Regularização em MLGs

Carlos de Moura

2025-12-01

Índice

Sobre

“Eu tô te explicando Pra te confundir Eu tô te confundindo Pra te esclarecer” Tom Zé

Este é o material auxiliar da apresentação do trabalho final do curso de Modelos Lineares Generalizados (MLGs) (DEST-UFMG, 2025/2). O tema é regularização em MLGs, em específico os métodos de regularização ridge, lasso e elastic net e este quarto book está dividido da seguinte forma:

- Definição de regularização e apresentação dos métodos de shrinkage, com exemplos de regressão normal;
- Estimação dos parâmetros em MLG com penalização;
- Tunagem dos parâmetros de regularização;
- Exemplo prático com dados reais no R.

Referências bibliográficas importantes que foram usadas para a feitura desse trabalho são citadas ao final do documento.

Alguns pacotes que usaremos estão abaixo listados.

```r
if (!{"pak" %in% rownames(installed.packages())}) install.packages("pak")

pak::pak(c("matrixcalc", "glmnet"))
```

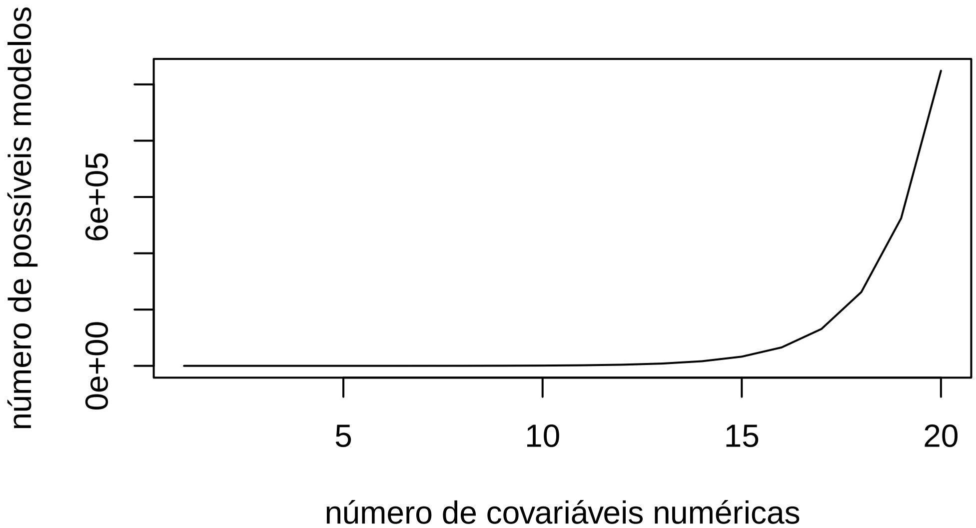
1 Regularização

1.1 Seleção de variáveis naïve

Stepwise

```r
f = function(n) {
  s = 0
  for (i in 1:n) {
    s = s + choose(n, i)
  }
  s
}

seq(1, 20, 1) |>
  lapply(f) |>
  unlist() |>
  plot(type = 'l',
       main = "",
       xlab = "número de covariáveis numéricas",
       ylab = "número de possíveis modelos"
       )
```



1.2 Ridge

A técnica Ridge foi a primeira das três técnicas a surgir, no trabalho de Hoerl & Kennard (1970). Originalmente, os autores buscavam entender como superar problemas em que a matriz de covariáveis X estava mal-especificada. Relembrando que na regressão linear normal, temos que

$$Y = X\beta + \epsilon, \qquad \epsilon \sim N_n(0, \sigma^2 I_n)$$

O estimador de máxima verossimilhança (EMV) de β será o mesmo estimador de mínimos quadrados (EMQ)

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

se:

- $X^\top X$ for inversível;
- A matriz de covariáveis é ortogonalizável, i.e., os dados foram coletados de maneira independente;
- há menos betas que observações, isto é `ncol(X)` « `nrow(X)`.

