Lasso : Algorithms

Chargé du cours
Prof. Mustapha Rachdi

Université Grenoble Alpes
UFR SHS, BP. 47
38040 Grenoble cedex 09
France
Bureau provisoire : 07 au BSHM
e-mail : mustapha.rachdi@univ-grenoble-alpes.fr

Introduction

- Dans le cours précédent, nous avons présenté le lasso et nous avons dérivé les conditions nécessaires et suffisantes que $\widehat{\beta}$ doit satisfaire pour minimiser la fonction objective du lasso.
- Cependant, ces conditions nous permettent seulement de vérifier une solution. Ils ne nous aident pas nécessairement à trouver la solution en premier lieu
- Aujourd'hui, nous discuterons deux algorithmes de résolution pour $\widehat{\beta}$. Les algorithmes sont, bien sûr, une nécessité pratique mais apportent également un éclairage considérable sur la nature du lasso en tant que méthode statistique

Lasso vs. sélection ascendante : pénalisation L^0

- Comme nous l'avons vu précédemment, le lasso peut être considéré comme une version multivariée du seuillage souple
- La version multivariée du seuillage rigide est la pénalisation, $L0$, dans laquelle nous minimisons la fonction objective :

$$\frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_0$$

où $\|\beta\|_0 = \sum_j I_{(\beta_j \neq 0)}$

- Dans le cas orthonormé, la solution est donnée par

$$\widehat{\beta}_j = H(\widehat{\beta}_j^{\text{MCO}}, \sqrt{2\lambda})$$

- L'estimation de β, de cette manière, est équivalente à la sélection de sous-ensembles, et les critères de sélection de modèle tels que AIC et BIC sont simplement des cas spéciaux correspondant à différentes valeurs de λ

- Ainsi, le lasso peut être considéré comme une relaxation "souple" de la régression pénalisée l_0
- Cette relaxation a deux avantages importants :
  - Les estimations sont continues par rapport à λ et les données
  - La fonction objective du lasso est convexe
- Ces faits permettent d'optimiser très efficacement l'optimisation de la régression avec la L^1-pénalisation, comme nous le verrons. En comparaison, la régression L^0-pénalisée est impossible à calculer lorsque p est grand

