Apprentissage Statistique : Partie III

Magalie Fromont (Université Rennes 2)

Ensai, 2015 - 2016

Plan du cours

Plan du cours

Modèle de régression linéaire

Rappel des notations

Données observées de type entrée-sortie : $d_1^n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ avec $x_i \in \mathbb{R}^p$), $y_i \in \mathbb{R}$ pour $i = 1 \ldots n$.

Objectif : prédire la sortie y associée à une nouvelle entrée x, sur la base de d_1^n (problème de régression réelle).

Modèle non paramétrique :
On suppose que d_1^n est l'observation d'un n-échantillon $D_1^n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ d'une loi conjointe P sur $\mathbb{R}^p \times \mathbb{R}$, totalement inconnue.
On suppose que x est une observation de la variable X, (X, Y) étant un couple aléatoire de loi conjointe P indépendant de D_1^n.

Modèle de régression linéaire

Modèle de régression linéaire : $x_i = (x_i^1, \ldots, x_i^p)'$. On suppose que :

$$Y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \varepsilon_i, \ 1 \leq i \leq n,$$

où

- $\beta_0, \beta_1, \ldots, \beta_p$ sont les paramètres du modèle (à estimer),
- les variables ε_i vérifient :
  $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0 \ \forall i \neq j$, $\mathrm{var}(\varepsilon_i) = \sigma^2$,
- les variables sont supposées gaussiennes pour l'inférence statistique (intervalles de confiance, tests, p valeurs...)

Modèle de régression linéaire

Modèle de régression linéaire sous forme matricielle :

$$Y = \mathbb{X}\beta + \varepsilon,$$

où

$$\mathbb{X} = \begin{pmatrix} x_1^0 & x_1^1 & x_1^2 & . & x_1^p \\ x_2^0 & x_2^1 & x_2^2 & . & x_2^p \\ . & . & . & . & . \\ x_n^0 & x_n^1 & x_n^2 & . & x_n^p \end{pmatrix}, \ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_p \end{pmatrix}, \ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{pmatrix},$$

$$x^0 = \begin{pmatrix} x_1^0 \\ x_2^0 \\ . \\ . \\ x_n^0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ . \\ . \\ 1 \end{pmatrix}, \ x^j = \begin{pmatrix} x_1^j \\ x_2^j \\ . \\ . \\ x_n^j \end{pmatrix} \ (j = 1 \dots p).$$

Estimation des paramètres par MCO

β est estimé par la méthode des **moindres carrés ordinaires** : l'estimateur $\hat{\beta}$ minimise pour $\beta \in \mathbb{R}^{p+1}$ le critère empirique :

$$\sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 x_i^1 - \beta_2 x_i^2 - \ldots - \beta_p x_i^p \right)^2 .$$

$\hookrightarrow$ minimisation de risque empirique pour la perte quadratique, sur l'ensemble des règles de régression linéaires, de la forme $f(x) = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \ldots + \beta_p x^p$.

- **Estimation des paramètres** : $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$
- **Vecteur des valeurs ajustées** : $\hat{Y} = \mathbb{X}\hat{\beta} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$
- **Vecteur des résidus** : $\hat{\varepsilon} = Y - \hat{Y}$
- **Somme des carrés résiduelle** : $SCR = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$
- **Somme des carrés totale** : $SCT = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$
- **Somme des carrés expliquée** : $SCE = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$
- **Équation d'analyse de la variance** : $SCT = SCE + SCR$
- **Coefficient de détermination** : $R^2 = 1 - SCR/SCT$

Sélection de variables

Une question :

Quel choix pour la matrice de plan d'expérience $\mathbb{X}$?
Multicolinéarité des variables explicatives, principe de parcimonie
$\Rightarrow$ critères de sélection de variables.
Impossibilité d'étudier TOUS les modèles linéaires possibles
$\Rightarrow$ méthodes algorithmiques de sélection.

