Chapitre 2 : Sélection de variables et pénalisations

Agathe Guilloux
Professeure au LaMME - Université d'Évry - Paris Saclay

Avant de commencer

- Les documents du cours sont disponibles ici : `http://www.math-evry.cnrs.fr/members/aguilloux/enseignements/m2upmc`
- Bibliographie (pour ce chapitre) :
- Contrôle des connaissances : 2 TP rendus, 1 projet (type concours de DataScience), 1 examen court.
- Pré-requis : cours de Statistique de base `http://www.proba.jussieu.fr/pageperso/rebafka/StatBase_poly_partie2.pdf`.

Chapitre 2 : Sélection de variables et pénalisations

Agathe Guilloux
Professeure au LaMME - Université d'Évry - Paris Saclay

Avant de commencer

- Se connecter à ma page `http://www.lsta.upmc.fr/guilloux.php?main=enseignement` et récupérer les transparents ici.
- Bibliographie :
  Jean-Marc Azaïs et Jean-Marc Bardet

Les données `swiss` du package `faraway`

Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.

A data frame with 47 observations on 6 variables, each of which is in percent, i.e., in [0, 100].

- `Fertility` common standardized fertility measure
- `Agriculture` % of males involved in agriculture as occupation
- `Examination` % draftees receiving highest mark on army examination
- `Education` % education beyond primary school for draftees.
- `Catholic` % 'catholic' (as opposed to 'protestant').
- `Infant.Mortality` live births who live less than 1 year.

All variables but `Fertility` give proportions of the population.

Switzerland, in 1888, was entering a period known as the demographic transition; i.e., its fertility was beginning to fall from the high level typical of underdeveloped countries.

The data collected are for 47 French-speaking "provinces" at about 1888.

Here, all variables are scaled to [0, 100], where in the original, all but "Catholic" were scaled to [0, 1].

Sélection de variables ℓ_0

Tests de Student et de Fisher

On se place dans le modèle

$$Y = X\beta + \epsilon$$

$$\text{avec } \epsilon \sim \mathcal{N}((0,\dots,0)^\top, \sigma^2 I_n) \text{ et } X\beta \in V = \text{vect}(\mathbf{1}, X^1, \dots, X^p).$$

On veut **sélectionner** les variables qui ont une influence sur Y, i.e. déterminer les indices $k \in \{1,\dots,p\}$ pour lesquels $\beta_k \neq 0$.

- Si on veut tester $H_0 : \beta_k = 0$, on peut utiliser le test de Student de niveau

$$\mathbb{P}_{H_0}\Big(\Big|\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 (X^\top X)^{-1}_{jj}}}\Big| > F^{-1}_{\mathcal{T}(n-p-1)}(1-\alpha/2)\Big) = \alpha$$

- Si on veut tester $H_0 : \beta_{k_1} = \ldots = \beta_{k_l} = 0$, on pourrait utiliser les statistiques de Student mais le niveau pour un seuil s est alors donné par :

$$\mathbb{P}_{H_0}(\bar{H}_0) = \mathbb{P}\Big(\Big|\frac{\hat{\beta}_{k_1}}{\sqrt{\hat{\sigma}^2(X^\top X)^{-1}_{k_1 k_1}}}\Big| > s \cup \ldots \cup \Big|\frac{\hat{\beta}_{k_l}}{\sqrt{\hat{\sigma}^2(X^\top X)^{-1}_{k_l k_l}}}\Big| > s\Big)$$

$$\leq \mathbb{P}\Big(\Big|\frac{\hat{\beta}_{k_1}}{\sqrt{\hat{\sigma}^2(X^\top X)^{-1}_{k_1 k_1}}}\Big| > s\Big) + \ldots + \mathbb{P}\Big(\Big|\frac{\hat{\beta}_{k_l}}{\sqrt{\hat{\sigma}^2(X^\top X)^{-1}_{k_l k_l}}}\Big| > s\Big).$$

Pour garantir un niveau α, il faut prendre $s = F^{-1}_{\mathcal{T}(n-p-1)}(1-\alpha/(2l))$ pour avoir :

$$\mathbb{P}_{H_0}(\bar{H}_0) \leq \frac{\alpha}{l} + \ldots + \frac{\alpha}{l} = \alpha.$$

Mais la puissance de chaque test est alors assez mauvaise car

$$F^{-1}_{\mathcal{T}(n-p-1)}(1-\alpha/(2l)) >> F^{-1}_{\mathcal{T}(n-p-1)}(1-\alpha/(2)) \text{ si } l >> 1.$$

Test pour modèles emboîtés

Pour tester $H_0 : \beta_{k_1} = \ldots = \beta_{k_l} = 0$, on utilise le test de Fisher.

