# Regularización

## Lasso, Ridge, ElasticNet

Carlos Pita

26 de octubre de 2023

# Trade-off Sesgo / Varianza

Recordemos el problema

$$y = f(x) + \epsilon = \beta \cdot x + \epsilon$$

con $\beta \in \mathbb{R}^p$ y $\epsilon$ aleatorio tal que $E(\epsilon) = 0$. Queremos encontrar una $\hat{f}$ (*ie.* un vector $\hat{\beta}$) que haga pequeño

$$\underbrace{ECM(\hat{f}(x))}_{\substack{\text{Error de} \\ \text{generalización}}} = E((y - \hat{f}(x))^2)$$

$$= \underbrace{Sesgo(\hat{f}(x))^2 + Var(\hat{f}(x))}_{\text{Trade-off sesgo/varianza}} + \underbrace{\sigma^2}_{\substack{\text{Error} \\ \text{irreductible}}}$$

# Complejidad

- Si el modelo es demasiado simple (*ie.* tiene pocos grados de libertad), entonces $E(\hat{f}(x)) \neq E(f(x))$ no importa cuán grande sea la muestra: tenemos sesgo o error sistemático.
- Si el modelo es demasiado complejo (*ie.* tiene demasiados grados de libertad), entonces el estimador puede ajustarse a regularidades espurias de la muestra incrementando $Var(\hat{f}(x))$: tenemos sobre-ajuste.
- Por lo tanto, el modelo no debe ser ni muy simple ni muy complejo: sería deseable explorar una familia de soluciones $\{\hat{f}_\lambda\}$ indexadas por un índice de complejidad $\lambda$ y quedarse con el $\lambda^\star$ que alcance el mejor trade-off entre sesgo y varianza.

# Colinealidad

- Si los regresores $x_i$ y $x_j$ están muy correlacionados, pequeñas fluctuaciones en la muestra pueden resultar en amplias variaciones de los coeficientes $\hat{\beta}_i$ y $\hat{\beta}_j$, incrementando así $Var(\hat{f}(x))$: el problema está mal condicionado.
- En el extremo, si $x_i$ y $x_j$ son perfectamente colineales, el problema ni siquiera tiene una solución única: está mal planteado.
- Querríamos evitar problemas mal condicionados / planteados repartiendo más suavemente los pesos entre los regresores muy correlacionados.

# Regularización

- La regularización produce una familia de problemas relacionados con la minimización de pérdida original a través de un término regulable de contracción (*shrinkage*).
- La familia está indexada por hiperparámetros (en general uno, que llamaremos $\lambda$) que regulan la complejidad del modelo disminuyendo o aumentando el término de contracción.
- Distintos tipos de regularización producen soluciones con diferentes características deseables (parsimoniosas —*sparse*—, bien condicionadas, etc.).
- Desde un punto de vista bayesiano se puede pensar la regularización como un supuesto *a priori* sobre las posibles soluciones.

# Regularización

Si el problema original consistía en minimizar la pérdida $L$

$$\hat{f} = \underset{f}{\text{argmin}}\, L(y, f(x))$$

el problema regularizado será

$$\hat{f}_R = \underset{f}{\text{argmin}}\, L(y, f(x)) + R(f)$$

con

- $R(f) = \lambda||\beta||_2^2$ para regularización ridge.
- $R(f) = \lambda||\beta||_1$ para regularización lasso.
- $R(f) = \lambda_1||\beta||_1 + \lambda_2||\beta||_2^2$ para regularización elasticnet.

donde, recordemos, $f(x) = \beta \cdot x + \epsilon$. En todos los casos, $R(f)$ penaliza valores mayores de $\beta$, tanto más cuanto mayor sea $\lambda$.

# Ridge (*aka. Tíjonov*)

## Término de regularización

$$\lambda||\beta||_2^2 = \lambda\beta_1^2 + \cdots + \lambda\beta_p^2$$

- $\lambda$ regula la complejidad del modelo comprimiendo los $\beta_i$ de manera aproximadamente proporcional.
- Aumentando $\lambda$ el problema se vuelve mejor condicionado, repartiendo suavemente el peso dentro de clusters de regresores correlacionados.
- Desde un punto de vista bayesiano, es equivalente a suponer un *a priori* gaussiano $\beta \sim N(0, \lambda I)$.

# Lasso

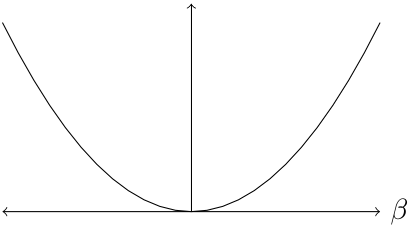
## Término de regularización

$$\lambda||\beta||_1 = \lambda|\beta_1| + \cdots + \lambda|\beta_k|$$

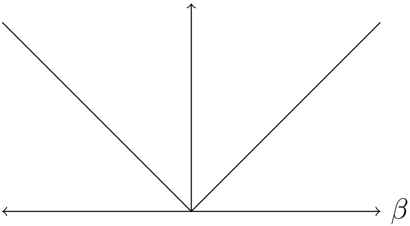
- $\lambda$ regula la complejidad del modelo comprimiendo los $\beta_i$ de manera aproximadamente constante sin cruzar el cero (*soft thresholding*).
- Aumentando $\lambda$ más $\beta_i$ se vuelven exactamente 0, por lo que es posible considerar lasso como un método de selección de variables para obtener modelos parsimoniosos (*sparse*).
- Desde un punto de vista bayesiano, es equivalente a suponer un *a priori* laplaciano (doble exponencial).