Notations :

Pour un sous-ensemble ξ de $\{0, 1, \ldots, p\}$, on note $|\xi|$ son cardinal,
$\mathbb{X}_\xi$ la matrice dont les vecteurs colonnes sont les x^j pour $j \in \xi$,
β_ξ le vecteur composé des β_j, pour $j \in \xi$,
$[\hat{\beta}]_\xi$ le vecteur composé des $\hat{\beta}_j$ pour $j \in \xi$,
$\widehat{\beta_\xi} = (\mathbb{X}'_\xi \mathbb{X}_\xi)^{-1} \mathbb{X}'_\xi Y$ l'estimateur des MCO de β_ξ dans

$$(M_\xi)\ Y = \mathbb{X}_\xi \beta_\xi + \eta.$$

Sélection de variables

Qualité de l'estimation, erreur quadratique moyenne (EQM)

En général, $\widehat{\beta}_\xi (\neq [\hat{\beta}]_\xi)$ est un estimateur biaisé de β_ξ.
Mais $\mathrm{Var}([\hat{\beta}]_\xi) - \mathrm{Var}(\widehat{\beta}_\xi)$ et $\mathrm{Var}(\mathbb{X}\hat{\beta}) - \mathrm{Var}(\mathbb{X}_\xi \widehat{\beta}_\xi)$ semi-déf. ≥ 0

Définition

L'erreur quadratique moyenne associée à (M_ξ) est définie par

$$\mathrm{EQM}(\xi) = \mathbb{E}\left[\|X\beta - \mathbb{X}_\xi \widehat{\beta}_\xi\|^2\right].$$

Proposition

$$\begin{aligned}\mathrm{EQM}(\xi) &= \|X\beta - \mathbb{E}[\mathbb{X}_\xi \widehat{\beta}_\xi]\|^2 + \mathbb{E}\left[\|\mathbb{X}_\xi \widehat{\beta}_\xi - \mathbb{E}[\mathbb{X}_\xi \widehat{\beta}_\xi]\|^2\right] \\ &= \|(I - \Pi_{\mathbb{X}_\xi})X\beta\|^2 + |\xi|\sigma^2.\end{aligned}$$

Compromis biais-variance idéal $\Rightarrow \xi^*$ idéal (oracle) inconnu !

Sélection de variables

Qualité de la prédiction, erreur quadratique moyenne de prédiction (EQMP)

Soit $x_{n+1,\xi}$ une nouvelle valeur des variables explicatives de (M_ξ).

Définition

L'erreur quadratique moyenne de prédiction associée à (M_ξ) est définie par

$$\mathrm{EQMP}(\xi) = \mathbb{E}\left[(Y_{n+1} - x_{n+1,\xi}\widehat{\beta}_\xi)^2\right].$$

Problème : $\mathrm{EQM}(\xi)$ et $\mathrm{EQMP}(\xi)$ sont inconnues $\Rightarrow$ Nécessité de construire des critères de sélection de ξ accessibles, ou d'estimer $\mathrm{EQM}(\xi)$ ou $\mathrm{EQMP}(\xi)$.

Sélection de variables

Cadre général (conditions standards)

Somme des carrés résiduelle, cœfficient de détermination R^2

Notation : $R^2(\xi) = 1 - \frac{SCR(\xi)}{SCT}$ associé à (M_ξ)

Proposition

Si $\xi^- \subset \xi$, tous deux contenant ou pas l'indice de la constante, alors

$$R^2(\xi) \geq R^2(\xi^-)$$

$\hookrightarrow$ Minimiser la SCR ou maximiser le R^2 conduit à choisir le modèle le plus complexe : la SCR et le R^2 sont de mauvais critères de sélection de variables (mais ils restent utiles pour choisir entre deux modèles avec le même nombre de variables explicatives).

$\hookrightarrow$ Deux possibilités : corriger le R^2 ou décider si l'augmentation du R^2 est statistiquement significative (tests sous hypothèse gaussienne).

Cœfficient de détermination ajusté R_a^2

Définition

Le cœfficient de détermination ajusté de (M_ξ) est défini par

$$R_a^2(\xi) = 1 - \frac{SCR(\xi)/(n-|\xi|)}{SCT/(n-1)} = 1 - \frac{n-1}{n-|\xi|}\left(1 - R^2(\xi)\right).$$

Remarque : $R_a^2(\xi) < R^2(\xi)$ dès que $|\xi| \geq 2$ ($R_a^2(\xi)$ peut même être négatif).