- On note $W = \text{vect}\{(\mathbf{1}, X_1, \ldots, X_p)/(X_{k_1}, \ldots, X_{k_l})\}$ de dimension $p + 1 - l$.
- On note $X\tilde{\beta} = \text{Proj}_W^\top(Y)$ et on a toujours $X\hat{\beta} = \text{Proj}_V^\top(Y)$.

Loi sous H_0

Sous H_0, on a donc

$$Y - X\hat{\beta} = \epsilon - \text{Proj}_V^\perp(\epsilon) = \text{Proj}_{V^\perp}^\perp(\epsilon) \text{ et}$$
$$Y - X\tilde{\beta} = \epsilon - \text{Proj}_W^\perp(\epsilon) = \text{Proj}_{W^\perp}^\perp(\epsilon).$$

On remarque que $W \subset V$ et donc

$$W^{\perp} = V^{\perp} \overset{\perp}{\bigoplus} W^{\perp_V}.$$

Statistique de Fisher

On a

$$\|Y - X\tilde{\beta}\|^2 - \|Y - X\hat{\beta}\|^2 = \|\operatorname{Proj}^{\perp}_{W^{\perp_V}}(\epsilon)\|^2.$$

donc

$$\frac{(n-p-1)\left(\|Y - X\tilde{\beta}\|^2 - \|Y - X\hat{\beta}\|^2\right)}{l\|Y - X\hat{\beta}\|^2} \underset{H_0}{\sim} \mathcal{F}(l, n-p-1).$$

Si le modèle de départ a p variables, il y a $\#\mathcal{P}\{1,\ldots,p\} = 2^p$ sous-modèles à explorer. Il faut des algorithmes efficaces :

- **forward** : on part du modèle avec seulement l'intercept et on ajoute les variables une à une. A chaque pas, on ajoute celle qui a la plus grande statistique de Fisher. On s'arrête quand la p-value associée devient > 0.1 (seuil arbitraire)
- **backward** : on part du modèle avec toutes les variables et on retire les variables une à une. A chaque pas, on retireee celle qui a la plus petite statistique de Fisher. On s'arrête quand la p-value associée devient < 0.1 (seuil arbitraire)
- **stepwise** mixte des deux premières (on ajoute, puis on permet une élimination, etc)

Modèles, vrai modèle

On se donne une famille de modèles $\mathcal{M}$, par exemple $\mathcal{M} = \mathcal{P}\{1, \ldots, p\}$. On suppose qu'il existe un vrai modèle $m^* \in \mathcal{M}$ tel que :

$$Y = X^{(m^*)}\beta^{(m^*)} + \epsilon^* \quad \text{avec } \epsilon^* \sim \mathcal{N}((0, \ldots, 0)^\top, (\sigma^{m^*})^2 I_n).$$

On veut retrouver m^*.

- **Attention :** le R^2 n'est pas un bon critère pour ce problème car il choisira toujours le modèle complet (avec toutes les covariables)
- On note m_{tot} le modèle complet.

Estimation dans le modèle m

Dans le modèle m, on note $|m|$ le nombre de covariables qu'il contient et

$$\hat{\beta}^{(m)} = \left((X^{(m)})^\top X^{(m)}\right)^{-1}(X^{(m)})^\top Y$$

$$\hat{Y}^{(m)} = X^{(m)}\hat{\beta}^{(m)}$$

$$\widehat{(\sigma^m)^2} = \frac{\|Y - \hat{Y}^{(m)}\|_2^2}{n - |m|}$$

Risque quadratique

Le risque quadratique de $\hat{Y}^{(m)}$ pour l'estimation de $X^*\beta^*$ est donné par

$$\mathbb{E}_{m^*}(\|\hat{Y}^{(m)} - X^*\beta^*\|^2) = \underbrace{\mathbb{E}_{m^*}(\|\hat{Y}^{(m)} - X^{(m)}\beta^{(m)}\|^2)}_{\text{variance}} + \underbrace{\|X^{(m)}\beta^{(m)} - X^*\beta^*\|^2}_{\text{biais}^2}$$

$$= \sigma^2|m| + \|X^{(m)}\beta^{(m)} - X^*\beta^*\|^2$$

où $X^{(m)}\beta^{(m)}$ est la projection de $X^*\beta^*$ sur vect($X^{(m)}$).

