



# Méthode de type LASSO

**Elaboré par:**
**KHENISSI Nour**
**TELMINI Mohamed Fyras**

**Année universitaire :2021/2022**
**Classe: 2A**

# Introduction

Une régression linéaire cherche à prédire et expliquer une variable quantitative Y à partir de p variables explicatives. Nous avons donc à notre disposition des variables candidates et nous cherchons à sélectionner celles qui sont potentiellement explicatives. L'étude des $2^p$ modèles possibles n'est pas envisageable dès que p est grand. C'est le cas de la régression en grande dimension où le nombre des variables explicatives est proche ou dépasse le nombre des observations ce qui induit naturellement de la corrélation ou quasi-colinéarité et donc une variance de prédiction très grande. Cela dit, la méthode du Lasso propose dans certains cas une solution à ce problème. C'est le principe de la régression pénalisée qui contraint les valeurs des paramètres qui peuvent prendre des valeurs erratiques en présence de corrélation et ainsi réduire la variance.
Afin d'étudier cette méthode plus en détails, nous reprenons dans une première partie les notions vues lors de la séance de la modélisation statistique présentée par Mr Mohamed Hebiri, puis nous essayons de résumer un article scientifique intitulé " Efficient Smoothed Concomitant Lasso Estimation for High Dimensional Regression" qui traite une nouvelle méthode d'estimation type LASSO.

# 1 Résumé de la séance: Méthodes de type LASSO pour la prévision de température

Les prévisions météorologiques sont une partie essentielle de notre vie quotidienne. Dans ce contexte, nous avons étudié le cas particulier de la température dans l'objectif d'arriver à la prédire pour un jour t donné. Cette prédiction est basée sur des données fournies par le NCDC (National Climatic Data Center) qui sont 2538 observations journalières de la température à Paris de 2003 à 2008. Chaque ligne dans ce jeu de données couvre un total de 17 variables tel que le jour, les incréments de températures hebdomadaires, la vitesse du vent etc...

La variable à prédire est donc $Y_t$ qui représente la différence des températures entre les jours t et t-1. L'objectif de la première partie de ce séminaire était de savoir comment caractériser la régression de $Y_t$ sur les variables explicatives présents dans la base de données.

## 1.1 Le modèle de régression hétéroscédastique

Ce modèle est défini comme:

$(x_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$ t=1,..,T tel que $y_t = b^*(x_t) + s^*(x_t)\xi_t$

- Espérance conditionnelle: $b^* : \mathbb{R}^d \to \mathbb{R}$ telle que $E[y_t|x_t] = b^*(x_t)$
- Variance conditionnelle: $s^{*2} : \mathbb{R}^d \to \mathbb{R}$ telle que $Var[y_t|x_t] = s^{*2}(x_t)$
- Erreurs normalisés $\xi_t \sim N(0, 1)$

Avant de pouvoir estimer la fonction $b^*(.)$, il faut tout d'abord connaître sa forme. Ce qu'on peut intuitivement dire est que la dépendance de cette fonction en l'entrée $x_t$ est polynomiale et qu'il y'a un coefficient sinusoïdal qui reflète de caractère oscillatoire de la température au cours du temps. L'idée est donc de prendre en compte ces deux phénomènes en construisant des groupes de fonctions qui correspondent aux différentes combinaisons possibles des paramètres de $x_t$. (slides 7 et 8)

Ainsi le problème se transforme en un problème de recherche d'un paramètre $\beta$ qui caractérise l'impact que chaque fonction des groupes de fonctions construits a sur le résultat $Y_t$. En appliquant cette méthodologie à notre cas, nous obtenons que $b^*(.)$ est de dimension p = 2176. Cela découle de l'hypothèse intuitive que la forme de $b^*(.)$ est polynomiale en les paramètres de $x_t$ (t exclu) et d'un coefficient de température.

Nous cherchons par la suite à caractériser le bruit $s^*(.)$. Pour pouvoir bien définir notre modèle de régression hétéroscédastique, il est naturel de dire que $s^*(.)$ et $Y_t$ sont de même dimension T. Le modèle est donc (slide 12)

## 1.2 Résolution du problème avec variance connue

La variance $\sigma^*$ peut être connue ou inconnue. Nous commencerons par étudier le cas où elle est connue.