Sélection de variables

Cadre général (conditions standards)

Critère du C_p de Mallows

$$\mathrm{EQM}(\xi) = \mathbb{E}[SCR(\xi)] - n\sigma^2 + 2|\xi|\sigma^2.$$

Définition (Mallows, 1973)

Le critère du C_p est défini par $C_p(\xi) = SCR(\xi)/\widehat{\sigma^2} - n + 2|\xi|$.

Proposition

Si ξ ne dépend pas de Y, $\widehat{\sigma^2} C_p(\xi)$ estime sans biais $\mathrm{EQM}(\xi)$.

$\hookrightarrow$ À utiliser sur un jeu de données "isolé".

$\hookrightarrow$ Règle usuelle : on retient (M_ξ) si $C_p(\xi) \leq |\xi|$.

$\hookrightarrow$ Défaut : biais de sélection

Estimation de l'EQMP par validation croisée ou bootstrap

Définition

L'estimateur de $\mathrm{EQMP}(\xi)$ par validation Leave One Out est défini par $\mathrm{PRESS} = \sum_{i=1}^{n} \left(Y_i - x_{i,\xi} \widehat{\beta}_{\xi(i)} \right)^2$, où $\widehat{\beta}_{\xi(i)}$ est l'estimateur des MCO de β_ξ obtenu après suppression de la $i$ème donnée.

Autres estimations possibles :

- bootstrap par paires
- (wild) bootstrap des résidus.

Tests de validité de sous-modèles

Choix entre deux sous-modèles emboîtés l'un dans l'autre, l'un (M_ξ) défini par ξ de cardinal $|\xi| \geq 2$, validé, et l'autre, (M_{ξ^-}), défini par ξ^- tel que $\xi^- \subset \xi$ et $|\xi^-| = |\xi| - 1$.

Statistique de test : $F(Y) = \frac{SCR(\xi^-) - SCR(\xi)}{SCR(\xi)/(n-|\xi|)}$.

Sous l'hypothèse de validité de (M_{ξ^-}), $F(Y) \sim \mathcal{F}(1, n - |\xi|)$.

On rejette donc (M_{ξ^-}) au profit de (M_ξ) au niveau α si

$$F(y) \geq f_{1,n-|\xi|,(1-\alpha)}$$

Variante : réduction par la variance estimée dans le modèle complet (plus pratique pour une méthode ascendante !).

Critères de vraisemblance pénalisée

Sous l'hypothèse gaussienne, la log-vraisemblance est donnée par
$\ln \mathcal{L}(Y, \beta, \sigma^2) = -\frac{n}{2} \ln \sigma^2 - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \| Y - \mathbb{X}\beta \|^2$.

Estimateurs du maximum de vraisemblance dans (M_ξ) :
$\tilde{\beta}_\xi = \widehat{\beta}_\xi$ et $\widetilde{\sigma^2}_\xi = SCR(\xi)/n$.

$$\ln \mathcal{L}(Y, \widehat{\beta}_\xi, \widetilde{\sigma^2}_\xi) = -\frac{n}{2} \ln \frac{SCR(\xi)}{n} - \frac{n}{2}(1 + \ln(2\pi)).$$

Choisir un modèle maximisant la vraisemblance reviendrait donc à choisir un modèle (M_ξ) ayant la plus petite $SCR(\xi)$ donc ayant le plus de variables
$\Rightarrow$ nécessité de pénaliser les modèles ayant un grand nombre de variables

Sélection de variables

Cadre gaussien

Critères de vraisemblance pénalisée

On cherche un modèle (M_ξ) minimisant un critère de la forme :

$$-2\ln \mathcal{L}(Y, \widehat{\beta}_\xi, \widetilde{\sigma^2}_\xi) + |\xi| f(n) = n \ln \frac{SCR(\xi)}{n} + n(1 + \ln(2\pi)) + |\xi| f(n),$$

où $(|\xi| + 1) f(n)$ est un terme de pénalité positif (croissant avec n).