Além disso, $\hat{\beta}$ é não viciado, consistente e tem matriz de covariâncias dada por

$$cov(\hat{\beta}) = \sigma^2 (X^\top X)^{-1}.$$

Uma vez que temos a distribuição do estimador, fica fácil fazer inferência via intervalos de confiança, por exemplo.

$$\hat{\beta} \sim N_p\left(\beta,\ \sigma^2(X^\top X)^{-1}\right).$$

Nesse sentido, se temos uma matriz de dados problemática - no sentido em que $(X^\top X)^{-1}$ não está bem definida, teremos problema de estimação via EMQ.

Veja o exemplo numérico abaixo.

```
set.seed(12345)

n = 10
beta = c(1, 0)
X = cbind(1:n, 2*(1:n))
Y = X %*% beta + rnorm(n)
```

Ver matriz X

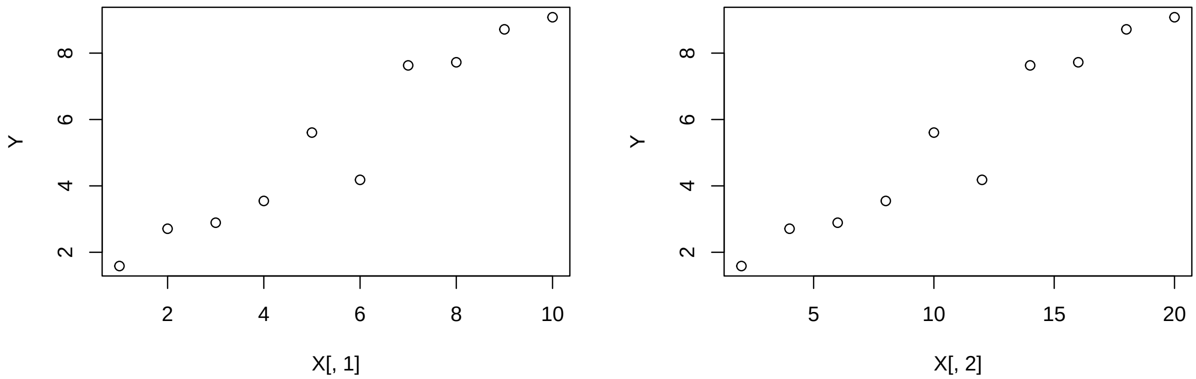
```
head(X)
```

```
     [,1] [,2]
[1,]    1    2
[2,]    2    4
[3,]    3    6
[4,]    4    8
[5,]    5   10
[6,]    6   12
```

```
matrixcalc::is.singular.matrix(t(X)%*%X)
```

```
[1] TRUE
```

```
par(mfrow=c(1,2))
plot(Y~X[,1])
plot(Y~X[,2])
```




```r
lm(Y~0+X)
```

```

Call:
lm(formula = Y ~ 0 + X)

Coefficients:
    X1      X2  
0.9544      NA  
```

E se a regressão estivesse na outra covariável?

```r
set.seed(12345)

beta = c(0, 1)
X = cbind(1:n, 2*(1:n))
Y = X %*% beta + rnorm(n)

lm(Y~0+X)
```

```

Call:
lm(formula = Y ~ 0 + X)

Coefficients:
   X1     X2  
1.954     NA  
```

Detalhes

Comentários sobre o modelo não saber selecionar variáveis.

```r
set.seed(12345)

beta = c(1, 0)
X = cbind(1:n, 2*(1:n))
ruido = cbind(rep(0,n), rnorm(n,0,.1))
X = X + ruido
Y = X %*% beta + rnorm(n)
```

Ver matriz X

```r
head(X)
```

```
     [,1]      [,2]
[1,]    1  2.058553
[2,]    2  4.070947
[3,]    3  5.989070
[4,]    4  7.954650
[5,]    5 10.060589
[6,]    6 11.818204
```

```r
matrixcalc::is.singular.matrix(t(X)%*%X)
```

```
[1] FALSE
```

```r
eigen(t(X)%*%X)$values
```

```
[1] 1.918025e+03 1.070613e-02
```

```r
lm(Y~0+X)
```

```
Call:
lm(formula = Y ~ 0 + X)

Coefficients:
    X1      X2
 6.658  -2.820
```

Comentários sobre combinações lineares, parcimônia e inflação da variância na inferência.

