Le Lasso
son chemin de réularisation

S.Canu
MLA ITI /INSA de Rouen Normandie

12 octobre 2026

Le Lasso

- Données
  - $X \in \mathbb{R}^{n \times p}$ les p variables observées n fois
  - $y \in \mathbb{R}^n$ la variable réponse
- Inconnues
  - $\beta \in \mathbb{R}^p$
- hypothèses : pas de biais (pas de terme constant)
  - $mean(X) = 0$
  - $\|X_j\|^2 = 1$
  - $mean(y) = 0$

le cout pénalisé

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

La forme contrainte

$$\min_{\beta \in \mathbb{R}^p} \; \tfrac{1}{2}\|y - X\beta\|^2 \quad \text{avec} \quad \|\beta\|_1 \leq t$$

Roadmap

1. Le chemin de régularisation désigne l'ensemble des solutions $C = \{\widehat{\beta}_\lambda | \lambda \in R^+\}$
   - Quand $\lambda = 0$, $\widehat{\beta}_0 = (X^T X)^{-1} X^T y$
   - Il existe un λ_{max} tel que $\widehat{\beta}_{max} = 0$
2. La solution $\widehat{\beta}_\lambda$ dépend de deux ensembles I_β et I_0
   - connaissant I_β, il existe une forme fermée pour calculer $\widehat{\beta}_\lambda$
3. Le chemin de régularisation du lasso est linéaire par morceau
   - Quand λ varie et que les deux ensembles I_β et I_0 demeurent inchangé, $\widehat{\beta}_\lambda$ varie linéairement
   - Le chemin est définit par une liste ordonnée $(\lambda_0 = 0, \lambda_1, \ldots, \lambda_k, \ldots, \lambda_{max})$

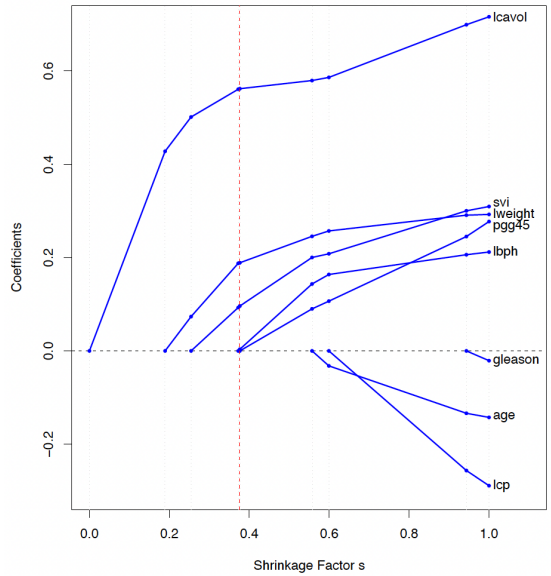
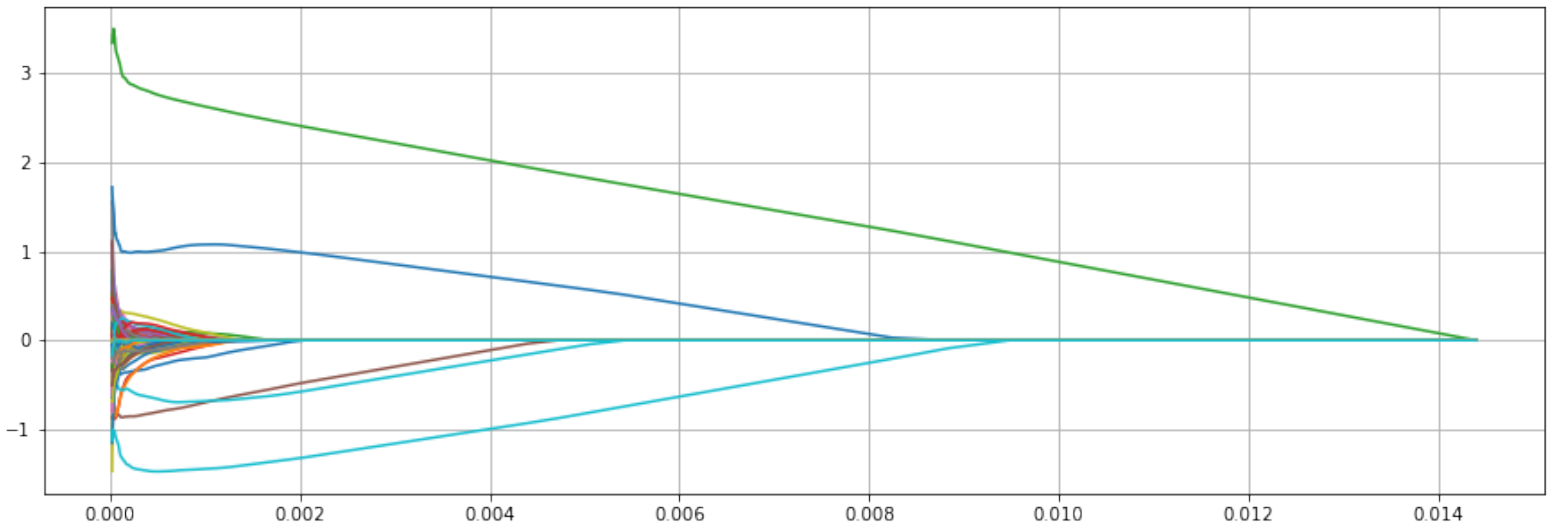