Définitions (Akaike,1973) (Schwarz, 1978)

Le critère AIC (Akaike's Information Criterion) correspond à $f(n) = 2$:
$AIC(\xi) = n \ln \frac{SCR(\xi)}{n} + n(1 + \ln(2\pi)) + 2|\xi|$.
Le critère BIC (Bayesian Information Criterion) correspond à $f(n) = \ln n$:
$BIC(\xi) = n \ln \frac{SCR(\xi)}{n} + n(1 + \ln(2\pi)) + |\xi| \ln n$.

- Si $n \geq 8$, la pénalité du critère BIC est plus lourde que celle de l'AIC : les modèles choisis par le BIC ont moins de variables explicatives.
- Lorsque $n \to +\infty$, la probabilité de sélectionner un modèle exact par minimisation du critère BIC tend vers 1.

Sélection de variables

Cadre gaussien

Liens avec les tests de validité de sous-modèles

Critère du R_a^2

$R_a^2(\xi^-) < R_a^2(\xi) \Leftrightarrow (n - |\xi|)\frac{SCR(\xi^-) - SCR(\xi)}{SCR(\xi)} > 1 \Leftrightarrow F(Y) > 1.$

$\hookrightarrow$ Test avec valeur critique 1 au lieu de $f_{1,n-|\xi|,(1-\alpha)}$ (qui est > 3.84).

Critère du C_p

Si $\tilde{F}(Y)$ est la statistique de test de validité de sous-modèle variante (réduction par $\widehat{\sigma^2}$), $C_p(\xi^-) > C_p(\xi) \Leftrightarrow \tilde{F}(Y) > 2$.

Critères de vraisemblance pénalisée

$-2 \ln \mathcal{L}(Y, \hat{\beta}_{\xi^-} \widetilde{\sigma^2}_{\xi^-}) + |\xi^-| f(n) > -2 \ln \mathcal{L}(Y, \hat{\beta}_{\xi} \widetilde{\sigma^2}_{\xi}) + |\xi| f(n) \Leftrightarrow F(Y) > (n - |\xi|)\left(e^{2f(n)/n} - 1\right)$.

Sélection de variables

Méthodes algorithmiques de sélection

Méthode de recherche exhaustive

Nombre de sous-modèles possibles pour $p+1$ variables explicatives : $2^{p+1} \Rightarrow$ coût algorithmique prohibitif pour p grand, même modéré.
Aucun sens pour l'utilisation des tests de validité de sous-modèles.

Méthode de recherche ascendante (forward) On part du modèle constant (aucune variable explicative), et on ajoute de manière séquentielle les variables qui permettent d'optimiser le critère (maximiser le R_a^2, minimiser les autres critères), ou d'accepter la validité du modèle (test).

Méthode de recherche descendante (backward) Même principe, mais on part du modèle complet et on élimine une à une les variables.

Méthode de recherche progressive (stepwise) Algorithme mixte qui ajoute ou supprime une variable du modèle à chaque étape.

$\hookrightarrow$ Sélection et donc estimation finalement peu satisfaisantes, instables, surtout inadaptées au cas $p \geq n$...

Plan du cours

Régression linéaire régularisée

Préambule : les variables explicatives x^j sont standardisées, puis on suppose que $\beta = 0$ et $\bar{Y} = 0$ (modèle sans constante). Concrètement, β_0 est estimé par $\bar{Y}$ et Y_i remplacé par $Y_i - \bar{Y}$.

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

$$\begin{aligned} \hat{\beta}_\lambda &\in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (Y_i - \beta_1 x_i^1 - \beta_2 x_i^2 - \ldots - \beta_p x_i^p)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \\ &\in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|Y - \mathbb{X}\beta\|_2^2 + \lambda \|\beta\|_q^q, \end{aligned}$$

λ étant un paramètre positif, appelé paramètre de régularisation.