Les termes de biais et de variance sont inconnus, il faut donc les estimer.

Estimations des termes de biais et de variance

On montre que

$$\mathbb{E}_{m^*}(\|\hat{Y}^{(m)} - Y\|^2) = (n - |m|)\sigma^2 + \|X^{(m)}\beta^{(m)} - X^*\beta^*\|^2.$$

Finalement

$$\begin{aligned}\mathbb{E}_{m^*}(\|\hat{Y}^{(m)} - X^*\beta^*\|^2) &= \sigma^2|m| - (n - |m|)\sigma^2 + \mathbb{E}(\|\hat{Y}^{(m)} - Y\|^2) \\ &= 2\sigma^2|m| + \mathbb{E}_{m^*}(\|\hat{Y}^{(m)} - Y\|^2) - n\sigma^2.\end{aligned}$$

On approxime $\mathbb{E}_{m^*}(\|\hat{Y}^{(m)} - Y\|^2)$ par

$$\|\hat{Y}^{(m)} - Y\|^2 = (n - |m|)\hat{\sigma}^2_{(m)}.$$

Cp de Mallows

On choisit $\hat{m}_{Cp} \in \mathcal{M}$ tel que :

$$\hat{m}_{Cp} = \underset{m \in \mathcal{M}}{\text{argmin}}\ Cp(m),$$

avec

$$Cp(m) = \frac{\hat{\sigma}^2_{(m)}}{\hat{\sigma}^2_{(m)_{\text{tot}}}} + 2\frac{|m|}{n}$$

Divergence de Kullback (1)

La divergence de Kullback entre deux densités f et f^* s'écrit (sous les bonnes conditions)

$$\mathcal{K}(f, f^*) = \int f^* \log\left(\frac{f^*}{f}\right).$$

Si les deux densités sont deux gaussiennes sur $\mathbb{R}^n$: $\mathcal{N}(\mu, \sigma^2)$ et $\mathcal{N}(\mu_*, \sigma_*^2)$, on obtient

$$\mathcal{K}(f, f^*) = \frac{n}{2} \log \frac{\sigma^2}{\sigma_*^2} - \frac{n}{2} + \frac{n}{2} \frac{\sigma_*^2}{\sigma^2} + \frac{1}{2\sigma^2} \|\mu_* - \mu\|^2.$$

Divergence de Kullback (2)

En régression, on veut minimiser

$$\mathbb{E}_{m^*}\left(\log\frac{\hat{\sigma}^2_{(m)}}{\sigma^2_*} - 1 + \frac{\sigma^2_*}{\hat{\sigma}^2_{(m)}} + \frac{1}{n\hat{\sigma}^2_{(m)}}\|\mu_* - X^{(m)}\hat{\beta}^{(m)}\|^2\right)$$

$$= \mathbb{E}_{m^*}\left(\log\frac{\hat{\sigma}^2_{(m)}}{\sigma^2_*}\right) - 1 + \frac{n-|m|}{n-|m|-2} + \frac{n}{n}\frac{n-|m|}{n-|m|-2}.$$

Après quelques approximations, on obtient le critère AIC.

AIC/BIC - divergence de Kullback

On choisit $\hat{m}_{AIC} \in \mathcal{M}$ et $\hat{m}_{BIC} \in \mathcal{M}$tel que :

$$\hat{m}_{AIC} = \underset{m \in \mathcal{M}}{\text{argmin}}\, AIC(m),$$
$$\hat{m}_{BIC} = \underset{m \in \mathcal{M}}{\text{argmin}}\, BIC(m),$$

avec

$$AIC(m) = \log\left(\hat{\sigma}^2_{(m)}\right) + 2\frac{|m|+1}{n}$$
$$BIC(m) = \log\left(\hat{\sigma}^2_{(m)}\right) + \frac{\log(n)|m|}{n}$$

