Regularización

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

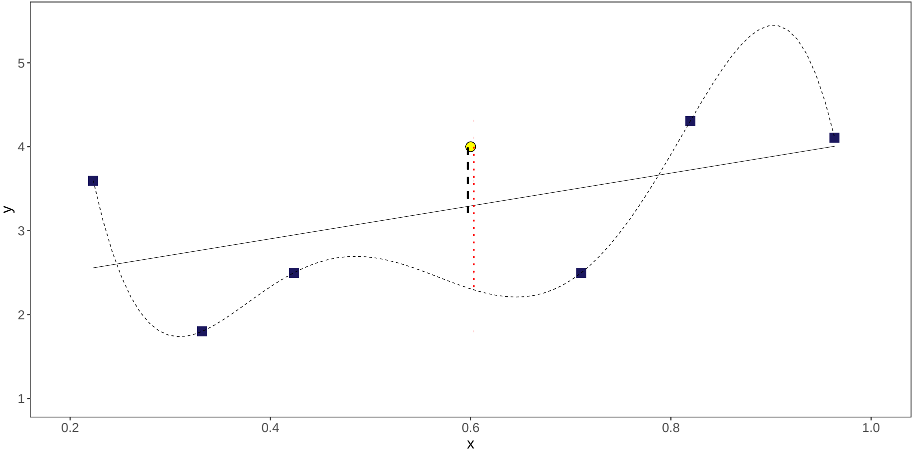
Universidad de los Andes

Agenda

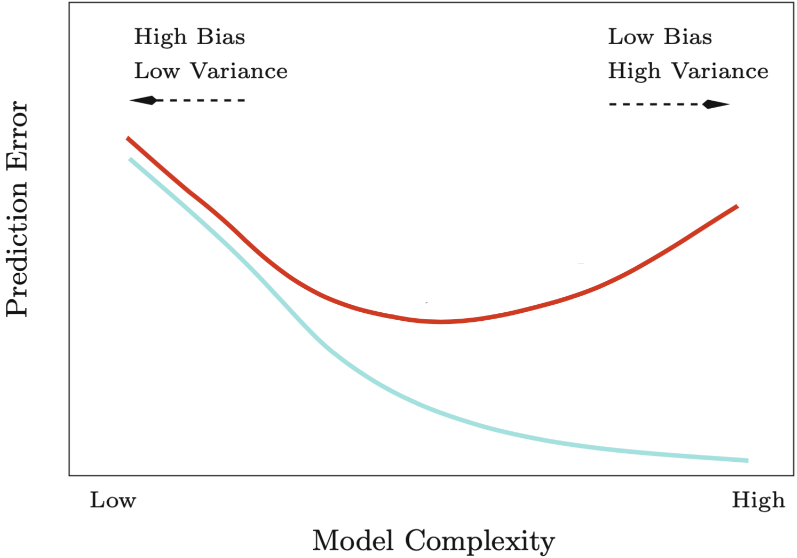
Agenda

1. Recap: Predicción y Overfit
2. Regularización
   - Recap: OLS Mechanics
   - Ridge
   - Lasso
   - $k > n$
   - Ridge and Lasso: Pros and Cons
   - Elastic Net

Overfit y Predicción fuera de Muestra



Overfit y Predicción fuera de Muestra



Overfit y Predicción fuera de Muestra

- ML nos interesa la predicción fuera de muestra
- Overfit: modelos complejos predicen bien dentro de muestra, pero tienden a hacer un mal trabajo fuera de muestra
- Hay que elegir el modelo que "mejor" prediga fuera de muestra (out-of-sample)
  - Penalización ex-post: AIC, BIC, etc
  - Métodos de Remuestreo
    - Enfoque del conjunto de validación
    - LOOCV
    - Validación cruzada en K-partes (5 o 10)

Agenda

Agenda

Regularización: Motivación

- Las técnicas econometricas estándar no están optimizadas para la predicción.
- OLS minimiza el error *"dentro de muestra"*, eligiendo βs de forma tal que

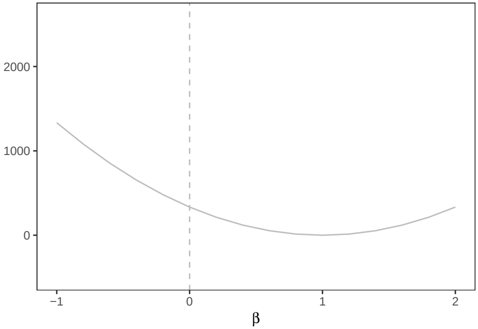
$$min_\beta E(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 \tag{1}$$