- $q = 2 \Rightarrow$ régression ridge
- $q = 1 \Rightarrow$ régression LASSO

Régression ridge

Définition et premières propriétés

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

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

Proposition

- Minimiser $\| Y - \mathbb{X}\beta \|_2^2 + \lambda \|\beta\|_2^2$ en $\beta \in \mathbb{R}^p$ est équivalent à minimiser $\| Y - \mathbb{X}\beta \|_2^2$ sous la contrainte $\|\beta\|_2^2 \leq r(\lambda)$, où r est bijective.
- La matrice $(\mathbb{X}'\mathbb{X} + \lambda I_p)$ est toujours définie positive, donc inversible, et $\hat{\beta}_\lambda^{ridge} = (\mathbb{X}'\mathbb{X} + \lambda I_p)^{-1}\mathbb{X}'Y$.

L'estimateur ridge est biaisé, son biais est égal à $-\lambda(\mathbb{X}'\mathbb{X} + \lambda I_p)^{-1}\beta$, sa variance à $\sigma^2(\mathbb{X}'\mathbb{X} + \lambda I_p)^{-1}\mathbb{X}'\mathbb{X}(\mathbb{X}'\mathbb{X} + \lambda I_p)^{-1}$.

Noter que les valeurs propres de $(\mathbb{X}'\mathbb{X} + \lambda I_p)$ sont plus élevées que celles de $\mathbb{X}'\mathbb{X}$, donc la variance de $\hat{\beta}_\lambda^{ridge}$ sera plus faible que celle de $\hat{\beta}$.

Régression ridge

Ajustement du paramètre de régularisation

- Lorsque $\lambda = 0$, $\hat{\beta}_\lambda^{ridge} = \hat{\beta}$.
- Lorsque $\lambda \to +\infty$, $\hat{\beta}_\lambda^{ridge} = 0$.
- Lorsque λ augmente, le biais de $\hat{\beta}_\lambda^{ridge}$ a tendance à augmenter et la variance à diminuer $\Rightarrow$ compromis ?

Ajustement de λ

On appelle chemin de régularisation de la régression ridge l'ensemble des fonctions $\lambda \mapsto (\hat{\beta}_\lambda^{ridge})_j$ pour $j = 1, \ldots, p$.

En pratique, on constate que le chemin de régularisation de la régression ridge est continu, ne permettant pas un ajustement aisé de λ.

$\hookrightarrow$ Choix usuel par validation croisée V fold sur une grille finie de valeurs de $\lambda > 0$.

Régression ridge

Rétrécissement

Décomposition en valeurs singulières (SVD) de $\mathbb{X}$ ($p \geq n$) : $\mathbb{X} = UDV'$, où

- U de taille $n \times n$, de vecteurs colonnes $u^1, \ldots, u^n$,
- D est une matrice "diagonale" de taille $n \times p$ dont tous les éléments vérifient $d_1 \geq \ldots \geq d_p \geq 0$,
- V est de taille $p \times p$, de vecteurs colonnes $v^1, \ldots, v^p$,
- U et V sont orthogonales : $UU' = U'U = I_n$, $VV' = V'V = I_p$.

On a alors $\mathbb{X}\hat{\beta}_\lambda^{ridge} = UD(D'D + \lambda I_p)^{-1}D'U'Y = \sum_{j=1}^{p} \left(\frac{d_j^2}{d_j^2+\lambda} \right) \langle u^j, Y \rangle u^j$.

Pour l'estimateur des MCO ($\lambda = 0$) : $\mathbb{X}\hat{\beta} = \sum_{j=1}^{p} u^j \langle u^j, Y \rangle$.

$\hookrightarrow \langle u^j, Y \rangle$ correspond à la $j$ième composante de Y dans la base des u^j.

$\hookrightarrow$ La régression ridge multiplie cette composante par $d_j^2 / \left(d_j^2 + \lambda \right) \in]0, 1[$: on dit qu'elle est "rétrécie" (shrinked).

Régression ridge

Rétrécissement et ACP

Les x^j étant centrées, $\mathbb{X}'\mathbb{X}/n = VD'DV'/n$ en est la matrice de variance-covariance empirique.