### 1.2.1 Estimateur MCO

Nous estimons le coefficient $\beta$ par l'estimateur des moindres carrés défini comme suit :
$\hat{\beta} \in argmin_{\beta \in \mathbb{R}^p} ||X\beta - Y||^2$ avec $||.||$ est la norme empirique. Pour étudier la convergence du MCO nous calculons $E[\frac{1}{T}|X\hat{\beta} - X\beta^*|^2]$ qui est égale à $\frac{(\sigma^*)^2 rang(X)}{T}$. Selon les valeurs de p et T il y a deux cas possibles

- si $p < T$ on a $rang(X) \sim p$ et la borne $\frac{(\sigma^*)^2 rang(X))}{T}$ converge vers 0 lorsque T tend vers l'infini
- si $p > T$ on a $rang(X) \sim T$ et la borne $\frac{(\sigma^*)^2 rang(X))}{T}$ se simplifie en $\sigma^{*2}$ et donc l'estimateur MCO ne converge pas

### 1.2.2 Sparsité

L'hypothèse de sparsité s'énonce comme suit: Peu de variables dans X contribuent à expliquer y
$|\beta^*|_0 = p_0 \ll min(p, T)$ où $|\beta^*|_0 = card(j, \beta_j \neq 0) \forall \beta \in \mathbb{R}^p$
Cette hypothèse est applicable à notre modèle car la construction qu'on a fait pour $\beta$ moyennant les groupes de fonctions arbitraires fait que très peu de ces fonctions expliquent réellement $Y_t$. Il suffirait ainsi de connaître ces fonctions et d'enlever les restes (à coefficient nul) pour pouvoir satisfaire l'hypothèse de sparsité et utiliser l'estimateur des MCO.

### 1.2.3 Critères de pénalisation AIC - BIC et LASSO

Par définition, un critère d'information permet de se prémunir contre la surparamétrisation du modèle en introduisant un coût à l'introduction de chaque paramètre supplémentaire. Parmi ces critère on cite le AIC et le BIC qui sont définis pénalisent la régression linéaire par la norme $l_0$:
$\beta_\lambda = argmin_{\beta \in \mathbb{R}^p}(||X\beta - Y||^2 + \lambda|\beta|_0)$ avec $\lambda$ de l'ordre de $\frac{c\sigma^{*2} log(p)}{T}$ où $c > 0$
En effet le paramètre $\lambda$ contrôle l'importance de la pénalité. Plus particulièrement, pour $\lambda = 0$ aucune pénalité n'est appliquée, on retrouve le cadre de l'estimateur des moindres carrés classique et pour $\lambda \to \infty$, toutes les variables sont associées à un coefficient estimé nul. Cependant dans le cas de la pénalisation $l_0$ l'estimation effective du $\beta$ nécessite $2^p$ modèles à tester ce qui n'est pas envisageable en pratique.
Pour remédier à ce problème, on change la pénalisation $l_0$ par une pénalisation de type $l_1$ qui a l'avantage d'être convexe et de garantir l'unicité de la solution. La formulation mathématique est la suivante:
$\tilde{\beta_\lambda} = argmin_{\beta \in \mathbb{R}^p}(||X\beta - Y||^2 + \lambda|\beta|_1)$
Théoriquement la valeur de $\lambda$ est de l'ordre de $2\sigma^{*2}\sqrt{\frac{log(\frac{p}{\eta})}{T}}$ avec $\eta \in ]0, 1[$. Mais en pratique cette valeur n'aboutit pas à de bons résultats. Ainsi, des méthodes empiriques comme la validation croisée ont été mises en place pour calibrer $\lambda$.

### 1.2.4 Algorithme PCO pour LASSO

Il existe un grand nombre de méthodes pour calculer le LASSO en pratique (méthodes de points intérieurs, LARS, algorithmes PCO,...). Dans cette partie nous intéréssons à la méthode d'optimisation coordonnée par coordonnée (ou Pathwise Coordinate Optimization, PCO dans la suite) qui semble être la plus simple.

**Algorithme:**

On part de $\beta \leftarrow (0, ..., 0)'$.

Puis, jusqu''a ce que $\beta$ se stabilise, on parcourt les coordonnées $(j = 1, ..., p)$ en prenant

$\beta_j = Sgn(a_j)(|a_j| - \frac{\lambda \xi_j^{\frac{1}{2}}(M)}{M_{j,j}})$

où $M = (X'X)$ et $a_j$ est défini par $a_j = \frac{(Y'X(X^{\tau}X)^{-1}M) - \sum_{l \neq j} \beta_l M_{i,j}}{M_{j,j}}$

## 1.3 Résolution du problème avec variance inconnue

Dans le cas où $\sigma^*$ est inconnu, nous devons procéder autrement. Ceci correspond au cas pratique dans lequel nous nous situons mais aussi la majorité des cas réels puisque l'hypothèse de la variance connue est assez forte. La formulation matricielle est la suivante:

$Y = X\beta^* + \sigma^* \xi$

- Réponse: $Y = [y_1, ..., y_T]^T \in \mathbb{R}^T$
- Bruit: $\xi = [\xi_1, ..., \xi_T]^T \in \mathbb{R}^T$
- Matrice construite $X_{t,j} = [f_j(x_t)] \in \mathbb{R}$
- Coefficients $\beta = [\beta_1, ..., \beta_p]^T \in \mathbb{R}^p$
- Ecart type $s^*(x_t) = \sigma^* \in \mathbb{R}_*^+$