A proposta de Hoerl & Kennard (1970)

1.3 Lasso

A regressão lasso surgiu com o artigo de Tibshirani (1996).

1.4 Comparação das técnicas

"You Can't Always Get What You Want"

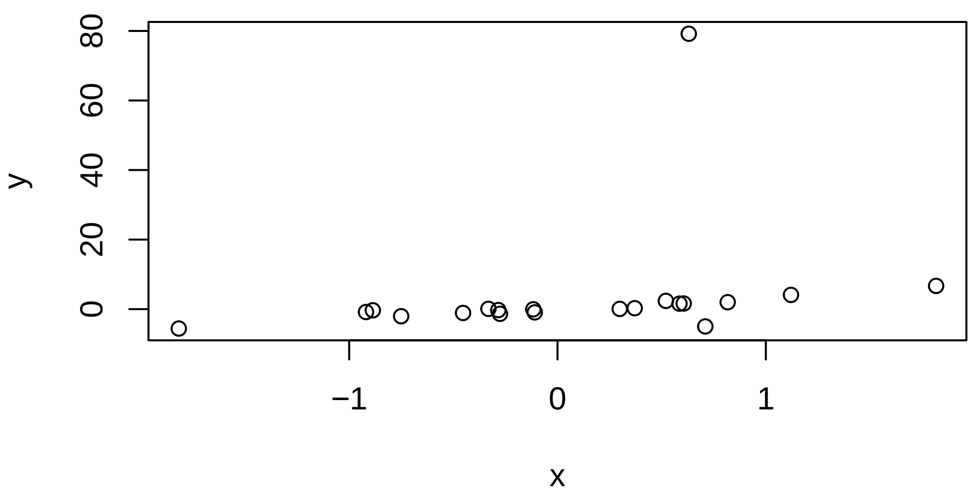
1.5 Elastic net

Can you?

O artigo Zou & Hastie (2005) introduz o elastic net, que nada mais é que uma mistura das técnicas ridge e lasso.

1.6 Lidando com outliers

```r
set.seed(12345); x = rnorm(20); e = rt(20, 1); y = 2*x + e; plot(y~x); lm(y~0+x)
```




```
Call:
lm(formula = y ~ 0 + x)
```

```
Coefficients:
    x
6.136
```

2 Estimação

"Se isto for possível, Pois, me contem, Como escrever de novo, Um jornal de ontem" Tom Zé

2.1 Regularização como uma restrição do espaço paramétrico

É possível entender cada um dos processos de regularização descritos anteriormente como uma restrição do espaço paramétrico dos coeficientes de regressão. Se não fazer seleção é considerer que $\beta \in \mathbb{R}^d$, é possível mostrar que - escolhidos os parâmetros de shrinkage - então minimar a soma de quadrados do resíduo penalizada é a mesma coisa que minimizar a soma de quadrado da regressão num espaço paramétrico menor (que depende dos parâmetros de shrinkage escolhidos).

Seja $SQRes = (Y - X\beta)^\top (Y - X\beta) = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij}\right)^2$, então (na regressão linear normal) vale que:

$$\hat{\beta}_{ridge} = \arg\min_{\beta} \{SQRes\} \quad \text{com } \beta \text{ tal que} \quad \sum_{j=1}^p \beta_j^2 \leq t_{ridge},$$