Rappel : si v est un vecteur de $\mathbb{R}^p$ de norme 1, $\mathrm{Var}(\mathbb{X}v) = v'\mathbb{X}'\mathbb{X}v$ est maximale pour $v = v^1$ et vaut $d_1^2 \Rightarrow z^1 = \mathbb{X}v^1$ est la première composante principale de $\mathbb{X}$.
Les vecteurs propres orthogonaux $v^1, \ldots, v^p$ sont les directions principales (ou directions de Karhunen Loeve) de $\mathbb{X}$.
Les $z^j = \mathbb{X}v^j = UDV'v^j = d_j u^j$ sont les composantes principales de $\mathbb{X}$.

$\hookrightarrow$ La régression ridge rétrécit peu les premières composantes principales (pour lesquelles d_j est grand), et davantage les dernières.

Régression LASSO

Définition et premières propriétés

L'estimateur LASSO (Least Absolute Selection and Shrinkage Operator) (Tibshirani 1996) est défini pour $\lambda > 0$ par

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

La fonction $\mathcal{L} : \beta \mapsto \|Y - \mathbb{X}\beta\|_2^2 + \lambda\|\beta\|_1$ est convexe, non différentiable.
La solution du problème de minimisation peut ne pas être unique.
Le vecteur des valeurs ajustées en résultant $\mathbb{X}\hat{\beta}_\lambda^{lasso}$, lui, est toujours unique.

Proposition

Minimiser $\|Y - \mathbb{X}\beta\|_2^2 + \lambda\|\beta\|_1$ en $\beta \in \mathbb{R}^p$ est équivalent à minimiser $\|Y - \mathbb{X}\beta\|_2^2$ sous une contrainte de la forme $\|\beta\|_1 \leq \hat{R}_\lambda$ ($\hat{R}_\lambda = \|\hat{\beta}_\lambda^{lasso}\|_1$).

Régression LASSO

Parcimonie et rétrécissement

L'estimateur LASSO a l'avantage d'avoir un certain nombre de cœfficients nuls lorsque λ est suffisamment grand
$\Rightarrow$ Estimateur parcimonieux / sélection de variables induite.

Explication graphique

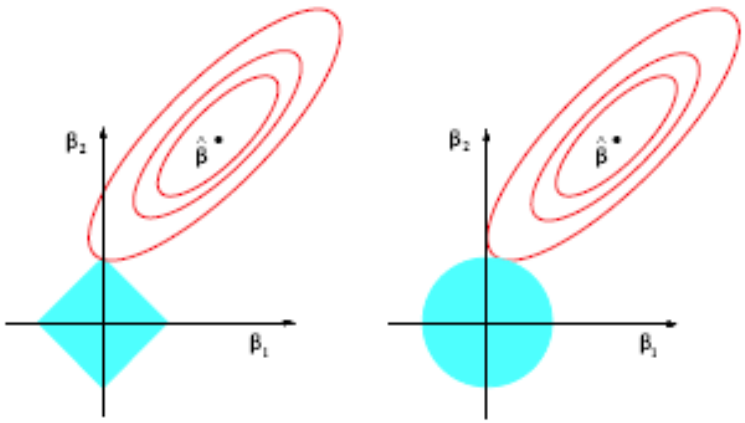

Figure 3.12: *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$, *respectively, while the red ellipses are the contours of the least squares error function.*

Régression LASSO

Parcimonie et rétrécissement

Proposition : condition d'optimalité du premier ordre

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

Cas orthonormal : x^j orthonormés, $\mathbb{X}'\mathbb{X} = I_p$ (et $p \leq n$).
Pour $\left[\hat{\beta}_\lambda^{lasso}\right]_j \neq 0$, $\left[\hat{\beta}_\lambda^{lasso}\right]_j = (x^j)'Y - \lambda signe\left(\left[\hat{\beta}_\lambda^{lasso}\right]_j\right)/2$, d'où
$signe\left(\left[\hat{\beta}_\lambda^{lasso}\right]_j\right) = signe\left((x^j)'Y\right)$ et $\left[\hat{\beta}_\lambda^{lasso}\right]_j = (x^j)'Y - \lambda signe\left((x^j)'Y\right)/2$.
$\Rightarrow$ Si $\left|(x^j)'Y\right| \leq \lambda/2$, $\left[\hat{\beta}_\lambda^{lasso}\right]_j = 0$.

