Universidade Federal do Paraná

Laboratório de Estatística e Geoinformação - LEG

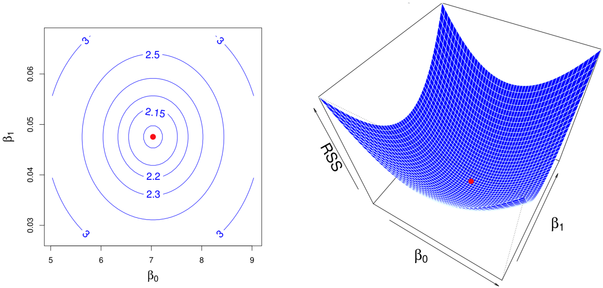
Regularização

Eduardo Vargas Ferreira

Estimando os coeficientes

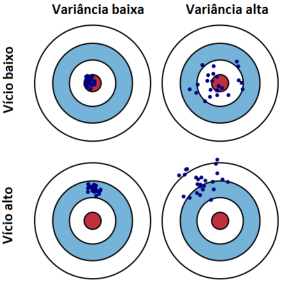
- Quando estimamos os parâmetros da regressão por **mínimos quadrados ordinários**, estamos interessados na seguinte minimização:

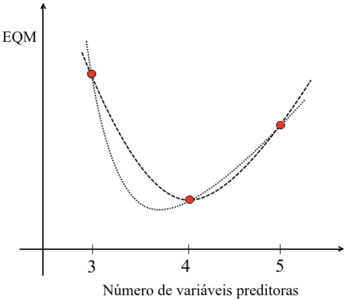
$$min\{J[y_i, h(\boldsymbol{x})]\} = min\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\}$$



Bias-Variance Trade-Off

- Uma solução é flexibilizar o modelo admitindo certo vício das estimativas.



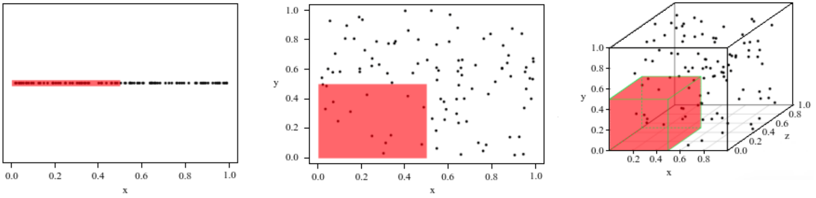


$$\mathrm{E}\left[y_0 - h(\boldsymbol{x}_0)\right]^2 = \mathrm{V}ar\left[h(\boldsymbol{x}_0)\right] + \left[\text{Vício}(h(\boldsymbol{x}_0)\right]^2 + \mathrm{V}ar(\varepsilon).$$

Problemas dos mínimos quadrados ordinários

- Em situações com "*small n, large p*" a maioria dos métodos modernos de análise de dados falha, por diferentes razões, p. ex.:
  - ⋆ **Modelos Lineares Generalizados:** falham, pois a matriz do modelo não tem posto completo;
  - ⋆ **Random Forests** falham, pois a probabilidade de selecionar variáveis importantes diminui muito.
  - ⋆ **Análise de Clusters** e métodos baseados em distâncias no plano cartesiano falham devido à "maldição da dimensionalidade".

42% dos dados capturados | 14% dos dados capturados | 7% dos dados capturados



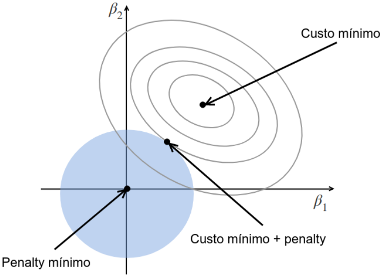
Regularização

Uma solução é a Regularização

- Regulamos o número de variáveis, impondo um custo ao algoritmo:

$$\underset{\boldsymbol{\beta}}{min} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \textbf{ sujeito a } g(\boldsymbol{\beta}) < t, \ \text{com } t > 0$$

- $g(\boldsymbol{\beta})$ representa a **função penalty** (*shrinkage penalty*).



Como fazemos isso?

