Introdução à inferência estatística em alta dimensão

X Encontro Baiano de Estatística (EBEST)

Rodney Fonseca

Departamento de Estatística
Universidade Federal da Bahia

09/06/26

Introdução

Apresentação

- Nome: Rodney Fonseca
- Formação: graduação (UFC), mestrado (UFPE) e doutorado (Unicamp) em Estatística
  Pós-doc no dept. de matemática aplicada e computação do Instituto Weizmann de Ciências em Israel
- Pesquisa: modelos de regressão, métodos não-paramétricos, séries temporais, dados de alta dimensão

Conteúdo do minicurso

- Modelagem estatística em alta dimensão
  - Alta dimensionalidade e o conceito de esparsidade
  - Lasso para modelos lineares
- Inferência estatística com o estimador lasso
  - Bootstrap
  - Inferência via lasso deviesado
- Material disponível no site `rodneyfv.github.io`

Referências

- Referência principal
  - Hastie, T., R. Tibshirani, and M. Wainwright (2015). Statistical Learning with Sparsity: The Lasso and Generalizations. Boca Raton: CRC Press.
  - Disponível neste link: `https://hastie.su.domains/StatLearnSparsity/`
- Referências complementares
  - Izbicki, R. and T. M. dos Santos (2020). Aprendizado de máquina: uma abordagem estatística.
  - Fan, J., R. Li, C.-H. Zhang, and H. Zou (2020). Statistical Foundations of Data Science. Boca Raton: CRC press.
  - Wainwright, M. J. (2019). High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge University Press.

O problema de alta dimensionalidade e o conceito de esparsidade

Coleta constante de grandes volumes de dados

Primeiros data centers de IA no Brasil podem consumir mesma energia de 16 milhões de casas; conheça os projetos

Mercado cresceu com o surgimento de aplicativos como o ChatGPT, que demandam mais energia e refrigeração. Empresas anunciaram projetos para o Rio de Janeiro (RJ), Eldorado do Sul (RS), Maringá (PR) e Uberlândia (MG).

Por Victor Hugo Silva, g1

03/08/2025 05h00 · Atualizado há uma semana

Figura: https://g1.globo.com/inovacao/noticia/2025/08/03/primeiros-data-centers-de-ia-no-brasil-podem-consumir-mesma-energia-de-16-milhoes-de-casas-conheca-os-projetos.ghtml

Alguns motivos para coleta massiva de dados

- Descobrir indicadores para desenvolvimento de doenças através de transcrições genéticas
- Identificar biomarcadores para o tempo de vida de pacientes
- Encontrar conexões entre regiões cerebrais associadas com certa característica

Formação em Estatística e CD

3. EIXO DE FORMAÇÃO: CIÊNCIAS DE DADOS E GRANDES BASES DE DADOS

No exercício da profissão é necessária a manipulação e análise de grandes massas de dados, oriundos de diferentes fontes e com foco na extração de co extração, transformação e carga de grandes bases de dados, modelagem, construção e avaliação de algoritmos descritivos e preditivos, bem como imple ambientes de produção para a tomada de decisão.

COMPETÊNCIA: Analisar grandes bases de dados estruturados e não estruturados, considerando:

- Compreender bancos de dados estruturados e não estruturados Saber como consultar, modificar tais bases.
- Conhecer ambientes computacionais específicos para análise de grandes bases de dados.
- Conhecer métodos de estimação, inferência e modelagem para dados de alta dimensionalidade.
- Selecionar métodos de aprendizado de máquina, supervisionados e não supervisionados para tais bases.
- Gerenciar projetos de análise, incluindo a definição de escopo, coleta e preparação de dados, modelagem, avaliação e implementação de soluções.

Figura: `https://www.gov.br/participamaisbrasil/proposta-diretrizes-cursos-de-graduacao-estatistica-de-dados`

“Estamos nos afogando em informação e famintos por conhecimento” (Hastie et al., 2015, p. 1)

- Necessidade de extrair informações úteis de grandes volumes de dados

"Estamos nos afogando em informação e famintos por conhecimento" (Hastie et al., 2015, p. 1)

- Necessidade de extrair informações úteis de grandes volumes de dados
- **Ideia:** Supor que o mundo não é tão complexo quanto parece
  - Dentre milhares de genes, alguns estão diretamente ligados à doença
  - Alguns fatores afetam mais o tempo de vida (idade, peso, fuma?, etc.)
  - Algumas regiões de interesse se ativam mais durante atividades específicas

- O número de preditores disponível pode ser grande, mas somente alguns são relevantes