Proposition

Si les x^j sont orthonormés, $\left[\hat{\beta}_\lambda^{lasso}\right]_j = (x^j)'Y\left(1 - \lambda/\left(2\left|(x^j)'Y\right|\right)\right)_+$.

$\hookrightarrow$ parcimonie et rétrécissement des MCO dans le cas orthogonal, mais attention : pas de formule explicite dans le cas non orthogonal !

Régression LASSO

Sélection de modèles

Pour $\beta \in \mathbb{R}^{p_n}$, $S_n(\beta) = \{j \in \{1, \dots, p_n\} \ / \ \beta_j \neq 0\}$ de cardinal q_n.
$\mathbb{X}_S$ = matrice composée des x^j, $j \in S$.
On suppose que $\|\mathbb{X}'_{S_n(\beta)^c}\mathbb{X}_{S_n(\beta)}(\mathbb{X}'_{S_n(\beta)}\mathbb{X}_{S_n(\beta)})^{-1} sign(\beta_{S_n(\beta)})\|_\infty \leq 1 - \gamma$ pour $\gamma \in]0, 1[$ (irreprésentabilité), que $\eta' \mathbb{X}'_{S_n(\beta)}\mathbb{X}_{S_n(\beta)}\eta \geq M_1 \ \forall \|\eta\|_2 = 1$, $q_n = O(n^{c_1})$, $n^{\frac{1-c_2}{2}} min_{i \in S_n(\beta)}|\beta_i| \geq M_2$, et $\mathbb{E}[\varepsilon_i^{2k}] < +\infty$, pour $0 \leq c_1 < c_2 \leq 1$, $M_1, M_2 > 0$, $k \in \mathbb{N}^*$.

Théorème (Zhao and Yu, 2006)

Si les x^j sont normés et $\lambda_n/\sqrt{n} = o(n^{\frac{c_2-c_1}{2}})$ et $(\lambda_n/\sqrt{n})^{2k}/p_n \to +\infty$, alors $P(sign(\hat{\beta}^{lasso}_{\lambda_n}) = sign(\beta)) \geq 1 - O(p_n n^k/\lambda_n^{2k}) \to 1$.

$\hookrightarrow$ Condition d'irreprésentabilité = les variables non influentes ($\beta_j = 0$) ne sont pas trop corrélées à celles qui le sont.

$\hookrightarrow$ Sélection en général d'un trop grand nombre de variables $\Rightarrow$ Bolasso ?

$\hookrightarrow$ Attention : si $\frac{q_n}{n} \ln\left(\frac{p_n}{q_n}\right) > \frac{1}{2}$, aucune sélection pertinente possible.

Régression LASSO

Borne de risque / inégalité oracle

Théorème

Si les x^j sont normés, et $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Pour tout $L > 0$, si $\lambda = 3\sigma\sqrt{2\ln p + 2L}$, avec probabilité au moins égale à $1 - e^{-L}$,

$$\left\|\mathbb{X}\left(\hat{\beta}_\lambda^{lasso} - \beta\right)\right\|_2^2 \leq inf_{\beta' \neq 0} \left\{ \left\|\mathbb{X}\left(\beta' - \beta\right)\right\|_2^2 + \frac{18\sigma^2(L + \ln p)}{\kappa^2(\beta')} \sum_{j=1}^{p} \mathbb{1}_{\beta'_j \neq 0} \right\},$$

où $\kappa(\beta')$ est ce qu'on appelle une constante de compatibilité, mesurant le manque d'orthogonalité des x^j, dépendant de β'.

Régression LASSO

Ajustement du paramètre de régularisation

- Lorsque $\lambda = 0$, $\hat{\beta}_\lambda^{lasso} = \hat{\beta}$.
- Lorsque $\lambda \to +\infty$, $\hat{\beta}_\lambda^{lasso} = 0$.

