Le chemin des solutions de l'estimateur SLOPE

Patrick J C Tardivel, Xavier Dupuis

Le chemin des solutions de l'estimateur SLOPE

Patrick Tardivel [1] Xavier Dupuis [1]

[1] *Institut de Mathématiques de Bourgogne UMR 5584, CNRS Université de Bourgogne, F-21000 Dijon, France, Patrick.Tardivel@u-bourgogne.fr*

Résumé. L'estimateur SLOPE a la particularité d'avoir des composantes nulles (parcimonie) et des composantes égales en valeur absolue (appariement). Le nombre de groupes d'appariement dépend du taux de régularisation de cet estimateur. Avoir un petit nombre de groupes facilite l'interprétation de l'estimateur SLOPE. Ainsi, le taux de régularisation peut être choisi de façon à avoir un bon compromis entre un faible nombre de groupes d'appariement et un estimateur précis. Trouver un tel compromis nécessite de calculer l'estimateur SLOPE en fonction du taux de régularisation ; c'est le problème que nous allons aborder durant cette présentation.

Mots-clés. Chemin des solutions, SLOPE, parcimonie, appariement.

Abstract. SLOPE estimator has the particularity of having null components (sparsity) and components that are equal in absolute value (clustering). The number of clusters depends on the regularization parameter of this estimator. Having a small number of clusters makes SLOPE estimator easy to interpret. Thus, the regularization parameter can be chosen to have a good trade-off between a small number of clusters and an accurate estimator. Finding such a compromise requires calculating the SLOPE estimator as a function of the regularization parameter; this is the issue that we will address during this presentation.

Keywords. Solutions path, SLOPE, sparsity, clustering.

1 Introduction

L'estimateur SLOPE (de l'anglais Sorted L One Penalized Estimator (Bogdan et al. (2015), Zeng et Figueiredo (2014)) est défini comme une solution du problème d'optimisation des moindres carrés pénalisés décrit ci-dessous :

$$\underset{b\in\mathbb{R}^p}{\text{minimiser}} \quad \frac{1}{2}\|y - Xb\|_2^2 + \gamma \sum_{i=1}^{p} \lambda_i |b|_{\downarrow i}. \tag{1}$$

Dans l'équation précédente, $\lambda_1 > 0$, $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ est une famille donnée de paramètres de pénalité, $\gamma > 0$ est le taux de régularisation et $|b|_{\downarrow 1} \geq \cdots \geq |b|_{\downarrow p} \geq 0$ sont les composantes de b décroissantes en valeur absolue. L'estimateur SLOPE généralise l'estimateur LASSO (de l'anglais Least Absolute Shrinkage and Selection Operator (Tibshirani (1996))) ainsi que l'estimateur OSCAR (de l'anglais Octagonal Shrinkage and Clustering Algorithm for

Regression (Bondell et Reich (2008))). En effet, $\lambda_1 = \cdots = \lambda_p = 1$ pour l'estimateur LASSO et la suite $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ est arithmétique pour l'estimateur OSCAR.

L'estimateur SLOPE gagne en popularité chez les statisticiens dû aux propriétés pertinentes de contrôle du taux de faux positifs (Bogdan et al. (2015)) et de réduction de la dimension du modèle de régression. Cette dernière propriété résulte de la structure des solutions du problème d'optimisation (1) qui ont des composantes nulles (parcimonie) ainsi que des composantes égales en valeur absolue (appariement) (voir Schneider et Tardivel (2022) lorsque X est une matrice quelconque[1]). Lorsque y représente la réponse aléatoire d'un modèle de régression linéaire, la parcimonie a une interprétation statistique bien connue : identifier les variables explicatives pertinentes. L'appariement a également une interprétation statistique lorsque les variables explicatives sont aléatoires et ont la même variance : les variables explicatives ayant le même coefficient de régression en valeur absolue ont, à un signe près, la même corrélation conditionnelle avec la réponse (*i.e* lorsque les autres variables explicatives sont fixes). Par ailleurs, sans restriction sur les variances, pour une variable catégorielle ayant différents niveaux, les coefficient de régression égaux représentent des niveaux pouvant être regroupés (Stockell et al. (2021)). Par conséquent, l'estimateur SLOPE peut identifier les variables explicatives pertinentes, regrouper les variables explicatives ayant le même impact sur la réponse et, plus généralement, réduire la dimension du modèle de régression.

