LASSO Clássico e Bayesiano

Kévin Allan Sales Rodrigues

kevin.asr@outlook.com

Universidade de São Paulo
IME - Instituto de Matemática e Estatística
Programa de Pós-Graduação em Estatística

31 de outubro de 2018

Sumário

IME-USP

Introdução

O que significa LASSO?

Least Absolute Shrinkage and Selection Operator.

Em tradução livre, significa operador de seleção e redução de variáveis/coeficientes via norma L_1.

O Lasso foi proposto por Tibshirani (1996, JRSS) e sua versão bayesiana foi descrita e aprofundada em Park e Casella (2008, JASA).

IME-USP

Os Autores do LASSO e do BLASSO

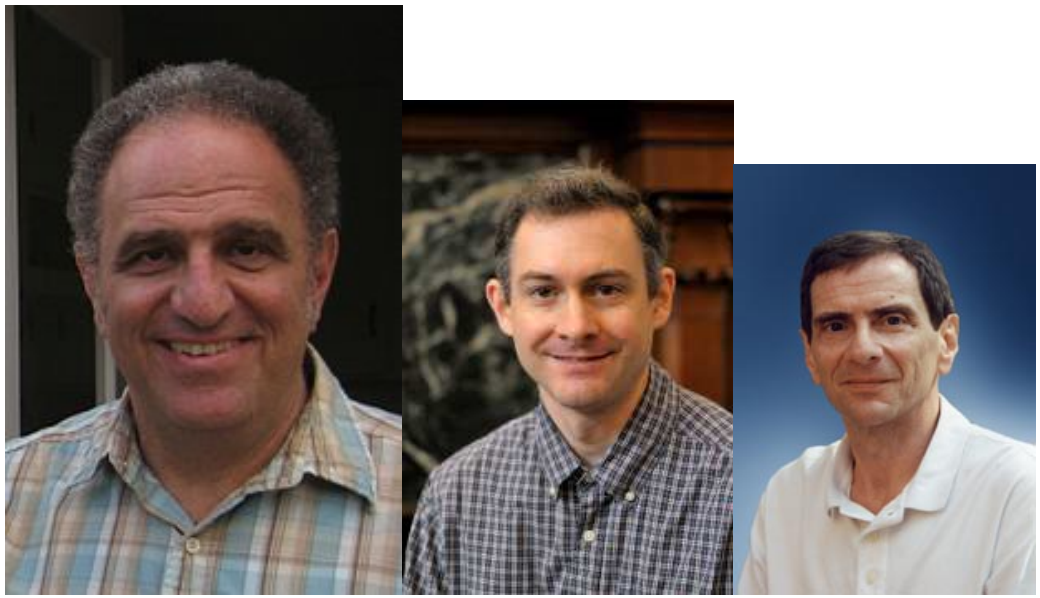

Figura: Tibshirani (LASSO), Park (BLASSO) e Casella (BLASSO).

IME-USP

Introdução

Porque utilizar o LASSO?

- Pode ser usado quando $p > n$ (alta dimensionalidade);
- Pode estimar alguns β_j como 0;
- Evita a associação espúria;
- É um método parcimonioso;
- Predição: diminui a variância e aumenta viés.

Obs: as estimativas do LASSO não são invariantes por escala, por isso as covariáveis, $\boldsymbol{X}$, devem ser padronizadas antes de usar o LASSO.

IME-USP

Sumário

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.

IME-USP

Podemos expressar o modelo de forma mais compacta, como:

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

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

IME-USP

Modelo Linear

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

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

se $p + 1 \leq n$ então a matriz $\boldsymbol{X}$ terá posto culuna completo, consequentemente $\boldsymbol{X}^\top \boldsymbol{X}$ tem posto completo e sua inversa existe. Então obtemos a conhecida solução de mínimos quadrados

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

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 $p + 1 > n$ não temos como ajustar todos os modelos (no sentido de obter ajuste único para cada modelo) pois nem sempre existirá a inversa de $\boldsymbol{X}^\top \boldsymbol{X}$.

IME-USP

Sumário

Estimador via LASSO

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

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

ou equivalentemente, por

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

IME-USP

Algumas Perguntas

O que ocorre no estimador se na Equação (2) $\lambda = 0$ e se $\lambda \to \infty$?