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed; see Section 3.4.4 for details.*

La notion de chemin de régularisation

$\beta(\lambda)$

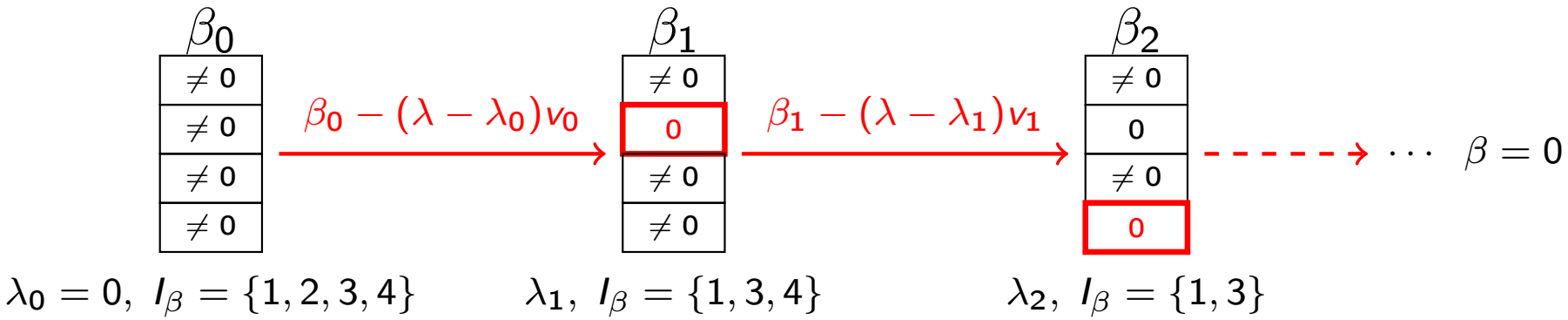


λ large $\Rightarrow \beta = 0$ so that $|X^\top y| \leq \lambda$

the first non zero component for vector β appears for $\lambda = \max |X^\top y|$

the regularization path is piecewise linear

Chemin de régularisation



Il existe un λ_{min} tel que $\widehat{\beta}_{min} \neq 0$

Quand $\lambda = 0$, $\widehat{\beta}_0 = (X^T X)^{-1} X^T y$

Pour λ tout petit

$$\nabla_\beta J_\lambda(\beta) = X^T(X\beta - y) + \lambda \, \text{sign}(\beta)$$

or

$$sign(\beta) = sign(\beta_0)$$

$$\begin{array}{lll} X^T(X\beta_\lambda - y) & +\lambda \, \text{sign}(\beta_0) & = 0 \\ X^T(X\beta_0 - y) & & = 0 \end{array}$$

et donc

$$\beta_\lambda = \beta_0 - \lambda \, (X^T X)^{-1} \text{sign}(\beta_0)$$

$$\lambda_{min} = \min_{j \in 1,\ldots,p} \left| \frac{(\beta_0)_j}{v_j} \right| \qquad \text{avec } v = (X^T X)^{-1} \text{sign}(\beta_0)$$