- o en forma matricial

$$min_\beta E(\beta) = (y - X\beta)'(y - X\beta) \tag{2}$$

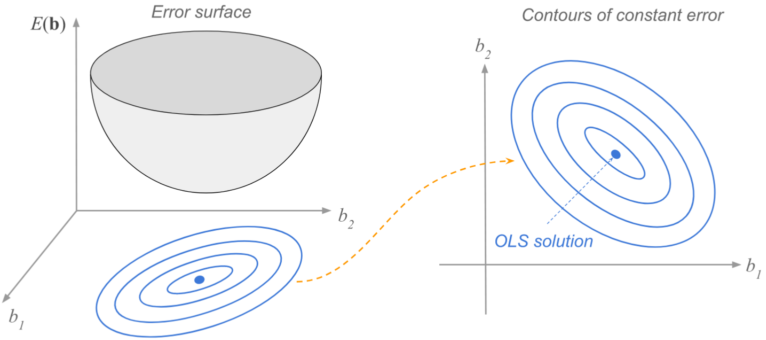
OLS 1 Dimension

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 \tag{3}$$



Intuición en 2 Dimensiones (OLS)

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i2}\beta_2)^2 \tag{4}$$



Fuente: https://allmodelsarewrong.github.io

Regularización: Motivación

- Las técnicas econometricas estándar no están optimizadas para la predicción.

- OLS minimiza el error *"dentro de muestra"*, eligiendo βs de forma tal que

$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 \tag{1}$$

- o en forma matricial

$$min_{\beta}E(\beta) = (y - X\beta)'(y - X\beta) \tag{2}$$

- Predicción queremos hacer un buen trabajo **fuera de muestra**

Regularización

- Asegurar cero sesgo dentro de muestra crea problemas fuera de muestra: trade-off Sesgo-Varianza
- Las técnicas de machine learning fueron desarrolladas para hacer este trade-off de forma empírica.
- Vamos a proponer modelos del estilo

$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda\sum_{j=1}^{p}R(\beta_j) \tag{5}$$

- donde R es un regularizador que penaliza funciones que crean varianza

Agenda

Ridge

- Para un $\lambda \geq 0$ dado, consideremos ahora el siguiente problema de optimización

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} (\beta_j)^2 \tag{6}$$

- o en forma matricial

$$min_{\beta} E(\beta) = (y - X\beta)'(y - X\beta) + \lambda \beta' \beta \tag{7}$$

Ridge: Intuición en 1 Dimension

- 1 predictor estandarizado
- El problema:

$$min_\beta E(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\beta^2 \quad (8)$$

- La solución?

Intuición en 2 Dimensiones (Ridge)

- En 2 dim

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i2}\beta_2 + \lambda \left(\beta_1^2 + \beta_2^2\right) \tag{9}$$

- el dual es

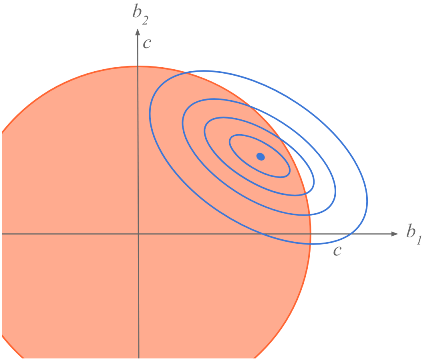
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \tag{10}$$

sujeto a

$$\left((\beta_1)^2 + (\beta_2)^2\right) \leq c$$

Intuición en 2 Dimensiones (Ridge)

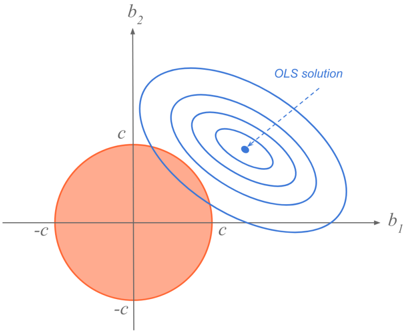
$$min_\beta E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \quad (11)$$



Fuente: https://allmodelsarewrong.github.io

Intuición en 2 Dimensiones (Ridge)