Illustration d'après Azaïs et Bardet

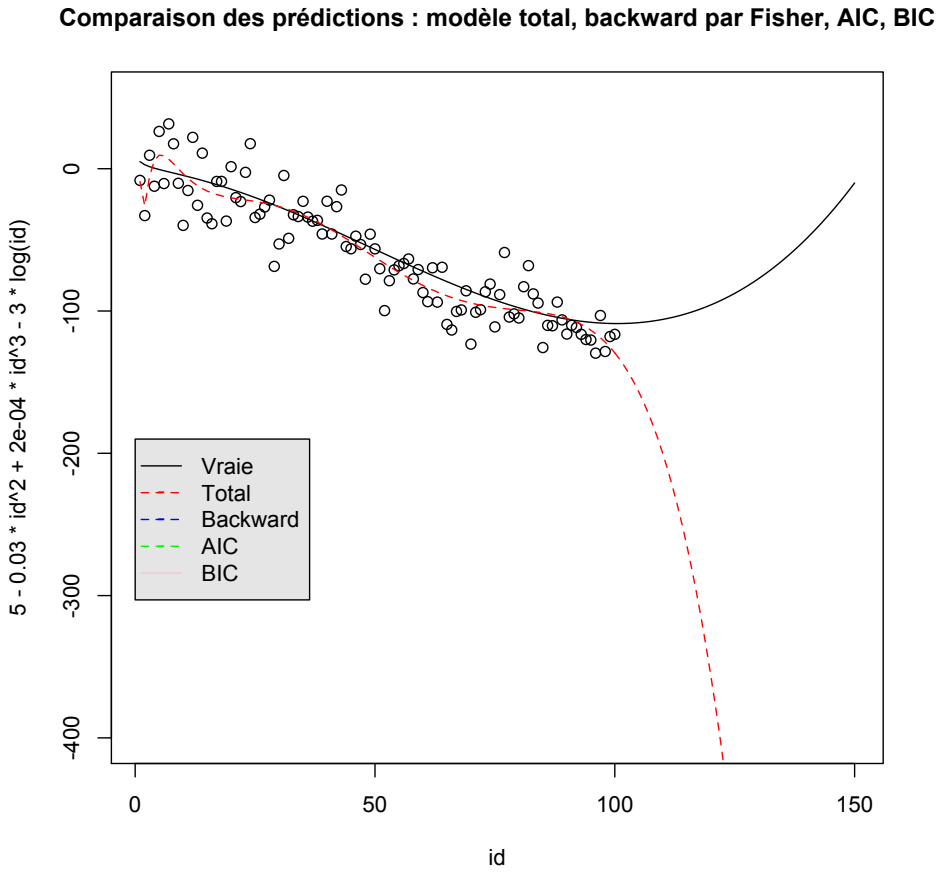
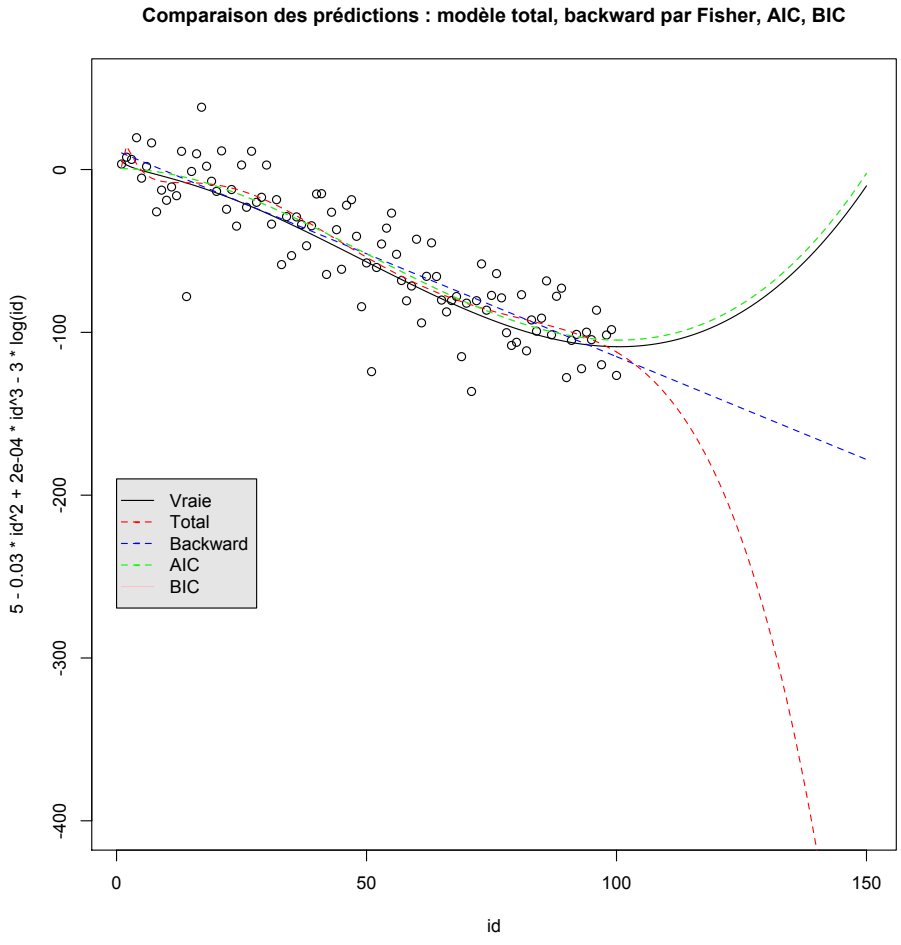