O que ocorre no estimador se na Equação (3) $s = 0$ e se $s \geq \sum_{j=1}^{p} \hat{\beta}_j$ ($\hat{\beta}_j,\ j = 1, \ldots, p$ são os estimadores de mínimos quadrados) ?

Então como escolher o valor de λ ?

IME-USP

Estimador via Ridge Regression

O estimador via Ridge regression, $\hat{\boldsymbol{\beta}}_R$, é dado por

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

ou equivalentemente, por

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

IME-USP

LASSO e Ridge Regression

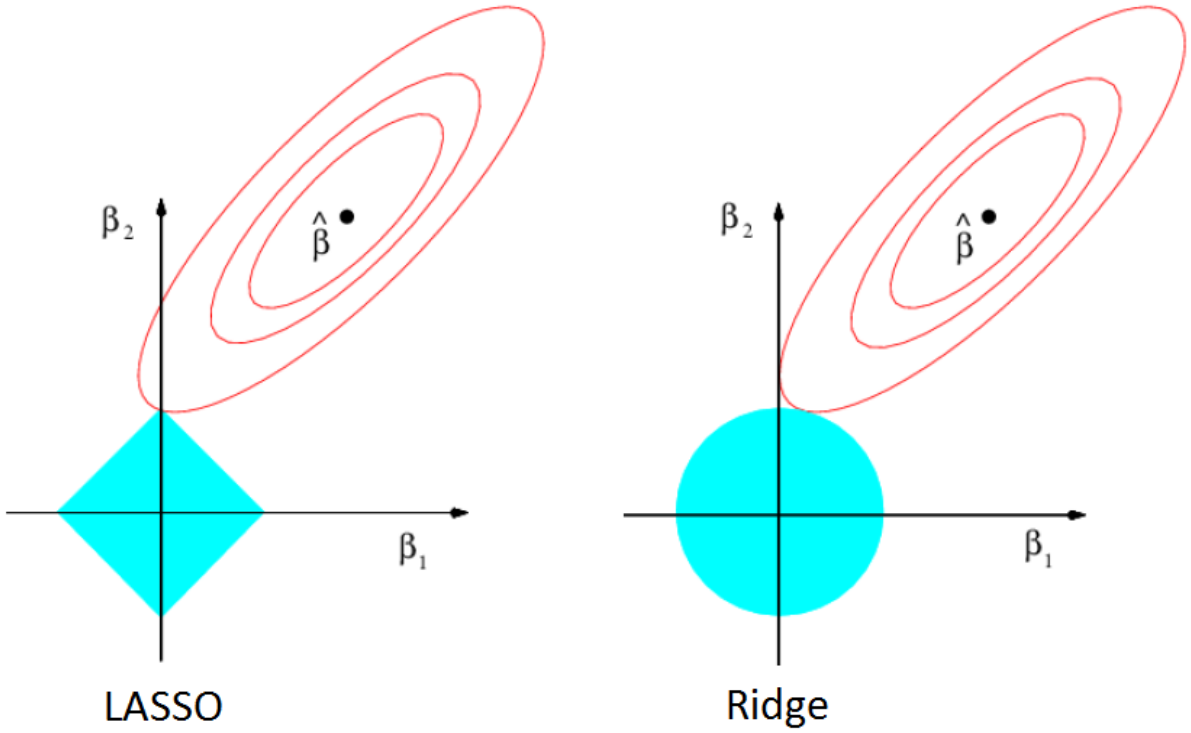

Figura: Esquema gráfico do LASSO e Ridge regression no caso em que existem apenas duas variáveis explicativas.

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

Sumário

IME-USP

Escolhendo o Valor de λ

Escolher um bom valor de λ é fundamental para que o método funcione adequadamente, 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, calcularemos a média dos k erros de predição. Esse procedimento será feito para vários valores λ_0 distintos. E escolheremos o valor λ_0 que minimize o erros de predição médio.

IME-USP

Escolhendo o Valor de λ

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 divido 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 pora contornar esses problemas.

IME-USP

Escolhendo o Valor de λ

O erro de predição relacionado à 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.

IME-USP

Escolhendo 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 nós 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).

IME-USP

Sumário

IME-USP

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 outras coisas, como 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?