Il existe un λ_{max} tel que $\widehat{\beta}_{max} = 0$

- C'est la multiplicateur de Lagrange de la solution contrainte avec $t = 0$
- Nous verrons d'autres façons de la retrouver (mais il nous faudra travailler avec la sous différentielle...)

$$\lambda_{max} = \max_{j \in 1,\dots,p} (|X^\top y|)_j$$

Connaissant I_β, il existe une forme fermée pour calculer $\widehat{\beta}_\lambda$

$$\forall j \in I_\beta \qquad \partial_{\beta_j} J(\beta) = 0$$

$$\left(X_{I_\beta}^\top X_{I_\beta} \quad X_{I_\beta}^\top X_{I_0}\right) \begin{pmatrix} \beta_{I_\beta} \\ 0 \end{pmatrix} - \left(X_{I_\beta}^\top y\right) + \lambda \left(\text{sign}(\beta_{I_\beta})\right) = 0$$

$$\widehat{\beta}_\lambda = \begin{cases} \beta_{I_\beta} & = (X_{I_\beta}^\top X_{I_\beta})^{-1}(X_{I_\beta}^\top y - \lambda \, \text{sign}(X_{I_\beta}^\top y)) \\ \beta_{I_0} & = 0 \end{cases}$$

Sous-gradient et sous-différentiel

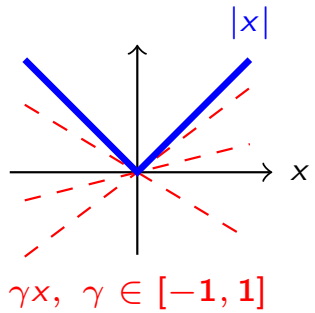
Exemple : $J_1 : \mathbb{R} \to \mathbb{R},\ x \mapsto J_1(x) = |x|$

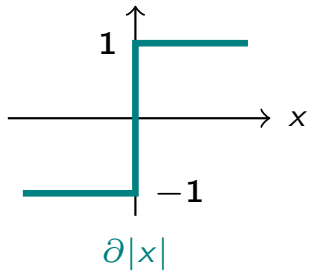
γ sous-gradient de J_1 en 0 : $\quad \forall x \in \mathcal{V}(0)$

$$J_1(x) \geq J_1(0) + \gamma\,(x - 0) \quad \Leftrightarrow \quad |x| \geq \gamma\, x$$

si $x > 0 : \gamma \leq 1 \qquad$ si $x < 0 : \gamma \geq -1$

$$\Rightarrow \quad \partial_x J_1(0) = [-1, 1]$$





Définition

$J : \mathbb{R}^n \to \mathbb{R}$ convexe. $\gamma \in \mathbb{R}^n$ est un sous-gradient de J au point x_0 si

$$J(x) \geq J(x_0) + \gamma^\top (x - x_0) \qquad \forall x \in \mathcal{V}(x_0)$$

Le sous-différentiel de J en x_0 est l'ensemble de ses sous-gradients :

$$\partial_x J(x_0) = \left\{ \gamma \in \mathbb{R}^n \;\middle|\; J(x) \geq J(x_0) + \gamma^\top (x - x_0),\ \forall x \in \mathcal{V}(x_0) \right\}$$

- J convexe : $\mathcal{V}(x_0)$ peut être remplacé par $\mathbb{R}^n$ (local = global)
- J différentiable en x_0 : $\partial J(x_0) = \{\nabla J(x_0)\}$ (ex. $\partial|x| = \{\text{sign}(x)\},\ x \neq 0$)
- règle de Fermat : $x^\star \in \arg\min J \ \Leftrightarrow\ 0 \in \partial J(x^\star)$

Rockafellar, *Convex Analysis*, 1970 – Hiriart-Urruty & Lemaréchal, *Fundamentals of Convex Analysis*, 2001

Conditions d'optimalité du Lasso : I_β et I_0

$$J(\beta) = \tfrac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1, \qquad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$$

$$I_\beta = \{j \mid \beta_j \neq 0\} \text{ (variables actives)}, \qquad I_0 = \{j \mid \beta_j = 0\}, \qquad X_I = X(:, I)$$

$$\partial_\beta J(\beta) = -X^\top y + X^\top X\beta + \lambda s \qquad \text{avec} \quad s_j = \begin{cases} \operatorname{sign}(\beta_j) & \text{si } \beta_j \neq 0 \\ \alpha_j \in [-1, 1] & \text{si } \beta_j = 0 \end{cases}$$