- Estamos diante de uma otimização com restrição:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \textbf{ sujeito a } g(\boldsymbol{\beta}) < t, \ \text{com } t > 0$$

- Fazemos isso aumentando a função objetivo:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda g(\boldsymbol{\beta}), \ \text{com } \lambda > 0$$

- **t e λ são inversamente proporcionais**.

Função penalty

- Temos a função objetivo aumentanda:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda g(\boldsymbol{\beta}), \ \text{com } \lambda > 0$$

- Utilizaremos a **família das potências** para penalizar o modelo:

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

- λ é o **tuning parameter**, determinado separadamente. Ele controla o impacto do *penalty*, $g(\boldsymbol{\beta})$, nas estimativas dos parâmetros.

Padronização dos preditores

- Os coeficientes obtidos por mínimos quadrados ordinários são equivariantes por transformação de escala;
- Na regressão penalizada, $\mathbf{X}_j \hat{\beta}_{j,\lambda}^{restrito}$ depende não somente de λ, mas da escala do j-ésimo preditor.

$$\underset{\boldsymbol{\beta}}{min} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q$$

- Diante disso, deve-se padronizar os preditores na regressão penalizada:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}}.$$

Penalização Ridge

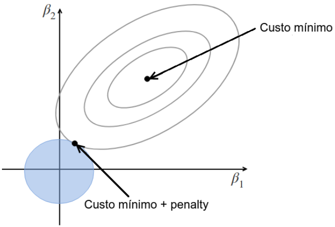
$q = 2$ - Penalização Ridge

- Neste caso, o problema de otimização utiliza o *penalty* ℓ^2:

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

- No caso de vetores em $\boldsymbol{X}$ ortonormais

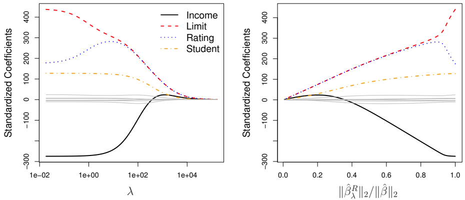
$$\hat{\beta}_\lambda^R = \frac{\hat{\beta}^{OLS}}{1+\lambda}.$$



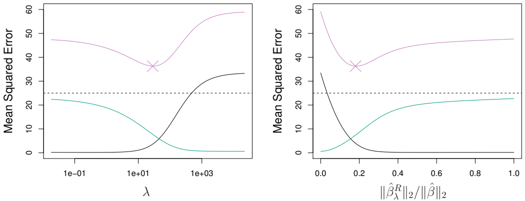
- Os $\beta's$ podem ser próximos de zero, mas não assumem esse valor.

Exemplos

- **Inadimplência no cartão de crédito:** o objetivo é prever se um cliente será ou não inadimplente no próximo mês.



- **Exemplo simulado:** foi gerado utilizando 45 preditores relacionados à resposta, em que nenhum dos verdadeiros $\beta_1, \beta_2, \ldots, \beta_{45}$ eram zero.



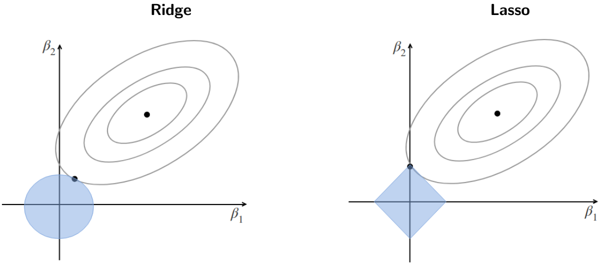
Penalização Lasso

$q = 1$ - Penalização Lasso

- Os coeficientes *Lasso*, $\hat{\beta}^L_\lambda$, minimizam a quantidade

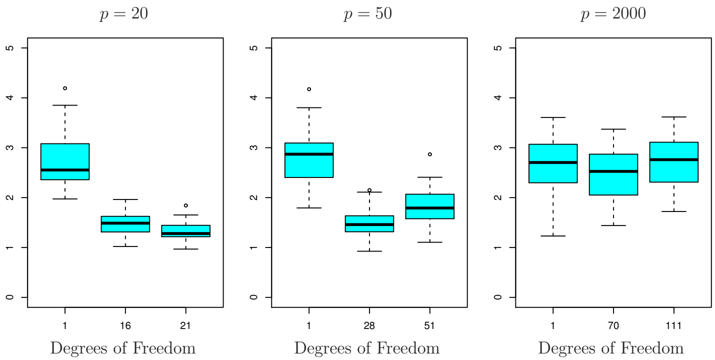
$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- O *penalty* ℓ^1 funciona também como um **selecionador de variáveis**.