### 1.3.1 Méthode du scaled LASSO

l'algorithme du *scaled LASSO* est une descente de gradient dans une minimisation convexe d'une fonction de perte pénalisée conjointement pour le niveau de bruit et les coefficients de régression. Sous certaines conditions de régularité, nous pourrons démontrer que le *scaled LASSO* produit simultanément un estimateur du niveau de bruit et un vecteur de coefficients estimés satisfaisant certaines inégalités d'oracle pour la prédiction, l'estimation du niveau de bruit et les coefficients de régression. Ces inégalités fournissent des conditions suffisantes pour la convergence et la normalité asymptotique de l'estimateur du niveau de bruit, y compris dans le cas où $p < n$.

### 1.3.2 Méthode du square-root LASSO

Une autre version modifiée du LASSO est la méthode du *square-root LASSO*. Celle-ci ne repose pas sur la connaissance de * et ne nécessite pas de le préestimer. En outre, elle ne dépend pas de la normalité du bruit. Cette méthode atteint une performance proche de l'oracle, avec un taux de convergence similaire à celui atteint par le lasso dans le cas où $\sigma$ connu. Ces résultats de performance sont valables pour les erreurs gaussiennes et non gaussiennes.

### 1.3.3 Reformulation du problème

Nous rappelons que dans un cadre de régression hétéroscédastique avec un niveau de bruit inconnu, notre objectif est l'estimation conjointe de l'espérance conditionnelle $b^*$ et de la volatilité conditonnelle $s^*$. La formulation mathématique est la suivante:
$r^*(x_t)y_t = f^*(x_t) + \xi_t$ avec $r^*(x) = \frac{1}{s^*(x)}$ et $f^*(x) = \frac{b^*(x)}{s^*(x)}$

### 1.3.4 Hypothèses sur le modèle

La résolution de ce problème nécessite la considération de deux hypothèses:

- **Sparcité par groupe**: Pour une famille $G_1...G_K$ de sous ensembles disjoints de 1,..,p, il existe un vecteur $\phi^* \in \mathbb{R}^p$ tel que $[f(x_1)...f(x_T)]^T = X\phi^*$ $Card(k : |\phi^*_{G_k}| \neq 0) \ll K$
- **Volatilité de Faible dimension**: Etant données q fonctions $r_1$,...,$r_q$ de $\mathbb{R}^d$ dans $\mathbb{R}_+$, il existe un vecteur $\alpha^* \in \mathbb{R}^q$ tel que $r^*(x) = \sum_{l=1}^q \alpha_l^* r_l(x)$ pour presque tout $x \in \mathbb{R}^d$, et alors on peut écrire $[r^*(x_1)...r^*(x_T)]^T = R^*$

### 1.3.5 Group-Lasso

L'idée du Group-Lasso est d'avoir une méthode fournissant une sélection parcimonieuse de groupes (fournis a priori) et non de variables. Nous avons recours à la formulation log-vraisemblance pénalisée pour estimer le Group-Lasso. Nous cherchons donc à minimiser:
$PL(\phi, \alpha) = -\sum_{t=1}^T log(R_{t,:}\alpha) + \sum_{t=1}^T \frac{1}{2}(y_t R_{t,:}\alpha - X_{t,:}\phi) + \sum_{k=1}^K \lambda_k |X_{:,G_k}\phi_{G_k}|$ sous la contrainte $min_t R_{t,:}\alpha > 0$
Nous rappelons que dans notre cas R est une matrice de taille T*q et qui représente un bruit sinusoïdale.
Par la méthode de Scaled Heteroscedastic Dantzig selector (ScHeDs) nous déterminons la paire $(\hat{\phi}, \hat{\alpha})$ associée au minimiseur $(\hat{\phi}, \hat{\alpha}, \hat{v})$ par rapport à $(\phi, \alpha, v)$ de la fonction de perte
$\sum_{k=1}^K \lambda_k |X_{:,G_k}\phi_{G_k}|$ sous les contraintes:

- $|\Pi_{G_k}(diag(Y)R\alpha - X\phi)|_2 \leq \lambda_k$, $\forall k \in (1, .., K)$
- $R^T v \leq R^T diag(Y)(diag(Y)R\alpha - X\phi)$
- $0 \leq \frac{1}{v_t} \leq R_{t,:}\alpha$ $\forall t \in (1, ..., T)$

avec $\Pi_{G_k}$ est le projecteur sur Span(X:,$G_k$ ).
Les contraintes correspondent aux conditions du premier ordre.
En effet, Le ScHeDs est toujours bien défini dans le sens où l'ensemble solution du problème de minimisation est non vide et il contient le minimiseur de la log-vraisemblance pénalisée.