Le chemin des solutions donne la solution d'un problème d'optimisation pénalisé en fonction du taux de régularisation $\gamma > 0$. Pour le LASSO, ce chemin montre que le nombre de variables explicatives sélectionnées par cet estimateur a tendance à diminuer lorsque le taux de régularisation devient grand. L'ajustement du paramètre de régularisation γ permet un compromis entre sélectionner un petit nombre de variables explicatives et construire un estimateur précis. De façon similaire, la construction du chemin des solutions du SLOPE est utile pour ajuster le taux de régularisation pour avoir un bon compromis entre sélectionner un faible nombre de groupes de variables explicatives et avoir un estimateur précis.

Durant la présentation, pour une famille de paramètres de pénalité $\lambda_1 > 0$ et $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ spécifiée, nous donnerons quelques propriétés théoriques sur le chemin des solutions de l'estimateur SLOPE lorsque le taux de régularisation γ varie dans $]0, +\infty[$.

1.1 Notions essentielles liées à l'estimateur SLOPE

Soit $\Lambda = (\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^p$ tel que $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ avec $\lambda_1 > 0$. La norme ℓ_1 ordonnée est définie par :

$$J_\Lambda(b) = \sum_{i=1}^{p} \lambda_i |b|_{\downarrow i}, \quad b \in \mathbb{R}^p.$$

[1]Dans le cas particulier où X est une matrice orthogonale, la solution du problème (1) est explicite et il est facile de vérifier les propriétés de parcimonie et d'appariement de cette solution (Bogdan et al (2015), Tardivel et al. 2020 ou Dupuis et Tardivel (2022)).

Le dual de cette norme a une expression explicite (Negrinho et Martins (2014)) rappelée ci-dessous :

$$J_\Lambda^*(b) = \max\left\{\frac{\|b\|_{(1)}}{\lambda_1}, \frac{\|b\|_{(2)}}{\sum_{i=1}^2 \lambda_i}, \ldots, \frac{\|b\|_{(p)}}{\sum_{i=1}^p \lambda_i}\right\}, \quad b \in \mathbb{R}^p,$$

où $\|b\|_{(k)} = |b|_{\downarrow 1} + \cdots + |b|_{\downarrow k}$ est la k-norme : la somme des k plus grandes composantes de b en valeur absolue. La notation $S_{X,\gamma J_\Lambda}(y)$ représente l'ensemble des solutions du problème (1). Notons que $S_{X,\gamma J_\Lambda}(y)$ est un ensemble convexe compact (non vide)[2] qui dans dans certains cas pathologiques peut ne pas être réduit à un singleton (Schneider et Tardivel (2022)).

2 Chemin des valeurs ajustées et chemin des solutions

La Proposition 1 donne quelques propriétés du chemin des solutions et des valeurs ajustées du SLOPE.

Proposition 1 *Soit $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$, $\Lambda = (\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^p$ tel que $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ avec $\lambda_1 > 0$ et $\gamma > 0$. Les propriétés suivantes sont satisfaites :*