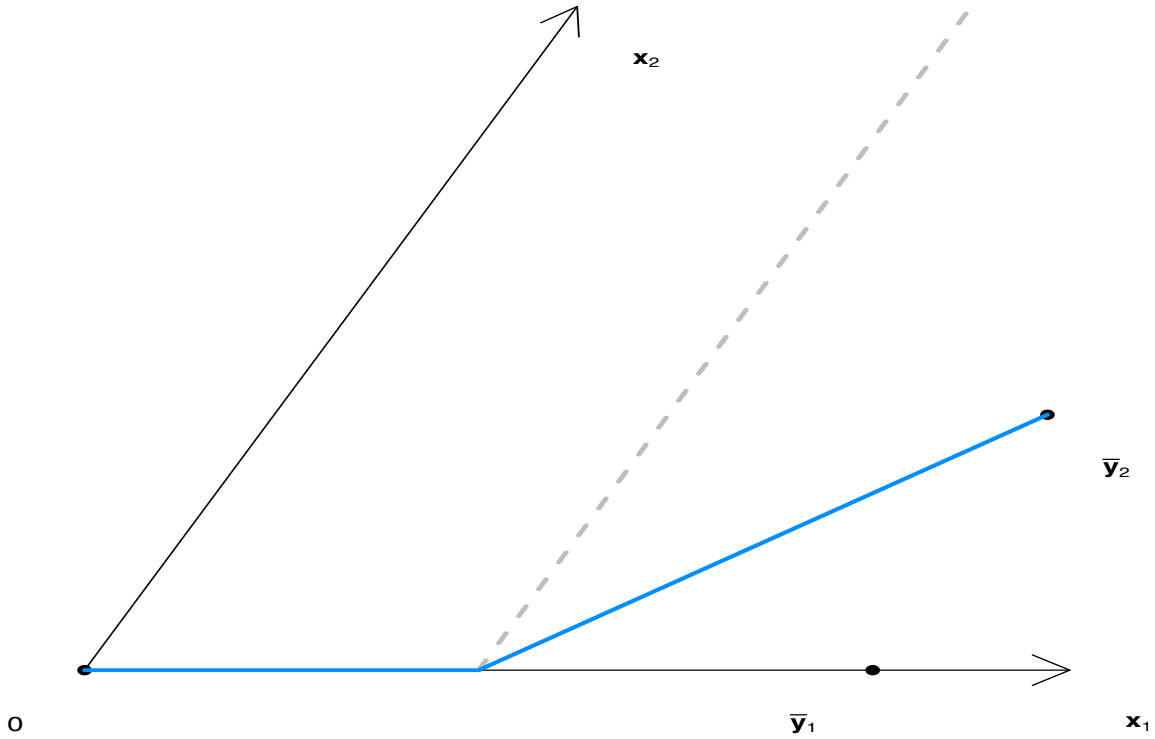
Lasso vs. la sélection ascendante

- Pour éviter la difficulté de trouver le meilleur sous-ensemble possible, une approche commune consiste à utiliser l'algorithme glouton/greedy connu sous le nom de sélection ascendante.
- Comme pour la sélection ascendante, le lasso permettra à plus de variables d'entrer dans le modèle lorsque λ est petit.
- Cependant, le lasso effectue une version continue de la sélection des variables et est moins enthousiaste à l'idée d'autoriser des variables à entrer dans le modèle.

Les chemins (paths) Lasso vs. la sélection ascendante

- Considérons les chemins des coefficients de la régression lasso et de la sélection ascendante (pénalisation respectives L^1 et L^0) lorsque nous abaissons λ, à partir de $\lambda_{\max}$ où $\widehat{\beta} = 0$
- Comme λ est abaissé en dessous de $\lambda_{\max}$, les deux approches trouvent le prédicteur le plus fortement corrélé à la réponse (supposons que x_j est ce prédicteur), et définissons $\widehat{\beta}_j \neq 0$:
  - Avec la sélection ascendante, l'estimation saute de $\widehat{\beta}_j = 0$ vers $\widehat{\beta}_j = {}^t x_j y / n$
  - La solution de lasso $\widehat{\beta}_j = 0$ va également dans cette direction, mais procède avec plus de prudence en avançant progressivement vers $\widehat{\beta}_j = {}^t x_j y / n$ quand nous diminuons λ

Chemins du Lasso vs. la sélection ascendante



Lasso vs. la sélection ascendante : Remarques

- La solution de lasso fonctionne de cette manière jusqu'à atteindre le point où un nouveau prédicteur, x_k, qui est corrélé, de façon égale, avec le résidu $r(\lambda) = y - X\widehat{\beta}(\lambda)$
- A partir de ce point, la solution lasso contiendra à la fois x_1 et x_2, et procédera dans la direction équiangulaire entre les deux prédicteurs
- Le lasso avance toujours dans une direction telle que chaque prédicteur actif (c'est-à-dire, avec $\widehat{\beta}_j \neq 0$) qui a la même corrélation avec le résidu $r(\lambda)$, ce qui peut également être observé à partir des conditions KKT
- La géométrie du lasso illustre clairement la "gourmandise" de la sélection ascendante
- En continuant le long du chemin allant de y à $\bar{y}_1$ après le point de corrélation "égale, la sélection ascendante continue à exclure x_2 du modèle, même lorsque x_2 est plus étroitement corrélé aux résidus que x_1
- Le lasso, quant à lui, autorise les prédicteurs qui sont les plus fortement corrélés avec les résidus à entrer dans le modèle, mais seulement de façon progressive, jusqu'à ce que le prochain prédicteur soit également utile pour expliquer la réponse (outcome)

LARS

- Ces informations géométriques ont été la clé pour développer le premier algorithme efficace de recherche des estimations lasso $\widehat{\beta}(\lambda)$
- L'approche, appelée régression par le moindre angle, ou l'algorithme *least angle regression* (LARS), offre un moyen élégant d'effectuer une estimation lasso.
- L'idée derrière l'algorithme est de
  1. Projeter les résidus sur les variables actives
  2. Calculer jusqu'où on peut avancer dans cette direction avant quune autre variable atteint le niveau de corrélation nécessaire avec les résidus

  puis on l'ajoute à l'ensemble des variables actives, et on répète (1) et (2), etc.

