sigma^2\tau_j^2}\right) \prod_{j=1}^p \frac{\lambda^2}{2} \exp\left(-\tau_j^2 \frac{\lambda^2}{2}\right) d\boldsymbol{\tau}$$
$$= \prod_{j=1}^p \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2\tau_j^2}} \exp\left(-\frac{\beta_j^2}{2\sigma^2\tau_j^2}\right) \frac{\lambda^2}{2} \exp\left(-\tau_j^2 \frac{\lambda^2}{2}\right) d\tau_j^2.$$

Perceba que em cada integral podemos usar (6) com $\beta_j|\sigma^2, \tau_j^2 = \tau_j^* \sim N(0, \sigma^2\tau_j^2)$ e $\tau_j^2 \sim \text{Exp}(\lambda^2/2)$.

$$\pi\left(\boldsymbol{\beta}|\sigma^2\right) = \prod_{j=1}^p \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{\lambda}{\sqrt{\sigma^2}}|\beta_j|\right).$$

Note que

$$\log\left(\pi\left(\boldsymbol{\beta}|\sigma^2, \tilde{\boldsymbol{y}}\right)\right) = \log\left(f\left(\tilde{\boldsymbol{y}}|\boldsymbol{\beta}, \sigma^2\right)\right) + \log\left(\pi\left(\boldsymbol{\beta}|\sigma^2\right)\right) + C$$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^n \left(\tilde{\boldsymbol{y}} - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2 - \frac{\lambda}{\sqrt{\sigma^2}} \sum_{j=1}^p |\beta_j| + C.$$

Queremos maximizar a posteriori para encontrar a moda. Logo,

$$\boldsymbol{\beta}_{BLASSO} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n \left(\tilde{\boldsymbol{y}} - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2 + 2\sqrt{\sigma^2}\lambda \sum_{j=1}^p |\beta_j|.$$

Note que nesta última equação a quantidade $2\sqrt{\sigma^2}\lambda$ está fazendo o mesmo papel que λ do caso clássico.

Para a construção do amostrador de Gibbs precisamos saber as distribuições a posteriori dos parâmetros do modelo. Se atribuirmos a distribuição $\pi\left(\sigma^2\right) = 1/\sigma^2$ a priori para σ^2. Estas distribuições são dadas por

$$\boldsymbol{\beta}|\tilde{\boldsymbol{y}} \sim N\left(\boldsymbol{A}^{-1}\boldsymbol{X}^\top\tilde{\boldsymbol{y}}, \sigma^2\boldsymbol{A}^{-1}\right), \ \boldsymbol{A} = \boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{D}_\tau^{-1},$$

$$\sigma^2|\tilde{\boldsymbol{y}} \sim GI\left(a, b\right), \ a = \frac{n+p-1}{2} \text{ e } b = \frac{1}{2}\sum_{i=1}^n \left(\tilde{y}_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}\right)^2 + \frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{D}_\tau^{-1}\boldsymbol{\beta},$$

$$\frac{1}{\tau_j^2} \mid \tilde{\boldsymbol{y}} \overset{iid}{\sim} NI\left(\mu', \lambda'\right), \ \mu' = \sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}} \text{ e } \lambda' = \lambda^2,$$

em que a densidade de $NI\left(\mu', \lambda'\right)$ é dada por

$$f(w) = \sqrt{\frac{\lambda'}{2\pi}} w^{-3/2} \exp\left[-\frac{\lambda'(w-\mu')^2}{2(\mu')^2 w}\right], w > 0.$$

O BLASSO pode ser facilmente aplicado ao conjunto de dados Hitters usando os códigos do Apêndice B, por isso não apresentaremos a análise do ponto de vista bayesiano.

7.1 Como Escolher λ

Existem duas abordagens propostas para "escolher λ":

- Usar Bayes empírico via máxima verossimilhança marginal (algoritmo MCEM);
- Atribuir uma priori gama "não informativa" para λ^2 (não λ !)
  É interessante atribuir priori gama para λ^2 devido à conjugação resultante.

  Observe-se que não se deve usar priori imprópria como $\pi(\lambda^2) = 1/\lambda^2$, pois isto acarreta posteriori imprópria.