$\beta^\star_\lambda$ optimal $\Leftrightarrow 0 \in \partial_\beta J(\beta^\star_\lambda) \Leftrightarrow \exists\, \alpha_j \in [-1, 1],\ j \in I_0$, tels que

(variables réordonnées)

$$\begin{pmatrix} X_{I_\beta}^\top X_{I_\beta} & X_{I_\beta}^\top X_{I_0} \\ X_{I_0}^\top X_{I_\beta} & X_{I_0}^\top X_{I_0} \end{pmatrix} \begin{pmatrix} \beta_{I_\beta} \\ 0 \end{pmatrix} + \lambda \begin{pmatrix} \operatorname{sign}(\beta_{I_\beta}) \\ \alpha_{I_0} \end{pmatrix} - \begin{pmatrix} X_{I_\beta}^\top y \\ X_{I_0}^\top y \end{pmatrix} = 0 \qquad \begin{matrix} \text{①}\ I_\beta \\ \text{②}\ I_0 \end{matrix}$$

① $X_{I_\beta}^\top X_{I_\beta}\, \beta_{I_\beta} = X_{I_\beta}^\top y - \lambda \operatorname{sign}(\beta_{I_\beta})$ — un système linéaire en β_{I_β}

② $\left|X_j^\top (y - X_{I_\beta} \beta_{I_\beta})\right| \leq \lambda \quad \forall j \in I_0$ — car $|\alpha_j| \leq 1$

À I_β et signes fixés, $\beta_{I_\beta} = (X_{I_\beta}^\top X_{I_\beta})^{-1}(X_{I_\beta}^\top y - \lambda \operatorname{sign}(\beta_{I_\beta}))$ est affine en λ

Osborne, Presnell & Turlach, *J. Comput. Graph. Stat.*, 2000 – Fuchs, *IEEE Trans. Inf. Theory*, 2004

Retour au chemin de régularisation

$$\min_{\beta \in \mathbb{R}^p} J_\lambda(\beta) \quad \text{avec} \quad J_\lambda(\beta) = \frac{1}{2}\|X\beta - y\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

$$\partial J(\beta) = X^\top(X\beta - y) + \lambda s(\beta) \qquad \text{avec } s_j(\beta) = \begin{cases} 1 & \text{si } \beta_j > 0 \\ [-1,1] & \text{si } \beta_j = 0 \\ -1 & \text{si } \beta_j < 0 \end{cases}$$

β optimal si $0 \in \partial J(\beta)$ c'est-à-dire si pour $g(\beta) = X^\top(X\beta - y)$

$$\begin{cases} g_j(\beta) + \lambda = 0 & \text{si } \beta_j > 0 \\ -\lambda \leq g \leq \lambda & \text{si } \beta_j = 0 \\ g_j(\beta) - \lambda = 0 & \text{si } \beta_j < 0 \end{cases}$$

Un cas intéressant :

$$\text{si } \lambda \geq \max_j (X^\top y)_j \qquad \text{alors} \qquad \beta = 0$$

Pénalisé ou contraint ? Vers le chemin de régularisation

forme pénalisée

$$\beta^\star_\lambda = \underset{\beta\in\mathbb{R}^p}{\arg\min}\ \tfrac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1$$