1. *Si $\widehat{\beta}, \overline{\beta} \in S_{X,\gamma J_\Lambda}(y)$ alors $X\widehat{\beta} = X\overline{\beta}$. On note désormais $\widehat{\text{va}}(\gamma) = X\widehat{\beta}$ qui ne dépend pas de l'élément $\widehat{\beta}$ appartenant à $S_{X,\gamma J_\Lambda}(y)$.*
2. *Le chemin des valeurs ajustées $\gamma > 0 \mapsto \widehat{\text{va}}(\gamma)$ est continue et affine par morceaux.*
3. *Supposons que, pour tout $\gamma > 0$, $S_{X,\gamma J_\Lambda}(\gamma)$ soit un singleton et notons $\widehat{\beta}(\gamma)$ l'unique élément de cet ensemble. Alors, le chemin des solutions $\gamma > 0 \mapsto \widehat{\beta}(\gamma)$ est continue et affine par morceaux.*

Notons pour la propriété 3) que l'ensemble

$$\{X \in \mathbb{R}^{n\times p} : \exists y \in \mathbb{R}^n \ \exists \gamma > 0 \text{ pour lequel } S_{X,\gamma J_\Lambda}(y) \text{ n'est pas singleton}\}$$

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

2.1 Chemin du gradient et groupes d'appariement

Une solution du problème d'optimisation SLOPE est caractérisée par les deux conditions suivantes

$$\widehat{\beta} \in S_{X,\gamma J_\Lambda}(y) \Leftrightarrow \begin{cases} J_\Lambda^*(X^T(y - X\widehat{\beta})) \leq \gamma \\ \widehat{\beta}^T X^T(y - X\widehat{\beta}) = \gamma J_\Lambda(\widehat{\beta}) \end{cases}$$

Remarquons que $X^T(y - X\widehat{\beta}) = X^T(y - \widehat{\text{va}}(\gamma))$ est le gradient en $\widehat{\beta}$ de la somme des carrés résiduels $b \mapsto \frac{1}{2}\|y - Xb\|_2^2$. Par la suite, on appelle chemin du gradient l'expression $\gamma > 0 \mapsto$

[2]La fonction objective du problème d'optimisation (1) est convexe et coercive donc l'ensemble des solutions de ce problème est convexe et compact.

$X^T(y - \widehat{\text{vâ}}(\gamma))$. L'ensemble des inégalités décrivant la boule de rayon γ pour la norme ℓ_1 ordonnée duale[3] qui sont saturées par le gradient est :

$$\mathcal{A}(\gamma) := \left\{ i \in [p] : \frac{\|X^T(y - \widehat{\text{vâ}}(\gamma))\|_{(i)}}{\sum_{j=1}^{i} \lambda_j} = \gamma \right\}.$$

D'après la Proposition 2, l'ensemble $\mathcal{A}(\gamma)$ fournit à la fois le nombre de groupes d'appariement non nuls, la taille de ces groupes et le nombre de composantes non nulles.

Proposition 2 *Soit $\Lambda = (\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^p$ tel que $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ avec $\lambda_1 > 0$, $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$, $\gamma > 0$ et $\widehat{\beta} \in S_{X,\gamma J_\Lambda}(y)$.*

1. *Soit $1 \leq k_1 \leq \cdots \leq k_l \leq p$ une subdivision telle que[4] :*

   $\text{card}(\{i \in [p] : \widehat{\beta}_i \neq 0\}) = k_l$ et $|\widehat{\beta}|_{\downarrow 1} = \cdots = |\widehat{\beta}|_{\downarrow k_1} > \cdots > |\widehat{\beta}|_{\downarrow k_{l-1}+1} = \cdots = |\widehat{\beta}|_{\downarrow k_l} > 0$

   alors, $\{k_1, \ldots, k_l\} \subset \mathcal{A}(\gamma)$.

2. *Inversement, si $\{k_1, \ldots, k_l\} = \mathcal{A}(\gamma)$ alors[5]*

$$|\widehat{\beta}|_{\downarrow 1} = \cdots = |\widehat{\beta}|_{\downarrow k_1} \geq \cdots \geq |\widehat{\beta}|_{\downarrow k_{l-1}+1} = \cdots = |\widehat{\beta}|_{\downarrow k_l} \geq |\widehat{\beta}|_{\downarrow k_l+1} = \cdots = |\widehat{\beta}|_{\downarrow p} = 0.$$

Dans l'exemple suivant nous calculons les chemins des solutions et du gradient du SLOPE et nous illustrons que le nombre de groupes d'appariement dépend du cardinal de $\mathcal{A}(\gamma)$.