## 1.4 Résultats sur les données réelles

Afin de tester la performance de la méthode ScHeDs, nous avons choisi de la comparer en première partie au square-root LASSO moyennant des données simulées avec un bruit homoscédastique. Cette simulation

contient 8 variations de 500 échantillons différents de $\mathbf{Y}$. Une variation correspond à des valeurs différents de (T, p, $i^*$, $\sigma^*$). Le tableau ci dessous résume les résultats de ce test.

| **ScHeDs** | $\lvert\widehat{\beta} - \beta^*\rvert_2$ | | $\lvert\widehat{i} - i^*\rvert$ | | $10\lvert\widehat{\sigma} - \sigma^*\rvert$ | |
|---|---|---|---|---|---|---|
| $(T,\ p,\ i^*,\ \sigma^*)$ | Ave | StD | Ave | StD | Ave | StD |
| $(100, 100, 2, .5)$ | **.06** | .03 | **.00** | .00 | **.29** | .21 |
| $(100, 100, 5, .5)$ | **.11** | .08 | **.01** | .12 | **.32** | .37 |
| $(100, 100, 2,\ 1)$ | **.13** | .07 | **.03** | .16 | **.57** | .46 |
| $(100, 100, 5,\ 1)$ | **.28** | .23 | **.10** | .33 | .77 | .68 |
| $(200, 100, 5, .5)$ | **.08** | .02 | **.00** | .00 | **.23** | .16 |
| $(200, 100, 5,\ 1)$ | **.16** | .05 | **.00** | .01 | **.09** | .29 |
| $(200, 500, 8, .5)$ | **.09** | .03 | **.00** | .00 | **.22** | .16 |
| $(200, 500, 8,\ 1)$ | .21 | .11 | **.03** | .17 | .48 | .43 |

| **Square-root Lasso** | $\lvert\widehat{\beta} - \beta^*\rvert_2$ | | $\lvert\widehat{i} - i^*\rvert$ | | $10\lvert\widehat{\sigma} - \sigma^*\rvert$ | |
|---|---|---|---|---|---|---|
| $(T,\ p,\ i^*,\ \sigma^*)$ | Ave | StD | Ave | StD | Ave | StD |
| $(100, 100, 2, .5)$ | .08 | .06 | .19 | .44 | .32 | .23 |
| $(100, 100, 5, .5)$ | .12 | .04 | .18 | .42 | .33 | .24 |
| $(100, 100, 2,\ 1)$ | .16 | .10 | .19 | .44 | .59 | .48 |
| $(100, 100, 5,\ 1)$ | .25 | .16 | .21 | .43 | **.68** | .47 |
| $(200, 100, 5, .5)$ | .09 | .03 | .21 | .45 | .24 | .17 |
| $(200, 100, 5,\ 1)$ | .18 | .07 | .21 | .48 | .48 | .32 |
| $(200, 500, 8, .5)$ | .10 | .03 | .14 | .38 | .23 | .17 |
| $(200, 500, 8,\ 1)$ | .21 | .07 | .18 | .40 | **.46** | .34 |

**Figure 1:** Résultats avec correction de biais pour les deux méthodes

Nous observons clairement que la méthode ScHeDs donne quasiment tout le temps plus des estimations plus précises sur les variables désirées. Avec la méthode squared-root qui la dépasse seulement deux fois sur 48. La méthode ScHeDs est ainsi clairement beaucoup mieux performante.

Un deuxième test est effectué sur les données réelles du NCDC. En utilisant cette fois les données de 2003 à 2007 (2172 valeurs) pour l'apprentissage et les 366 valeurs de 2008 pour le test. Cela nous a permis de conclure quant à la question de la sparsité puisqu'on a observé qu'il y'avait une réduction de dimensionalité du problème de 2176 à 26 variables. 62% des estimations du signal étaient de signe correcte. La volatilité estimée nous a fait conclure aussi que les oscillations de la température durant la période entre Mai et Juillet sont significativement plus grands qu'en Mars, Septembre et Octobre. Les graphiques ci-dessous montrent les incréments de températures de 2008 réels et prédits et leurs volatilité estimée.