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \{SQRes\} \quad \text{com } \beta \text{ tal que} \quad \sum_{j=1}^p |\beta_j| \leq t_{lasso},$$

$$\hat{\beta}_{elastic\ net} = \arg\min_{\beta} \{SQRes\} \quad \text{com } \beta \text{ tal que} \quad (1-\alpha) \sum_{j=1}^p |\beta_j| + \alpha \sum_{j=1}^p \beta_j^2 \leq t_{elastic\ net}.$$

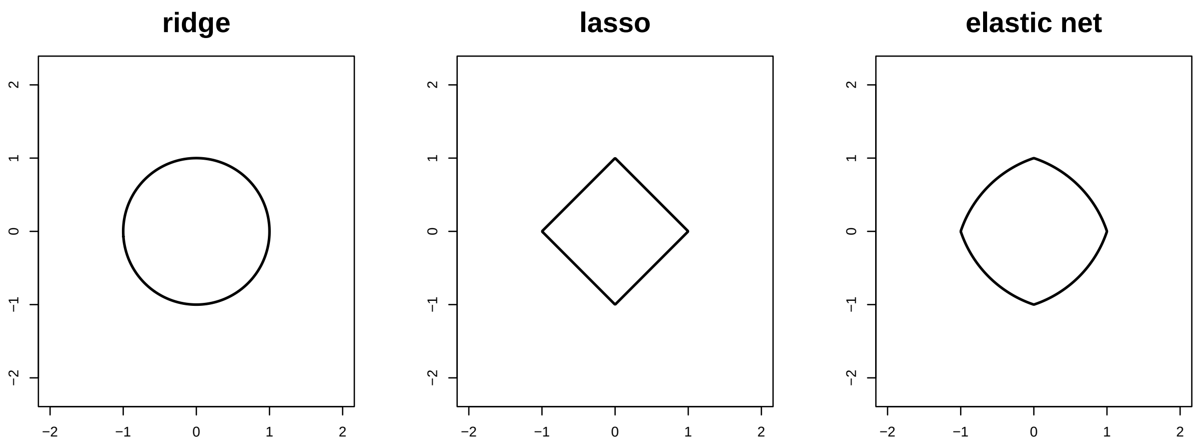
```r
desenhar_espaço_paramétrico = function(alpha, t, título = NULL) {
  stopifnot(
    "alpha deve estar entre 0 e 1" = all(alpha >= 0, alpha <= 1),
    "t deve ser positivo"           = t > 0
  )

  F = function(x, y) alpha*(x^2 + y^2 - t) + (1-alpha)*(abs(x) + abs(y) - t)
```

```
  x = seq(-2, 2, length = 400)
  y = seq(-2, 2, length = 400)
  g = outer(x, y, F)

  contour(x, y, g,
          levels = 0,
          drawlabels = FALSE,
          lwd = 2, asp = 1,
          main = título,
          cex.main = 2
          )
}


par(mfrow = c(1,3))
desenhar_espaço_paramétrico(1,   1, "ridge")
desenhar_espaço_paramétrico(0,   1, "lasso")
desenhar_espaço_paramétrico(1/2, 1, "elastic net")
```



2.2 Regularização nos MLGs

FS vs shrinkage

```
1+ 1
```

```
[1] 2
```

3 Tunagem

Validação cruzada

```
1
```

```
[1] 1
```

4 Exemplo prático

"Minha jangada vai sair por mar"

Esparsidade, multicolinearidade e outliers

```
1
```

```
[1] 1
```

References

Friedman, J., Hastie, T., Tibshirani, R., et al. (2025). *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*. Stanford University. Obtido de https://glmnet.stanford.edu/

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67. Obtido de https://www.jstor.org/stable/1267351

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer Texts em Statistics. Springer. Obtido de https://link.springer.com/book/10.1007/978-1-0716-1418-1

Kuhn, M., & Silge, J. (2022). *Tidy Modeling with R*. O'Reilly Media. Obtido de https://www.tmwr.org/

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Obtido de https://www.r-project.org/

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. Obtido de https://www.jstor.org/stable/2346178

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *67*(2), 301–320. Obtido de https://www.jstor.org/stable/3647580