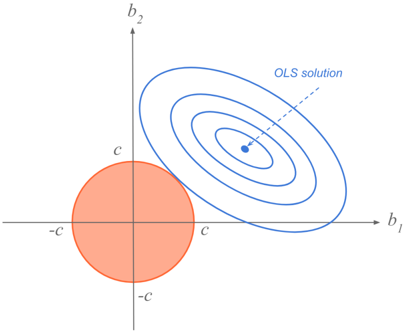
$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \tag{12}$$



Fuente: https://allmodelsarewrong.github.io

Intuición en 2 Dimensiones (Ridge)

$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \qquad (13)$$



Fuente: https://allmodelsarewrong.github.io

Términos generales

- En regresión multiple (X es una matriz $n \times k$)
- Regresión: $y = X\beta + u$
- OLS

$$\hat{\beta}_{ols} = (X'X)^{-1}X'y$$

- Ridge

$$\hat{\beta}_{ridge} = (X'X + \lambda I)^{-1}X'y$$

Ridge vs OLS

- Ridge es sesgado $E(\hat{\beta}_{ridge}) \neq \beta$
- Pero la varianza es menor que la de OLS
- Para ciertos valores del parámetro $\lambda \Rightarrow MSE_{OLS} > MSE_{ridge}$

Example

photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

Agenda

Lasso

- Para un $\lambda \geq 0$ dado, consideremos el siguiente problema de optimización

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{14}$$

- o en forma matricial

$$min_{\beta} E(\beta) = (y - X\beta)'(y - X\beta) + \lambda ||\beta||_1 \tag{15}$$

Lasso Intuición en 1 Dimension

$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda|\beta| \tag{16}$$

- Un solo predictor, un solo coeficiente
- Si $\lambda = 0$

$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 \tag{17}$$

- y la solución es

$$\hat{\beta}_{OLS} \tag{18}$$

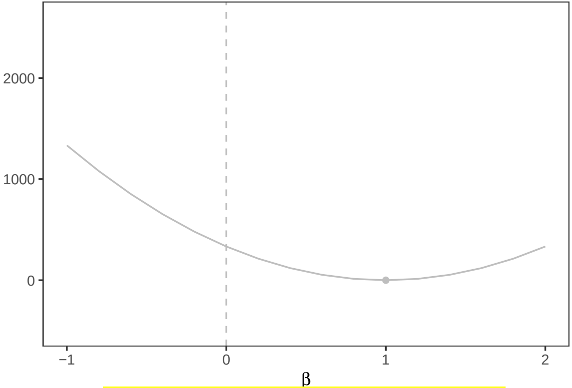
Intuición en 1 Dimension

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta| \tag{19}$$

Intuición en 1 Dimensión

$\hat{\beta} > 0$

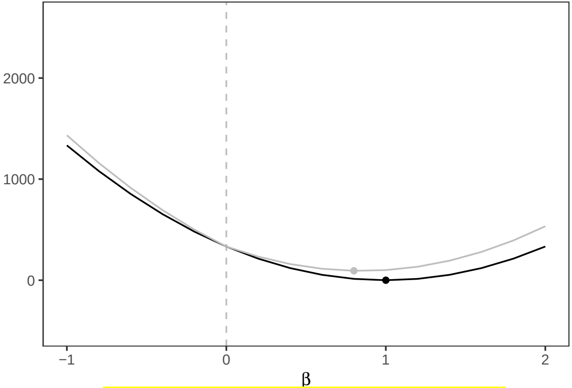
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{20}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

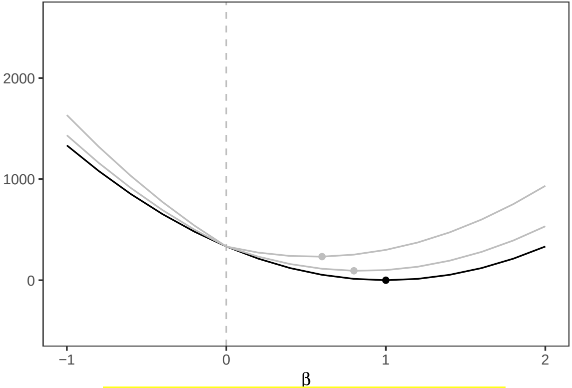
$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\beta \quad (21)$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