IME-USP

Inferência no LASSO

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

IME-USP

Ideia por trás da inferência após seleção

Após escolher um modelo $\widehat{M}$ baseado nos dados queremos testar uma hipótese $\widehat{H}_0$. Note que $\widehat{H}_0$ é aleatória, pois $\widehat{H}_0$ varia em função do modelo selecionado, $\widehat{M}$, que por sua vez varia em função da amostra.

Neste contexto temos que controlar o erro seletivo do tipo 1 que é dado por

$$P(T(\boldsymbol{Y}) \in R | \widehat{M}, \widehat{H}_0) \leq \alpha,$$

em que a região de rejeição é dada pelo evento $\{T(\boldsymbol{Y}) \in R\}$.

IME-USP

Sumário

Software para realizar LASSO/BLASSO

Software *R* (R Core Team, 2018).
Pacotes:

- *glmnet*, ajusta MLGs com o LASSO (ridge regression também).
- *monomvn*, oferece ajuste com LASSO, BLASSO, Ridge reg., LAR, entre outros.
- *selectiveInference*, dá suporte para realizar inferência em modelos em que o LASSO foi aplicado.

Um exemplo no R

Para usar o LASSO no R usaremos o pacote *glmnet* 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.

IME-USP

Um exemplo no R

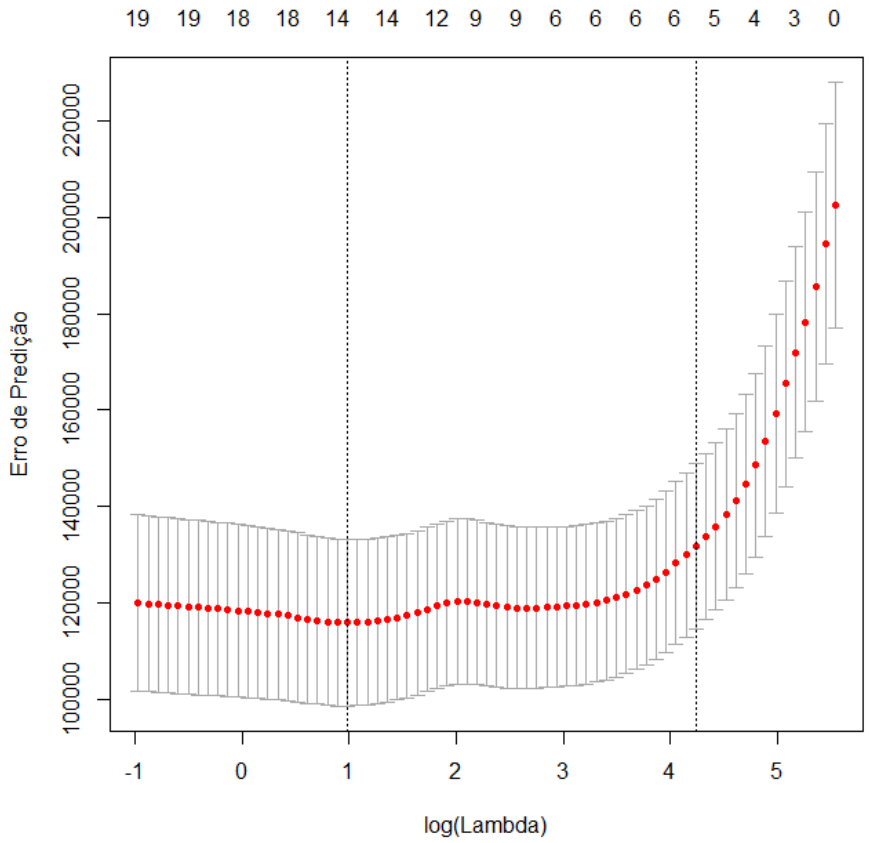

Figura: Gráfico da validação cruzada com $k = 10$ para o conjunto de dado Hitters na escala logarítmica. Com retas verticais no ponto $\lambda = 2,674375$ que minimiza o erro de predição e o ponto a um desvio padrão a mais de distância $\lambda = 69,40069$.