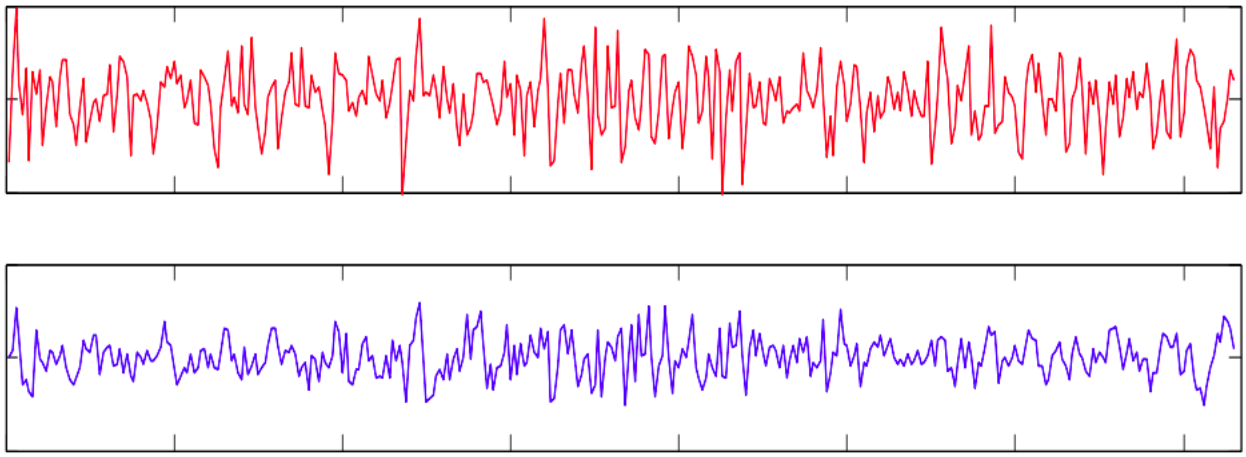

**Figure 2:** Incréments de température réels en 2008 (en rouge) et incréments de température estimés pour la même période (en bleu)

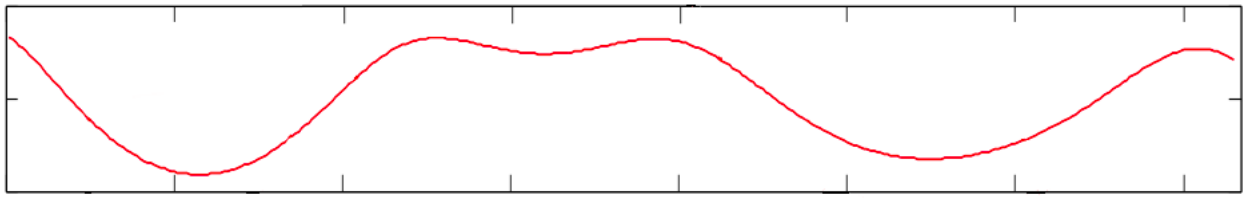

**Figure 3:** volatilité de la température estimée

# 2 Résumé de l'article scientifique: Efficient Smoothed Concomitant Lasso Estimation for High Dimensional Regression

Dans la deuxième partie de ce rapport, nous avons choisi de résumer l'article[1] qui nous a été fourni par notre encadrant. Cette article est assez technique et traite une autre méthode d'estimation type LASSO qui est utile dans des problèmes qui ressemblent à celui que nous avons traité dans la partie 1. Cette nouvelle méthode d'estimation s'intitule Smoothed Concomitant Lasso et est une formulation plus numériquement stable des travaux faits par d'autres chercheurs[3][2].

## 2.1 Formulation théorique

L'estimateur Concomitant Lasso[4] $\hat{\beta}^{(\lambda)}$ est défini pour $\lambda$ un paramètre donné comme la solution du problème d'optimisation suivant:

$$(\hat{\beta}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in argmin_{\beta \in \mathbb{R}^p, \sigma > 0} P_\lambda(\beta, \sigma) := \frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1$$

Une observation des valeurs critiques de cet estimateur montre que, d'une part $\hat{\beta}^{(\lambda)} = 0$ pour des valeurs de $\lambda$ très grandes ($\lambda \geqslant \lambda_{max} := \left\|\mathbf{X}^T y\right\|_\infty \setminus (\|y\| \sqrt{n})$). D'autre part, pour des valeurs de lambda plus petits qu'un certain seuil $\lambda_{min}$, $\hat{\sigma}^{(\lambda)} = 0$. Chose qui ne pose pas problème au niveau de la définition de l'estimateur mais fait que son calcul numérique devient de plus en plus long. La résolution de ce problème a motivé la formalisation de la méthode étudiée par cette article appelée Smoothed Concomitent LASSO en ajoutant une contrainte sur $\sigma$ qui est $\sigma_0$ la limite inférieure du seuil de bruit. Cet estimateur $\hat{\beta}^{(\lambda, \sigma_0)}$ est défini pour $\lambda$ et $\sigma_0$ donnés comme la solution du problème d'optimisation suivant:

$$argmin_{\beta \in \mathbb{R}^p, \sigma_0 > 0} P_\lambda(\beta, \sigma) := \frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1 + \mathbb{1}_{[\sigma_0, +\infty]}(\sigma)$$

Cet estimateur présente aussi un problème pour les valeurs de $\lambda$ très grandes, puisque $\hat{\beta}^{(\lambda, \sigma_0)} = 0$ pour $\lambda_{max} := \left\|\mathbf{X}^T y\right\|_\infty \setminus (n(max(\sigma_0, \|y\| \sqrt{n})))$.
Par ailleurs, voulant avoir une meilleure efficacité en terme de calcul, les auteurs ont utilisé implémenté un test de validité des paramètres à l'intérieur de l'algorithme. Ce test vérifie la nullité des j-èmes composantes de $\hat{\beta}^{(\lambda, \sigma_0)}$ et les enlève de la boucle d'optimisation dès que cette nullité est vérifiée.