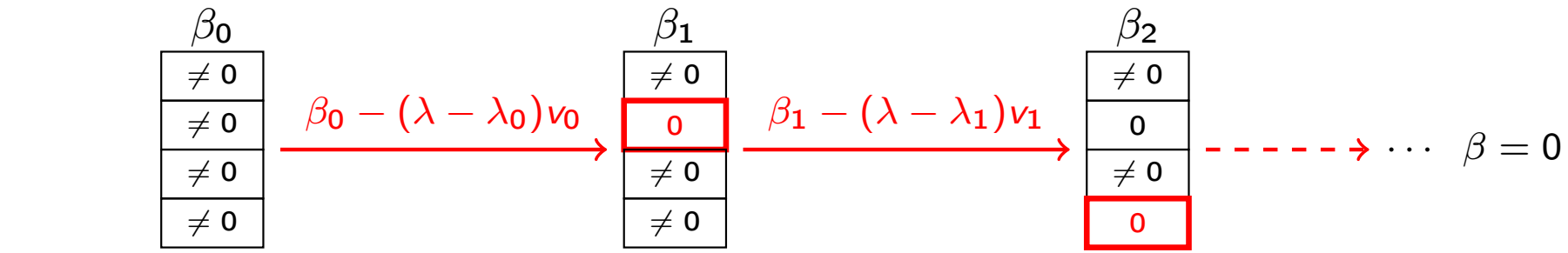
forme contrainte

$$\tilde\beta_k = \underset{\beta\in\mathbb{R}^p}{\arg\min}\ \tfrac{1}{2}\|y - X\beta\|^2 \ \text{ t.q. } \ \|\beta\|_1 \le k$$

Équivalence

$\forall k\ \exists\lambda$ tel que $\beta^\star_\lambda = \tilde\beta_k$ (λ = multiplicateur de Lagrange de la contrainte)
et réciproquement $k = \|\beta^\star_\lambda\|_1$; $\lambda = 0 \leftrightarrow$ moindres carrés ; $\lambda \ge \|X^\top y\|_\infty \leftrightarrow k = 0$

? $\beta^\star(\lambda) = \beta_0 - (\lambda - \lambda_0)\, v$ pour $\lambda \in \mathcal{V}(\lambda_0)$? (chemin localement affine ?)



Oui (transparents suivants) : à I_β et signes fixés, chemin affine par morceaux ; cassures $\lambda_0 < \lambda_1 < \lambda_2 < \dots$

Tibshirani, *J. R. Stat. Soc. B*, 1996 – Rosset & Zhu, *Ann. Statist.*, 2007

Calcul du chemin (1) : partir des moindres carrés, $\lambda_0 = 0$

① $\lambda = 0$: $J(\beta) = \frac{1}{2}\|y - X\beta\|^2$

A $\quad X^\top X\,\hat\beta_0 = X^\top y$

$(\hat\beta_0)_j \neq 0 \quad \forall j$

$I_\beta = \{1, 2, \ldots, p\}$ (toutes)

$I_0 = \emptyset$

② $\lambda \in \mathcal{V}(0)$, $\lambda = \varepsilon$ tout petit : I_β et les signes ne changent pas

B $\quad X^\top X\,\beta_\lambda = X^\top y - \lambda\,\mathrm{sign}(\hat\beta_0)$

B − A : $\quad X^\top X\,(\beta_\lambda - \hat\beta_0) = -\lambda\,\mathrm{sign}(\hat\beta_0) \quad \Rightarrow \quad \beta_\lambda = \hat\beta_0 - \lambda \underbrace{(X^\top X)^{-1}\mathrm{sign}(\hat\beta_0)}_{v}$

Jusqu'où ? La composante j s'annule quand $(\hat\beta_0)_j - \lambda\, v_j = 0 \quad \Leftrightarrow \quad \boxed{\lambda^{(j)} = \dfrac{(\hat\beta_0)_j}{v_j}}$

Premier point de cassure

$$\lambda_1 = \min_{j\,:\,(\hat\beta_0)_j/v_j > 0} \frac{(\hat\beta_0)_j}{v_j} \qquad \beta_1 = \hat\beta_0 - \lambda_1 v \qquad I_\beta \leftarrow I_\beta \setminus \{j_1\},\ \ j_1 = \text{l'argmin}$$

hypothèse : $X^\top X$ inversible ($n \geq p$, X de rang p)

Osborne, Presnell & Turlach (homotopie), *IMA J. Numer. Anal.*, 2000

Calcul du chemin (2) : de λ_k à λ_{k+1}

λ_k, β_k associés à I_β et I_0 pour $\lambda > \lambda_k$, $\lambda \in \mathcal{V}(\lambda_k)$ $X_I = X(:, I_\beta)$

A $X_I^\top X_I \, \beta_\lambda = X_I^\top y - \lambda \, \mathrm{sign}(\beta_k)$

