

Le chemin des solutions du SLOPE

Patrick Tardivel (collaboration avec Xavier Dupuis)

Journées de Statistiques et Optimisation en Occitanie

Institut de Mathématiques de Bourgogne, Dijon

3 avril 2025



Le modèle de régression linéaire

Modèle de régression linéaire : $Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$,
 $i = 1, \dots, n$ noté de façon matricielle $Y = X\beta + \varepsilon$.

- ▶ $X = (x_{ij})_{1 \leq i \leq n, 0 \leq j \leq p}$ avec $x_{i0} = 1$: matrice de planification.
- ▶ $\beta = (\beta_0, \dots, \beta_p) \in \mathbb{R}^{p+1}$ coefficients de régression.
- ▶ $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n) \in \mathbb{R}^n$ résidus aléatoires.

n individus goûtent et notent du vin ; on cherche à expliquer la note moyenne d'un vin en fonction de mesures physico-chimiques

$$\text{note}_i = \beta_0 + \beta_1 \times \text{alcool}_i + \beta_2 \times \text{sucres}_i + \dots + \beta_{11} \times \text{pH}_i + \varepsilon_i,$$

pour $i = 1, \dots, n$.

L'estimateur des moindres carrés

L'estimateur des moindres carrés est la solution du problème

$$\min_{b \in \mathbb{R}^p} \|Y - Xb\|_2^2.$$

Solution unique ssi $\ker(X) = \{0\}$: $\hat{\beta}^{mc} = (X^T X)^{-1} X^T Y$.

L'estimateur des moindres carrés

L'estimateur des moindres carrés est la solution du problème

$$\min_{b \in \mathbb{R}^p} \|Y - Xb\|_2^2.$$

Solution unique ssi $\ker(X) = \{0\}$: $\hat{\beta}^{mc} = (X^T X)^{-1} X^T Y$.

Cet estimateur n'est pas bien défini lorsque $n \leq p$.

Si $n > p$ la matrice de variance $\sigma^2 (X^T X)^{-1}$ de l'estimateur $\hat{\beta}^{mc}$ peut avoir des composantes très grandes.

Estimateur OSCAR

L'estimateur OSCAR (Octagonal Shrinkage and Clustering Algorithm for Regression (Bondell et Reich, 2008)) est la solution du problème

$$\min_{b \in \mathbb{R}^p} \left\{ \frac{1}{2} \|Y - Xb\|_2^2 + \lambda_1 \sum_{i=1}^p |b_i| + \lambda_2 \sum_{1 \leq i < j \leq p} \underbrace{\frac{|b_i - b_j|}{2} + \frac{|b_i + b_j|}{2}}_{=\max\{|b_i|, |b_j|\}} \right\}.$$

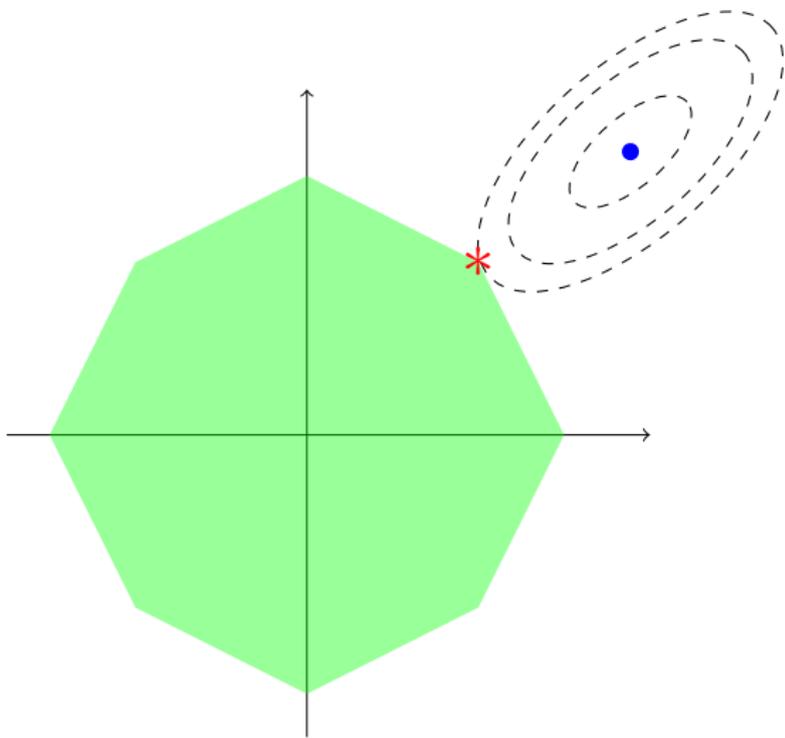
Si $\lambda_2 = 0$ alors OSCAR coïncide avec l'estimateur LASSO.