Exemplo simulado

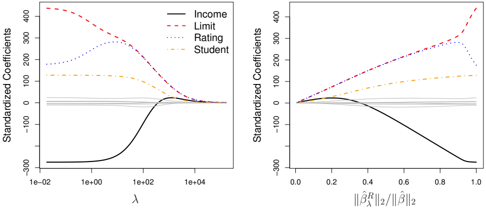
- No exemplo abaixo temos 100 observações com $p = 20, 50$ e 2000, das quais apenas 20 são relacionadas com a resposta;



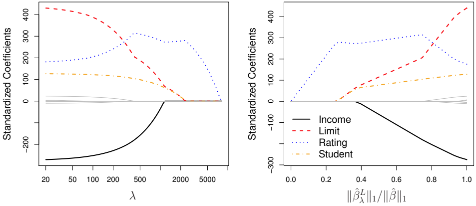
- Os graus de liberdade, substituem o λ, e representam o número de parâmetros estimados não nulos.

Exemplo: Inadimplência no cartão de crédito

- **Regressão Ridge**

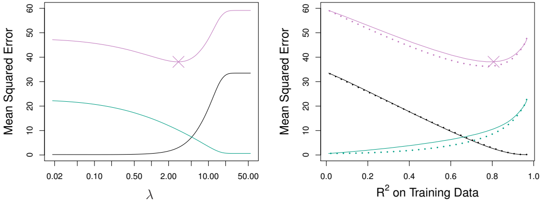


- **Regressão Lasso**

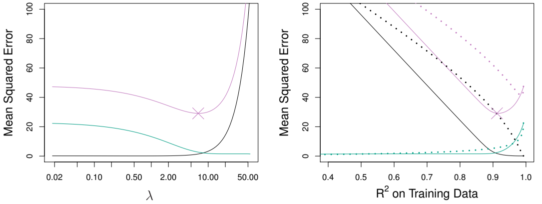


Comparação entre Ridge e Lasso

- Este exemplo foi gerado utilizando 45 preditores relacionados à resposta, em que nenhum dos verdadeiros coeficientes $\beta_1, \beta_2, \ldots, \beta_{45}$ eram zero;



- No exemplo seguinte, a resposta é função apenas de 2 dos 45 preditores. I.e., 43 dos verdadeiros coeficientes $\beta_1, \beta_2, \ldots, \beta_{45}$ eram zero.

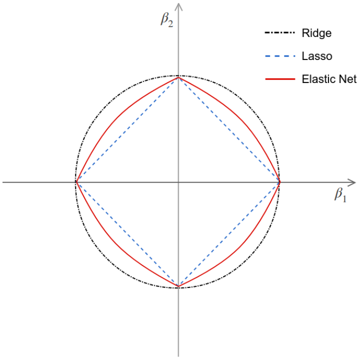


Penalização Elastic net

Elastic net

- **Elastic net** é um compromisso entre a regressão *Ridge* e *Lasso*. Os coeficientes elastic net, $\hat{\beta}^{E}_{\lambda}$, minimizam a quantidade

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

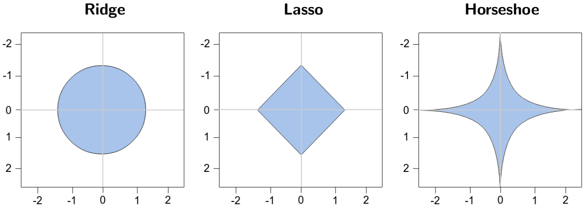


Penalização horseshoe

$q < 1$ - Penalização horseshoe

- Ela favorece mais ainda a presença de 0's (maior esparsidade);

- Ou seja, tende a encontrar as elipses geradas pelos mínimos quadrados em cima dos eixos com mais frequência que *Ridge* e *Lasso*;



- E quando $q = 0$ voltamos ao **Best subset selection**.