Casella (2001) propôs um algoritmo MCEM que complementa o amostrador de Gibbs e dá estimativas de máxima verossimilhança dos hiperparâmetros.O algoritmo iterativo é dado por

$$\lambda^{(k)} = \sqrt{\frac{2p}{\sum_{j=1}^{p} E_{\lambda^{(k-1)}}\left[\tau_j^2 | \tilde{\boldsymbol{y}}\right]}},$$

em que o valor inicial (sugerido) é

$$\lambda^{(0)} = p\sqrt{\hat{\sigma}^2_{MQ}} / \sum_{j=1}^{p} \left|\hat{\beta}_j^{MQ}\right|$$

e a esperança condicional é substituida pela média do amostrador de Gibbs.

7.2 LASSO *versus* BLASSO

1. Em geral, os resultados provenientes do LASSO e do BLASSO são muito similares.
2. Embora o BLASSO seja computacionalmente intensivo ele é mais fácil de implementar e também tem a vantagem de gerar estimativas intervalares (e erro padrão) automaticamente durante o processo (afinal teremos as distribuições a posteriori dos parâmetros).
3. Não há como calcular os erros padrão analiticamente via LASSO (exceto aproximações analíticas). Por isso para obter os erros padrão e assim poder construir estimativas intervalares é necessário utilizar Bootstrap.
4. Além das formas de escolher λ no caso do LASSO (validação cruzada, dentre outros) o BLASSO oferece mais duas formas de escolher λ.

8 Extensões do LASSO

Existem várias extensões do LASSO, dentre elas a *Bridge Regression* e o LASSO "Huberizado", que é uma versão robusta do estimador via LASSO. A seguir são apontados o processo de estimação dos dois métodos.

- *Bridge Regression*.

$$\hat{\boldsymbol{\beta}}_{Bridge} = \arg\min_{\boldsymbol{\beta}} \left[(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})^{\top} (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|^q \right], \ \lambda \geq 0 \text{ (fixo)},$$

  priori adequada

$$\pi\left(\boldsymbol{\beta}|\sigma^2\right) \propto \prod_{j=1}^{p} \exp\left[-\lambda \left(|\beta_j|/\sqrt{\sigma^2}\right)^q\right],$$

  isto é, exponenciais potência estão fazendo o papel da Laplace neste caso.

- LASSO "Huberizado" (LASSO Robusto).

$$\hat{\boldsymbol{\beta}}_{H} = \arg\min_{\boldsymbol{\beta}} \left[L\left(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}\right) + \lambda \sum_{j=1}^{p} |\beta_j| \right], \ \lambda \geq 0 \text{ (fixo)},$$

  em que $L(\cdot)$ é uma função de perda do tipo Huber que é quadrática na vizinhança de zero e linear crescente fora da vizinhança de zero.

9 Observações Finais

De forma geral, o LASSO pode ser aplicado em qualquer modelo estatístico cujo método de estimação envolva um processo de otimização (maximização ou minimização). Para isso basta acrescentar a penalidade vista na equação (2). No caso de problemas de maximização (como no contexto dos modelos lineares generalizados onde geralmente se usam os estimadores de máxima verossimilhança) a penalização deve ser subtraída da função de verossimilhança.

Para o ajuste do LASSO bayesiano no R (R Core Team, 2018) um ótimo pacote é o *monomvn*. Este pacote além de ajustar o BLASSO ainda tem funções para ajustar o LASSO clássico e a *ridge regression*.

Muitos artigos sobre o LASSO foram publicados nas últimas duas décadas. Ainda assim o LASSO se mostra um terreno fértil para pesquisa tanto estatística quanto em computação. Para quem quiser entender a teoria do LASSO rigorosamente é recomendada a leitura de Hastie *et al.* (2008). Outros dois livros para iniciar o estudo do LASSO são James *et al.* (2014) e Hastie *et al.* (2015).