Esparsidade

Suposição de que o fenômeno de interesse pode ser bem explicado por um modelo estatístico com um número pequeno de parâmetros/preditores

Regressão linear

- Considere que temos n observações de uma variável resposta y_i e p preditores associados $x_i = (x_{i1}, \ldots, x_{ip})^\top$
- O modelo de regressão linear é dado por

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_i, \quad i = 1, \ldots, n,$$

em que β_0 e $\beta = (\beta_1, \ldots, \beta_p)^\top$ são parâmetros desconhecidos e e_i é um erro aleatório

- Estimativas por **mínimos quadrados** (MQ)

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

- Tipicamente, todas as estimativas são diferentes diferentes de zero, ou seja, todos os preditores tem alguma importância

- Estimativas por **mínimos quadrados** (MQ)

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

- Tipicamente, todas as estimativas são diferentes diferentes de zero, ou seja, todos os preditores tem alguma importância
- Problemas quando p é muito grande
  - interpretação mais difícil
  - Sobreajuste (overfitting)
  - soluções de MQ não são únicas

Alternativa: regularização/penalização

- Estimativas via **lasso** ou com regularização ℓ_1:

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \quad \text{sujeito a} \quad \|\beta\|_1 \leq t,$$

  em que $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ é a norma ℓ_1 de β e $t \geq 0$ é um parâmetro de ajuste (tuning parameter)

- O parâmetro t controla a quantidade de preditores que parecem relevantes

Por que o lasso é especial?

- lasso pode gerar estimativas esparsas, ou seja, somente alguns β_j estimados diferentes de zero
  - Isso não ocorre com penalizações ℓ_q para $q > 1$, como penalização **ridge** ($q = 2$)
- lasso é obtido de um problema convexo, o que simplifica a otimização da função objetivo
  - Isso não ocorre com penalizações ℓ_q para $q < 1$, como **escolher a melhor base** ($q = 0$)

Aposta na esparsidade

- Se $p \gg n$ e o modelo verdadeiro não é esparso, então a amostra não é suficiente para estimar bem os parâmetros desconhecidos
  - Modelo não é identificável (o problema assume inúmeras soluções)

Aposta na esparsidade

- Se $p \gg n$ e o modelo verdadeiro não é esparso, então a amostra não é suficiente para estimar bem os parâmetros desconhecidos
  - Modelo não é identificável (o problema assume inúmeras soluções)
- Se o modelo verdadeiro é esparso, só com $k < n$ parâmetros diferentes de zero, então é possível estimar esses parâmetros
  - Métodos esparsos, como o lasso, são capazes de estimar tais modelos mesmo sem sabermos exatamente quais são os k parâmetros importantes

Exemplo

Figura: Dado em formato de uma imagem 512x512

- A imagem foi vetorizada e foi aplicada uma transformada de wavelet
- São os mesmos dados, só que numa base diferente
- A image ao lado é uma amostra de 5.000 dos $512^2 = 262.144$ coeficientes

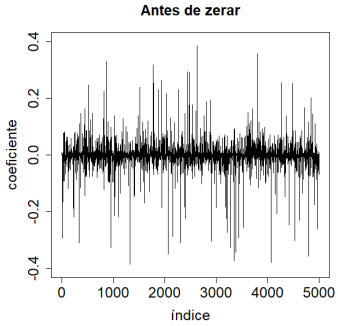

Figura: Coeficientes de wavelet

- Será que esse dado admite uma representação esparsa?
- Para testar, os 80% menores coeficientes (em valor absoluto) foram encolhidos para ter valor zero

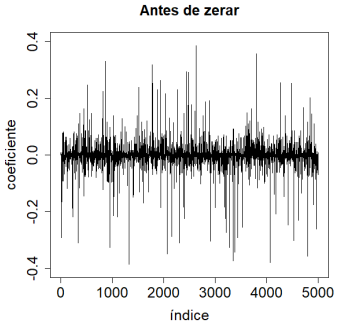

Figura: Coeficientes de wavelet

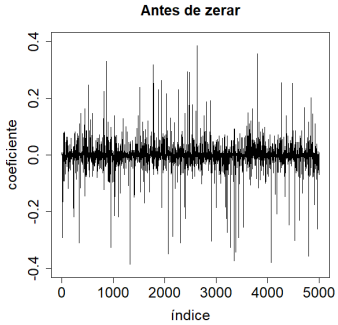

Figura: Coeficientes de wavelet (975 zeros)