B $X_I^\top X_I \, \beta_k = X_I^\top y - \lambda_k \, \mathrm{sign}(\beta_k)$ vraie pour λ_k

A − B $X_I^\top X_I \, (\beta_\lambda - \beta_k) = -(\lambda - \lambda_k) \, \mathrm{sign}(\beta_k)$

$\beta_\lambda = \beta_k - (\lambda - \lambda_k) \, v_k$ $v_k = (X_I^\top X_I)^{-1} \mathrm{sign}(\beta_k) \in \mathbb{R}^{|I_\beta|}$

sortie de i : $(\beta_\lambda)_i = 0 \Leftrightarrow (\beta_k)_i - (\lambda - \lambda_k)(v_k)_i = 0 \Leftrightarrow \lambda = \lambda_k + \dfrac{(\beta_k)_i}{(v_k)_i}$

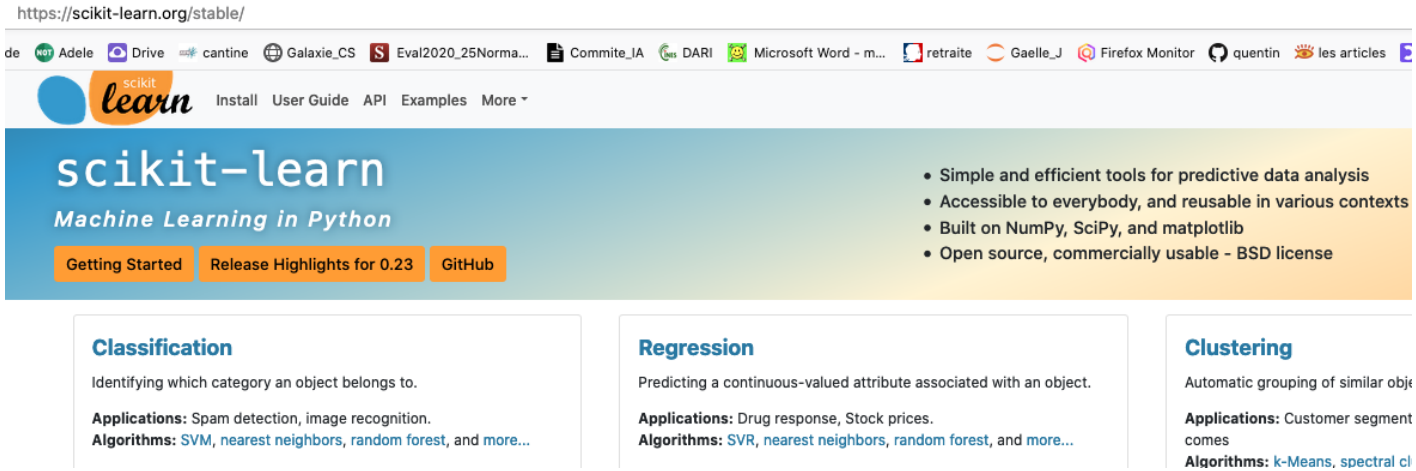
$$\lambda_{k+1} = \lambda_k + \min_{i \in I_\beta \, : \, (\beta_k)_i/(v_k)_i > 0} \frac{(\beta_k)_i}{(v_k)_i}$$

Pour un algorithme exact (homotopie)

- entrée : $j \in I_0$ entre si $|X_j^\top (y - X_I \beta_\lambda)| = \lambda$; $\lambda_{k+1} = 1^{\text{er}}$ des 2 événements
- arrêt : $\beta = 0$ dès $\lambda \geq \|X^\top y\|_\infty$; en sens inverse (λ décroissant) : LARS-Lasso
- nombre de segments : souvent $\approx \min(n, p)$, exponentiel en p dans le pire cas

Efron, Hastie, Johnstone & Tibshirani, Least angle regression, *Ann. Statist.*, 2004 – Mairal & Yu, *ICML*, 2012

Sklearn et le Lasso



See also

`lars_path_gram`
Compute LARS path.

`lasso_path`
Compute Lasso path with coordinate descent.

`LassoLars`
Lasso model fit with Least Angle Regression a.k.a. Lars.

`Lars`
Least Angle Regression model a.k.a. LAR.