## 2.2 Algorithme de résolution

L'algorithme numérique du calcul de $\hat{\beta}^{(\lambda, \sigma_0)}$ tel qu'il est énoncé dans l'article scientifique est:

**Algorithm 1:** CD4SCL – Coordinate Descent for the Smoothed Concomitant Lasso with Gap Safe screening

**Input** : $X, y, \varepsilon, K, f^{\text{ce}}(= 10), \lambda, \sigma_0, \beta, \sigma$
$\mathcal{A} \leftarrow [p]$
**for** $k \in [K]$ **do**
  **if** $k \mod f^{\text{ce}} = 1$ **then**
    Compute $\theta$ as in Proposition 6
    **if** $G_{\lambda,\sigma_0}(\beta, \sigma, \theta) = P_{\lambda_t,\sigma_0}(\beta, \sigma) - D_{\lambda_t,\sigma_0}(\theta) \leq \varepsilon$. **then** // Stopping criterion
      **break**
    Update $\mathcal{A}$ thanks to Proposition 5 // Screening test
  **for** $j \in \mathcal{A}$ **do** // Loop over coordinates
    $\beta_j \leftarrow \mathcal{S}_{n\sigma\lambda_t/\|X_j\|^2}\left(\beta_j - X_j^\top (X\beta - y)/\|X_j\|^2\right)$ // Soft-thresholding step
    $\sigma \leftarrow \sigma_0 \vee (\|y - X\beta\|/\sqrt{n}))$ // Noise estimation step
**Output**: $\beta$, $\sigma$, $\mathcal{A}$

**Figure 4:** Algorithme tel qu'il est décrit dans l'article scientifique

Nous estimons qu'il est important d'inclure l'algorithme en tant que figure dans ce résumé pour comprendre visuellement la structure des boucles. Voici la manière dont il fonctionne:

- les données en entrée sont: les données d'entrée **X**, les sorties associées y, une précision $\varepsilon$, des points de départ pour $\hat{\beta}$ et $\hat{\sigma}$, un nombre maximal d'itérations et un ensemble de coordonnées à parcourir K.
- Une boucle sur l'ensemble des coordonnées commence. Il y'a vérification du critère d'arrêt qui est le calcul du saut de dualité du problème d'optimisation. Ceci reflète le fait que, si la précision donnée est atteinte par le saut de dualité, alors l'estimateur obtenu est optimal.
- Si le critère d'arrêt n'est pas vérifié, alors l'ensemble des paramètres $\mathcal{A}$ est mis à jour en annulant les j-èmes composantes de $\hat{\beta}^{(\lambda,\sigma_0)}$ (et donc en les éliminant du problème d'optimisation).
- Le calcul des composantes de l'estimateur de fait par l'algorithme de descente par coordonnées.

L'algorithme retourne l'estimateur recherché avec son ensemble de composantes non nulles.

## 2.3 Résultats numériques

La dernière partie de l'article concerne le test de la performance de l'algorithme établi précédemment par rapport à d'autres algorithmes de pointe. Ces algorithmes correspondent aux estimateurs présentés dans le tableau ci-dessous:

| Estimateur | formule |
|---|---|
| $\hat{\sigma}_{OR}$ | $\frac{\left\|y-P_{X,S^*}y\right\|}{\sqrt{n-\|S^*\|}}$ |
| $\hat{\sigma}_{M-LS}$ | $\frac{\left\|y-X\hat{\beta}_M^{\lambda_{cv}}\right\|}{\sqrt{n-\|\hat{S}_M^{\lambda_{cv}}\|}}$ |
| $\hat{\sigma}_{M-CV}$ | $\frac{\left\|y-P_{X,\hat{S}_M}y\right\|}{\sqrt{n-\|\hat{S}_M\|}}$ |
| $\hat{\sigma}_i$ | $\frac{\left\|y^{(i')}-P_{X^{(i')},\hat{S}_i}y^{(i')}\right\|}{\sqrt{n/2-\|\hat{S}_i\|}}$ |
| $\hat{\sigma}_{D2}$ | $\frac{(1+\frac{p\hat{m}_1^2}{(n+1)\hat{m}_2})\|y\|^2}{n} - \frac{\hat{m}_1\left\|X^Ty\right\|^2}{\sqrt{n(n+1))\hat{m}_2}}$ |

Un premier test a été effectué sur des données synthétiques qui sont comme suit: On cherche à prédire $y = X\beta^* + \sigma\varepsilon$ où $\varepsilon \sim \mathcal{N}(0, Id_n)$ et X $\in R^{n\times p}$ suit une distribution normale multivariée de covariance $\sum = (\rho^{|ij|})_{i,j\in[p]}$. $\beta^* = \alpha\beta$ où les coordonnées de $\beta$ sont tirées d'une distribution de Laplace avec un taux aléatoire s% de ses composants sont mis à zéro. Le scalaire $\alpha$ est choisi afin de satisfaire un rapport signal sur bruit noté snr donné. $\alpha$ est défini par: $\alpha = \sqrt{snr \times \sigma^2/\beta^T \sum \beta}$. La figure suivante montre le résultat en terme de performance des différents estimateurs par rapport à l'estimation de ce jeu de données synthétique avec les paramètres suivants: n = 100, p = 500, $\rho$ = 0.6, snr = 5, s = 0.9.