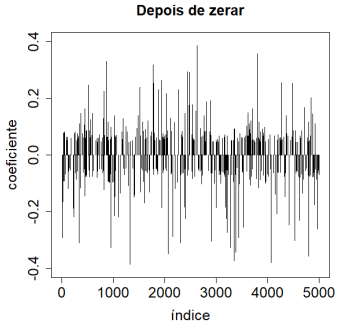

Figura: Versão esparsa desses coeficientes (4492 zeros)

Antes e depois das imagens

Figura: Versão original

Figura: Versão esparsa

O estimador lasso

Regressão linear

- O estimador de mínimos quadrados (MQ) é amplamente utilizado e tem boas propriedades[1]:
  - É não viesado
  - É o estimador linear não viesado de variância mínima (BLUE)
  - Tem distribuição assintótica normal

[1] Sob certas suposições no modelo

Regressão linear

- O estimador de mínimos quadrados (MQ) é amplamente utilizado e tem boas propriedades[1]:
  - É não viesado
  - É o estimador linear não viesado de variância mínima (BLUE)
  - Tem distribuição assintótica normal
- Alguns motivos para considerar alternativas:
  - Melhor generalização/previsão: permitindo algum viés, podemos achar estimadores com menor variabilidade em termos de previsão
  - Maior interpretabilidade: com poucas estimativas diferentes de zero, é mais simples explicar quais preditores são mais relevantes

[1] Sob certas suposições no modelo

- O lasso (*least absolute selection and shrinkage operator*) foi proposto por Tibshirani (1996)

Figura: `https://hastie.su.domains/StatLearnSparsity/images/jsm_booksigning_2015.jpg`

- Representaremos pares de resposta-preditor $\{(y_i, x_i)\}_{i=1}^n$ com um vetor $\mathbf{y} = (y_1, \ldots, y_n)$ de respostas e uma matriz $\mathbf{X} \in \mathbb{R}^{n \times p}$ cuja i-ésima linha contém a observação $x_i^\top \in \mathbb{R}^p$

- O estimador lasso acha a solução $(\hat{\beta}_0, \hat{\beta})$ da minimização com restrição

$$\min_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{y} - \beta_0 \mathbf{1} - \mathbf{X}\beta\|_2^2 \right\} \quad \text{suj.} \quad \|\beta\|_1 \leq t,$$

em que $\mathbf{1}$ é um vetor de n uns, $\|\cdot\|_1$ é a norma ℓ_1 e $\|\cdot\|_2$ é a norma Euclidiana

- Tipicamente as colunas de **X** são **padronizadas** para terem média zero ($\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$) e variância um ($\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$)
  - Evita estimativas serem afetadas pela unidade de medida

- Tipicamente as colunas de **X** são **padronizadas** para terem média zero ($\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$) e variância um ($\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$)
  - Evita estimativas serem afetadas pela unidade de medida
- A resposta também é centralizada para ter média zero ($\frac{1}{n}\sum_{i=1}^{n} y_i = 0$)
  - Podemos **omitir o intercepto** β_0 da otimização
  - O intercepto pode ser estimado como $\hat{\beta}_0 = \bar{y}$, a média original da variável resposta

- Uma versão alternativa da função objetivo do lasso é a sua **forma Lagrangiana**

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\},$$

em que $\lambda \geq 0$ é uma penalização

- As formas Lagrangiana e de minimização restrita são equivalentes, i.e., para cada λ existe um t que gera a mesma solução

- Valores de λ tipicamente são escolhidos via validação cruzada

Validação cruzada

- O valor da penalização λ controla a complexidade do modelo
  - λ baixo: melhor ajuste, mas risco de sobreajuste (overfitting)
  - λ alto: ajuste esparso e mais interpretável, mas risco de "perder o sinal"

- O valor da penalização λ controla a complexidade do modelo
  - λ baixo: melhor ajuste, mas risco de sobreajuste (overfitting)
  - λ alto: ajuste esparso e mais interpretável, mas risco de "perder o sinal"
- Intuito é achar λ com um bom balanço entre esses dois extremos
  - Qualidade medida em termos do erro quadrático médio (EQM) de previsão

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

em que $\hat{y}_i = x_i^\top \hat{\beta}$ é o valor predito para x_i

- Note que se usarmos somente este EQM como critério, poderíamos ter um ajuste “perfeito” tomando $\hat{y}_i = y_i$
  - Alternativa com baixa capacidade de **generalização**, ou seja, de predizer y's para x's não observados nos dados

- Note que se usarmos somente este EQM como critério, poderíamos ter um ajuste "perfeito" tomando $\hat{y}_i = y_i$
  - Alternativa com baixa capacidade de **generalização**, ou seja, de predizer y's para x's não observados nos dados

- Necessidade de estimar o EQM em **dados de teste**, que não foram usados como **dados de treino** (estimação) do modelo