Illustration d'après Azaïs et Bardet



Exercice

1. Reprendre l'exemple d'école et coder les différentes procédures de sélection de modèles.
2. Finir l'analyse du jeu de données https://archive.ics.uci.edu/ml/datasets/Student+Performance, sélection de variables incluse. Comparer les modèles choisis par les différentes procédures de sélection.

Régressions pénalisées et sur variables synthétiques

On observe Y et X de dimension $n \times p$. On fait l'hypothèse qu'il existe un vrai modèle m^* tel que

$$Y = X^{m^*}\beta^{m^*} + \epsilon^{m^*} = X^*\beta^* + \epsilon^*,$$

et que $|m^*|$ est petit devant p et n.

Quand p devient grand par rapport à n, il y a trois problèmes potentiels:

- $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ peut ne pas être défini car $X^\top X$ n'est alors plus inversible (c'est aussi le cas si des colonnes de X sont colinéaires)
- La variance d'estimation à partir de X devient trop grande. En effet, on a déjà montré que

$$\mathbb{E}(\|X\hat{\beta} - X^*\beta^*\|^2) = p\sigma^2 + \|X^{(m)}\beta^{(m)} - X^*\beta^*\|^2 = p\sigma^2.$$

- Les algorithmes de sélection de variables ℓ_0 ne peuvent plus être appliqués à cause des 2^p modèles à comparer.

Introduction

La régression ridge et PCR (principal components regression) s'attaquent aux deux premiers problèmes :

- la première en régularisant le problème des moindres carrés
- la seconde en travaillant sur des variables synthétiques issues d'une PCA (APC).

Les deux ont de bonnes performances en prédiction mais ont l'inconvénient de produire des modèles difficilement interprétables.

Régression Ridge

Hoerl et Kennard (1970) ont l'idée d'ajouter à $X^\top X$ une matrice diagonale $\lambda \mathrm{Id}_n$ pour retrouver l'inversibilité pour

$$X^\top X + \lambda \mathrm{Id}_p = PDP^{-1} + \lambda \mathrm{Id}_p = P(D + \lambda \mathrm{Id}_p)P^{-1}.$$

Pénalité ridge

$$\hat{\beta}_\lambda^{\text{ridge}} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \, \|Y - X\beta\|^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

Montrons que les deux définitions sont équivalentes.

Loi de l'estimateur ridge

Loi gaussienne

Sous l'hypothèse gaussienne $\hat{\beta}_\lambda^{\text{ridge}}$ reste gaussien mais a pour moments

$$\begin{aligned}
\mathbb{E}\hat{\beta}_\lambda^{\text{ridge}} &= (X^\top X + \lambda \mathrm{Id}_p)^{-1} X^\top Y = (X^\top X + \lambda \mathrm{Id}_p)^{-1} X^\top X \beta^* \\
\mathbb{V}\hat{\beta}_\lambda^{\text{ridge}} &= (X^\top X + \lambda \mathrm{Id}_p)^{-1} X^\top \mathbb{V}(\epsilon^*) X (X^\top X + \lambda \mathrm{Id}_p)^{-1} \\
&= \sigma^2 (X^\top X + \lambda \mathrm{Id}_p)^{-1} X^\top X (X^\top X + \lambda \mathrm{Id}_p)^{-1}.
\end{aligned}$$

Il est donc biaisé pour l'estimation de β^*, de même $X\hat{\beta}_\lambda^{\text{ridge}}$ est biaisé pour l'estimation de $X^*\beta^*$. On va montrer que, même si la régression ridge augmente le biais, elle peut diminuer le risque quadratique.