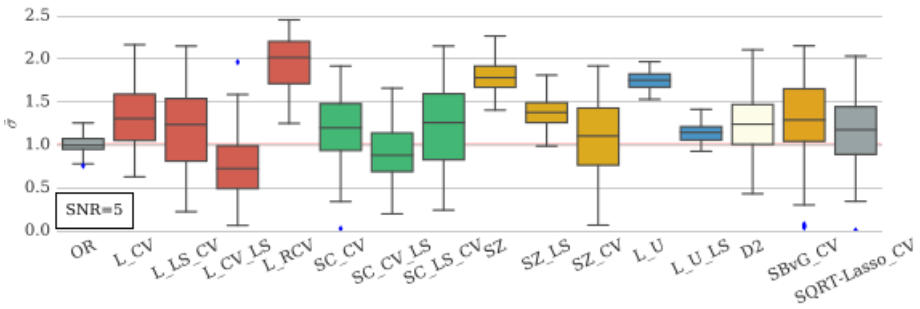


**Figure 5:** Performances sur le jeu de données synthétique

La performance par rapport au temps de calcul de ces estimateurs a aussi été calculée en utilisant ces mêmes données synthétiques:

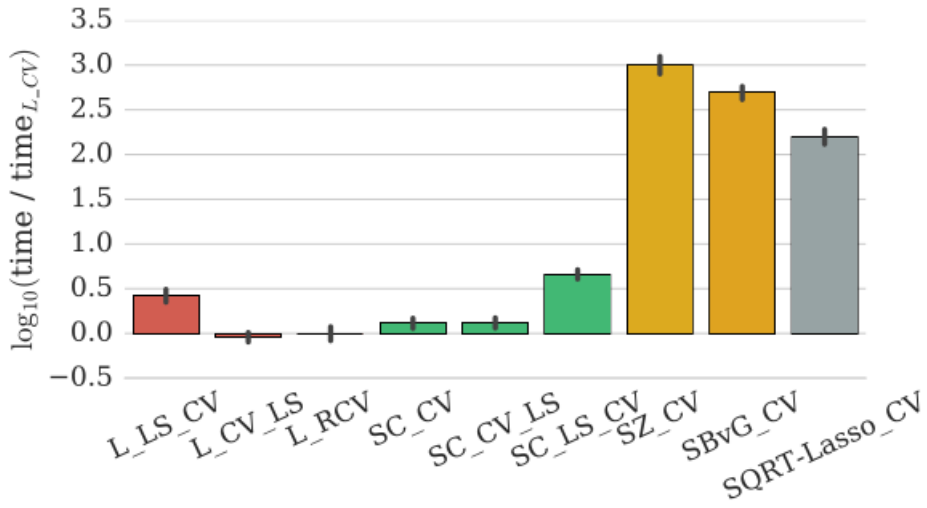


**Figure 6:** Temps de calcul des estimateurs, relativement au temps de calcul moyen d'un Lasso

Le temps pour atteindre la convergence, pour l'estimateur "Smoothed Concomitant Lasso", est illustré par la figure ci dessous. Elle montre bien l'utilité des différentes méthodes de vérification illustrés dans l'algorithme. Notamment la vérification du saut de dualité qui permet d'annuler des composantes de l'estimateur pour accélérer le calcul (l'élimination des composantes réduit la complexité des itérations). Le logarithme du saut du dualité illustre la distance initiale entre le point de départ de la déscente en coordonnées et la valeur de convergence de l'estimateur.