Exemple : Nous allons illustrer les chemins des solutions et du gradient du SLOPE (voir Figures 1 et 2) lorsque $y = (6, 2)^T \in \mathbb{R}^2$, $\Lambda = (4, 2) \in \mathbb{R}^2$ et $X \in \mathbb{R}^{2\times 2}$ est la matrice donnée ci-après

$$X = \begin{pmatrix} 1 & 0,5 \\ 0,5 & 1 \end{pmatrix}.$$

3 Conclusion et perspectives

La programmation d'algorithmes efficaces permettant la résolution numérique des chemins des solutions et du gradient du SLOPE ainsi que des illustrations de ces chemins sur des données réelles sont des travaux en cours.

[3] Cette boule est un polytope connu, en géométrie, sous le nom de permutoèdre signé.

[4] Cette chaîne d'inégalités signifie que $\widehat{\beta}$ a l groupes d'appariement non nuls, le groupe de la plus grande valeur a k_1 éléments, etc., et $\widehat{\beta}$ a k_l composantes non nulles.

[5] Cette chaîne d'inégalités signifie que le nombre de groupes d'appariement non nuls de $\widehat{\beta}$ est inférieur ou égal à l et que le nombre de composantes non nulles de $\widehat{\beta}$ est inférieur ou égal à k_l.

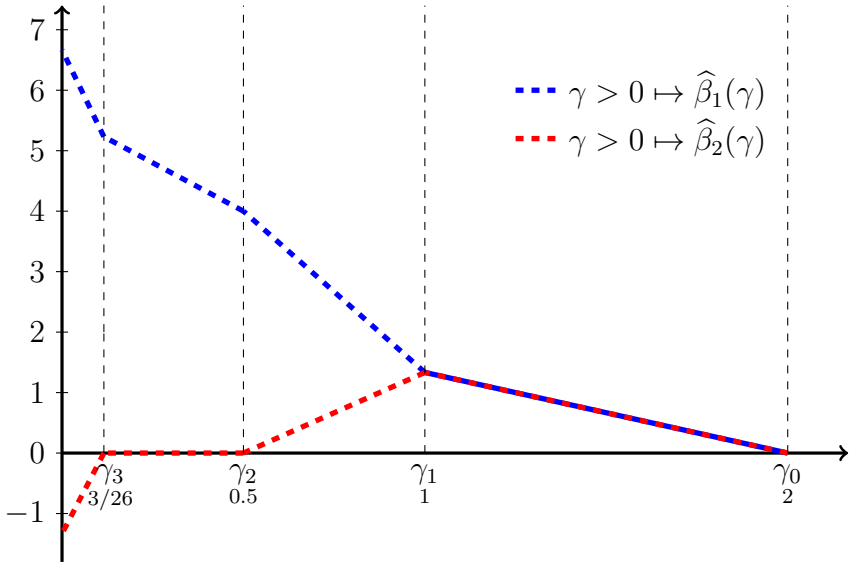

Figure 1: Cette figure fournit le chemin de la solution du SLOPE $\gamma > 0 \mapsto \widehat{\beta}(\gamma)$.