Um exemplo no R

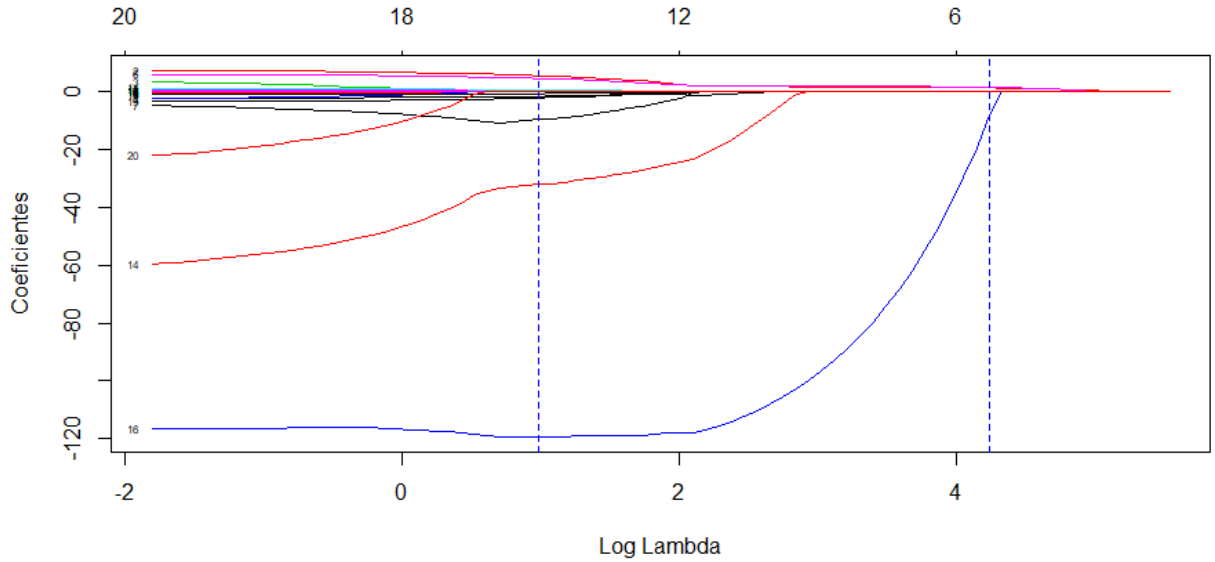

Figura: Gráfico da estimativa dos parâmetros via LASSO à medida que λ varia na escala logarítmica para o conjunto de dados Hitters. Com retas verticais no ponto $\lambda = 2,674375$ que minimiza o erro de predição e o ponto a um desvio padrão a mais de distância $\lambda = 69,40069$.

IME-USP

Um exemplo no R

Tabela: 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

IME-USP

Funções no R

```
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
```

IME-USP

Sumário

IME-USP

$$\boldsymbol{y} = \mu \mathbf{1}_n + \boldsymbol{X}\beta + \boldsymbol{\epsilon}$$

em que $\boldsymbol{y}_{n\times 1}$ é o vetor com as variáveis resposta, $\boldsymbol{X}_{n\times p}$ é a matriz com as variáveis explicativas (já padronizadas), μ é a média geral, $\beta_{p\times 1}$ é o vetor de parâmetros angulares e $\boldsymbol{\epsilon}_{n\times 1}$ é o vetor com as fontes de variação.

Sejam $\hat{\beta}$ o estimador de mínimos quadrados e $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \mu \mathbf{1}_n$.

$$\hat{\beta} = \arg\min_{\beta} \left(\tilde{\boldsymbol{y}} - \boldsymbol{X}\beta\right)^{\top} \left(\tilde{\boldsymbol{y}} - \boldsymbol{X}\beta\right)$$

IME-USP

BLASSO: LASSO Bayesiano

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

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

Obs: condicionar a σ^2 é importante, pois garante que a posteriori $\pi\left(\boldsymbol{\beta}, \sigma^2|\tilde{\boldsymbol{y}}\right)$ será unimodal.

Se a posteriori não fosse unimodal qual seria a moda a posteriori?

Outra consequência da falta de unimodalidade é a lenta convergência do amostrador de Gibbs.