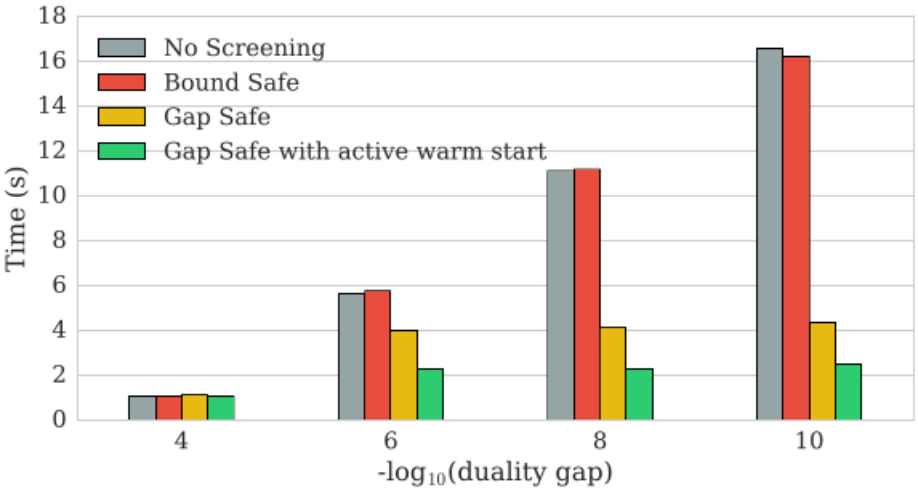


**Figure 7:** Impact des différentes méthodes d'optimisation de l'algorithme sur le temps de calcul

Finalement, on observe la distribution des valeurs optimates de $\lambda_{opt}$ des différents estimateurs pour différents niveaux de bruit $\sigma$ différents.

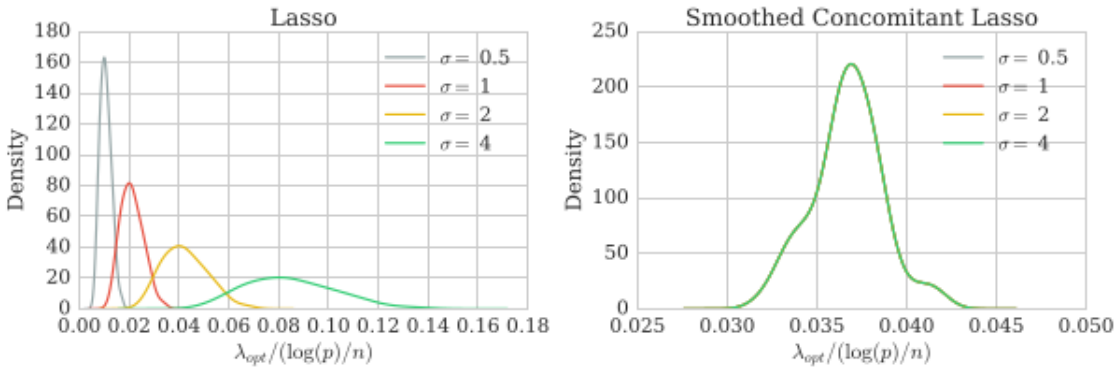


**Figure 8:** Distribution des lambda optimaux pour différents rapports de bruit pour les différents estimateurs étudiés

## 2.4 Résultats et observations:

D'après les auteurs de l'article, la méthode d'estimation "Smoothed Concomitant Lasso" propose une performance assez bonne tout au long des tests. Les méthodes de "gap screening" proposées dans l'algorithme permettent d'accélérer le temps de calcul pour les problèmes à haute complexité de départ. Un autre avantage de cet estimateur est la constance du $\lambda_{opt}$ qu'il estime par rapport au snr. La conclusion est que cet estimateur est en général aussi efficace en terme de temps de calcul et de précision qu'un estimateur LASSO usuel, tout en ayant l'avantage d'estimer le rapport de bruit.

# Conclusion générale

Au cours de ce séminaire, nous avons pu explorer les notions théoriques de l'estimation par méthode de LASSO en commencent tout d'abord par voir le contexte pratique de celles-ci, qui est l'estimation de température. Les problèmes à haute parcimonie étant très pertinents dans la statistique moderne. Nous avons observé les résultats de ces estimations directement sur l'exemple des prévisions météorologiques et nous avons conclu quant à la précision des estimations. La deuxième partie de ce projet nous a permis d'explorer un aspect plus algorithmique de l'estimation par LASSO en introduisant une autre construction de ce dernier qui prends en compte différents améliorations en terme de temps de calcul mais aussi de capacité à estimer simultanément le coefficient de corrélation et le rapport de bruits. Ce qui nous a montré que la théorie derrière ces méthodes d'estimation est toujours ouverte au progrès par petits incréments qui peuvent s'avérer utiles pour des problèmes particuliers.

# Bibliographie


[1] Eugene Ndiaye et al. “Efficient smoothed concomitant lasso estimation for high dimensional regression”. In: Journal of Physics: Conference Series. Vol. 904. 1. IOP Publishing. 2017, p. 012006.

[2] Art B. Owen. “A robust hybrid of lasso and ridge regression”. In: 2006.

[3] Stephen Reid, Robert Tibshirani, and Jerome H. Friedman. “A Study of Error Variance Estimation in Lasso Regression”. In: arXiv: Methodology (2013).

[4] Tingni Sun and Cun-Hui Zhang. “Scaled sparse linear regression”. In: Biometrika 99 (2011), pp. 879–898.