# Ridge vs. Lasso

$$||\beta||_2^2 = \beta^2$$

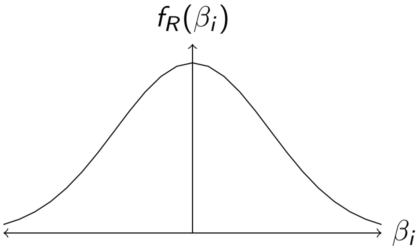



$$||\beta||_1 = |\beta|$$



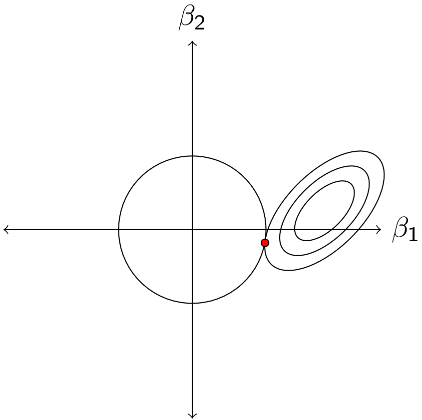

El término de *shrinkage* de Ridge es suave, mientras que el de Lasso no es diferenciable en el 0. Si bien ambas funciones son convexas, desde un punto de vista numérico es más difícil resolver el problema de Lasso.

# Ridge vs. Lasso



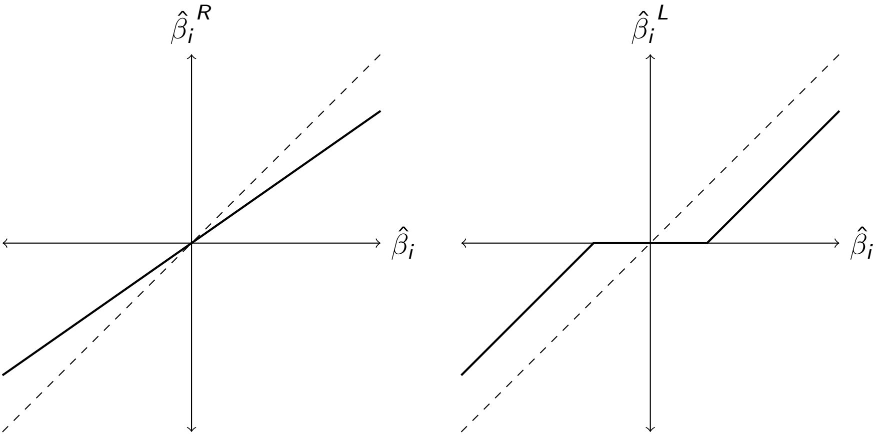

El problema de Ridge equivale a asumir un *a priori* gaussiano, mientras que el problema de Lasso asume uno laplaciano. Si bien ambas distribuciones concentran los valores más probables de $\beta_i$ alrededor del 0, el máximo es mucho más "suave" en el modelo gaussiano, mientras que el laplaciano promueve soluciones más *sparse*.

# Ridge vs. Lasso



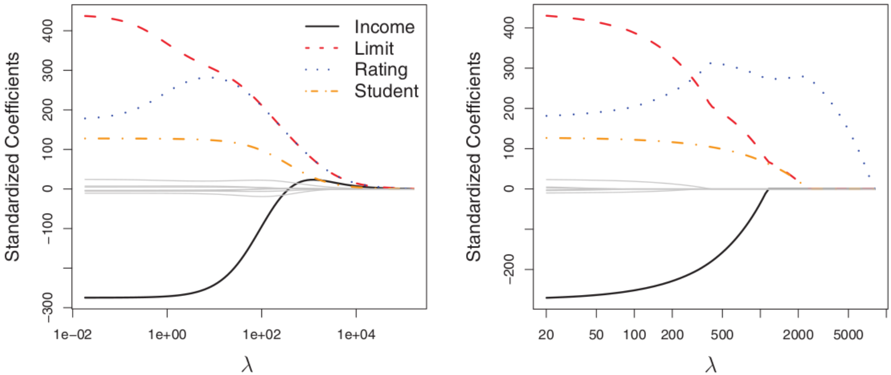

Las regularizaciones pueden pensarse como "restricciones presupuestarias" a los posibles valores de $\beta$. La curva de nivel de $L$ tangente a la restricción determina la solución. Lasso es más propenso a generar soluciones en las que alguno de los coeficientes es exactamente cero.

# Ridge vs. Lasso




Cuando $L = RSS$ y la matriz $X$ es ortonormal, Ridge comprime las estimaciones *OLS* de forma proporcional mientras que Lasso lo hace de forma constante sin cruzar el cero (*soft thresholding*). Si $X$ no es ortonormal, el análisis sigue valiendo de forma aproximada.

# Ridge vs. Lasso




El recorrdido de los coeficientes a medida que aumenta $\lambda$ muestra que para Ridge (izquierda) se acercan asintóticamente a cero, mientras que para Lasso (derecha) eventualmente alcanzan el cero de manera nada suave (ejemplo tomado de *Elements of Statistical Learning*).

# ElasticNet: Ridge ♡ Lasso

**Término de regularización**

$$\lambda_1||\beta||_1 + \lambda_2||\beta||_2^2 = \lambda(||\beta||_1 + \alpha||\beta||_2^2)$$

- ElasticNet combina (linealmente) lo mejor de ambos mundos.
- El parámetro $\lambda$ regula la complejidad del modelo.
- El parámetro $\alpha$ regula la importancia relativa de Lasso vs. Ridge.
- Es posible obtener soluciones parsimoniosas y bien condicionadas.
- No free lunch!: ahora hay que calibrar dos hiperparámetros.