Regressão com Componentes Principais

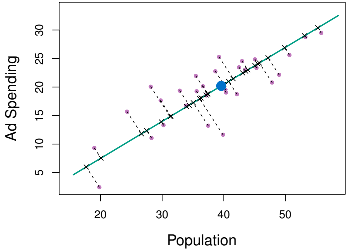
Análise de Componentes Principais (ACP)

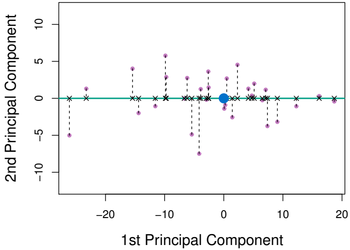
- Suponha que temos a informação sobre a população (`pop`) em 100 cidades dos EUA e gasto com propaganda de determinada empresa (`ad`);

- Podemos resumir essas variáveis em apenas uma, Z_1, da seguinte forma:

$$
\begin{aligned}
Z_1 &= \phi_{11}\left(\texttt{pop} - \overline{\texttt{pop}}\right) + \phi_{21}\left(\texttt{ad} - \overline{\texttt{ad}}\right) \\
&= 0.839 \times \left(\texttt{pop} - \overline{\texttt{pop}}\right) + 0.544 \times \left(\texttt{ad} - \overline{\texttt{ad}}\right).
\end{aligned}
$$

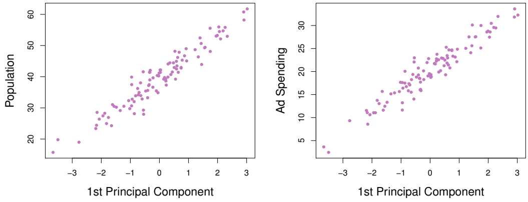
- Tal que $Var(Z_1)$ seja o máximo possível.



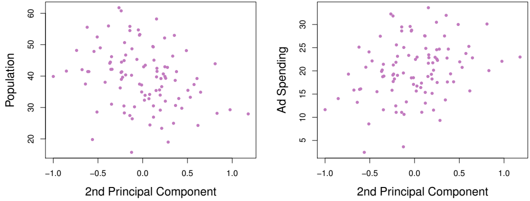


Análise de Componentes Principais (ACP)

$$Z_1 = 0.839 \times (\texttt{pop} - \overline{\texttt{pop}}) + 0.544 \times \left(\texttt{ad} - \overline{\texttt{ad}}\right)$$



$$Z_2 = 0.839 \times (\texttt{pop} - \overline{\texttt{pop}}) + 0.544 \times \left(\texttt{ad} - \overline{\texttt{ad}}\right)$$



Regressão com Componentes Principais

- Considere que $Z_1, Z_2, \ldots, Z_M$ $(M < p)$ represente as combinações lineares das variáveis originais X_j, $j = 1, \ldots, p$:

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j, \text{ com } m = 1, \ldots, M.$$

- Assumindo que direção de máxima variabilidade é a mesma associada ao Y, podemos construir uma regressão utilizando as C.P's, da forma:

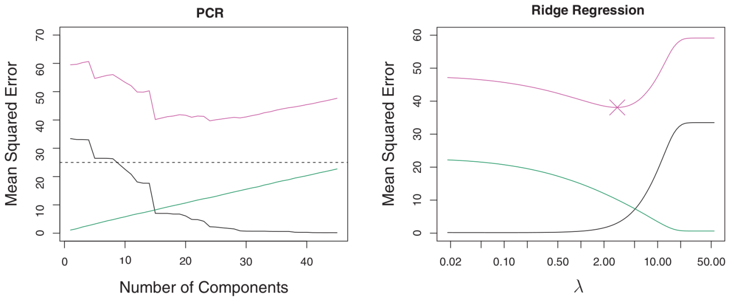
$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \varepsilon_i, \text{ com } i = 1, \ldots, n.$$

- Mas, note que

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{jm} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}.$$

Exemplos simulados

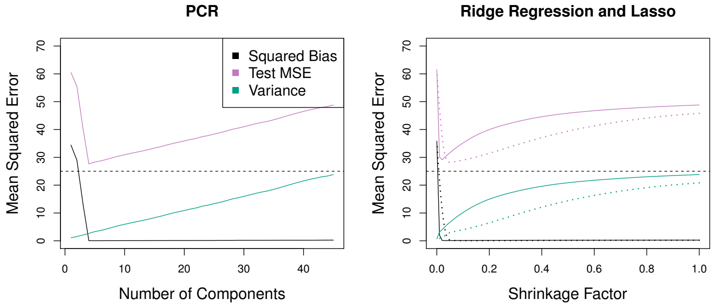
- Foi gerado utilizando 45 preditores relacionados à resposta, em que nenhum dos verdadeiros $\beta_1, \beta_2, \ldots, \beta_{45}$ eram zero.