IME-USP

Consequência de não condionar a distribuição de β

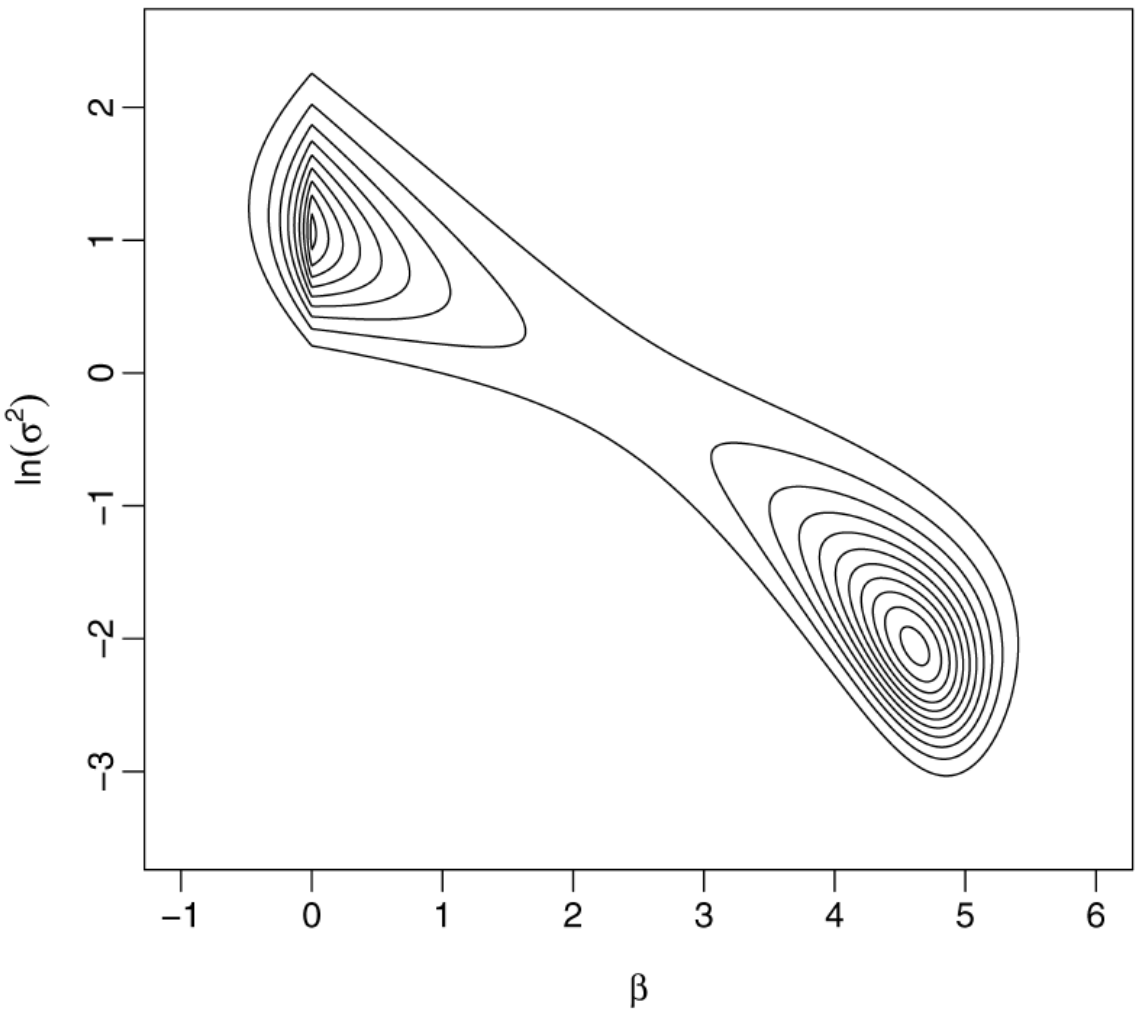

Figura: Fonte: Park e Casella (2008).

IME-USP

BLASSO: LASSO Bayesiano

Park e Casella (2008) propuseram o amostrador de Gibbs partindo da representação da Laplace pela 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 $S \sim \text{Exp}(a^2/2)$.