Solution unique si $\ker(X) = \{0\}$. On note S_{oscar} l'ensemble des solutions du problème OSCAR :

$$\{X \in \mathbb{R}^{n \times p} \mid \exists Y \in \mathbb{R}^n \text{ tel que } \text{Card}(S_{\text{oscar}}) > 1\}$$

est négligeable sur $\mathbb{R}^{n \times p}$ (Schneider et Tardivel 2022).

L'estimateur OSCAR a des composantes nulles et des composantes égales (Schneider et Tardivel 2022).



Estimateur SLOPE

Soit $\lambda_1 \geq \dots \geq \lambda_p \geq 0$ et $\lambda_1 > 0$. La norme ℓ_1 ordonnée J_λ est définie par

$$J_\lambda(b) = \sum_{i=1}^p \lambda_i |b|_{\downarrow i}$$

avec $|b|_{\downarrow 1} \geq \dots \geq |b|_{\downarrow p}$.

L'estimateur SLOPE (Sorted L One Penalized Estimation (Bogdan et al., 2015)) est la solution du problème

$$\min_{b \in \mathbb{R}^p} \left\{ \frac{1}{2} \|Y - Xb\|_2^2 + J_\lambda(b) \right\}.$$

SLOPE généralise OSCAR (et LASSO) : si λ est arithmétique alors

$$J_\lambda(b) = \lambda_p \sum_{i=1}^p |b_i| - \text{raison} \sum_{1 \leq i < j \leq p} \max\{|b_i|, |b_j|\}.$$

- ▶ La norme ℓ_1 ordonnée duale J_λ^* est donnée par

$$J_\lambda^*(b) = \max \left\{ \frac{|b|_{\downarrow 1}}{\lambda_1}, \dots, \frac{|b|_{\downarrow 1} + \dots + |b|_{\downarrow p}}{\lambda_1 + \dots + \lambda_p} \right\}$$

- ▶ Boule unité de J_λ^* est le permutoèdre signé :

$$\text{conv}\{(\pm\lambda_{\pi(1)}, \dots, \pm\lambda_{\pi(p)}) : \pi \text{ permutation de } \{1, \dots, p\}\}.$$

- ▶ On note S_{slope} l'ensemble des solutions du problème SLOPE alors :

$$\begin{aligned} \hat{\beta}^{\text{slope}} \in S_{\text{slope}} &\Leftrightarrow X^\top(Y - X\hat{\beta}^{\text{slope}}) \in \partial J_\lambda(\hat{\beta}^{\text{slope}}) \\ &\Leftrightarrow \begin{cases} J_\lambda^*(X^\top(Y - X\hat{\beta}^{\text{slope}})) \leq 1 \\ \hat{\beta}^{\text{slope}\top} X^\top(Y - X\hat{\beta}^{\text{slope}}) = J_\lambda(\hat{\beta}^{\text{slope}}) \end{cases} \end{aligned}$$

$$\hat{\beta}^{\text{slope}} = 0 \Leftrightarrow J_\lambda^*(X^\top Y) \leq 1$$

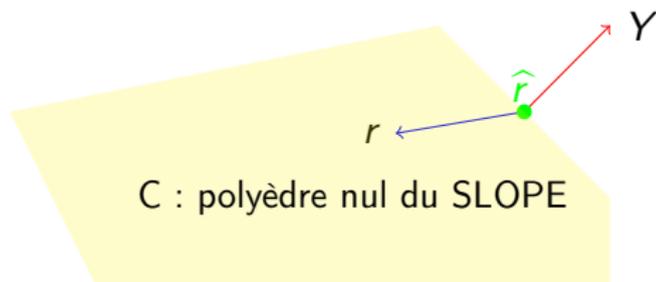
On note $C = \{r \in \mathbb{R}^n : J_\lambda^*(X^\top r) \leq 1\}$ le polyèdre nul du SLOPE