`LassoLarsCV`
Cross-validated Lasso, using the LARS algorithm.

`LarsCV`
Cross-validated Least Angle Regression model.

`sklearn.decomposition.sparse_encode`
Sparse coding.

RidgeCV
SGDRegressor
ElasticNet
ElasticNetCV
Lars
LarsCV
Lasso
LassoCV
LassoLars
LassoLarsCV
LassoLarsIC
OrthogonalMatchingPursuit
OrthogonalMatchingPursuitCV
ARDRegression
BayesianRidge
MultiTaskElasticNet
MultiTaskElasticNetCV
MultiTaskLasso
MultiTaskLassoCV
HuberRegressor
QuantileRegressor
RANSACRegressor
TheilSenRegressor
GammaRegressor
PoissonRegressor
TweedieRegressor
PassiveAggressiveRegressor
enet_path
lars_path
lars_path_gram
lasso_path

Lasso

class sklearn.linear_model.Lasso(*alpha=1.0, *, fit_intercept=True, precompute=False, copy_X=True, max_iter=1000, tol=0.0001, warm_start=False, positive=False, random_state=None, selection='cyclic'*) [source]

Linear Model trained with L1 prior as regularizer (aka the Lasso).

The optimization objective for Lasso is:

```
(1 / (2 * n_samples)) * ||y - Xw||^2_2 + alpha * ||w||_1
```

Technically the Lasso model is optimizing the same objective function as the Elastic Net with `l1_ratio=1.0` (no L2 penalty).

Read more in the User Guide.

Parameters:

alpha : *float, default=1.0*

Constant that multiplies the L1 term, controlling regularization strength. `alpha` must be a non-negative float i.e. in `[0, inf)`.

When `alpha = 0`, the objective is equivalent to ordinary least squares, solved by the `LinearRegression` object. For numerical reasons, using `alpha = 0` with the `Lasso` object is not advised. Instead, you should use the `LinearRegression` object.

fit_intercept : *bool, default=True*

Whether to calculate the intercept for this model. If set to False, no intercept will be used in calculations (i.e. data is expected to be centered).

precompute : *bool or array-like of shape (n_features, n_features), default=False*

Whether to use a precomputed Gram matrix to speed up calculations. The Gram matrix can also be passed as argument. For sparse input this option is always `False` to preserve sparsity.

copy_X : *bool, default=True*

If `True`, X will be copied; else, it may be overwritten.

max_iter : *int, default=1000*

The maximum number of iterations.

Pour aller plus loin : quelques références

Sous-différentiel et analyse convexe

- R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.
- J.-B. Hiriart-Urruty, C. Lemaréchal, *Fundamentals of Convex Analysis*, Springer, 2001.

Lasso et conditions d'optimalité

- R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. B*, 58(1), 267–288, 1996.
- M. R. Osborne, B. Presnell, B. A. Turlach, On the LASSO and its dual, *J. Comput. Graph. Stat.*, 9(2), 319–337, 2000.
- J.-J. Fuchs, On sparse representations in arbitrary redundant bases, *IEEE Trans. Inf. Theory*, 50(6), 1341–1344, 2004.
- R. J. Tibshirani, The lasso problem and uniqueness, *Electron. J. Stat.*, 7, 1456–1490, 2013.

Chemin de régularisation

- M. R. Osborne, B. Presnell, B. A. Turlach, A new approach to variable selection in least squares problems, *IMA J. Numer. Anal.*, 20(3), 389–403, 2000.
- B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, *Ann. Statist.*, 32(2), 407–499, 2004.
- S. Rosset, J. Zhu, Piecewise linear regularized solution paths, *Ann. Statist.*, 35(3), 1012–1030, 2007.
- J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, *J. Stat. Softw.*, 33(1), 1–22, 2010.
- J. Mairal, B. Yu, Complexity analysis of the Lasso regularization path, *ICML*, 2012.

Synthèses

- F. Bach, R. Jenatton, J. Mairal, G. Obozinski, Optimization with sparsity-inducing penalties, *Found. Trends Mach. Learn.*, 4(1), 1–106, 2012.
- T. Hastie, R. Tibshirani, M. Wainwright, *Statistical Learning with Sparsity : The Lasso and Generalizations*, CRC Press, 2015.