

Variantes du Lasso

S.Canu
MLA ITI /INSA de Rouen Normandie

20 octobre 2025

Reweighted least square (Lasso and Ridge)

$$J_{\lambda}(\beta) = \frac{1}{2} \|X\beta - y\|^2 + \lambda \sum_{j=1}^p |\beta_j| = \frac{\beta_j^2}{|\beta_j|},$$

the idea : iterate the weighted ridge towards a fixed point

$$\beta^{(k+1)} = \arg \min_{\beta} \frac{1}{2} \|X\beta - y\|^2 + \lambda \sum_{j=1}^p \frac{\beta_j^2}{|\beta_j^{(k)}|} = w_j \beta_j^2,$$

with $w_j = 1/|\beta_j^{(k)}|$, that is

$$\beta^{(k+1)} = (X^{\top}X + \lambda W)^{-1} X^{\top} y \quad \text{with } \text{diag}(W) = \frac{1}{|\beta_j^{(k)}|}$$

$W = \text{Identit }$

Tantque on n'a pas converg 

$$\beta = \arg \min \frac{1}{2} \|X\beta - y\|^2 + \lambda \beta^{\top} W \beta$$

$$W = \text{diag}(1/|\beta|)$$

fin de tantque

Cout et pénalité

- Ce que l'on veut : le modèle le plus simple qui s'ajuste aux données

$$\min_{\beta} \text{cout}(\beta) \text{ ET } \text{penalité}(\beta)$$

- Les deux critères sont contradictoires (le problème est mal posé)
- Le compromis que l'on doit faire (λ , t ou ε)

$$\min_{\beta} \text{cout}(\beta) + \lambda \text{ pénalité}(\beta)$$

$$\begin{array}{ll} \min_{\beta} & \text{cout}(\beta) \\ \text{avec} & \text{penalité}(\beta) \leq t \end{array}$$

$$\begin{array}{ll} \min_{\beta} & \text{penalité}(\beta) \\ \text{avec} & \text{cout}(\beta) \leq \varepsilon \end{array}$$

- exemple : le lasso

$$\text{cout}(\beta) = \frac{1}{2} \|X\beta - y'\|^2 \quad \text{penalité}(\beta) = \|\beta\|_1$$

Lasso et Elastic Net

Le Lasso

$$J_{\lambda}(\beta) = \frac{1}{2} \|X'\beta - y'\|^2 + \lambda \|\beta\|_1 ,$$

Elastic Net

$$J_{\text{el}}(\beta) = \frac{1}{2} \|X\beta - y\|^2 + \lambda \|\beta\|_1 + \frac{\gamma}{2} \|\beta\|_2^2 ,$$

Elastic Net = Lasso quand $\gamma = 0, X = X', y = y'$

Elastic Net = Ridge quand $\lambda = 0, X = X', y = y'$

Elastic Net = Lasso with $X' = (X^{\top}, \sqrt{\gamma}I_p)^{\top}, y' = (y^{\top}, 0_p)^{\top}$

$$\|X\beta - y\|_2^2 + \underbrace{\gamma \|\beta\|_2^2}_{\|\sqrt{\gamma}\beta\|_2^2} = \left\| \underbrace{\begin{bmatrix} X \\ \sqrt{\gamma}I_p \end{bmatrix}}_{=:X'} \beta - \underbrace{\begin{bmatrix} y \\ 0_p \end{bmatrix}}_{=:y'} \right\|_2^2 = \|X'\beta - y'\|_2^2 .$$

Mise en œuvre de l'Elastic Net

- c'est un QP : on peut utiliser CVX, ou un autre solveur
- il existe aussi un chemin de régularisation
- Component wise

$$\partial_{\beta} J_{el}^{(1d)}(\beta) = \mathbf{x}^{\top}(\mathbf{x}\beta - \mathbf{r}) + \gamma\beta + \begin{cases} \lambda\alpha & \text{if } \beta = 0 \\ \lambda\text{sign}(\beta) & \text{else.} \end{cases}$$

$$\beta = \frac{1}{\|\mathbf{x}\|^2 + \gamma} \text{sign}(c) \max(|c| - \lambda, 0) \quad c = \mathbf{x}^{\top} \mathbf{r}$$

- et le gradient proximal

$$t = \frac{2}{\|\mathbf{X}^{\top} \mathbf{X} + 2\gamma \mathbf{I}_p\|}$$

Tant qu'on n'a pas convergé....

$$\begin{aligned} \mathbf{v} &= \beta^k - t(\mathbf{X}^{\top}(\mathbf{X}\beta^k - \mathbf{y}) + \gamma\beta^k) \\ \beta^{k+1} &= \text{sign}(\mathbf{v}) \max(|\mathbf{v}| - t\lambda, 0) \end{aligned}$$

The adaptive Lasso

$$J_a(\beta) = \frac{1}{2} \|X\beta - y\|^2 + \lambda \sum_{j=1}^p w_j |\beta_j| ,$$

$$w_j = \frac{1}{|\hat{\beta}_j^{(ls)}|^\gamma} \quad \text{avec } \gamma = \frac{1}{2}$$

```
w = np.linalg.solve(X.T@X,X.T@y)
w = 1/np.sqrt(np.abs(w))
X_c = X/w
c = mon_lasso(X_c,y,lambda)
beta = c/w
```

The Adaptive Lasso and Its Oracle Properties, H. Zou, JASA, 2012

Mise en œuvre du Lasso adaptatif

- c'est un QP : on peut utiliser CVX, ou un autre solveur
- il existe aussi un chemin de régularisation
- Component wise

$$\partial_{\beta} J_a^{(1d)}(\beta) = \mathbf{x}^{\top}(\mathbf{x}\beta - \mathbf{r}) + \begin{cases} \lambda \mathbf{w} \alpha & \text{if } \beta = 0 \\ \lambda \mathbf{w} \text{sign}(\beta) & \text{else.} \end{cases}$$

$$\beta = \text{sign}(c) \max(|c| - \lambda \mathbf{w}, 0) \quad c = \frac{\mathbf{x}^{\top} \mathbf{r}}{\|\mathbf{x}\|^2}$$