Proposition (Schneider et Tardivel 2022)

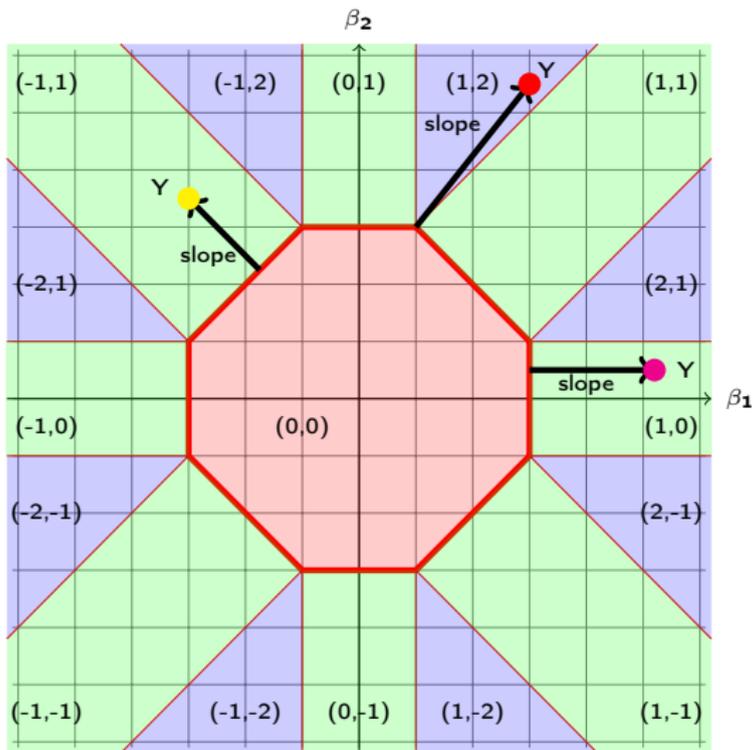
La projection de Y sur C vaut $\hat{r} = Y - X\hat{\beta}^{\text{slope}}$ où $\hat{\beta}^{\text{slope}} \in S_{\text{slope}}$

Esquisse de preuve :

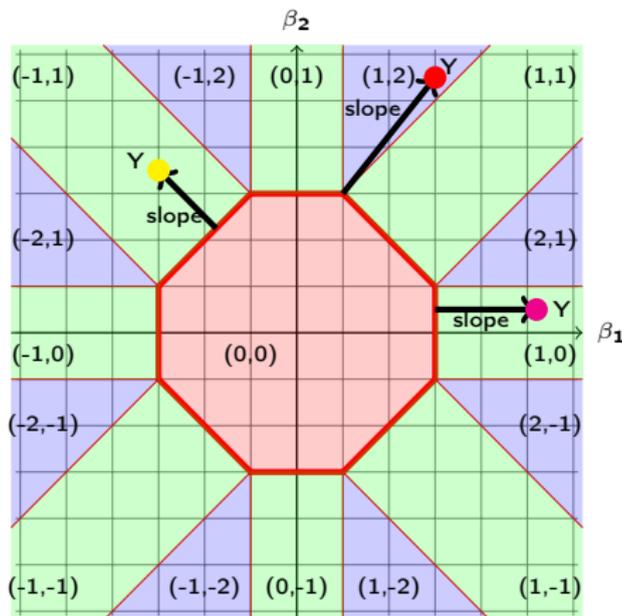
- ▶ On a $\hat{r} \in C$
- ▶ On vérifie que $(Y - \hat{r})^\top (r - \hat{r}) \leq 0$



$X = I_2$: SLOPE a une expression explicite (Dupuis et Tardivel 2022 ; Tardivel, Servien et Concordet 2020). Illustration avec $\lambda_1 = 3$ et $\lambda_2 = 1$.



$X = I_2$: SLOPE a une expression explicite (Dupuis et Tardivel 2022 ; Tardivel, Servien et Concordet 2020). Illustration avec $\lambda_1 = 3$ et $\lambda_2 = 1$.