Referências

Audrino, F. e Camponovo, L. (2013). Oracle properties and finite sample inference of the adaptive lasso for time series regression models. *Cornell University*, **Identificador: Arxiv ID: 1312.1473**. Citado na pág. 2

Casella, G. (2001). Empirical bayes gibbs sampling. *Bioinformatics*, **20**, 3423–3430. Citado na pág. 11

Friedman, J., Hastie, T. e Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**(1), 1–22. URL http://www.jstatsoft.org/v33/i01/. Citado na pág. 6

Hastie, T., Tibshirani, R. e Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining Inference, and Prediction*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA. Citado na pág. 12

Hastie, T., Tibshirani, R. e Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC. ISBN 1498712169, 9781498712163. Citado na pág. 12

James, G., Witten, D., Hastie, T. e Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R, 2nd ed.* Springer Publishing Company. ISBN 1461471370, 9781461471370. Citado na pág. 12

Lee, J. D., Sun, D. L., Sun, Y. e Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, **44**(3), 907–927. doi: 10.1214/15-AOS1371. URL https://doi.org/10.1214/15-AOS1371. Citado na pág. 6

Park, T. e Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, **103**, 681–686. Citado na pág. 9

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Citado na pág. 6, 12

Sang, H. e Sun, Y. (2013). Simultaneous sparse model selection and coefficient estimation for heavy-tailed autoregressive processes. *Cornell University*, **Identificador: Arxiv ID: 1112.2682**. Citado na pág. 2

Taylor, J. e Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, **112**(25), 7629–7634. ISSN 0027-8424. doi: 10.1073/pnas.1507583112. URL http://www.pnas.org/content/112/25/7629. Citado na pág. 6

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288. Citado na pág. 5, 9

Tibshirani, R., Tibshirani, R., Taylor, J., Loftus, J. e Reid, S. (2017). *selectiveInference: Tools for Post-Selection Inference*. URL https://CRAN.R-project.org/package=selectiveInference. R package version 1.2.4. Citado na pág. 6

Wang, H., Li, G. e Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, **25**, 347–355. Citado na pág. 2

A Códigos no R para LASSO

```
library(ISLR)
summary(Hitters)

Hitters = na.omit(Hitters)
with(Hitters,sum(is.na(Salary)))

library(glmnet) #carregando pacote
x = model.matrix(Salary~.-1,data=Hitters) #definindo matriz X
y = Hitters$Salary #definindo variável resposta

modelo = glmnet(x,y) #ajustando o modelo via LASSO
valCruz = cv.glmnet(x,y, nfolds=10) #validação cruzada para escolher lambda

plot(valCruz)  #gráfico da validação cruzada

coef(valCruz)  #extraindo coeficientes

plot(modelo)
#gráfico para checar o que acontece com as estimativas quando lambda aumenta
```

B Códigos no R para BLASSO

```
# Carregando os pacotes
library(monomvn); library(lars); library(glmnet); library(miscTools)
data(diabetes); attach(diabetes)

# definindo o número de iterações discartadas (burn-in),
#número de amostras do mcmc e valores iniciais
burnin <- 500
iter <- 1000
initial.beta <- rep(-500, dim(x2)[2])
# atribuindo um valor inicial extremo para todos os betas
initial.lambda2 <- 10
# atribuindo um valor inicial extremo para lambda (parâmetro de penalização)
initial.variance <- 500
# atribuindo um valor inicial extremo para o parâmetro de variância

# Iniciando o amostrador de Gibbs
lasso <- blasso(X = x2, # matriz de covariáveis 442 x 64
                y = y,  # variáveis resposta of 442
                T = iter, # número de iterações
                beta = initial.beta,
                lambda2 = initial.lambda2,
                s2 = initial.variance)
                #rd = c(1, 1.78))
```

```
                # hiperparâmetros sugeridos por Park e Casella (2008)

# extraindo valores de alguns parâmetros para visialização
coef.lasso <- as.data.frame(cbind(iter = seq(iter),
                                  beta1 = lasso$beta[, "b.1"],
                                  beta2 = lasso$beta[, "b.2"],
                                  variance = lasso$s2,
                                  lambda.square = lasso$lambda2))

colMedians(coef.lasso[-seq(burnin), -1])
#beta1       beta2        variance         lambda.square
#0.0000000  -172.3840906  2841.4410472     0.3031814


##############################lasso clássico

#Vamos comparar o LASSO (glmnet) com o BLASSO (monomvn)

fit.glmnet <-  glmnet(as.matrix(x2), y,
                          lambda=cv.glmnet(as.matrix(x2), y)$lambda.1se)
coef.glmnet <- coef(fit.glmnet)
sum(coef.glmnet == 0)
#53
sum(colMedians(lasso$beta[-seq(burnin), ]) == 0)
#56
```

O LASSO atribuiu o valor zero para 53 parâmetros. E o BLASSO atribuiu o valor zero para 56 parâmetros.