- et le gradient proximal

$$t = \frac{2}{\|X^t X + 2\gamma I_p\|}$$

Tant qu'on n'a pas convergé....

$$\begin{aligned} \mathbf{v} &= \beta^k - t(X^t(X\beta^k - \mathbf{y})) \\ \beta^{k+1} &= \text{sign}(\mathbf{v}) \max(|\mathbf{v}| - t\lambda \mathbf{w}, 0) \end{aligned}$$

The Grouped Lasso

Exemple : X code sous la forme de "one hot vector" (codage disjonctif complet) G variables catégorielles, chacune avec n_g catégories. I_g désigne l'ensemble des indices associés de sorte que $\beta_{I_g} = \{\beta_j | j \in I_g\}$.

$$J_\lambda(\beta) = \frac{1}{2} \|X\beta - y\|^2 + \lambda \sum_{g=1}^G \sqrt{n_g} \|\beta_{I_g}\|_2,$$

- si $n_g = 1$, on retrouve le lasso

Yuan and Lin, 2006 ; Bakin, 1999 ; Cai, 2001 ; Antoniadis and Fan, 2001

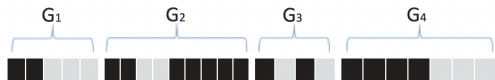
The Grouped Lasso

$$\min_{\beta} \text{cout}(\beta) \text{ ET } \text{penalité}(\beta) \quad \min_{\beta} \text{cout}(\beta) + \lambda \text{penalité}(\beta)$$

Le terme d'attache aux données

$$\text{cout}(\beta) = \frac{1}{2} \|X\beta - y\|^2$$

La pénalité



Lasso



Group
Lasso



Sparse Group
Lasso

$$\|\beta\|_1$$

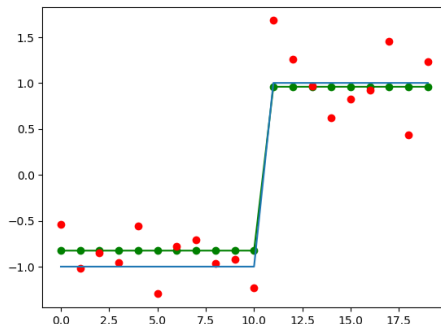
$$\sum_{g=1}^G \sqrt{n_g} \|\beta_{I_g}\|_2$$

$$\|\beta\|_1 + \gamma \sum_{g=1}^G \sqrt{n_g} \|\beta_{I_g}\|_2$$

The Fused Lasso

$$J_f(\beta) = \frac{1}{2} \|X\beta - y\|^2 + \mu \sum_{j=1}^{p-1} |\beta_j - \beta_{j+1}| \left(+ \lambda \|\beta\|_1 \right),$$

Exemple : recherche de composantes continues



$$n = p$$

$$X = I, \lambda = 0$$

- y les points rouges
- β les points verts

◀ (Tibshirani et al., 2005) 🔍 ↻

Exemples d'autres termes d'attache aux données

- The Dantzig selector

$$J_{\lambda}(\beta) = \|X^T(X\beta - y)\|_{\infty} + \lambda\|\beta\|_1 ,$$

Candès and Tao, 2007

- LAD lasso (least absolute deviation)

$$J_{\lambda}(\beta) = \|X\beta - y\|_1 + \lambda\|\beta\|_1 ,$$

Wang et al., 2007

- The logistic lasso

$$J_{\lambda}(\beta) = -\sum_{i=1}^n \left(y_i x_i^T \beta - \log(1 + \exp^{x_i^T \beta}) \right) + \lambda\|\beta\|_1 ,$$

The L1 penalty for logistic regression, Tibshirani, 1996

Huber loss et L1 L2 loss

Huber loss (continue et dérivable)

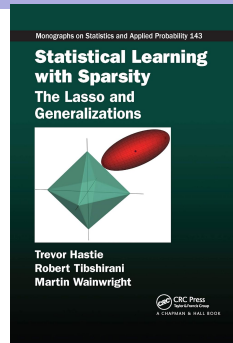
$$L_{21}(\beta) = \begin{cases} \frac{1}{2}(y - X\beta)^2 & \text{for } |y - X\beta| \leq \delta \\ \delta(|y - X\beta| - \frac{1}{2}\delta) & \text{otherwise} \end{cases}$$

L1 L2

$$L_{12}(\beta) = \begin{cases} |y - X\beta| & \text{for } |y - X\beta| \leq \delta \\ \frac{1}{\delta} \left(\frac{1}{2}(y - X\beta)^2 + \frac{1}{2}\delta^2 \right) & \text{otherwise} \end{cases}$$

Conclusion

- machine learning : optimisation bi critère
 - ▶ terme d'attache aux données (vraisemblance)
 - ▶ pénalité (a priori)
- Sélection de variables
 - ▶ Parcimonie = singularité
 - ▶ Parametric QP (pénalité convexe)
- Lasso when p is very large : Sceening
- Sélection d'individus ?



Variant	Key Feature	Best Use Case
Elastic Net	Combines L1 and L2 penalties	Highly correlated predictors
Adaptive Lasso	Weighted penalties for coefficients	High-dimensional data
Group Lasso	Selects groups of variables	Grouped predictors
Fused Lasso	Encourages sparsity in changes	Time-series or spatial data
Dantzig selec	L_∞ data related loss	Highly correlated