Propriété sur le risque quadratique

On connaît le risque quadratique de l'estimateur des moindres carrés

$$\mathbb{E}(\|X\hat{\beta} - X\beta^*\|^2) = p\sigma^2 + \|X^{(m)}\beta^{(m)} - X\beta^*\|^2 = p\sigma^2.$$

Calculons celui de pour la régression ridge

$$\begin{aligned}
&\mathbb{E}(\|X\hat{\beta}_\lambda^{\text{ridge}} - X\beta^*\|^2) \\
&= \mathbb{E}(\|X(X^\top X + \lambda \text{Id}_p)^{-1}X^\top X\beta^* - X\beta^*\|^2) + \mathbb{E}(\|X(X^\top X + \lambda \text{Id}_p)^{-1}X^\top \epsilon\|^2)
\end{aligned}$$

Design orthogonal

Dans le cas d'un design orthogonal, le risque quadratique s'écrit

$$\mathbb{E}(\|X\hat{\beta}_\lambda^{\text{ridge}} - X\beta^*\|^2) = \sigma^2 p \frac{1}{(\lambda+1)^2} + \|\beta^*\|^2 \frac{\lambda^2}{(1+\lambda)^2}.$$

Il existe donc un $\lambda^* > 0$ tel que $\mathbb{E}(\|X\hat{\beta}_{\lambda^*}^{\text{ridge}} - X\beta^*\|^2)$ est miminale et inférieure $p\sigma^2$.

Valeur des coefficients

Effet de λ sur les estimateurs

Toujours en design orthogonal, on écrit simplement

$$\hat{\beta}_{\lambda}^{\text{ridge}} = \frac{1}{1+\lambda}\hat{\beta}.$$

En pratique

Pour éviter les problèmes d'échelle, on ajoute cette matrice diagonale $\lambda \mathrm{Id}_n$ à la matrice X dont les colonnes ont été au préalable renormalisées et pour un signal Y recentré.

Exercice

On note

- $Y^c = Y - \bar{Y}$
- pour tout $j = 2, \ldots, p$ $X^{j,S} = \frac{X^j - \bar{(}X^j)}{\mathrm{sd}(X^j)}$ et $X^S = (X^{2,S}, \ldots, X^{p,S})$.

Ecrire la correspondance les moindres carrés classiques $\underset{\beta \in \mathbb{R}^p}{\mathrm{argmin}} \| Y - X\beta \|^2$ et les moindres carrés obtenus sur les variables renormalisées.

$$\underset{\gamma \in \mathbb{R}^{p-1}}{\mathrm{argmin}} \| Y^c - X^S \gamma \|^2$$

Dans l'estimateur des moindres carrés obtenus sur les variables renormalisées, on remplace $(X^S)^\top X^S$ par $((X^S)^\top X^S + \lambda \mathrm{Id}_n)$. C'est équivalent à considérer

Pénalité ridge

$$\hat{\gamma}_\lambda^{\text{ridge}} = \underset{\gamma \in \mathbb{R}^p}{\text{argmin}} \, \| Y^c - X^S \gamma \|^2 + \lambda \sum_{j=2}^{p} \gamma_j^2 .$$

Exercice

Que valent alors les estimateurs des coefficients de β^* ?

Cross-validation pour λ

On a montré (dans le cas d'un design orthogonal) qu'il existe une valeur λ^* du paramètre de régularisation qui minimise le risque quadratique de l'estimateur ridge. Cette valeur idéale dépend de quantités inconnues, il faut donc l'estimer.

Si on avait à disposition d'autres données, on aurait

- des données d'apprentissage (training, learning set)
  $\mathcal{S}_L = \{(Y_1, X_i), \ldots, (Y_n, X_n)\}$
- des données de validation, de test (testing, validation set)
  $\mathcal{S}_T = \{(Y_{+,1}, X_{+,1}), \ldots, (Y_{+,n'}, X_{+,n'})\}$ avec $Y_+ = X_+\beta^* + \epsilon_+$ **ET** ϵ et ϵ_+ indépendants.

Erreur de généralisation

On choisirait alors le modèle qui minimise l'erreur de généralisation.