Cette figure illustre que l'estimateur SLOPE est parcimonieux et a des groupes d'appariement.

Chemin des solutions du SLOPE

$\lambda_1 > \dots > \lambda_p > 0$ et $\gamma > 0$.

$$\mathcal{S}_{\text{slope}}(\gamma) = \arg \min_{b \in \mathbb{R}^p} \left\{ \frac{1}{2} \|Y - Xb\|_2^2 + \gamma J_\lambda(b) \right\}.$$

Si $\text{Card}(\mathcal{S}_{\text{slope}}(\gamma)) = 1$ on note $\hat{\beta}^{\text{slope}}(\gamma)$ l'unique élément $\mathcal{S}_{\text{slope}}(\gamma)$.

- ▶ Si $\ker(X) = \{0\}$ l'application $\gamma \mapsto \hat{\beta}^{\text{slope}}(\gamma)$ est bien définie.
- ▶ L'ensemble

$$\{X \in \mathbb{R}^{n \times p} \mid \exists Y \in \mathbb{R}^n \exists \gamma > 0 \text{ tel que } \text{Card}(\mathcal{S}_{\text{slope}}(\gamma)) > 1\}$$

est négligeable sur $\mathbb{R}^{n \times p}$ (Schneider et Tardivel 2022). Sauf cas pathologique, $\gamma \mapsto \hat{\beta}^{\text{slope}}(\gamma)$ est bien définie.

Définition (Schéma du SLOPE)

Le schéma du SLOPE $\text{schm}(b) \in \mathbb{Z}^p$ de $b \in \mathbb{R}^p$ est défini par

$$\text{schm}(b)_i = \text{signe}(b_i) \text{rang}(|b|)_i, \quad i \in \{1, \dots, p\},$$

- ▶ $\text{rang}(|b|)_i \in \{0, 1, \dots, k\}$.
- ▶ k est le nombre de composantes distinctes de $\{|b_1|, \dots, |b_p|\}$.
- ▶ $\text{rang}(|b|)_i = 0$ si et seulement si $b_i = 0$.
- ▶ $\text{rang}(|b|)_i < \text{rang}(|b|)_j$ si $|b_i| < |b_j|$.

$b = (4.2, -1.3, 0, 1.3, 4.2)$ alors $\text{schm}(b) = (2, -1, 0, 1, 2)$.

On pose $\mathcal{P}_p^{\text{slope}} = \text{schm}(\mathbb{R}^p)$ l'ensemble des schémas du SLOPE.

Le chemin des solutions du SLOPE est affine par morceau

Théorème (Dupuis et Tardivel 2024)

$X \in \mathbb{R}^{n \times p}$, $Y \in \mathbb{R}^n$, $m \in \mathcal{P}_p^{\text{slope}}$ et $\lambda_1 > \dots > \lambda_p > 0$.

1. Si $\hat{\beta}, \bar{\beta} \in \mathcal{S}^{\text{slope}}$ alors $X\hat{\beta} = X\bar{\beta}$. On note $\hat{v}_a(\gamma) = X\hat{\beta}$ qui ne dépend pas de l'élément $\hat{\beta} \in \mathcal{S}^{\text{slope}}$.
2. $I_m = \{\gamma > 0 \mid \exists \hat{\beta} \in \mathcal{S}^{\text{slope}}(\gamma) \text{ tel que } \text{schm}(\hat{\beta}) = m\}$ est un intervalle.
3. Le chemin des valeurs ajustées $\gamma > 0 \mapsto \hat{v}_a(\gamma)$ est continue sur $]0, +\infty[$ et la restriction à I_m est affine.
4. Si $\text{Card}(\mathcal{S}^{\text{slope}}(\gamma)) = 1$ pour tout $\gamma > 0$ alors $\gamma > 0 \mapsto \hat{\beta}^{\text{slope}}(\gamma)$ est continue sur $]0, +\infty[$ et la restriction à I_m est affine.