IME-USP

BLASSO: LASSO Bayesiano

A estrutura hierárquica do modelo é dada por

$$\boldsymbol{y}|\mu, \boldsymbol{X}, \beta, \sigma^2 \sim N_p\left(\mu \mathbf{1}_n + \boldsymbol{X}\beta, \sigma^2 \mathbf{I}_n\right),$$

$$\beta|\sigma^2, \tau_1^2, \ldots, \tau_p^2 \sim N_p\left(\mathbf{0}_p, \sigma^2 \boldsymbol{D}_\tau\right), \boldsymbol{D}_\tau = \text{diag}\left(\tau_1^2, \ldots, \tau_p^2\right)$$

$$\sigma^2 \sim \pi\left(\sigma^2\right) \perp \tau_j^2 \overset{iid}{\sim} \text{Exp}\left(\frac{\lambda^2}{2}\right)$$

Note que para obter a distribuição de $\beta|\sigma^2$ 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 $\tau_1^2, \ldots, \tau_p^2$.

IME-USP

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

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

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 $\tau_j^2 \sim \text{Exp}(\lambda^2/2)$.

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

IME-USP

Porque o BLASSO é obtido a partir da Moda a Posteriori?

Note que

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

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

queremos maximizar a posteriori para encontrar a moda, logo

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

IME-USP

Distribuição dos Parâmetros a Posteriori $(\pi(\sigma^2) = 1/\sigma^2)$

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

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

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

cuja densidade é dada por

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

IME-USP

Escolhendo o λ

Existem duas abordagens propostas para "escolher o 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 a conjugação resultante.

  Obs: não usar priori imprópria como $\pi\left(\lambda^2\right) = 1/\lambda^2$, pois isto acarreta posteriori imprópria.

IME-USP

Escolhendo o λ:

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

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

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 substituída pela média do amostrador de Gibbs.

IME-USP

LASSO versus BLASSO

- Em geral, os resultados provenientes do LASSO e do BLASSO são muito similares.
- 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 parâmetro).
- 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.
- Além das formas de escolher o λ no caso do LASSO (validação cruzada, dentre outros) o BLASSO oferece mais duas formas de escolher o λ.

IME-USP

Sumário

Extensões do BLASSO/LASSO

- Bridge Regression.

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

priori adequada

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

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

IME-USP

Extensões do BLASSO/LASSO

- LASSO "Huberizado" (LASSO Robusto).

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

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.

IME-USP

Usando os Pacotes monomvn e glmnet

```
# 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
```

IME-USP

Usando os Pacotes monomvn e glmnet

```
##############################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.

IME-USP

Referências I

Casella, G. (2001).

Empirical Bayes gibbs sampling.

Bioinformatics, **20**, 3423-3430.

Friedman, J.; Hastie, T.; Tibshirani, R. (2010).

Regularization Paths for Generalized Linear Models via Coordinate Descent .

Journal of Statistical Software, **33**(1), 1-22.

Hastie, T.; Tibshirani, R.; Friedman, J. (2008).

The Elements of Statistical Learning: Data Mining, Inference, and Prediction 2ed .

New York: Springer-Verlag.

Hastie, T.; Tibshirani, R.; Wainwright, M. (2015).

Statistical Learning with Sparsity: The Lasso and Generalizations.

Chapman & Hall/CRC.

IME-USP

Referências II

James, G.; Witten, D; Hastie, T.; Tibshirani, R. (2014).

An Introduction to Statistical Learning: With Applications in R, 2nd ed..

pringer Publishing Company.

Lee, J. D.; Sun, D. L.; Sun, Y.; Taylor, J. E. (2016).

Exact post-selection inference, with application to the lasso

The Annals of Statistics, **44**(3), 907-927.

Park, T.; Casella, G. (2008).

The Bayesian Lasso.

Journal of the American Statistical Association, **103**, 681-686.

R Core Team (2018).

R: A language and environment for statistical computing.

R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

IME-USP

Referências III

Taylor, J.; Tibshirani, R. J. (2015).

Statistical learning and selective inference.

Proceedings of the National Academy of Sciences. **112**(25), 7629-7634.

Tibshirani, R. (1996).

Regression Shrinkage and Selection via the LASSO.

Journal of the Royal Statistical Society. Series B (Methodological), **58**, 267-288.

Tibshirani, R.; Tibshirani, R.; Taylor, J.; Loftus, J.; Reid, S. (2017).

selectiveInference: Tools for Post-Selection Inference (R package version 1.2.4)

url = https://CRAN.R-project.org/package=selectiveInference

IME-USP

Obrigado pela atenção!!!