Generalization error, erreur de généralisation

$$\mathbb{E}(\|Y_+ - \hat{Y}_+^{(\lambda)}\|^2) = \mathbb{E}(\|Y_+ - X_+\hat{\beta}^{(\lambda)}\|^2) = n'\sigma_\star^2 + \mathbb{E}(\|X_+\hat{\beta}^{(\lambda)} - X_+\beta^*\|^2).$$

Petit rappel, le risque quadratique est défini par

$$\mathbb{E}(\|\hat{Y}^{(\lambda)} - X\beta^*\|^2) = \mathbb{E}(\|X\hat{\beta}^{(\lambda)} - X\beta^*\|^2).$$

Estimation de l'erreur de généralisation

On estime l'erreur de généralisation $\mathbb{E}(\|Y_+ - \hat{Y}_+^{(\lambda)}\|^2) = \mathbb{E}(\|Y_+ - X_+\hat{\beta}^{(\lambda)}\|^2)$ à partir de l'échantillon $\mathcal{S}_T = \{(Y_{+,1}, X_{+,1}), \ldots, (Y_{+,n'}, X_{+,n'})\}$ par

$$\frac{1}{n'}\sum_{i=1}^{n'}(Y_{+,i} - X_{+,i}\hat{\beta}^{(\lambda)})^2,$$

où $\hat{\beta}^{(\lambda)}$ a été calculé sur l'échantillon d'apprentissage $\mathcal{S}_L$ et pour la valeur λ du paramètre de régularisation.

En pratique

Même en l'absence de données de validation (situation fréquente en pratique), on peut vouloir créer des données qui "ressemblent" à des données de test pour appliquer ce qui précède. Il y a deux grandes classes de méthodes : la cross-validation et le bootstrap.

Leave-one-out (jackknife)

Chaque observation joue à tour de rôle le rôle d'échantillon de validation.

Estimation de l'erreur de généralisation par leave-one-out

$$\mathbb{E}(\|\widehat{Y_+ - \hat{Y}_+^{(\lambda)}}\|^2)_{loo} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i\hat{\beta}_{(-i)}^{(\lambda)})^2,$$

où $\hat{\beta}_{(-i)}^{(\lambda)}$ a été calculé sur l'échantillon $\mathcal{S}_L \backslash (Y_i, X_i)$ et pour la valeur λ du paramètre.

K-fold cross-validation

On découpe l'échantillon initial en K sous-ensembles pour obtenir la partition $\mathcal{S}_L = \mathcal{S}_{L,1} \cup \ldots \cup \mathcal{S}_{L,K}$. Dans le cas, où $n = Kn_K$, on tire aléatoirement et sans remise dans $\mathcal{S}_L$ pour former les $\mathcal{S}_{L,k}$.

Estimation de l'erreur de généralisation par K-fold cross-validation

$$\widehat{eg(\lambda)}_{Kfold-cv} = \mathbb{E}(\| \widehat{Y_+ - \hat{Y}_+^{(\lambda)}} \|^2)_{Kfold-cv} = \frac{1}{n_K K} \sum_{k=1}^{K} \sum_{i=1}^{n_K} (Y_{k,i} - X_{k,i} \hat{\beta}_{(-k)}^{(\lambda)})^2,$$

où $\hat{\beta}_{(-k)}^{(\lambda)}$ a été calculé sur l'échantillon $\mathcal{S}_L \backslash \mathcal{S}_{L,k}$ et pour la valeur λ du paramètre.

On choisit alors

$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \widehat{eg(\lambda)}_{Kfold-cv}$$

Propriété sur les variables corrélées

Théorème - Zou et Hastie (2005)

On suppose que toutes les variables ont été renormalisées. Pour deux variables X^j et $X^{j'}$, on a

$$\frac{1}{\|Y\|}|\hat{\beta}_\lambda^{j,\text{ridge}} - \hat{\beta}_\lambda^{j',\text{ridge}}| \leq \frac{1}{\lambda}\sqrt{2(1 - cor(X^j, X^{j'}))}.$$

Régression sur PC

En considérant X^S, on peut décrire l'ACP via la formule

$$(X^S)^\top X^S = VD^2V^\top$$

les colonnes de V sont les vecteurs propres de X^S et également les composantes principales. La PCR consiste à régresser Y sur les $m < p$ premières composantes principales.