Le chemin des solutions du SLOPE est affine par morceau

Théorème (Dupuis et Tardivel 2024)

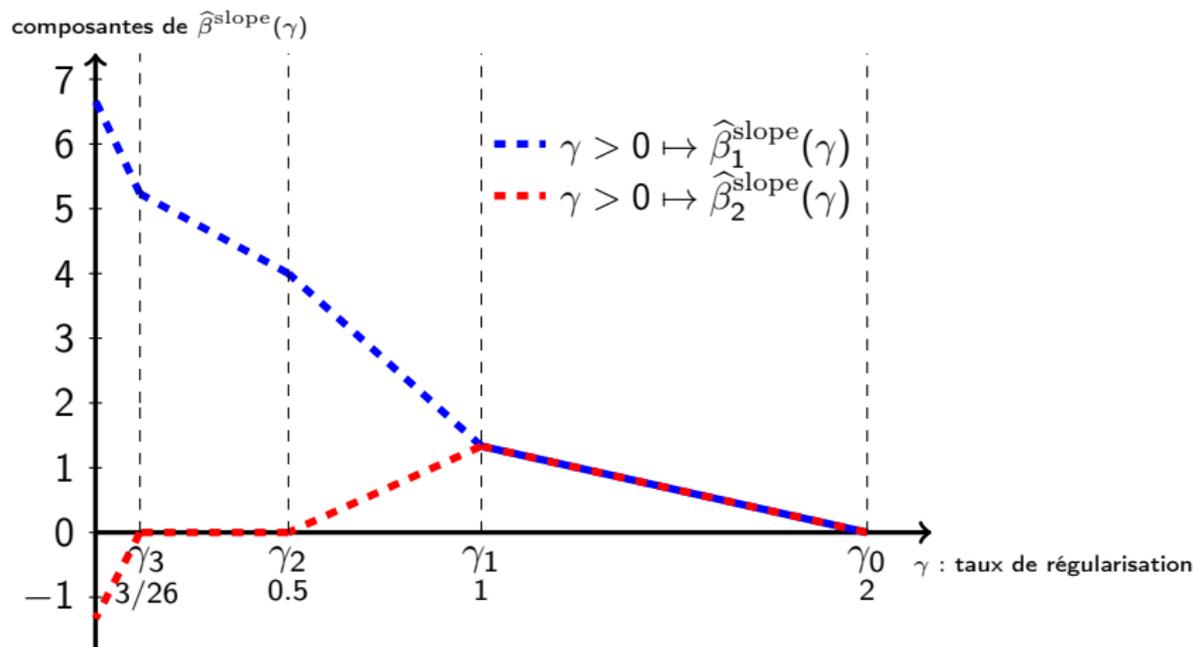
$X \in \mathbb{R}^{n \times p}$, $Y \in \mathbb{R}^n$, $m \in \mathcal{P}_p^{\text{slope}}$ et $\lambda_1 > \dots > \lambda_p > 0$.

1. Si $\hat{\beta}, \bar{\beta} \in \mathcal{S}^{\text{slope}}$ alors $X\hat{\beta} = X\bar{\beta}$. On note $\hat{v}a(\gamma) = X\hat{\beta}$ qui ne dépend pas de l'élément $\hat{\beta} \in \mathcal{S}^{\text{slope}}$.
2. $I_m = \{\gamma > 0 \mid \exists \hat{\beta} \in \mathcal{S}^{\text{slope}}(\gamma) \text{ tel que } \text{schm}(\hat{\beta}) = m\}$ est un intervalle.
3. Le chemin des valeurs ajustées $\gamma > 0 \mapsto \hat{v}a(\gamma)$ est continue sur $]0, +\infty[$ et la restriction à I_m est affine.
4. Si $\text{Card}(\mathcal{S}^{\text{slope}}(\gamma)) = 1$ pour tout $\gamma > 0$ alors $\gamma > 0 \mapsto \hat{\beta}^{\text{slope}}(\gamma)$ est continue sur $]0, +\infty[$ et la restriction à I_m est affine.

Sachant que $\text{schm}(\hat{\beta}) = m$ on calcule explicitement $\hat{\beta}$.