- Note que se usarmos somente este EQM como critério, poderíamos ter um ajuste "perfeito" tomando $\hat{y}_i = y_i$
  - Alternativa com baixa capacidade de **generalização**, ou seja, de predizer y's para x's não observados nos dados
- Necessidade de estimar o EQM em **dados de teste**, que não foram usados como **dados de treino** (estimação) do modelo
- **Validação cruzada** (cross-validation) é uma forma de estimar o EQM em dados de teste, que pode ser usado então para escolher λ

Validação cruzada para λ fixo

- Fixamos um valor de λ e os dados são divididos aleatoriamente em $K > 1$ grupos (tipicamente 5 ou 10)
  1. Um dos grupos é tomado como dados de teste e os demais $K - 1$ são tomados como dados de treino
  2. lasso é ajustado nos dados de treino e o EMQ de previsão é calculado com os dados de teste, obtendo digamos $\widetilde{EQM}_1$
- Processo é repetido $K - 1$ vezes, tomando cada um dos outros grupos como dados de teste por vez, obtendo $\widetilde{EQM}_2, \ldots, \widetilde{EQM}_K$
- O EQM de previsão estimado para λ é $\widehat{EQM}(\lambda) = \frac{1}{K} \sum_{j=1}^{K} \widetilde{EQM}_j$

Escolha de λ via validação cruzada

- Definimos um grid de valores e calculamos $\widehat{EQM}$ para cada um deles
- Escolhemos λ que produz o menor $\widehat{EQM}$

Escolha de λ via validação cruzada

- Definimos um grid de valores e calculamos $\widehat{EQM}$ para cada um deles
- Escolhemos λ que produz o menor $\widehat{EQM}$
- A mesma metodologia pode ser aplicada em outros problemas
  - Ex.: escolha de parâmetros de ajuste, escolha de modelos, etc.

Escolha de λ via validação cruzada

- Definimos um grid de valores e calculamos $\widehat{EQM}$ para cada um deles
- Escolhemos λ que produz o menor $\widehat{EQM}$
- A mesma metodologia pode ser aplicada em outros problemas
  - Ex.: escolha de parâmetros de ajuste, escolha de modelos, etc.
  - **Atenção:** metodologia válida para dados **i.i.d.**, mas para dados dependentes pode levar a uma sub-estimação do EQM de previsão[2]

[2](Bergmeir et al., 2018; Bates et al., 2024)

Exemplo

Exemplo

- Dados sobre taxa de crime nos EUA ($n = 50$ e $p = 5$)
- A variável resposta é a taxa de crimes por milhão de habitantes nos 50 estados dos EUA
- Os preditores são:
  - `policia`: financiamento anual de polícia
  - `em`: percentual com > 25 anos e > 4 anos de ensino médio
  - `fora_em`: percentual entre 16 e 19 anos e fora do ensino médio
  - `facul`: percentual entre 18 e 24 anos na faculdade
  - `facul4`: percentual com > 25 anos e > 4 anos de faculdade
- Preditores foram padronizados e a resposta foi centralizada

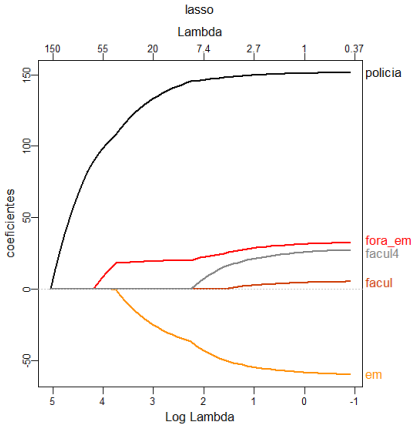

Figura: Trajetória das estimativas do lasso para diferentes penalizações λ

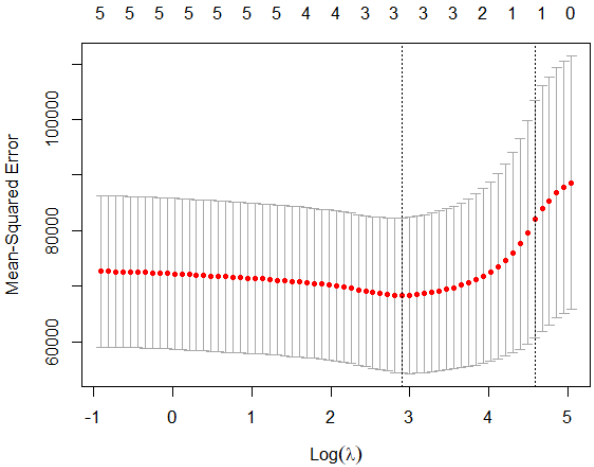

Figura: Estimativas do EQM de previsão via validação cruzada. Barras verticais são bandas de confiança de ± 1 erro padrão.