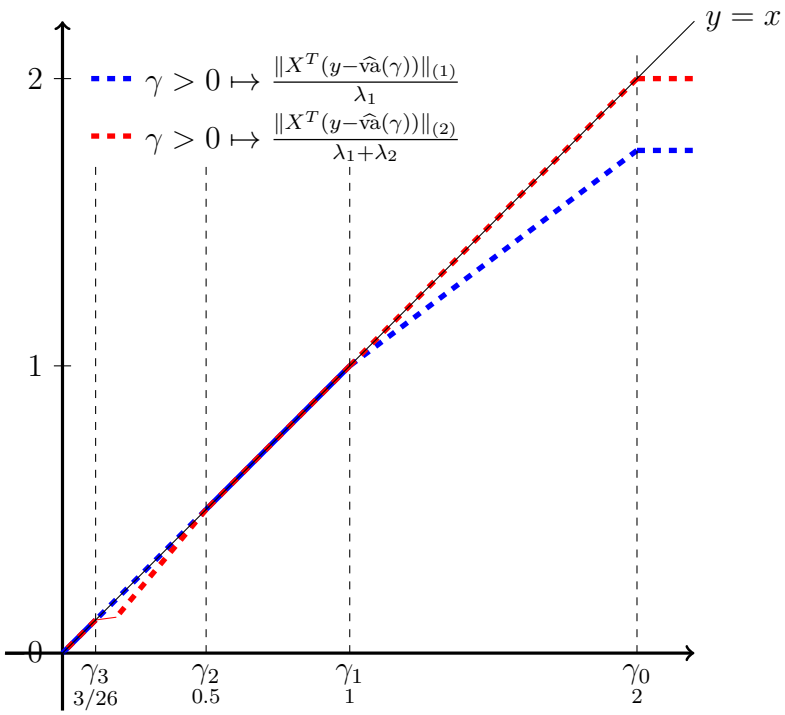

Figure 2: Cette figure fournit le chemin du gradient $\gamma > 0 \mapsto X'(y - \widehat{\text{va}}(\gamma))$. On remarque, d'après cette figure et la précédente (Figure 1), que pour $\gamma \in]1; 2]$ on a $\mathcal{A}(\gamma) = \{2\}$ de plus $\widehat{\beta}(\gamma)$ a au maximum un groupe d'appariement et deux composantes non-nulles. Pour $\gamma \in]0; 3/26] \cup [0,5; 1]$ on a $\mathcal{A}(\gamma) = \{1, 2\}$ de plus, $\widehat{\beta}(\gamma)$ a au maximum deux groupes d'appariement et deux composantes non-nulles. Pour $\gamma \in]3/26; 0,5[$ on a $\mathcal{A}(\gamma) = \{1\}$ de plus, $\widehat{\beta}(\gamma)$ a un groupe d'appariement et une composante non-nulle.

Remerciements

L'Institut de Mathématiques de Bourgogne (IMB) bénéficie du soutien de l'EIPHI Graduate School (contrat ANR-17-EURE-0002).

Bibliographie

Bogdan, M. and Van Den Berg, E. and Sabatti, C. and Su, W. and Candès E. J. (2015), Slope—adaptive variable selection via convex optimization. *The annals of applied statistics*, 9(3), p. 1103.

Bondell, H. D. and Reich, B. J. (2008), Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1), pp. 115–123.

Dupuis, X. and Tardivel, P. J. C. (2022), Proximal operator for the sorted ℓ_1 norm: Application to testing procedures based on slope. *Journal of Statistical Planning and Inference*, 221, pp. 1–8.

Negrinho, R. and Martins, A. (2014), Orbit regularization. *Advances in neural information processing systems*, 27, pp. 3221–3229.

Schneider, U. and Tardivel, P. (2022), The geometry of uniqueness, sparsity and clustering in penalized estimation. *Journal of Machine Learning Research*, 23(331), pp. 1–36.

Stokell, B. G. and Shah, R. D. and Tibshirani, R. J. (2021), Modelling high-dimensional categorical data using nonconvex fusion penalties. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(3), pp. 579–611.

Tardivel, P. J. C. and Servien, R. and Concordet, D. (2020), Simple expressions of the lasso and slope estimators in low-dimension. *Statistics*, 54(2), pp. 340–352.

Tibshirani, R. (1996), Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), pp. 267–288, 1996.

Zeng, X. and Figueiredo, M. A. T. (2014), Decreasing weighted sorted l1 regularization. *IEEE Signal Processing Letters*, 21(10), pp. 1240–1244.