Exemple jouet du chemin du SLOPE

$$X = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, Y = \begin{pmatrix} 6 \\ 2 \end{pmatrix} \text{ et } \lambda = \begin{pmatrix} 4 \\ 2 \end{pmatrix}.$$



$$\hat{\beta}^{\text{slope}}(\gamma) = 0 \Leftrightarrow \gamma \geq J_{\lambda}^*(X^{\top} Y) \quad (J_{\lambda}^*(7, 5) = \max\{7/4, 12/6\} = 2)$$

Sous-différentiel et schéma du SLOPE

Le sous-différentiel d'une norme vérifie

$$\partial\|\cdot\|(b) = \left\{ v \in \mathbb{R}^p : \|v\|^* \leq 1 \text{ et } b^\top v = \|b\| \right\}.$$

Pour $\lambda_1 > \dots > \lambda_p > 0$, $m \in \mathcal{P}_p^{\text{slope}} \mapsto \partial J_\lambda(m)$ est une bijection entre $\mathcal{P}_p^{\text{slope}}$ et les faces de la boule unité de J_λ^* (permutoèdre signé) (Schneider et Tardivel 2022).

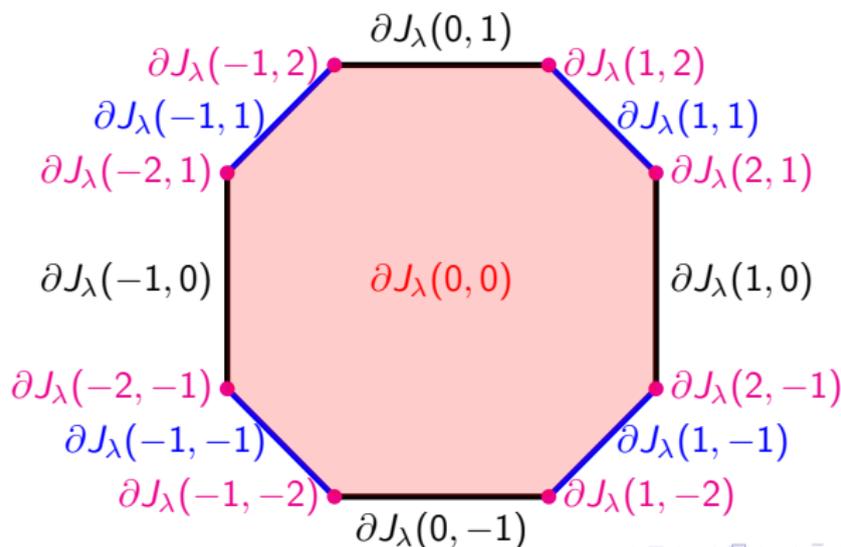
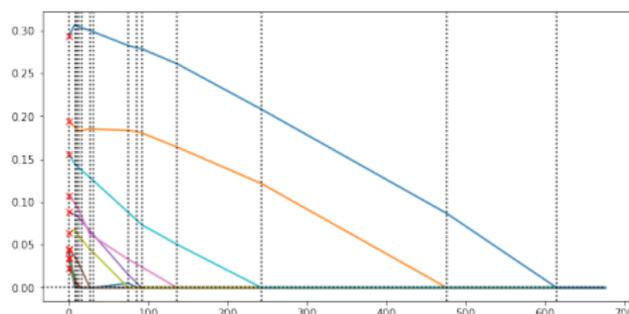
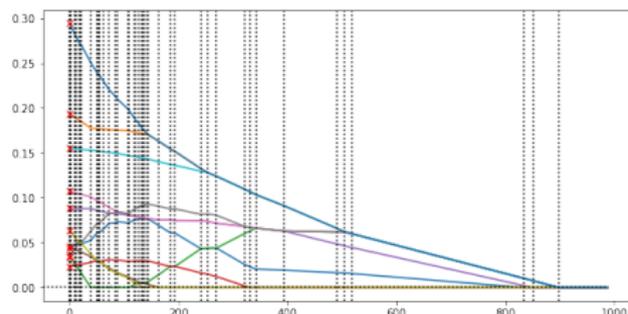


Illustration sur des données réelles

Jeu de données sur les vins rouges « Vinho Verde ».

- ▶ $X \in \mathbb{R}^{1599 \times 11}$: mesures physico-chimiques telles que le degrés d'alcool, taux de sucre, le pH, la densité,...
- ▶ $Y \in \mathbb{R}^{1599}$: note du vin 0 (mauvais) et 10 (excellent).
- ▶ Figure de gauche $\lambda = (1, \sqrt{2} - 1, \dots, \sqrt{11} - \sqrt{10})$. Figure de droite $\lambda = (1, \dots, 1)$ (LASSO).