Tabela: Estimativas e erros padrão utilizando mínimos quadrados ordinários (MQO), lasso e MQO aplicado nas variáveis selecionadas pelo lasso (MQO+lasso)

	MQO	ep	lasso	ep[3]	lasso+MQO	ep
policia	151.76	42.06	135.43	49.24	155.98	39.64
em	-60.67	64.48	-27.12	31.85	-47.44	44.66
fora_em	33.01	59.88	19.60	27.66	20.73	46.66
facul	5.58	64.65	0	19.90	0	0
facul4	28.37	70.14	0	14.94	0	0

[3]Calculado via bootstrap

Alguns detalhes da teoria

- Considere o modelo $\mathbf{y} = \mathbf{X}\beta^* + \varepsilon$, em que β^* é desconhecido e desejamos estimá-lo
- Fixado λ, será que $\hat{\beta} = \hat{\beta}(\lambda, \mathbf{X}, \mathbf{y})$ é um estimador **consistente** de β^* (em algum sentido)?

- Considere o modelo $\mathbf{y} = \mathbf{X}\beta^* + \varepsilon$, em que β^* é desconhecido e desejamos estimá-lo
- Fixado λ, será que $\hat{\beta} = \hat{\beta}(\lambda, \mathbf{X}, \mathbf{y})$ é um estimador **consistente** de β^* (em algum sentido)?
  - Diferentes tipos de consistência: estimação de β^*, previsão de $\mathbf{X}\beta^*$, seleção de variáveis $\{j; \beta_j^* \neq 0\}$

- Considere o modelo $\mathbf{y} = \mathbf{X}\beta^* + \varepsilon$, em que β^* é desconhecido e desejamos estimá-lo
- Fixado λ, será que $\hat{\beta} = \hat{\beta}(\lambda, \mathbf{X}, \mathbf{y})$ é um estimador **consistente** de β^* (em algum sentido)?
  - Diferentes tipos de consistência: estimação de β^*, previsão de $\mathbf{X}\beta^*$, seleção de variáveis $\{j; \beta_j^* \neq 0\}$
  - Diferentes formas de assintótica: permite $n, p \to \infty$

- Considere o modelo $\mathbf{y} = \mathbf{X}\beta^* + \varepsilon$, em que β^* é desconhecido e desejamos estimá-lo
- Fixado λ, será que $\hat{\beta} = \hat{\beta}(\lambda, \mathbf{X}, \mathbf{y})$ é um estimador **consistente** de β^* (em algum sentido)?
  - Diferentes tipos de consistência: estimação de β^*, previsão de $\mathbf{X}\beta^*$, seleção de variáveis $\{j; \beta_j^* \neq 0\}$
  - Diferentes formas de assintótica: permite $n, p \to \infty$
  - Resultados **não-assintóticos**: $P(\hat{\beta} \approx \beta^*) > 1 - \epsilon$, em que $\epsilon \equiv \epsilon(n, p)$ é pequeno

- Sob condições apropriadas:

$$\frac{1}{n}\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta^*\|_2^2 = O_P\left(\frac{K \log p}{n}\right)$$

$$\|\hat{\beta} - \beta^*\|_2 = O_P\left(\sqrt{\frac{K \log p}{n}}\right), \quad \text{quando } n, p \to \infty$$

- Sob condições apropriadas:

$$\frac{1}{n}\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta^*\|_2^2 = O_P\left(\frac{K \log p}{n}\right)$$

$$\|\hat{\beta} - \beta^*\|_2 = O_P\left(\sqrt{\frac{K \log p}{n}}\right), \quad \text{quando } n, p \to \infty$$

- Sendo $\mathcal{S} = \{j; \beta_j^* \neq 0\}$ o **suporte** de β^*, se $\min_{i \in \mathcal{S}} |\beta^*|$ é suficientemente alto (**condição beta-min**)

$$P(\hat{\mathcal{S}} \supseteq \mathcal{S}) \to 1 \text{ quando } n, p \to \infty,$$

em que $\hat{\mathcal{S}} = \{j; \hat{\beta}_j \neq 0\}$

Referências

S. Bates, T. Hastie, and R. Tibshirani. Cross-validation: what does it estimate and how well does it do it? *Journal of the American Statistical Association*, 119(546): 1434–1445, 2024.

C. Bergmeir, R. J. Hyndman, and B. Koo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83, 2018.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, Boca Raton, 2015.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.