

Sommaire

1 Introduction : Apprentissage Statistique

- Formalisme de l'apprentissage supervisé
- Fléau de la dimension
- Modèles Linéaires

2 Sélection de variables par critères AIC/BIC et validation Croisée

3 Lasso

- Préambule : Ridge
- Définition du Lasso
- Lasso et Grande dimension
- Aspects Algorithmiques

4 Extensions du Lasso

- Lasso et Modèles linéaires Généralisés
- Lasso et Estimation Structurée

Notations vectorielles

Dans la suite, on notera l'échantillon D_1^n du modèle de régression linéaire sous la forme :

$$\mathbf{Y} = \mathbf{X}\beta^* + \varepsilon ,$$

$$\text{où } \mathbf{Y} = \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix}, \quad \mathbf{X}_{i,j} = X_i^{(j)}, \quad \beta^* = \begin{pmatrix} \beta_1^* \\ \dots \\ \beta_p^* \end{pmatrix} \text{ et } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

ATTENTION : Pour simplifier la présentation des méthodes, on supposera parfois que $\beta_0^* = 0$.

On peut facilement ajouter ce paramètre en ajoutant la colonne constante $\begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix}$ à la matrice de design \mathbf{X} . Dans les packages R décrits dans ce cours, le coefficient d'ordonnée à l'origine β_0^* est toujours estimé.

Critère des moindres carrés

Le modèle linéaire est généralement ajusté par le critère des moindres carrés. Si on note $l(\mathbf{y}, \mathbf{y}') = (\mathbf{y} - \mathbf{y}')^2$ la perte quadratique,

$$\hat{\beta} \in \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^n l(Y_i, \sum_{j=1}^p X_i^{(j)} \beta_j)$$

Expression alternative en notation matricielles

$$\hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2$$

Si $\mathbf{X}^T \mathbf{X}$ est inversible, alors $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$.

Pourquoi considérer des alternatives aux moindres carrés ?

Dans ce cours, nous étudierons des alternatives aux critères des moindres carrés (et plus généralement de minimisation du risque empirique) :

- ▶ **Pour améliorer le risque (diminuer la variance)** : en particulier lorsque $p > n$.
- ▶ **Pour l'interprétation des modèles** : en supprimant les covariables inutiles, c'est-à-dire en annulant les coefficients correspondants, on obtient un modèle qui s'interprète plus facilement.

Que ce soit pour augmenter la précision ou pour améliorer l'interprétabilité, nous allons chercher à sélectionner des variables.

Deux classes de méthodes

- ▶ **Selection d'un sous-ensemble.** Nous identifions un sous-ensemble des p prédicteurs pour lesquels nous pensons qu'ils sont en lien avec la réponse. Nous ajustons ensuite un modèle par moindres carrés sur le sous-ensemble réduit.
- ▶ **Régularisation.** Nous ajustons un modèle sur l'ensemble complet des p prédicteurs, mais les coefficients estimés sont tirés vers 0 par rapport à l'estimateur des moindres carrés. Cette méthode de **régularisation** réduit la variance, et peut aussi aider à sélectionner les variables.

Sélection de modèles

Objectif : Sélectionner un sous-ensemble de variables explicatives qui explique au mieux Y .
Soit $m \subset \{1, \dots, p\}$ un sous-ensemble d'indices. On note

$$\hat{\beta}_m \in \arg \min_{\beta, \text{supp}(\beta) \subset m} \|Y - X\beta\|_2^2,$$

l'estimateur des moindres carrés sur des paramètres β dont toutes les coordonnées en dehors de m sont fixés à zéro.

Il correspond à l'estimateur des moindres carrés dans le modèle linéaire dont seules les variables $X^{(j)}$, $j \in m$ ont été gardée.

Le problème de la sélection de modèle est le suivant. Etant donnée une collection $\mathcal{M} = \{m_1, \dots, m_r\}$ de modèles, on veut sélectionner le modèle $m^* \in \mathcal{M}$ tel que

$$\mathbb{E}_{(Y,X)} \left[\left(Y - \sum_{j=1}^p X^{(j)} (\hat{\beta}_{m^*})_j \right)^2 \right] = R(\hat{f}_{m^*}) \text{ est le plus petit possible,}$$

où $\hat{f}_m(X) = \sum_{j=1}^p X^{(j)} (\hat{\beta}_m)_j$.

Deux exemples de problèmes de sélection de modèles

- 1 **Sélection ordonnée.** Supposons qu'il existe un ordre naturel sur les covariables $X^{(1)}, \dots, X^{(p)}$.

Exemple : régression polynomiale $X^{(1)} = X, X^{(2)} = X^2, \dots, X^{(k)} = X^k$.

L'objectif est de sélectionner le "meilleur" degré du polynôme pour prédire Y .

$$\mathcal{M} := \{\{1\}, \{1, 2\}, \{1, 2, 3\}, \dots, \}$$

- 2 **Sélection complète.** On veut choisir les "meilleurs" covariables $X^{(1)}, \dots, X^{(p)}$.

$\mathcal{M} = \mathcal{P}(\{1, \dots, p\})$, l'ensemble des parties de $\{1, \dots, p\}$.

Comment sélectionner un bon modèle

- ▶ Sélectionner le modèle qui minimise l'erreur d'entraînement

$$\widehat{R}_n(\widehat{f}_m) = \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\widehat{\beta}_m\|_2^2$$

est une mauvaise idée. Pourquoi ?

- ▶ L'objectif étant de choisir un modèle dont le risque $R(\widehat{f}_m)$ est le plus petit possible, il est naturel de vouloir estimer ce risque pour chaque \widehat{f}_m , $m \in \mathcal{M}$. Deux approches s'offrent à nous :
 - 1 Estimer le risque en *ajustant* l'erreur d'entraînement pour tenir compte du biais dû au sur-apprentissage
 - 2 Estimer *directement* l'erreur de test, par une approche de validation ou une approche de validation croisée

Pénalisation : AIC (C_p) et BIC

- ▶ Ces techniques corrigent l'erreur d'entraînement par la taille du modèle, et peuvent être utilisées pour sélectionner des modèles de dimension différentes.

$$\widehat{m} \in \arg \min_{m \in \mathcal{M}} \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\widehat{\beta}_m\|_2^2 + \text{pen}(m)$$

où $\text{pen} : \mathcal{M} \rightarrow \mathbb{R}^+$ est une pénalité qui va pénaliser les plus grands modèles. Dans la suite, on va voir deux (ou trois) fonctions de pénalités différentes.

$$\begin{aligned} \text{pen}_{\text{AIC}}(m) = \text{pen}_{C_p}(m) &= 2\widehat{\sigma}_m^2 \frac{|m|}{n} \\ \text{pen}_{\text{BIC}}(m) &= \log(n)\widehat{\sigma}_m^2 \frac{|m|}{n}, \end{aligned}$$

où $\widehat{\sigma}_m^2 = \|\mathbf{Y} - \mathbf{X}\widehat{\beta}_m\|_2^2/n$.

Remarque sur les pénalités AIC et BIC

$$\begin{aligned}\text{pen}_{\text{AIC}}(m) &= 2\hat{\sigma}_m^2 \frac{|m|}{n} \\ \text{pen}_{\text{BIC}}(m) &= \log(n)\hat{\sigma}_m^2 \frac{|m|}{n},\end{aligned}$$

- ▶ Comme les C_p , BIC a tendance à être petit lorsque le risque est petit, et on choisit donc généralement le modèle qui a la plus petite valeur de BIC.
- ▶ Notons que BIC remplace le $2\hat{\sigma}^2$ utilisé par C_p par un terme $\log(n)\hat{\sigma}^2$ où n est le nombre d'observations.
- ▶ Puisque $\log(n) > 2$ dès que $n > 7$, le critère BIC pénalise plus les modèles de grandes dimensions. Les modèles choisis avec ce critère seront donc de dimension plus petite.

Comparaison de ces critères

- ▶ AIC sont des critères qui réalisent un compromis biais-variance. Ils sont donc indiqués pour choisir un modèle que l'on souhaite utiliser pour prédire.
- ▶ BIC pénalise plus les modèles de grandes dimensions. C'est le seul critère à être consistant (i.e., à sélectionner le vrai modèle $\text{supp}(\beta^*)$ avec probabilité tendant vers 1 lorsque $n \rightarrow \infty$)
- ▶ BIC étant plus sélectif, on doit le préférer si l'on souhaite un modèle explicatif.
- ▶ Lorsque la taille de la base d'apprentissage est grande, préférer BIC (AIC fournit des modèles de trop grandes dimensions)
- ▶ **ATTENTION** : Lorsque p est grand (au moins de l'ordre de n), les pénalités BIC et AIC peuvent s'avérer trop petites et il faut recourir à d'autres pénalités.

Retour sur la sélection complète de variables

- 1 Pour chaque valeur de k entre 1 et p :
 - ▶ Choisir le meilleur parmi ces $\binom{p}{k}$ modèles et le noter \widehat{m}_k .
- 2 Choisir le meilleur modèle parmi $\widehat{m}_1, \dots, \widehat{m}_p$ en utilisant la validation croisée, ou AIC, ou BIC.

Cet algorithme est équivalent à la méthode présentée précédemment.

Sélection pas à pas

- ▶ Pour des raisons de calcul, la sélection du meilleur sous-ensemble ne peut pas être appliquée quand p est grand. *Pourquoi ?*
- ▶ Les méthodes *pas à pas*, qui n'explorent qu'une sous-partie de l'ensemble de tous les modèles possibles sont plus attirantes pour sélectionner le meilleur sous-ensemble.

Sélection pas à pas progressive

- ▶ La sélection progressive commence par le modèle nul et ajoute progressivement des prédicteurs au modèle, un par un, jusqu'à ce que l'on utilise tous les prédicteurs.
- ▶ En particulier, à chaque étape, la variable qui conduit à la meilleure amélioration du modèle est ajoutée.

Sélection pas à pas progressive

- 1 Noter \widehat{m}_0 le **modèle nul**, qui ne contient aucun prédicteurs. Ce modèle prédit simplement la réponse Y avec $\mathbb{E}(Y)$, ou plutôt la moyenne empirique \bar{Y} .
- 2 Pour chaque valeur de k entre 0 et $p - 1$:
 - 1 Considérer tous les $(p - k)$ modèles qui consistent à ajouter un prédicteur à \widehat{m}_k .
 - 2 Choisir le meilleur parmi ces $(p - k)$ modèles et noter le \widehat{m}_{k+1} . Ici, le **meilleur** modèle est celui qui minimise le critère des moindres carrés.
- 3 Choisir le meilleur modèle parmi $\widehat{m}_0, \widehat{m}_1, \dots, \widehat{m}_p$ en utilisant la validation croisée, ou les AIC ou BIC.

Sélection pas à pas progressive (suite)

- ▶ L'avantage en terme de temps de calcul par rapport à la méthode exhaustive du meilleur sous-ensemble est claire.
- ▶ Rien ne garantit de trouver le meilleur modèle possible parmi les 2^p modèles.

Exemple : jeu de données crédit

Nb covar	Meilleur sous-ensemble	Sélection progressive pas à pas
1	rating	rating
2	rating, income	rating, income
3	rating, income, student	rating, income, student
4	cards, income, student, limit	rating, income, student, limit

Les trois premiers modèles sont identiques, mais le dernier est différent de ce qu'on trouve par sélection complète.

Sélection pas à pas rétrograde

- ▶ Comme la sélection pas à pas progressive, la *sélection pas à pas rétrograde* propose une méthode efficace alternative au meilleur sous-ensemble.
- ▶ Cependant, contrairement à la méthode progressive, elle commence par le modèle complet, ajusté par moindres carrés, contenant les p prédicteurs, et les supprime un à un.

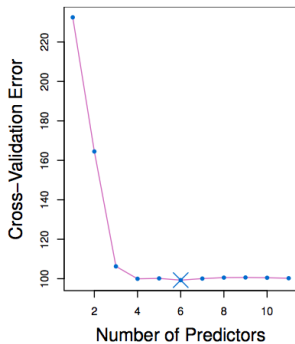
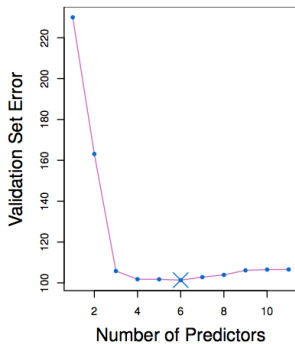
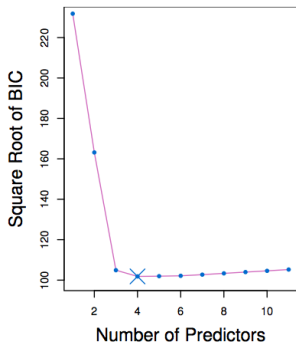
Sélection pas à pas rétrograde : détail

- 1 Noter \widehat{m}_p le **modèle complet**, qui contient tous les p prédicteurs
- 2 Pour chaque valeur de k allant de p à 1 :
 - 1 Considérer tous les k modèles qui consistent à supprimer un prédicteur à \widehat{m}_k .
 - 2 Choisir le meilleur parmi ces k modèles et noter le \widehat{m}_{k-1} . Le **meilleur** modèle est celui qui minimise le critère des moindres carrés.
- 3 Choisir le meilleur modèle parmi $\widehat{m}_0, \widehat{m}_1, \dots, \widehat{m}_p$ en utilisant la validation croisée, AIC, ou BIC.

Sélection pas à pas rétrograde (suite)

- ▶ Comme la méthode progressive, la méthode rétrograde ne visite que $1 + p(p + 1)/2$ modèles, et peut donc être appliquée dans des contextes où p est trop grand pour la méthode exhaustive.
- ▶ Comme la méthode progressive, la méthode rétrograde ne garantit pas de trouver le *meilleur* modèle.
- ▶ La méthode rétrograde suppose que *la taille de l'échantillon n est plus grande que le nombre de prédicteurs p* (pour pouvoir ajuster le modèle complet). En revanche, la méthode progressive peut s'arrêter à n covariables si $p > n$ et peut donc être utilisée dans un contexte plus large.

Exemple : jeu de données crédit



Sommaire

1 Introduction : Apprentissage Statistique

- Formalisme de l'apprentissage supervisé
- Fléau de la dimension
- Modèles Linéaires

2 Sélection de variables par critères AIC/BIC et validation Croisée

3 Lasso

- **Préambule : Ridge**
- Définition du Lasso
- Lasso et Grande dimension
- Aspects Algorithmiques

4 Extensions du Lasso

- Lasso et Modèles linéaires Généralisés
- Lasso et Estimation Structurée

Méthode de Régularisation

Régression ridge et Lasso

- ▶ Les méthodes précédentes de choix de sous-ensembles utilisent les moindres carrés pour ajuster chacun des modèles en compétition.
- ▶ Alternativement, on peut ajuster un modèle contenant toutes les p covariables en utilisant une technique qui *contraint* ou *régularise* les estimations des coefficients, ou de façon équivalente, pousse les coefficients vers 0.

Préambule

Les variables explicatives observées $x^{(j)} = (x_1^{(j)}, \dots, x_n^{(j)})$ sont centrées et standardisées (ie $\|x^{(j)}\|_2^2/n = 1$) et on suppose que $\beta_0^* = 0$ et $\bar{Y} = 0$ ce qui revient à estimer β_0^* par \bar{Y} et à remplacer Y_i par $Y_i - \bar{Y}$.

Remarque : Une fois les paramètres ajustés pour les variables centrées et standardisées, on peut facilement revenir au modèle initial en transformant linéairement les paramètres.

Pénalisation l_q

Un estimateur par minimisation du risque empirique régularisé (pour la perte quadratique) est dans le cadre de la régression linéaire défini par

$$\hat{\beta}_\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_q^q$$

λ étant un paramètre positif, appelé paramètre de régularisation.

- ▶ $q = 2 \rightsquigarrow$ régression ridge
- ▶ $q = 1 \rightsquigarrow$ régression lasso

Régression ridge

L'estimateur est défini par

$$\hat{\beta}_{\lambda}^{\text{ridge}} \in \arg \min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2$$

Proposition

- ▶ Minimiser $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2$ en $\beta \in \mathbb{R}^p$ est équivalent à minimiser $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2$ sous une contrainte de la forme $\|\beta\|_2^2 \leq r(\lambda)$.
- ▶ La matrice $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})$ est toujours définie positive, donc inversible et $\hat{\beta}_{\lambda}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$.

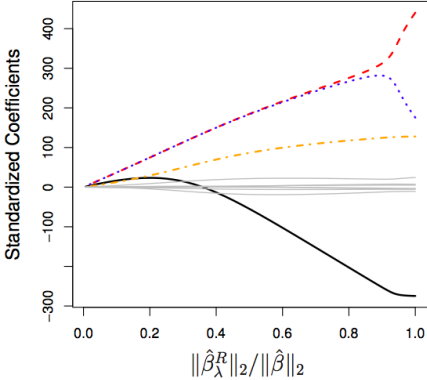
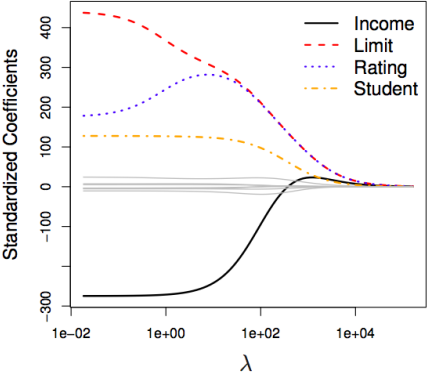
Remarque : L'estimateur $\hat{\beta}_{\lambda}^{\text{ridge}}$ est biaisé mais sa variance est plus faible que celle de l'estimateur des moindres carrés.

Rôle et ajustement du paramètre de régularisation

- ▶ Lorsque $\lambda = 0$, $\hat{\beta}_\lambda^{\text{ridge}}$ est l'estimateur des moindres carrés.
- ▶ Lorsque $\lambda \rightarrow \infty$, $\hat{\beta}_\lambda^{\text{ridge}}$ tend vers 0
- ▶ Lorsque λ augmente, le biais de $\hat{\beta}_\lambda^{\text{ridge}}$ a tendance à augmenter et la variance à diminuer
⇒ Recherche d'un compromis

~> Choix usuel de λ par validation croisée V fold sur une grille finie de valeur de $\lambda > 0$.

Exemple jeu de données crédit

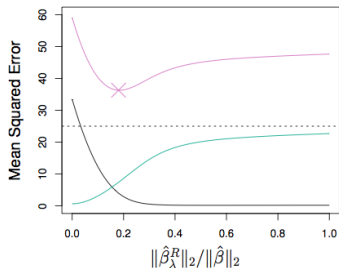
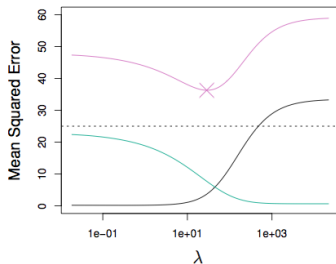


Commentaires

- ▶ À gauche, chaque courbe correspond à l'estimation des coefficients par régression ridge pour l'une des 10 variables, représentée en fonction de λ .
- ▶ À droite, l'axe des abscisses est maintenant le rapport entre la norme quadratique des coefficients estimés par régression ridge et les coefficients estimés par moindres carrés.

Pour la régression ridge ?

Compromis biais-variance



Données simulées : $n = 50$, $p = 45$, tous de coefficients non nuls. Biais au carré (en noir), variance (en vert) et erreur de test quadratique (en violet) pour la régression ridge. Droite horizontale : erreur minimale.

Sommaire

1 Introduction : Apprentissage Statistique

- Formalisme de l'apprentissage supervisé
- Fléau de la dimension
- Modèles Linéaires

2 Sélection de variables par critères AIC/BIC et validation Croisée

3 Lasso

- Préambule : Ridge
- **Définition du Lasso**
- Lasso et Grande dimension
- Aspects Algorithmiques

4 Extensions du Lasso

- Lasso et Modèles linéaires Généralisés
- Lasso et Estimation Structurée

La Régression Lasso

- ▶ La régression ridge a un inconvénient évident : contrairement à la sélection de variable, la régression ridge inclut tous les prédicteurs dans le modèle final.

L'estimateur LASSO (Least Absolute Selection and Shrinkage Operator) est défini pour $\lambda > 0$ par

$$\hat{\beta}_\lambda^{\text{lasso}} \in \arg \min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1$$

La fonction $\mathcal{L} : \beta \mapsto \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1$ est convexe, non différentiable. La solution du problème peut ne pas être unique.

Proposition

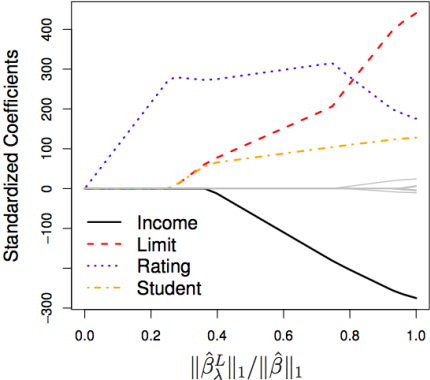
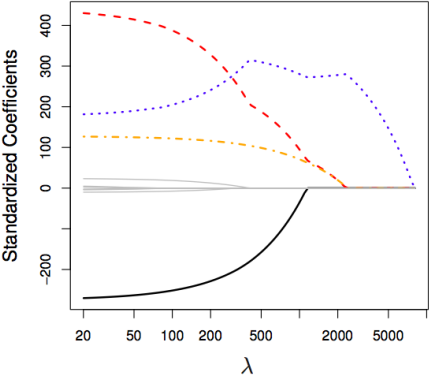
Minimiser $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1$ en $\beta \in \mathbb{R}^p$ est équivalent à minimiser $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2$ sous une contrainte de la forme $\|\beta\|_1 \leq R_\lambda$ pour une certaine quantité R_λ .

Preuve : Lagrangien

Le Lasso (suite)

- ▶ Comme pour la régression ridge, le Lasso tire les estimations des coefficients vers 0.
- ▶ Cependant, dans le cas du Lasso, la pénalité ℓ^1 a pour effet de forcer certains coefficients à s'annuler lorsque λ est suffisamment grand.
- ▶ Donc, le Lasso permet de faire de la *sélection de variable*.
- ▶ On parle de modèle creux (sparse), c'est-à-dire de modèles qui n'impliquent qu'un sous ensemble des variables.
- ▶ Comme pour la régression ridge, choisir une bonne valeur de λ est critique. Procéder par validation croisée.

Exemple : jeu de données crédit



Qu'est qui fait marcher le Lasso ?

Avec les multiplicateurs de Lagrange, on peut voir

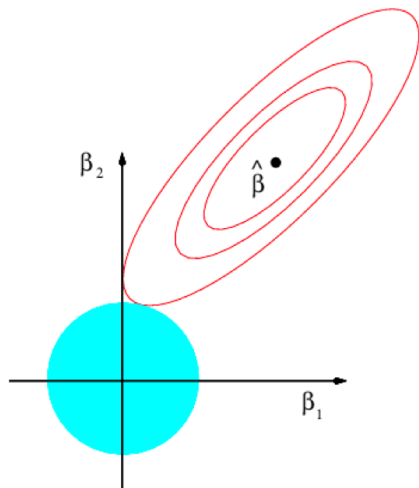
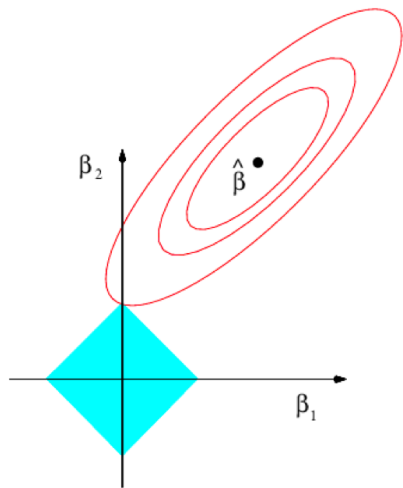
- ▶ La régression ridge comme

$$\text{minimise } \sum_{i=1}^n \left(Y_i - \sum_{j=1}^p \beta_j X_i^{(j)} \right)^2 \text{ sous la contrainte } \sum_{j=1}^p \beta_j^2 \leq s$$

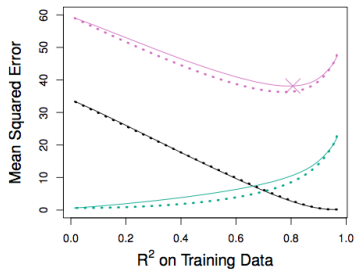
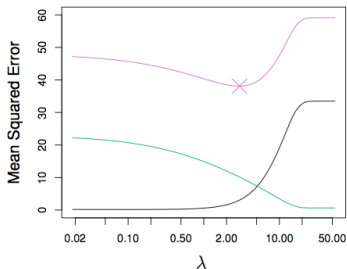
- ▶ Le Lasso comme

$$\text{minimise } \sum_{i=1}^n \left(Y_i - \sum_{j=1}^p \beta_j X_i^{(j)} \right)^2 \text{ sous la contrainte } \sum_{j=1}^p |\beta_j| \leq s$$

Le Lasso en image



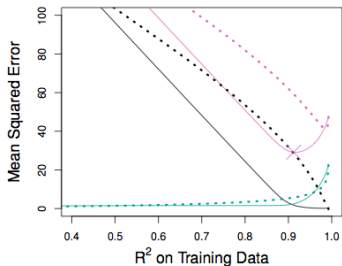
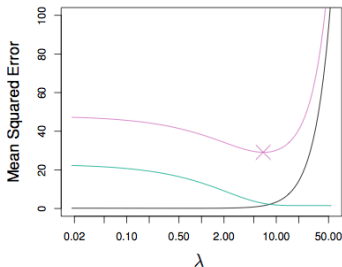
Comparaison du Lasso et de la régression ridge



À gauche, biais au carré (noir), variance (en vert) et erreur quadratique de test (violet) pour le Lasso sur données simulées.

À droite, comparaison du biais au carré, de la variance et de l'erreur de test quadratique pour le Lasso (traits plains) et la régression ridge (pointillés)

Comparaison du Lasso et de la régression ridge (suite)



À gauche, biais au carré (noir), variance (en vert) et erreur quadratique de test (violet) pour le Lasso sur données simulées (où seulement deux prédicteurs sont influents).

À droite, comparaison du biais au carré, de la variance et de l'erreur de test quadratique pour le Lasso (traits plats) et la régression ridge (pointillés)

Conclusions

- ▶ Ces deux exemples montrent qu'il n'y a pas de meilleur choix universel entre la régression ridge et le Lasso.
- ▶ En général, on s'attend à ce que le Lasso se comporte mieux lorsque la réponse est une fonction d'un nombre relativement faible de prédicteurs.
- ▶ Cependant, le nombre de prédicteurs reliés à la réponse n'est jamais connu *a priori* dans des cas concrets.
- ▶ Une technique comme la validation croisée permet de déterminer quelle est la meilleure approche.

Sommaire

1 Introduction : Apprentissage Statistique

- Formalisme de l'apprentissage supervisé
- Fléau de la dimension
- Modèles Linéaires

2 Sélection de variables par critères AIC/BIC et validation Croisée

3 Lasso

- Préambule : Ridge
- Définition du Lasso
- Lasso et Grande dimension
- **Aspects Algorithmiques**

4 Extensions du Lasso

- Lasso et Modèles linéaires Généralisés
- Lasso et Estimation Structurée

Coordinate descent

Proposition (Condition d'optimalité du premier ordre)

$\hat{\beta}_\lambda^{\text{lasso}}$ vérifie $\mathbf{X}^T \mathbf{X} \hat{\beta}_\lambda^{\text{lasso}} = \mathbf{X}^T \mathbf{Y} - \lambda \hat{\mathbf{Z}}/2$ avec $\hat{Z}_j \in [-1, 1]$ et $\hat{Z}_j = \text{signe}([\hat{\beta}_\lambda^{\text{lasso}}]_j)$ si $[\hat{\beta}_\lambda^{\text{lasso}}]_j \neq 0$.

Pas de solution explicite.

Une approche pour calculer l'estimateur lasso : la descente par coordonnées.

Proposition

La fonction $\beta_j \mapsto \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1$ est minimum en $\beta_j = R_j (1 - \lambda/(2|R_j|))_+ / n$ avec $R_j = (\mathbf{x}^{(j)})^T (\mathbf{Y} - \sum_{k \neq j} \beta_k \mathbf{x}^{(k)})$.

Cyclic Coordinate descent pour le lasso

Algorithme de Coordinate descent

```
input : X, Y,  $\lambda > 0$ 
begin
  Initialiser  $\beta = \beta_{in}$ ;
  while  $\beta$  n'a pas convergé do
    for  $j = 1, \dots, p$  do
      Calculer  $R_j = (x^{(j)})^T (Y - \sum_{k \neq j} \beta_k x^{(k)})$ ;
      Calculer  $\beta_j = R_j (1 - \lambda / (2|R_j|))_+ / n$ 
    end
  end
  output:  $\beta$ 
end
```

Sommaire

1 Introduction : Apprentissage Statistique

- Formalisme de l'apprentissage supervisé
- Fléau de la dimension
- Modèles Linéaires

2 Sélection de variables par critères AIC/BIC et validation Croisée

3 Lasso

- Préambule : Ridge
- Définition du Lasso
- Lasso et Grande dimension
- Aspects Algorithmiques

4 Extensions du Lasso

- Lasso et Modèles linéaires Généralisés
- **Lasso et Estimation Structurée**

Jusqu'ici on utilisait l'information a priori que β^* était potentiellement parcimonieux (la plupart de ces coefficients sont nuls).

En modifiant la pénalité $\|\beta\|_1$, on peut prendre en compte d'autres informations a priori.

Group Lasso

Cadre : Les variables $X^{(j)}$, $j = 1, \dots, p$ sont partitionnées en q groupes G_1, \dots, G_q de variables cohérentes.

Exemple pour des IRMf : Groupes G_1, \dots, G_q correspondent à des petite régions du cerveaux.

Hypothèse structurelle : Les coefficients β_i^* dans un même groupe G_j sont soit tous nul, soit tous non nuls.

Estimateur Group Lasso :

$$\hat{\beta} \in \arg \min_{\beta} \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^q \|\beta_{G_j}\|_2,$$

où β_{G_j} est la restriction de β aux variables dans G_j .

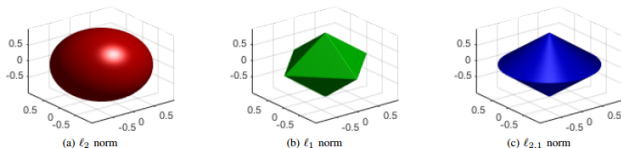


Figure 3. Isosurface for three different regularization terms, with $\mu_\lambda = 1$. (a) Standard squared ℓ_2 norm. (b) ℓ_1 norm enforcing sparsity. (c) $\ell_{2,1}$ norm applied to the groups $\{1,2\}$ and $\{3\}$ (without considering the scaling factors).

Fused Lasso

Cadre : Les variables $X^{(j)}$, $j = 1, \dots, p$ sont naturellement ordonnées.

Exemple une série temporelle ($p = n$) :

$$Y_t = \beta_t^* + \varepsilon_t, \quad t = 1, \dots, n$$

Hypothèse structurelle : Les coefficients β_t^* sont constants par morceaux
 \leadsto i.e. $\beta_t^* - \beta_{t-1}^*$ est souvent nul.

Estimateur Fused Lasso :

$$\hat{\beta} \in \arg \min_{\beta} \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^p \|\beta_j - \beta_{j+1}\|_1,$$

Bilan

- ▶ Les méthodes de sélection de modèles sont essentielles pour l'analyse de données, et l'apprentissage statistique, en particulier avec de gros jeu de données contenant de nombreux prédicteurs.
- ▶ En grande dimension, la prise en compte de structure (parcimonie, parcimonie par groupe,...) dans les paramètres joue un rôle central.