Erreur de prédiction et formule SURE

$$\text{note moyenne} = \underbrace{\beta_0 + \beta_1 \times \text{alcool} + \dots + \beta_{11} \times \text{pH}}_{X\beta}$$

$$\widehat{\text{note moyenne}}(\gamma) = \underbrace{\widehat{\beta}_0(\gamma) + \widehat{\beta}_1(\gamma) \times \text{alcool} + \dots + \widehat{\beta}_{11}(\gamma) \times \text{pH}}_{X\widehat{\beta}(\gamma)}$$

$$\text{Erreur de prédiction}(\gamma) = \mathbb{E}(\|X\widehat{\beta}(\gamma) - X\beta\|_2^2)$$

Formules SURE (Stein Unbiased Risk Estimate) :

$$\text{SURE}(\gamma) = \begin{cases} \|Y - X\widehat{\beta}(\gamma)\|_2^2 - n\sigma^2 + 2\sigma^2 \|\text{schm}(\widehat{\beta}(\gamma))\|_\infty & \text{SLOPE,} \\ \|Y - X\widehat{\beta}^{\text{lasso}}(\gamma)\|_2^2 - n\sigma^2 + 2\sigma^2 \|\widehat{\beta}^{\text{lasso}}(\gamma)\|_0 & \text{LASSO.} \end{cases}$$

SURE fournit un estimateur sans biais de l'erreur de prédiction :

$$\mathbb{E}(\text{SURE}(\gamma)) = \text{Erreur de prédiction}(\gamma)$$

	γ_{sure}	SURE	$\text{schm}(\hat{\beta}(\gamma_{\text{sure}}))$ or $\text{sign}(\hat{\beta}^{\text{lasso}}(\gamma_{\text{sure}}))$
SLOPE	18.6292	3.4641	(4, -8, -1, 2, -5, 3, -6, -4, -4, 7, 9)
LASSO	11.7602	4.1297	(0, -1, 0, 1, -1, 1, -1, 0, -1, 1, 1)

- ▶ L'estimateur SLOPE $\hat{\beta}(\gamma_{\text{sure}})$ apparie les variables « acidité », « densité » et « pH » (groupe d'appariement « 4 » : X_1, X_8, X_9).
- ▶ L'estimateur LASSO $\hat{\beta}^{\text{lasso}}(\gamma_{\text{sure}})$ sélectionne uniquement le « pH ».

Conclusion et perspectives

Conclusion :

- ▶ Le chemin des solutions du SLOPE est continu et affine par morceau.
- ▶ Algorithme pour calculer le chemin des solutions.

Perspectives :

- ▶ Choisir $\lambda \in \mathbb{R}^p$.
- ▶ Montrer que la notion d'appariement est pertinente pour le traitement de données.

Conclusion et perspectives

Conclusion :

- ▶ Le chemin des solutions du SLOPE est continu et affine par morceau.
- ▶ Caractérisation des morceaux (*i.e.* des intervalles I_m).
- ▶ Algorithme pour calculer le chemin des solutions.

Perspectives :

- ▶ Choisir $\lambda \in \mathbb{R}^p$.
- ▶ Montrer que la notion d'appariement est pertinente pour le traitement de données. **OSCAR ou SLOPE est pertinent pour votre application, faites-le savoir !**

Merci !

-  Dupuis, Xavier et Patrick Tardivel (2022). “Proximal operator for the sorted ℓ_1 norm : Application to testing procedures based on SLOPE”. In : *Journal of Statistical Planning and Inference*.
-  – (2024). “The solution path of slope”. In : *International Conference on Artificial Intelligence and Statistics*. PMLR, p. 775-783.
-  Schneider, Ulrike et Patrick Tardivel (2022). “The geometry of uniqueness, sparsity and clustering in penalized estimation”. In : *Journal of Machine Learning Research*.
-  Tardivel, Patrick, Rémi Servien et Didier Concordet (2020). “Simple expressions of the LASSO and SLOPE estimators in low-dimension”. In : *Statistics*.