Lasso et elastic-net

Introduction

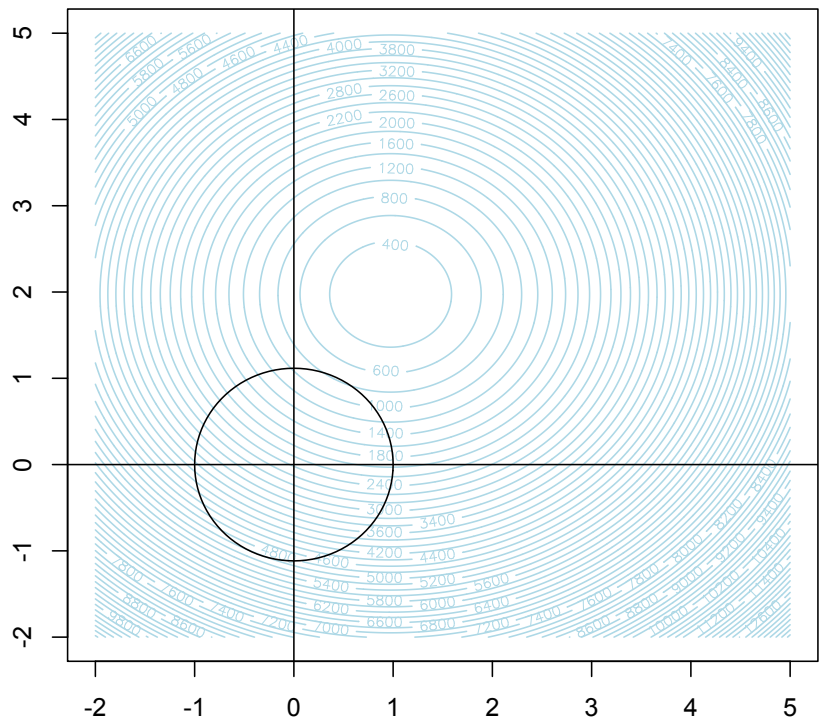
La régression ridge consiste à résoudre

$$\hat{\beta}_\lambda^{\text{ridge}} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \, \|Y - X\beta\|^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

ce qui est équivalent à

$$\hat{\beta}_\lambda^{\text{ridge}} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \, \|Y - X\beta\|^2 \text{ s.c. } \sum_{j=1}^{p} \beta_j^2 < t.$$

Introduction



Introduction

En reprenant les idées de la construction du Cp de Mallows, on montre que

$$\hat{m}_{\text{Mallows}} = \underset{m}{\text{argmin}}\, Cp(m) \simeq \hat{\beta}_{\lambda}^{\text{Mallows}} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \|Y - X\beta\|^2 + \lambda \|\beta\|_0.$$

C'est équivalent à

$$\hat{\beta}_{\lambda}^{\text{Mallows}} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \|Y - X\beta\|^2 \text{ s.c. } \|\beta\|_0 < t.$$

Le lasso

Introduit en 1996 par Tibshirani, la lasso peut être vu comme un intermédiaire entre la régression ridge et la sélection ℓ_0.

Pénalité lasso

$$\begin{aligned}\hat{\beta}_\lambda^{\text{lasso}} &= \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \, \|Y - X\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \\ &= \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \, \|Y - X\beta\|^2 + \lambda \|\beta\|_1.\end{aligned}$$

Le lasso

Solution lasso sur design orthogonal

Quand $X^\top X = Id_p$,

$$\hat{\beta}^{(\lambda)}_{\text{lasso}} = \mathcal{S}_{\lambda/2}(Y^\top X),$$

avec $\mathcal{S}_\tau(x) = sign(x)(|x| - \tau)_+$.

L'elastic net

Elastic net

$$\hat{\beta}_\lambda^{\text{en}} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \, \| Y - X\beta \|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$$

Attention dans `glmnet`, l'elastic-net est défini par

$$\hat{\beta}_\lambda^{\text{en}} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \, \| Y - X\beta \|^2 + \lambda \Big(\alpha \|\beta\|_1 + \frac{1-\alpha}{2} \lambda_2 \|\beta\|_2^2 \Big)$$

Comparaison sur simulation

On utilise la librairie `glmnet` On veut comparer sur le modèle

```
n=50
p=30
puis=c(0:9)
puis_petit=c(10:28)
beta=c(0.95^(puis),gamma^puis_petit)
X=cbind(matrix(rnorm(n*(p-1)),ncol=(p-1)))
epsilon=rnorm(n)
beta0=2
Y=beta0+X%*%beta+epsilon
```

les différentes procédures d'estimation (Cp de Mallows, ridge, lasso, elastic net) , en terme

- de pouvoir de sélection (taille du modèle sélectionné, sensibilité (TP/(TP+FN)), spécificité (TN/(FP+TN)))
- d'erreur d'estimation $\|\hat{\beta} - \beta^*\|$
- d'erreur de prévision $\|X\hat{\beta} - X\beta^*\|$

On peut faire varier la corrélation entre les covaraibles via la fonction "covariates.matrix".