Rôle historique de LARS

- L'algorithme LARS a joué un rôle important dans l'histoire du lasso
- Avant LARS, l'estimation lasso était lente et très informatisée. Le LARS, en revanche, nécessite uniquement des calculs $O(np^2)$, du même ordre de grandeur que MCO.
- Néanmoins, le LARS n'est plus très utilisé
- Au lieu de cela, l'approche la plus populaire pour l'ajustement lasso et d'autres modèles de régression pénalisées consiste à utiliser des algorithmes alternatifs de descente de coordonnées (coordinate descent), moins beaux mais plus simple et plus flexible

Algorithme

- L'idée sous-jacente à la descente de coordonnées consiste simplement à optimiser une fonction cible par rapport à un seul paramètre à la fois, en parcourant tous les paramètres de manière itérative jusqu'à ce que la convergence soit atteinte.
- La descente de coordonnées est particulièrement adaptée aux problèmes, comme le lasso, qui ont une solution de forme simple dans le cadre unidimensonnel mais qui fait défaut dans des dimensions plus élevées

Algorithme : notation CD

- Considérons le problème de minimisation de Q par rapport à β_j, tout en traitant temporairement les autres coefficients de régression β_{-j} comme étant fixes :

$$Q(\beta_j|\beta_{-j}) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \sum_{k\neq j} x_{ij}\beta_k - x_{ij}\beta_j\right)^2 + \lambda|\beta_j| + \text{Constante}$$

- Soit

$$\widetilde{r}_{ij} = y_i - \sum_{k\neq j} x_{ik}\widetilde{\beta}_k \quad \text{et} \quad \widetilde{z}_j = n^{-1}\sum_{i=1}^{n} x_{ik}\widetilde{r}_{ij}$$

où $\{\widetilde{r}_{ij}\}_{i=1}^n$ sont les résidus partiels par rapport au $j$ème prédicteur, et $\widetilde{z}_j$ est l'estimateur MCO basé sur $\{\widetilde{r}_{ij}, x_{ij}\}_{i=1}^n$

Algorithm : CD algorithm

- Nous avons déjà résolu le problème de trouver une solution lasso unidimensionnelle. Soit $\widehat{\beta}_j$ le minimiseur de $Q(\beta_j|\widetilde{\beta}_{-j})$,

$$\widetilde{\beta}_j = S(\widetilde{z}_j|\lambda)$$

- Ceci suggère l'algorithme suivant :

Répéter
pour $j = 1, 2, ..., p$

$$\widetilde{z}_j = n^{-1}\sum_{i=1}^{n} x_{ij} r_i + \widetilde{\beta}_j^{(s)}$$

$$\widetilde{\beta}_j^{(s+1)} \longleftarrow S(\widetilde{z}_j|\lambda)$$

$$r_i \longleftarrow r_i - (\widetilde{\beta}_j^{(s+1)} - \widetilde{\beta}_j^{(s)})x_{ij} \text{ pour tout } i$$

jusqu'à la convergence

Algorithme : Remarques

- L'algorithme de descente de coordonnées a le potentiel d'être assez efficace, dans la mesure où ses trois opérations ne nécessitent que $O(2n)$ opérations (pas de factorisations de matrice compliquées, ni même de multiplication de matrice, seulement deux produits internes)
- Ainsi, une itération complète peut être complétée à un coût de calcul de $0(2np)$ opérations
- Ainsi, la descente des coordonnées est linéaire en n et p, et augmente proportionnellement en grandes dimensions. Sa performance reste encore meilleure que celle de LARS, bien qu'il faille noter que la descente en coordonnées nécessite un nombre inconnu d'itérations, alors que LARS se termine en un nombre connu d'étapes.

Algorithme : Convergence : Remarques

- L'analyse numérique de problèmes d'optimisation de la forme :

$$Q(\beta) = L(\beta) + P(\beta)$$

a montré que les algorithmes de descente en coordonnées convergent vers une solution des équations de vraisemblance pénalisées à condition que la fonction de perte $L(\beta)$ soit différentiable et que la fonction de pénalité $P_\lambda(\beta)$ est *séparable*, ce qui signifie qu'elle peut être écrite ainsi :

$$P_\lambda(\beta) = \sum_j P_\lambda(\beta_j)$$

- La régression linéaire pénalisée par lasso satisfait à ces deux critères

Algorithme : Convergence

- De plus, comme la fonction objective de lasso est une fonction convexe, la séquence des fonctions objectives$\{Q(\widetilde{\beta}^{(s)})\}$ converge vers le minimum global
- Cependant, la fonction objective du lasso n'est pas strictement convexe. Il peut donc y avoir plusieurs solutions
- Dans de telles situations, la descente de coordonnées convergera vers l'une de ces solutions, mais la solution vers laquelle elle converge est essentiellement arbitraire, car elle dépend de l'ordre des co-variables.

Optimisation en fonction du chemin/trajet : optimisation de la descente en coordonnées et du chemin

- Comme nous l'avons vu avec la régression ridge, nous sommes généralement intéressés par la détermination de $\widehat{\beta}$ pour un intervalle de valeurs de λ, obtenant ainsi le chemin du coefficient
- En appliquant l'algorithme de descente de coordonnées pour déterminer le chemin du lasso, une stratégie efficace consiste à calculer des solutions pour les valeurs décroissantes de λ, à partir de

$$\lambda_{\max} = \max_{1 \leq j \leq p} |{}^t x_j y|/n,$$

  le point auquel tous les coefficients sont 0 (nuls)
- En continuant sur une grille décroissante de valeurs λ, nous pouvons utiliser les solutions $\widehat{\beta}(\lambda_k)$ comme valeurs initiales lors de la résolution de $\widehat{\beta}(\lambda_{k+1})$