- Os resultados via PCR e Ridge são muito próximos. Pode-se dizer até que Ridge é uma versão contínua de PCR.

Exemplos simulados

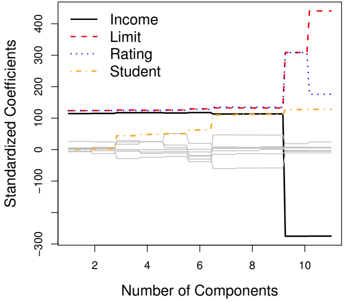
- Os dados foram gerados, tal que a resposta dependesse apenas das 5 primeiras componentes principais.

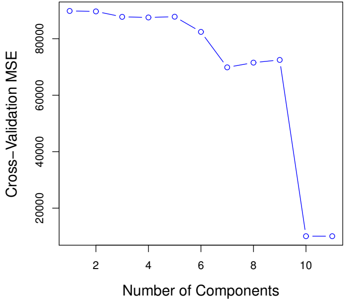


- A PCR se sai melhor quando poucas componentes são necessárias para explicar a maior parte da variabilidade dos dados.

Exemplo: Inadimplência no cartão de crédito

- Nestes dados o menor erro de validação foi obtido quando $M = 10$. Equivalente aos mínimos quadrados.





Selecionando o tuning parameter

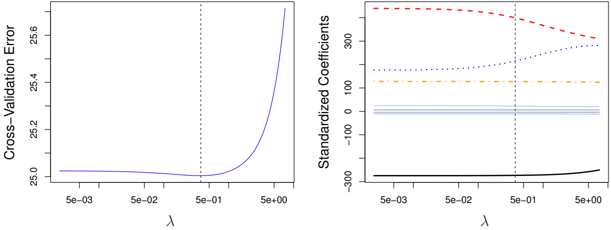
Selecionando o tuning parameter, λ

- **Validação cruzada** fornece uma maneira simples de resolver este problema:

  (a) A partir de uma grade de valores de λ, calculamos a taxa de erro de validação (para cada λ);

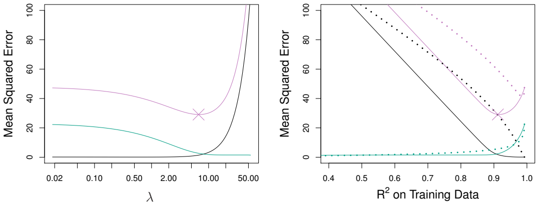
  (b) Escolhemos o valor de λ que fornece a menor taxa de erro;

  (c) Ajustamos novamente o modelo, utilizando todas as observações disponíveis, com o valor de λ encontrado anteriormente.

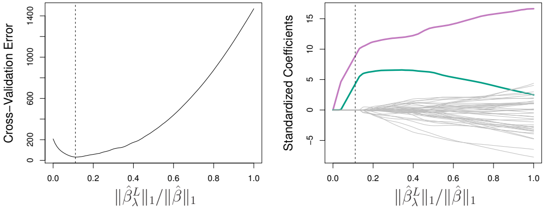


Exemplo simulado

- Este exemplo foi gerado utilizando 45 preditores, em que 43 dos verdadeiros coeficientes $\beta_1, \beta_2, \ldots, \beta_{45}$ eram zero.



- Aplicamos validação cruzada $10-$*fold* para selecionar o melhor λ.



Referências

- James, G., Witten, D., Hastie, T. e Tibshirani, An Introduction to Statistical Learning, 2013;
- Hastie, T., Tibshirani, R. e Friedman, J., The Elements of Statistical Learning, 2009;
- Lantz, B., Machine Learning with R, Packt Publishing, 2013;
- Tan, Steinbach, and Kumar, Introduction to Data Mining, Addison-Wesley, 2005;

- Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani