Sélection de modèles et régularisation

Joseph Salmon, Nicolas Verzelen

INRA / Université de Montpellier

Sélection de modèles linéaires et régularisation

- Rappelons que le modèle linéaire cherche à expliquer Y grâce à un modèle de la forme

$$\beta_0^* + \beta_1^* X^{(1)} + \cdots + \beta_p^* X^{(p)}$$

- Nous verrons dans la suite comment dépasser ce cas linéaire en construisant un modèle additif, mais non linéaire
- Certaines approches décrites dans ce chapitre s'étendent simplement au modèle de régression logistique

$$\text{logit}\left(\mathbb{P}\left[Y = 1 | (X^{(1)}, \ldots, X^{(p)})\right]\right) = \beta_0^* + \beta_1^* X^{(1)} + \cdots + \beta_p^* X^{(p)}$$

Notations vectorielles

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

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

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

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.

Défendons les modèles linéaires

- Malgré leur simplicité, le modèle linéaire a des avantages en termes d'**interprétabilité** et souvent il fournit de bon **performances prédictives**

Critère des moindres carrés

Nous avons vu que le modèle linéaire est généralement ajusté par le critère des moindres carré. Si on note $l(y, y') = (y - y')^2$ la perte quadratique,

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

Expression alternative en notation matricielles

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

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

Critères des moindres carrés et minimisation du risque empirique

Plus généralement, le critères moindres carrés fait partie de la famille des méthodes d'ajustement par **minimisation du risque empirique**. Soit $F \subset \mathcal{F}$ une collection de règle de prédictions

$$\widehat{f} \in \arg\min_{f \in F} \widehat{R}_n(F) \ ,$$

où $\widehat{R}_n(F) = \frac{1}{n} \sum_{i=1}^{n} l(Y_i, f(X_i))$.

Dans cette partie, nous étudierons des alternatives aux critères des moindres carrés (et plus généralement de minimisation du risque empirique).

Pourquoi considérer des alternatives aux moindres carrés ?

- **Pour améliorer la précision** : en particulier lorsque $p > n$ ou pour contrôler la variance.
- **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.

Nous présenterons des méthodes pour choisir les variables automatiquement.

Trois (ou 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 un estimateur des moindres carrés. Cette méthode de **régularisation** réduit la variance, et peut aussi aider à sélectionner les variables.

  Les deux premières approches rentrent dans le cadre de la minimisation du risque empirique pénalisée.
- **Réduction de la dimension**. Nous projetons les p prédicteurs dans un espace de dimension M, où $M < p$. Pour cela, nous devons calculer M différentes **combinaisons linéaires** ou **projections** des covariables. Ensuite, ces M projections sont utilisés comme prédicteurs pour ajuster un modèle de régression linéaire par moindres carrés.

Plan

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

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

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

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

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

Deux exemples de problèmes de sélection de modèles

1. **Sélection ordonnée**. Supposons qu'il existe un ordre naturel sur les covariables $X^{(1)}$, ... $X^{(p)}$.
   Exemple : régression polynomiale $X^{(1)} = X$, $X^{(2)} = X^2$, ..., $X^{(k)} = X^k$.
   L'objectif est de sélectionner le "meilleur" degré du polynôme pour prédire Y.
   $\mathcal{M} := \{\{1\}, \{1,2\}, \{1,2,3\}, \ldots,\}$
2. **Sélection compléte**. On veut choisir les "meilleurs" covariables $X^{(1)}, \ldots X^{(p)}$.
   $\mathcal{M} = \mathcal{P}(\{1, \ldots, p\})$, l'ensembles parties de $\{1, \ldots, p\}$.

Comment sélectionner un bon modèle

- Sélectionner le modèle qui minimise
$$\widehat{R}_n(\widehat{f}_m) = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\widehat{\beta}_m\|_2^2$$
est une mauvaisee idée. Pourquoi ?
- L'objectif étant de choisir un modèle dont le risque $R(\widehat{f}_m)$ est le plus petit possible, il 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'entrainement 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 (voir chapitre correspondant)

Pénalisation : AIC (C_P) et BIC

- Ces techniques corrigent l'erreur d'entrainement 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 $\text{pen} : \mathcal{M} \to \mathbb{R}^+$ est une pénalité qui va pénaliser les plus grands modèles.
Dans la suite, on va voir deux (ou trois) fonctions de pénalités différentes.

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

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

Justification de ces pénalités : heuristique de Mallows

Sortons un peu du cadre d'apprentissage statistique et considérons et le modèle de régression linéaire :

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

où $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ et $\mathbf{X}$ est supposé **déterministe**.
Considèrons le critère suivant

$$Crit_{C_p}(m) := \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\widehat{\beta}_m\|_2^2 + 2\frac{\sigma^2}{n}|m|$$

Notons β_m la meilleure approximation de β^* dans m :
$\beta_m \in \arg\min_{\beta,\ \mathrm{supp}(\beta)\subset m} \|\mathbf{X}\beta - \mathbf{X}\beta^*\|_2^2$.

Heuristique de Mallows (suite)

Proposition

Supposons que $X_m^T X_m$ est inversible. Alors,

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{Y} - \mathbf{X}\widehat{\beta}_m\|_2^2\big] &= \|\mathbf{X}\beta_m - \mathbf{X}\beta^*\|_n^2 + \sigma^2(n - |m|)\\ \mathbb{E}\big[\|\mathbf{X}\beta^* - \mathbf{X}\widehat{\beta}_m\|_2^2\big] &= \|\mathbf{X}\beta_m - \mathbf{X}\beta^*\|_n^2 + \sigma^2|m|\end{aligned}$$

Donc $\mathbb{E}[Crit_{C_p}(m)] = \mathbb{E}\big[\|\mathbf{X}\beta^* - \mathbf{X}\widehat{\beta}_m\|_2^2\big] + n\sigma^2$

La C_p de Mallows est un estimateur sans-biais du risque (à design fixe) de $\widehat{\beta}_m$!

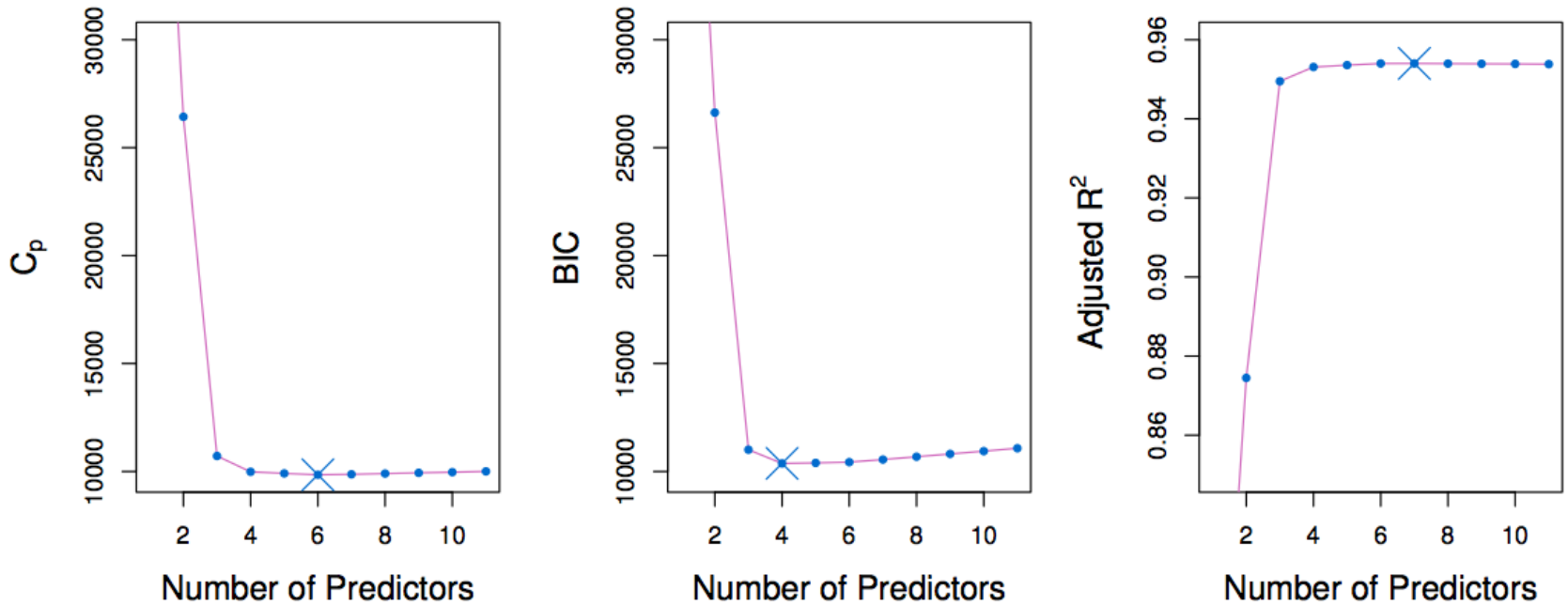
Remarque sur les pénalités AIC et BIC

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

- 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 $2\,d\,\hat{\sigma}^2$ utilisé par C_p par un terme $\log(n)\,d\,\hat{\sigma}^2$ 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.

Exemple : jeu de données de crédit

La figure suivante montre C_p et BIC pour le meilleur modèle de chaque dimension pour le jeu de données de crédit



Comparaison de ces critères

- AIC sont des critères qui réalisent un compris 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 consistent (i.e., à sélectionner le vrai modèle $\mathrm{supp}(\beta^*)$ avec probabilité tendant vers 1 lorsque $n \to \infty$)
- 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.

Extensions à des modèles ajustés par maximum de vraisemblance (ex : régression logistique)

- Le critère **AIC** (Akaike Information Criterion) est défini pour une large de modèles ajustés par maximum de vraisemblance par :

$$Crit_{\text{AIC}}(m) = -2 \log L(\widehat{\beta}_m) + 2\, d_m$$

où L est la vraisemblance maximale pour le modèle de dimension d_m considéré.

- Le critère **BIC** (Bayesian Information Criterion) est défini par

$$Crit_{\text{BIC}}(m) = -2 \log L(\widehat{\beta}_m) + \log(n)\, d_m$$

Validation et validation croisée

Une alternative à la pénalisation est d'estimer le risque de chaque estimateur $\widehat{\beta}_m$ par validation croisée.
Le modèle $\widehat{m}$ est choisie comme minimiseur du critère suivant

$$Crit_{CV}(m) = \widehat{R}^{CV}(\widehat{f}_m) \ ,$$

où $\widehat{R}^{CV}(\widehat{f}_m)$ est un estimateur par validation croisée du risque $R(\widehat{f}_m)$.

AVANTAGE : La sélection par validation croisée permet de sélectionner des estimateurs sans aucune hypothèse sur la distribution des données ou les procédures d'estimation.

INCONVENIENT : Plus coûteux en temps de calcul. Légérement moins efficace que la pénalisation lorsque la vraie distribution des données est celle d'un modèle linéaire gaussien.

Retour sur la sélection compléte 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 $\widehat{m}_1, \ldots \widehat{m}_p$ 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.

En détail

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 $p-1$:
   - 2.1 Considérer tous les $(p-k)$ modèles qui consistent à ajouter un prédicteur à $\widehat{m}_k$.
   - 2.2 Choisir le meilleur parmi ces $(p-k)$ modèles et noter le $\widehat{m}_{k+1}$. Ici, le *meilleur* modèle est celui qui minimise le critére des moindres carrés.
3. Choisir le meilleur modèle parmi $\widehat{m}_0$, $\widehat{m}_1$, $\ldots \widehat{m}_p$ 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-ens 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 compléte.

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.

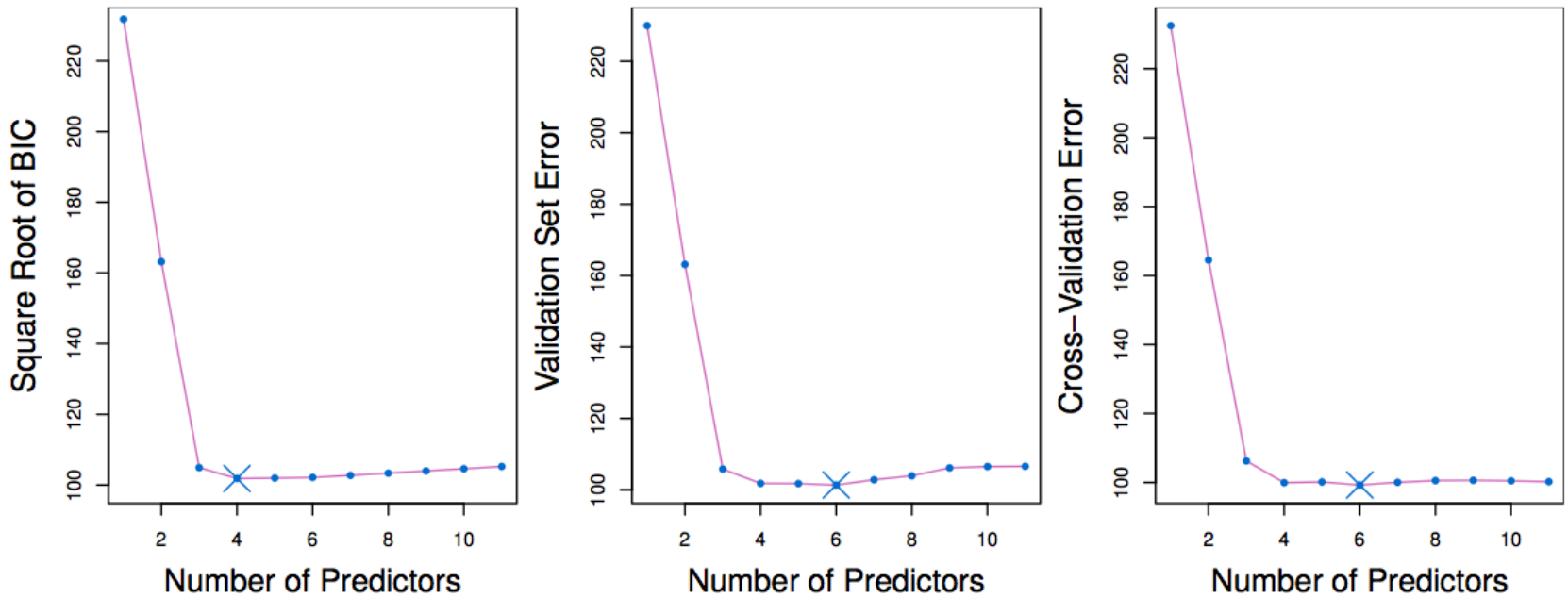
Sélection pas à pas rétrograde :détail

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 :
   - 2.1 Considérer tous les k modèles qui consistent à supprimer un prédicteur à $\widehat{m}_k$.
   - 2.2 Choisir le meilleur parmi ces k modèles et noter le $\widehat{m}_{k-1}$. Le *meilleur* modèle est celui qui minimise le critére des moindres carrés.
3. Choisir le meilleur modèle parmi $\widehat{m}_0$, $\widehat{m}_1$, $\ldots \widehat{m}_p$ 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



Commentaires

- L'erreur par validation a été estimée en mettant de côté un quart du jeu de données (tiré au hasard) pour valider. Les trois quarts restants servant à entrainer les modèles
- L'erreur de validation croisée a été calculée par une méthode à $k = 10$ blocs. Dans ce cas, ces deux méthodes renvoient un modèle à 6 variables (de dimension 7, pourquoi ?).
- Cependant, les trois approches suggèrent que les modèles à 4, 5 ou 6 variables sont à peu près équivalents en terme d'erreur de test.

Coin du UseR

`package leaps`

`regsubset(.~ .,)`
Calcul du meilleur modèle pour chaque dimension
`summary(regsubset(.~ .,))$cp`
Critére cp associé

`regsubset(.~ .,„method="forward")`
methode de sélection ascendante

`regsubset(.~ .,„method="backward")`
methode de sélection descendante

Autre commande : package `MASS`
`stepAIC()`

Plan

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 que *contraint* ou *régularise* les estimations des coefficients, ou de façon équivalente, pousse les coefficients vers 0.
- Il n'est pas évident de comprendre pourquoi de telles contraintes vont améliorer l'ajustement, mais il se trouve qu'elles réduisent la variance de l'estimation des coefficients.

Préambule

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

Remarque : Une fois les paramètres ajustés pour les variables centrées et standardisées, on peut facilement revenir au modèles initial en transformer linéairement les paramètres. le vérifier

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

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

λ étant un paramétre positif, appelé paramètre de régularisation.

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

Régression ridge

L'estimateur est défini par

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

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 $\widehat{\beta}_\lambda^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$.

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

Rôle et ajustement du paramétre de régularisation

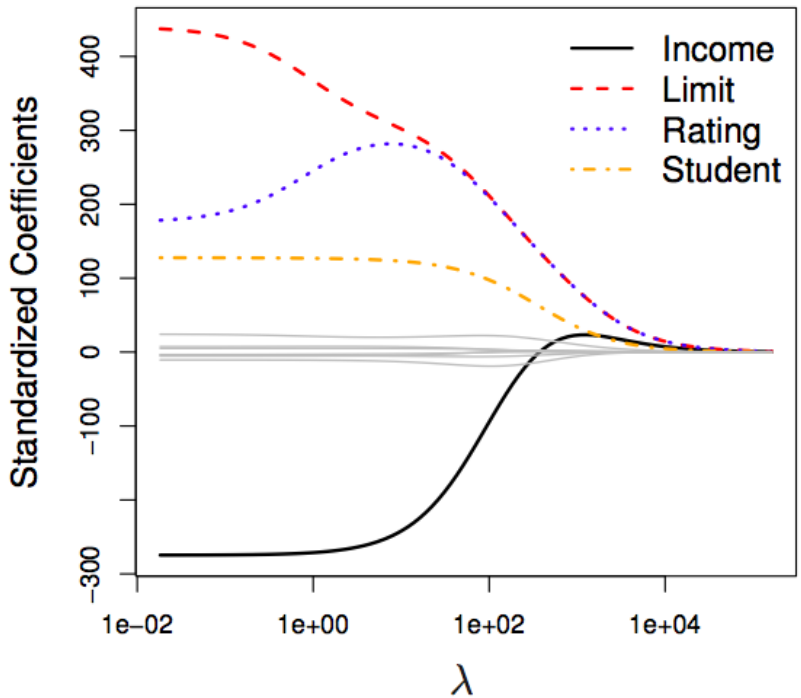
- Lorque $\lambda = 0$, $\widehat{\beta}_{\lambda}^{\text{ridge}}$ est l'estimateur des moindres carrés.
- Lorque $\lambda \to \infty$, $\widehat{\beta}_{\lambda}^{\text{ridge}}$ tend vers 0
- Lorque λ augmente, le biais de $\widehat{\beta}_{\lambda}^{\text{ridge}}$ a tendance à augmenter et la variance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

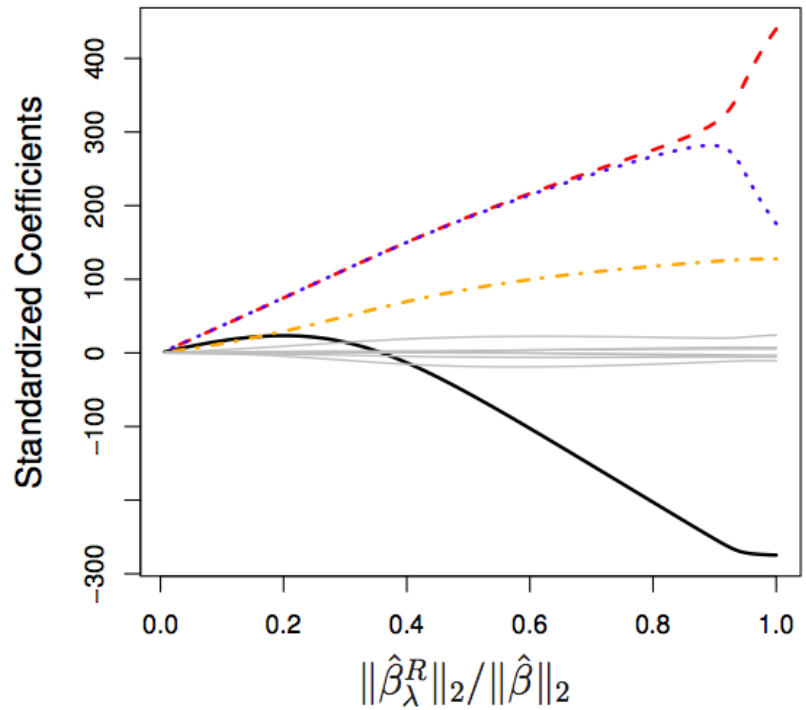
$\rightsquigarrow$ Choix usuel de λ par validation croisée V fold sur une grille finie de valeur de $\lambda > 0$.

Régression ridge (suite)

Exercice : Dans le cas très simple du plan d'expérience avec $n = p$ et $\mathbf{X} = n\mathbf{I}$, calculer la forme explicite de l'estimateur ridge et commenter.

Exemple jeu de données crédit



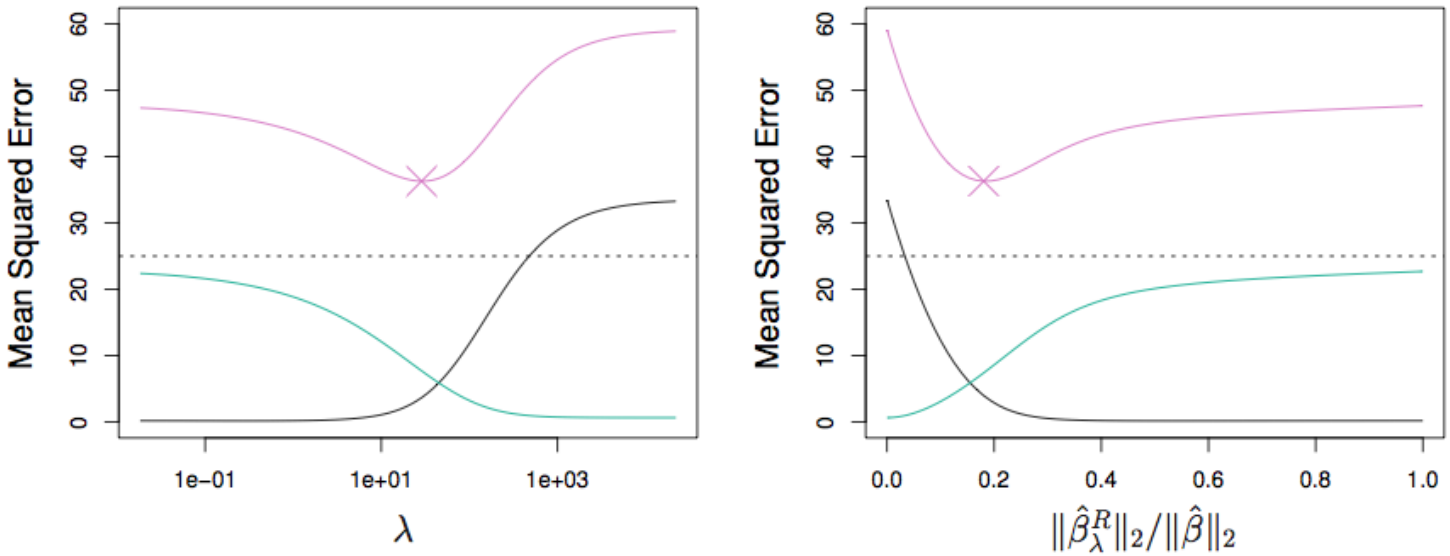


Commentaires

- 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.

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

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

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

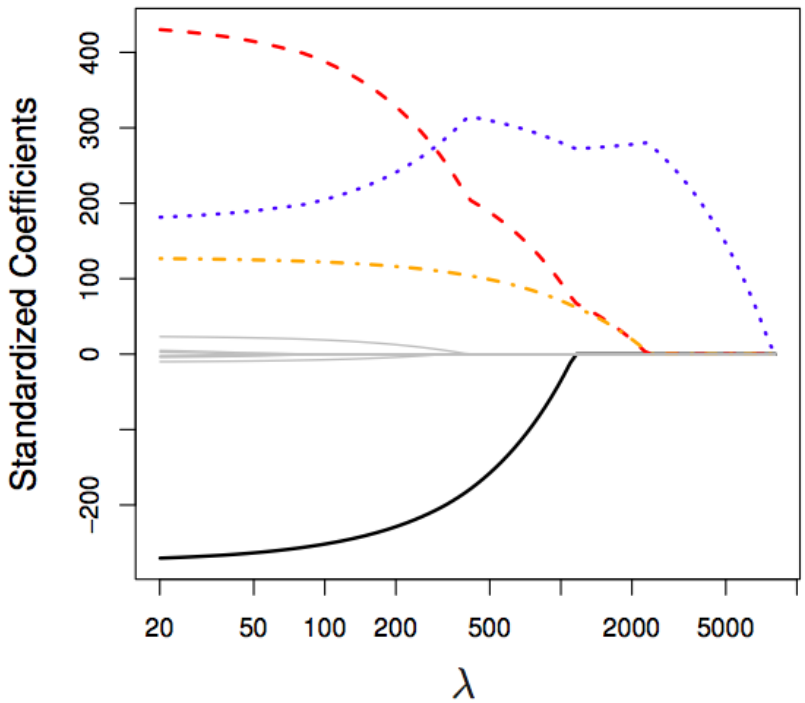
Minimiser $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1$ en $\beta \in \mathbb{R}^p$ est équivalent à minimiser $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2$ sous une contrainte de la forme $\|\beta\|_1 \leq R_\lambda$ pour une certaine quantité R_λ.

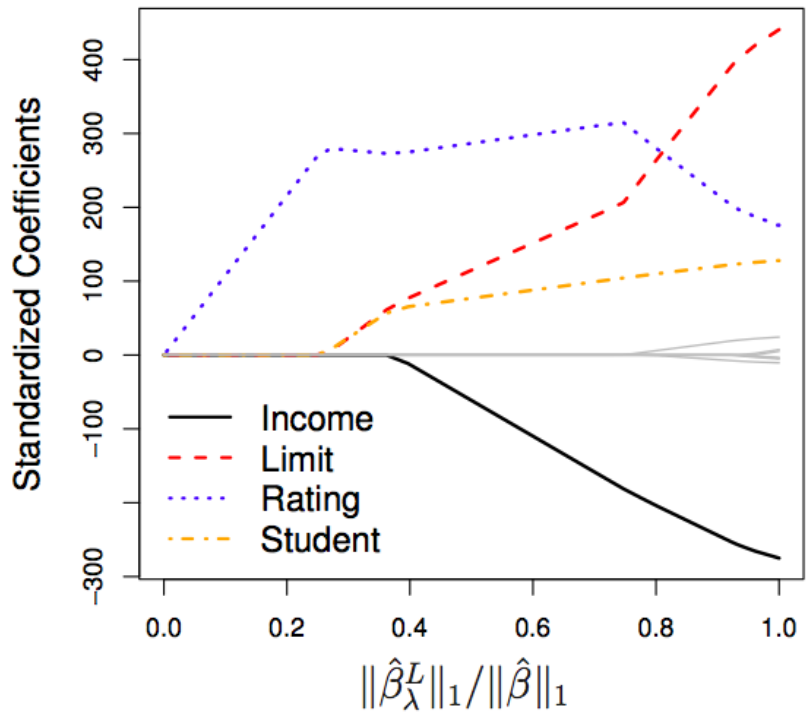
Preuve : Lagrangien

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



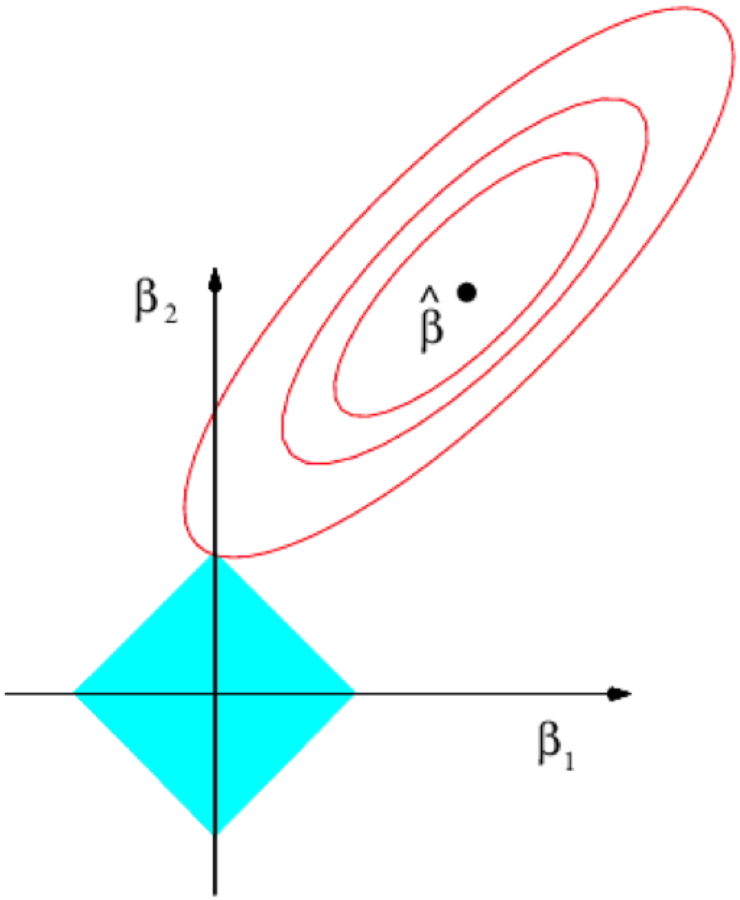


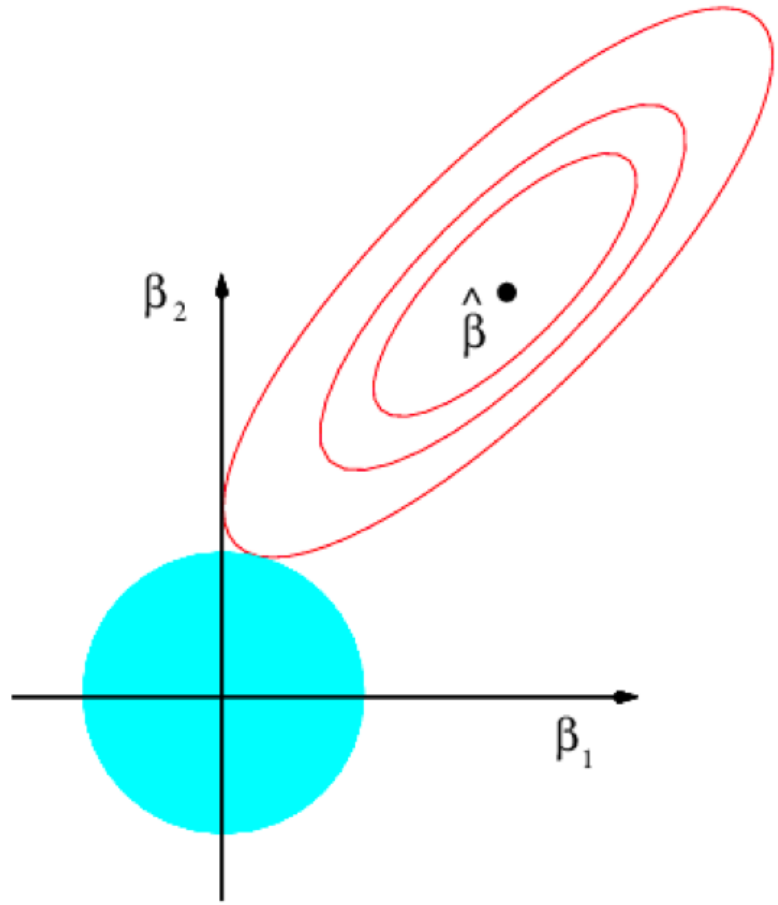
Qu'est qui fait marcher le Lasso ?

Avec les multiplicateurs de Lagrange, on peut voir

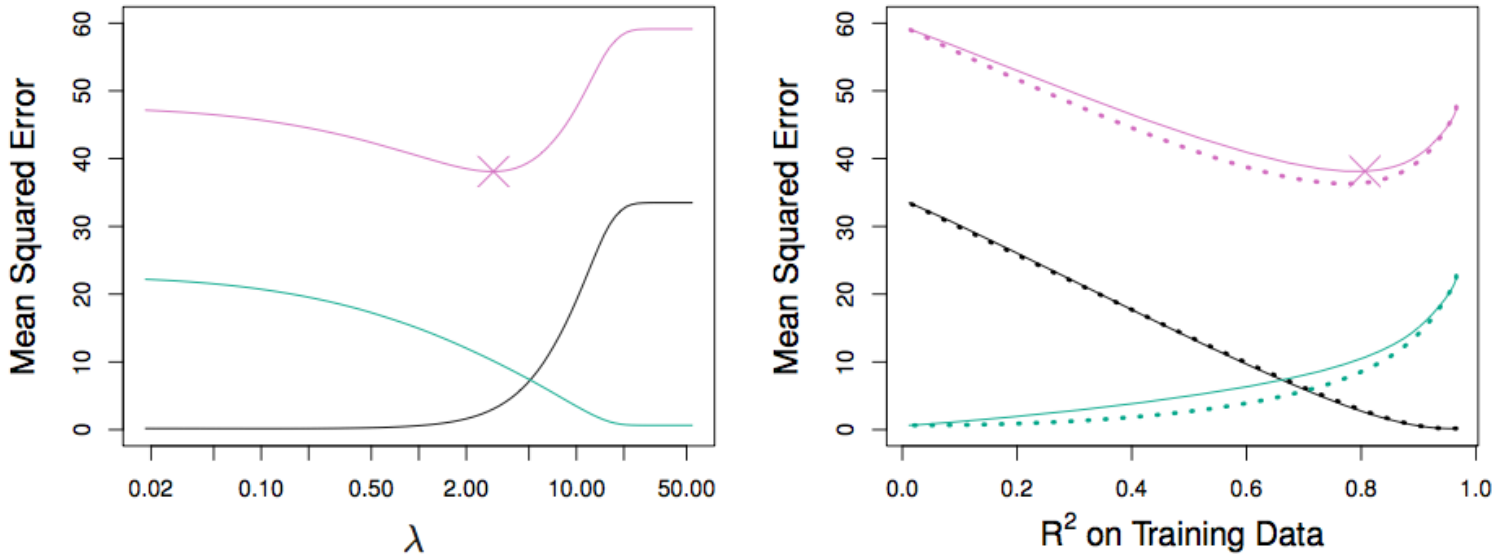
- La régression ridge comme
  minimise $\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}\beta_j X_i^{(j)}\right)^2$ sous la contrainte $\sum_{j=1}^{p}\beta_j^2 \leq s$
- Le Lasso comme
  minimise $\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}\beta_j X_i^{(j)}\right)^2$ sous la contrainte $\sum_{j=1}^{p}|\beta_j|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

Le Lasso en image





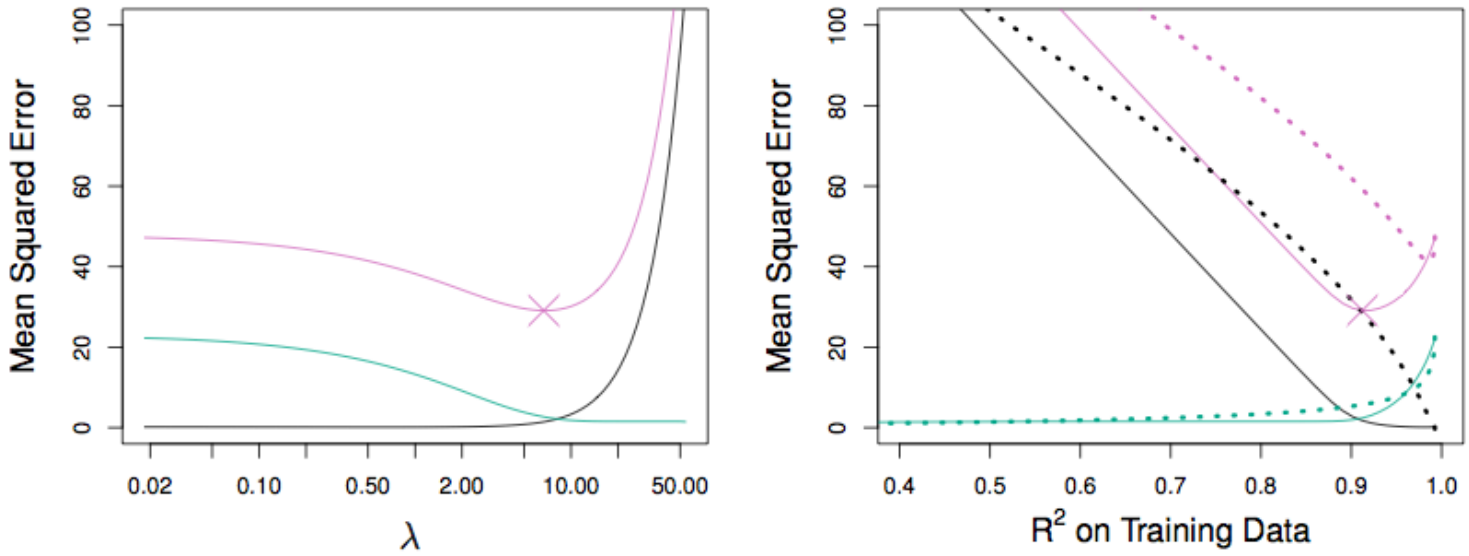
Comparaison du Lasso et de la régression ridge



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 plains) 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.

Bornes de risque

Notation $x^{(k)}$: k-ième colonne de $\mathbf{X}$.