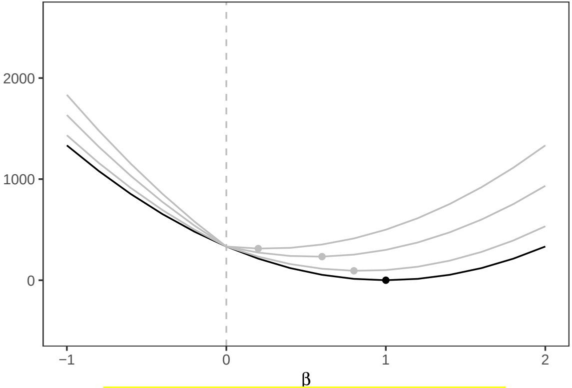
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i \beta)^2 + \lambda \beta \tag{22}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

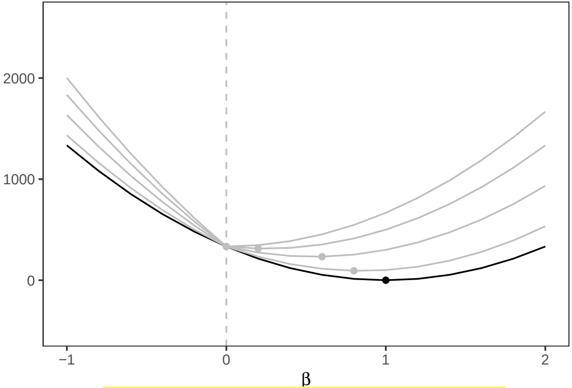
$$min_\beta E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \quad (23)$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

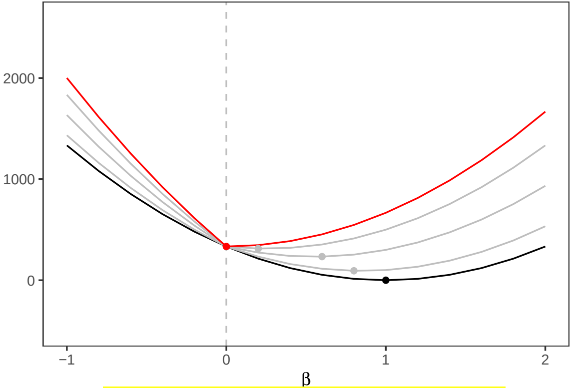
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{24}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

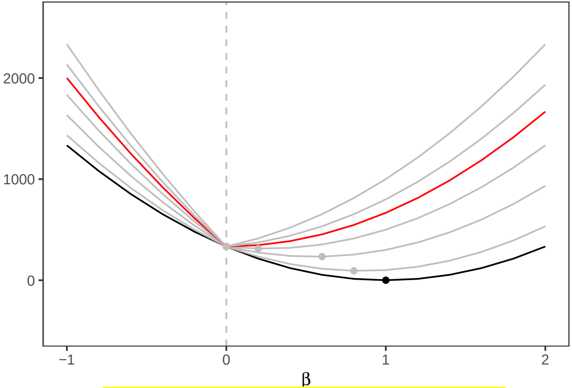
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{25}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{26}$$



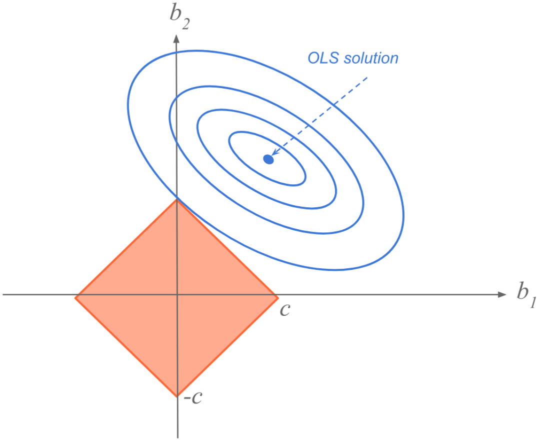
Intuición en 1 Dimension

Solución analitica

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta| \tag{27}$$

Intuición en 2 Dimensiones (Lasso)

$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } (|\beta_1| + |\beta_2|) \leq c \tag{28}$$



Example

photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

Sobre la implementación

- Importante para aplicación:
  - Escala los datos
  - Seleccion de λ

Selección de λ

- Asegurar cero sesgo dentro de muestra crea problemas fuera de muestra: trade-off Sesgo-Varianza
- Hacemos este trade-off de forma empírica.

$$min_\beta E(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} R(\beta_j) \tag{29}$$

- λ es el precio al que hacemos este trade off
- Como elegimos λ?

Agenda

More predictors than observations ($k > n$)

- What happens when we have more predictors than observations ($k > n$)?
  - OLS?
  - Ridge?
  - Lasso?