Optimisation de trajectoire : *démarrages à chaud*

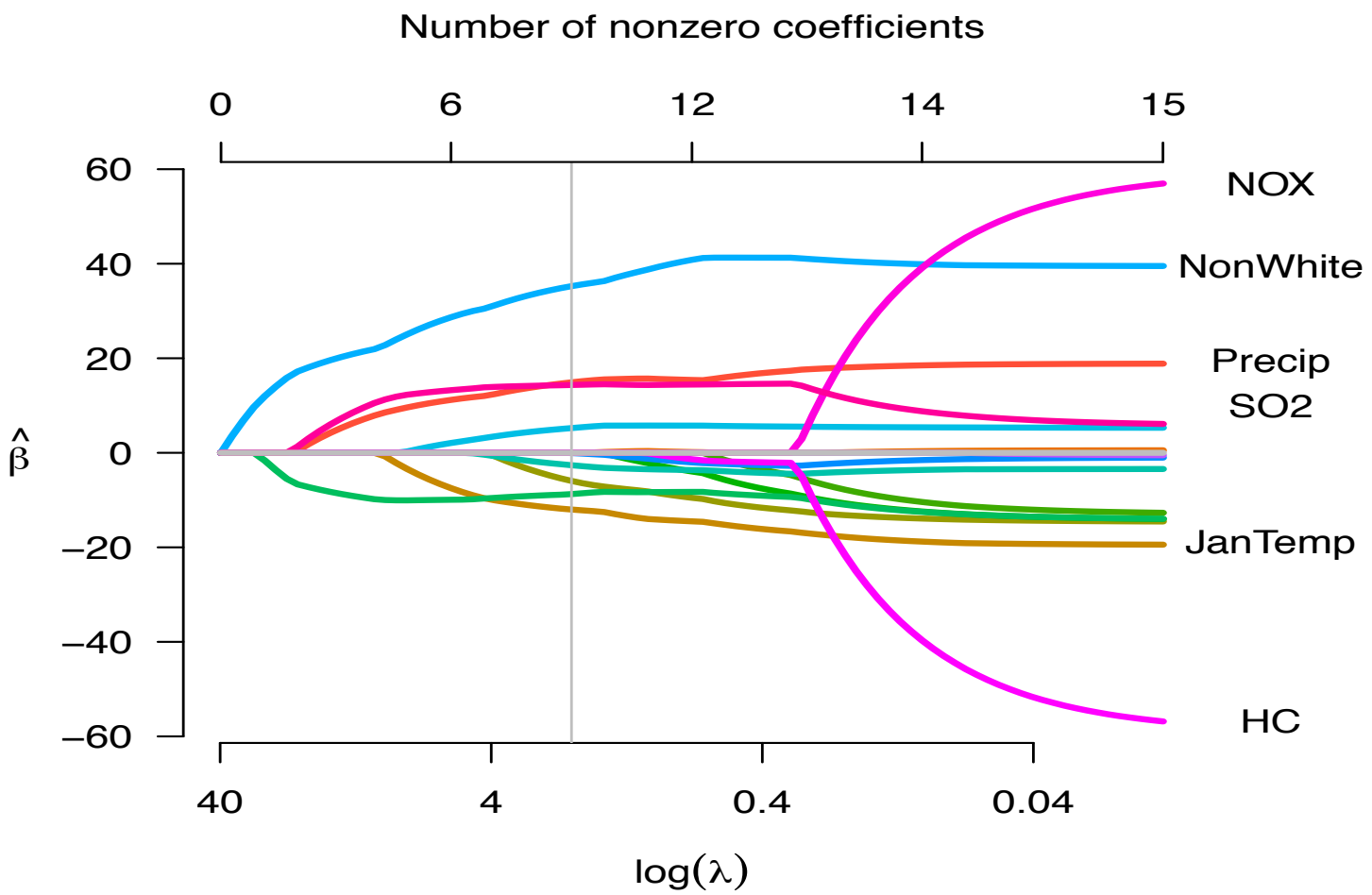
- Comme le chemin des coefficients est continu, cette opération fournit automatiquement de bonnes valeurs initiales pour la procédure d'optimisation itérative.
- Cette stratégie, connue sous le nom de *démarrage à chaud* (Warm starts), améliore considérablement l'efficacité de l'algorithme, car les valeurs initiales sont toujours assez proches de la solution finale
- On procède de cette manière jusqu'à une valeur minimale λ_{min}. Comme les solutions de lasso changent plus rapidement pour les valeurs petites/basses de λ, la grille de valeurs de λ est généralement choisie pour être uniformément espacée sur *l'échelle logarithmique* sur l'intervalle $[\lambda_{\text{max}}, \lambda_{\text{min}}]$.

Optimisation de trajectoire : glmnet

- Pour illustrer le chemin des coefficients du lasso, adaptons un modèle de lasso aux données de pollution que nous avons analysées précédemment au cours de la régression à l'aide de la régression ridge.
- L'algorithme de descente de coordonnées décrit dans cette section est implémenté dans le package `R glmnet`
- L'utilisation de base de `glmnet` est simple :

```
library(glmnet)
fit <- glmnet(X, y)
plot(fit)
```

Optimisation de trajectoires de lasso : Données de pollution



Optimisation de trajectoire : Remarques

- Comme dans le graphique correspondant pour la régression ridge,
  - Les estimations sont $\widehat{\beta} = 0$ (à gauche) et $\widehat{\beta} = \widehat{\beta}_{MCO}$ (à droite)
  - Les deux indiquent qu'il ne faut pas croire les fortes estimations des effets de MCO sur la pollution par les HC et les NOX
  - Les deux indiquent que le polluant ayant le plus grand effet sur la mortalité est le SO_2
- Cependant, le chemin de lasso est parcimonieux, avec les coefficients entrant dans le modèle un par un à mesure que λ diminue
- Par exemple, en $\lambda = 1.84$ (la valeur minimisant l'erreur de validation croisée), le modèle contient neuf variables - notamment, ceci ne comprend ni HC ni NOX : les variables avec les plus grands coefficients de régression MCO.

Optimisation de trajectoires : Remarques

- Une autre différence, plus subtile, est qu'avec le lasso, les coefficients deviennent plus grands plus rapidement qu'avec la régression ridge (c'est-à-dire qu'il y a une grande séparation entre les grands et les petits coefficients)