Proposition

[Koltchinski et al.(2010)] Supposons que les $x^{(j)}$ sont normés (ie $\|x^{(j)}\|_2^2 = n$) et $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$. Pour tout $L > 0$, si $\lambda = 3\sigma\sqrt{\frac{2}{n}[\log(p) + L]}$ avec probabilité au moins égale $1 - e^{-L}$,

$$\|\mathbf{X}(\widehat{\beta}_\lambda^{\text{Lasso}} - \beta^*)\|_2^2 \leq \inf_{\beta \neq 0} \left\{ \|\mathbf{X}(\beta - \beta^*)\|_2^2 + \frac{18\sigma^2(L + \log(p))}{\kappa^2(\beta)} \sum_{j=1}^{p} 1_{\beta_j \neq 0} \right\}$$

où $\kappa(\beta)$ est ce que l'on appelle une constante de compatibilité, mesurant le manque d'orthogonalité des colonnes de $\mathbf{X}$.

Coordinate descent

Condition d'optimalité du premier ordre

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

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 $R_j = (x^{(j)})^T(Y - \sum_{k\neq j}\beta_k x^{(k)})$.

Exercice : le prouver.

Coordinate descent pour la lasso

Algorithme de Coordinate descent

input : $\mathbf{X}$, $\mathbf{Y}$, $\lambda > 0$

begin

Initialiser $\beta = \beta_{in}$

while β *n'a pas convergé* **do**

for $j = 1, \ldots, p$ **do**

Calculer $R_j = (x^{(j)})^T (Y - \sum_{k \neq j} \beta_k x^{(k)})$

Calculer $\beta_j = R_j (1 - \lambda / (2|R_j|))_+ / n$

output: β

Utilisation du Lasso sous R

choix usuel par validation croisée V fold

Le coin du UseR : package `glmnet`

```
fit=glmnet(x,y,alpha=1)
plot(fit)
coef(fit,s=1) # affiche les coefficients pour un λ
predict(fit,newx=x[1:10,],s=1) # prédiction

fit2=cv.glmnet(x,y) # calcule de l'erreur par
validation croisée pour une collection de λ
```

Utilisation de Ridge sous R

choix usuel par validation croisée V fold

Le coin du UseR : package `glmnet`

```
fit=glmnet(x,y,alpha=0)
plot(fit)
coef(fit,s=1) # affiche les coefficients pour un λ
predict(fit,newx=x[1:10,],s=1) # prédiction

fit2=cv.glmnet(x,y) # calcule de l'erreur par
validation croisée pour une collection de λ
```

Régression logistique Lasso

Modèle de régression logistique : $p(x_i) = \mathbb{P}(Y_i = 1 | X_i = x_i)$.
On suppose que

$$\log \frac{p(x_i)}{1 - p(x_i)} = \sum_{j=1}^{p} \beta_j x_i^{(j)}$$

L'estimateur LASSO logistique est défini pour $\lambda > 0$ par

$$\widehat{\beta}_\lambda^{logL} \in \arg\min_{\beta \in \mathbb{R}^p} -2\mathcal{L}_n(Y_1, \ldots, Y_n; \beta) + \lambda \|\beta\|_1 \ ,$$

où
$\mathcal{L}_n(Y_1, \ldots, Y_n; \beta) = \sum_{i=1}^n Y_i (\sum_j \beta_j x_i^{(j)}) - \log\left(1 + \exp(\sum_j \beta_j x_i^{(j)})\right)$
est la log-vraisemblance.

Le coin du UseR : Package `glmnet`

```
fit=glmnet(x,y,family="binomial")
```

Plan

Méthodes de réduction de la dimension

- Les méthodes déjà présentées pour ajuster des modèles linéaires par moindres carrés, éventuellement pénalisés, utilisent les covariables originales $X^{(1)} \dots X^{(p)}$ (éventuellement standardisés)
- Nous allons maintenant présenter des méthodes qui *transforment* les prédicteurs, puis ajustent un modèle par moindres carrés sur ces variables transformées.

Méthode de réduction de la dimension (suite)

- Soit $Z_1, Z_2, \ldots, Z_M$ les M *combinaisons linéaires* ($M < p$) des prédicteurs originaux. C'est-à-dire

$$Z_m = \sum_{j=1}^{p} \phi_{mj} X^{(j)}$$

pour une matrice ϕ.

- On ajuste ensuite un modèle de régression linéaire

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \varepsilon_i, \quad i = 1, \ldots, n,$$

par moindres carrés ordinaires.

- Notons que dans ce dernier modèle, les coefficients de régression sont les θ_m et θ_0. Si la matrice ϕ est choisie sagement, de telles méthodes dont souvent meilleures que les méthodes des moindres carrés ordinaires.

- Notons que, si l'on reporte la définition des Z_m dans le modèle linéaire où ils interviennent, on obtient

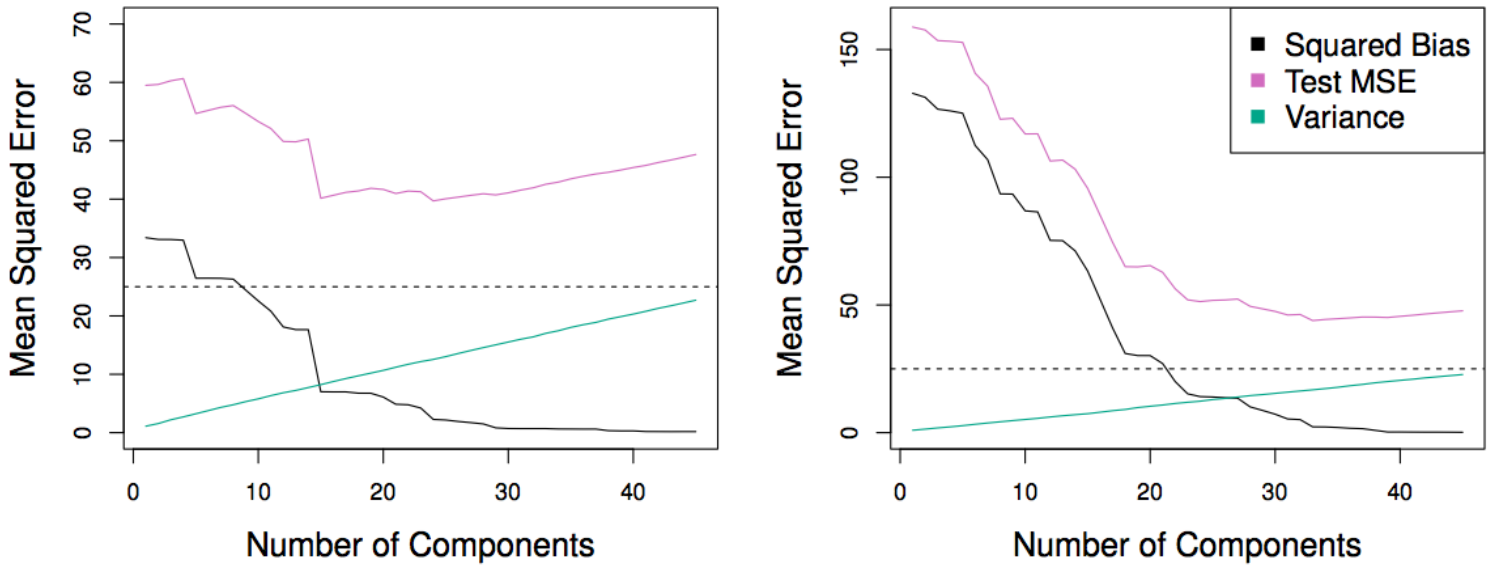
$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{mj}.$$