OLS when $k > n$

- Rank? Max number of rows or columns that are linearly independent
  - Implies $rank(X_{k \times n}) \leq min(k, n)$
- MCO we need $rank(X_{k \times n}) = k \implies k \leq n$
- If $rank(X_{k \times n}) = k$ then $rank(X'X) = k$
- If $k > n$, then $rank(X'X) \leq n < k$ then $(X'X)$ cannot be inverted
- Ridge and Lasso work when $k \geq n$

Ridge when $k > n$

$$min_{\beta} E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - \cdots - x_{ik}\beta_k)^2 + \lambda \sum_{j=1}^{k}(\beta_j)^2 \tag{30}$$

- Solution $\rightarrow$ data augmentation
- Intuition: Ridge “adds” k additional points.
- Allows us to “deal” with $k \geq n$

Ridge when $k > n$

Adding k additional points

Lasso when $k > n$

- In the $k > n$ case, the lasso selects at most n variables before it saturates,
- This is because because of the nature of the convex optimization problem.

Agenda

1. Recap: Predicción y Overfit

2. Regularización
   - Recap: OLS Mechanics
   - Ridge
   - Lasso
   - $k > n$
   - Ridge and Lasso: Pros and Cons
   - Elastic Net

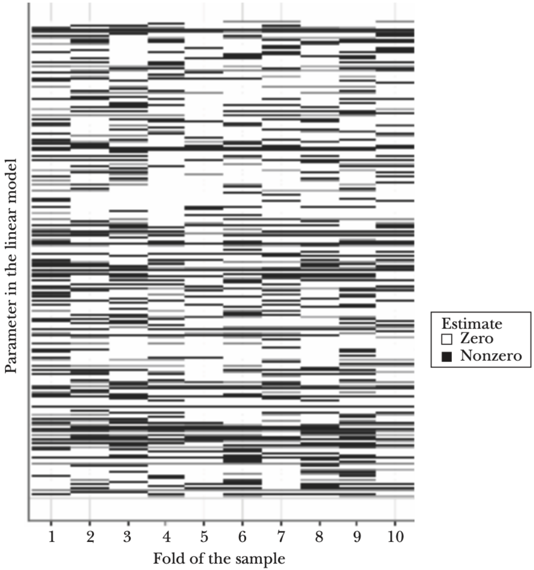
Ridge and Lasso: The good and the bad

- Objective 1: Accuracy
  - Minimize prediction error (in one step) $\rightarrow$ Ridge, Lasso
- Objective 2: Dimensionality
  - Reduce the predictor space $\rightarrow$ Lasso's free lunch
- More predictors than observations ($k > n$)
  - OLS fails
  - Ridge augments data
  - Lasso chooses at most n variables

Ridge and Lasso: The good and the bad

- When we have a group of highly correlated variables,
  - Lasso chooses only one.

Ridge and Lasso: The good and the bad



Ridge and Lasso: The good and the bad

- When we have a group of highly correlated variables,
  - Lasso chooses only one. Makes it unstable for prediction.
  - Ridge shrinks the coefficients of correlated variables toward each other.
  - For usual $n > k$ situations, if there are high correlations between predictors, it has been empirically observed that the prediction performance of the lasso is dominated by ridge regression (Tibshirani, 1996).
  - For $k > n$ situations?

Agenda

1. Recap: Predicción y Overfit

2. Regularización
    - Recap: OLS Mechanics
    - Ridge
    - Lasso
    - $k > n$
    - Ridge and Lasso: Pros and Cons
    - Elastic Net

Elastic net

$$min_\beta EN(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\left(\alpha\sum_{j=1}^{p}|\beta_j| + \frac{(1-\alpha)}{2}\sum_{j=1}^{p}(\beta_j)^2\right) \quad (31)$$

- Si $\alpha = 1$ Lasso
- Si $\alpha = 0$ Ridge

Elastic Net

- The elastic net simultaneously does automatic variable selection and continuous shrinkage, and it can select groups of correlated variables.
- It is like a stretchable fishing net that retains ‘all the big fish’.
- Simulation studies and real data examples show that the elastic net often outperforms the lasso in terms of prediction accuracy

$$min_{\beta} EN(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\left(\alpha\sum_{j=1}^{p}|\beta_j| + \frac{(1-\alpha)}{2}\sum_{j=1}^{p}(\beta_j)^2\right) \quad (32)$$

- Strict convexity part of the penalty (ridge) solves the grouping instability problem
- How to choose (λ, α)? $\rightarrow$ Bidimensional Crossvalidation

Example

photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/