- Le modèle sur variables projetées est donc un cas particulier de modèle linéaire.
- La réduction de dimension sert de contrainte sur les coefficients β_j estimés, pour qu'ils satisfassent les égalités ci-dessus.
- Cette méthode fournit souvent un bon compromis biais – variance.

Régression sur composantes principales

- Ici, nous appliquons une analyse en composantes principales (voir cours de M1) pour définir les Z_m comme combinaisons linéaires des variables initiales.
- Le premier axe est la combinaison linéaire (normalisée) qui a la plus grande variance.
- Le deuxième axe est la combinaison linéaire (normalisée) qui a la plus grande variance, parmi celles qui sont décorrélées du premier axe.
- Etc.
- On peut ainsi remplacer un grand ensemble de variables corrélées par des variables décorrélées qui capturent au mieux la variance jointe.

Application de la régression sur composantes principales

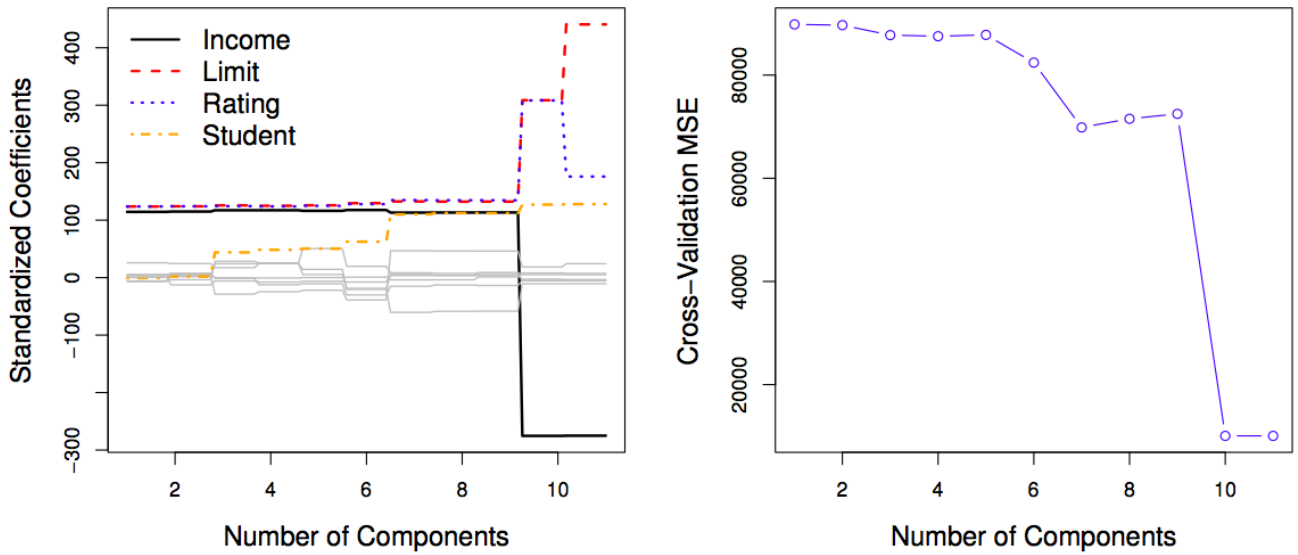


La PCR a été appliquée à deux jeux simulés. En noir, le biais au carré, en vert la variance et en violet l'erreur quadratique de test.

À gauche, sur un jeu où tous les régresseurs sont influents.

À droite, sur un jeu où seuls deux régresseurs sont influents.

Choix du nombre d'axes M



À gauche, l'estimation des coefficients standardisés sur le jeu de données crédit pour différentes valeurs de M.

À droite, l'erreur quadratique de test par validation croisée à 10 blocs, en fonction de M.

Moindres carrés partiels (Partial Least Square PLS)

- PCR identifie les combinaisons linéaires qui représente au mieux les covariables (en terme de variance)
- Ces axes sont obtenus avec une méthodes *non supervisée*, puisque la réponse Y n'influe pas sur le calcul de ces axes.
- Autrement dit, la réponse Y ne *supervise* pas les composantes principales.
- En conséquence, PCR souffre potentiellement d'un inconvénient : il n'y a aucune garantie que les axes principaux soient les meilleures pour prédire la réponse Y.

PLS (suite)

- Comme PCR, PLS est une méthode de réduction de la dimension, qui identifie un nouvel ensemble de variables Z_1, ..., Z_M, combinaisons linéaires des variables originales, et ajuste le modèle linéaire sur ces M nouvelles variables par moindres carrés.
- Mais, contrairement à PCR, PLS identifie ces nouvelles variables de façon supervisée — c'est-à-dire en utilisant la réponse Y pour les construire.
- Grosso modo, l'approche PLS tente de trouver les directions qui expliquent au mieux la réponse et les prédicteurs.

Détails sur PLS

- Après avoir standardisé les p prédicteurs, PLS calcule la première direction Z_1 en fixant tous les ϕ_{1j} par régression linéaire de Y sur $X^{(j)}$
- On peut montrer que ce coefficient est proportionnel à la corrélation entre Y et $X^{(j)}$.
- Donc, en calculant $Z_1 = \sum_{j=1}^{p} \phi_{1j} X^{(j)}$, PLS place les poids les plus fort sur les variables les plus corrélées avec la réponse Y.
- Les directions suivantes sont obtenues en prenant les résidus et en répétant la règle ci-dessus.

Le coin du UseR

Package `pls`

```
pcr(Y~.,validation="CV")
 # Ajustement d'une régression sur composantes princ.
par cv

plsr(Y~.,validation="CV")
 # Ajustement d'une régression pls par cv
```

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.
- Les questions de recherches qui donnent des solutions creuses (parcimonieuses, ou sparses), comme le Lasso, sont d'actualité.