On appelle chemin de régularisation de la régression lasso l'ensemble des fonctions $\lambda \mapsto (\hat{\beta}_\lambda^{lasso})_j$ pour $j = 1, \ldots, p$.

Ajustement de λ

Le chemin de régularisation de la régression lasso est linéaire par morceaux (c.f. Condition d'optimalité du premier ordre), avec changements en un nombre fini de points : les points de transition.

$\hookrightarrow$ Choix pour un nombre donné de cœfficients non nuls souhaités.

$\hookrightarrow$ Prédiction : choix sur les points de transition ? Choix usuel par validation croisée V fold sur une grille finie de valeurs de $\lambda > 0$.

$\hookrightarrow$ Base de l'algorithme LARS (package R `lars`), qui calcule les estimateurs LASSO en les points de transition et qui interpole linéairement

Régression LASSO

Algorithmes de calcul

Descente de coordonnées

Algorithme Coordinates descent

Variable

$\mathbb{X}$, Y, λ : *matrice* $n \times p$, *vecteur de taille* n, *réel* > 0

Début

Initialiser $\beta = \beta_{in}$

Répéter

Pour j **variant de** 1 **à** p

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

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

FinPour

Jusque Convergence de β

Retourner β

Fin

Régression LASSO

Algorithmes de calcul

Descente de coordonnées

Proposition

La fonction $\beta_j \mapsto \mathcal{L}(\beta)$ est minimum en $\beta_j = R_j\big(1 - \lambda/(2|R_j|)\big)_+$ avec $R_j = (x^j)'(Y - \sum_{k \neq j} \beta_k x^k)$.

Pour une évaluation sur une grille $\Lambda = \{\lambda_1, \ldots, \lambda_T\}$ de valeurs rangées par ordre décroissant, calculer $\hat{\beta}^{lasso}_{\lambda_1}$ avec $\beta_{in} = 0$, puis $\hat{\beta}^{lasso}_{\lambda_2}$ avec $\beta_{in} = \hat{\beta}^{lasso}_{\lambda_1}$...

Le coin du UseR : package `glmnet`

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

Elastic Net

La méthode Elastic-Net combine les atouts des méthodes ridge et LASSO.

En particulier, elle pallie le défaut de l'estimation LASSO lorsque les X^j sont fortement corrélées.

L'estimateur Elastic-Net (Zou, Hastie 2005) est défini pour $\lambda_1, \lambda_2 > 0$ par

$$\hat{\beta}^{EN}_{\lambda_1,\lambda_2} \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|Y - \mathbb{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2.$$

Le coin du UseR : packages `glmnet`, `elasticnet`.

Plan du cours

1. Régression linéaire : rappels
2. Régression linéaire régularisée : régressions ridge, LASSO, Elastic Net
3. Variantes du LASSO
   - Dantzig Selector
   - Adaptive LASSO
   - Group LASSO
   - Sparse-Group LASSO
   - Fused LASSO
   - Régression logistique (Group) LASSO
4. Méthodes à noyaux pour la régression non linéaire

Dantzig Selector

Le Dantzig Selector (Candès, Tao 2007) est défini pour $\lambda > 0$ par

$$\hat{\beta}_\lambda^{DS} \in \text{argmin}_{\beta \in \mathbb{R}^p} \|\beta\|_1 \text{ s.c. } \|\mathbb{X}'(Y - \mathbb{X}\beta)\|_\infty \leq \lambda.$$

Si $\{\beta \in \mathbb{R}^p, \ \|\mathbb{X}'(Y - \mathbb{X}\beta)\|_\infty \leq \lambda\}$ n'est pas "parallèle" à la boule unité pour la norme $\mathbb{L}^1$, la solution est unique.

Propriétés similaires à celles de l'estimateur LASSO
(Bickel, Ritov, Tsybakov 2007)

Le coin du UseR : package `flare`

Adaptive LASSO

Problème :
estimateur LASSO biaisé $\Rightarrow$ pondération "data-driven" de la pénalité ?

L'estimateur adaptive LASSO (Zou 2006) est défini pour $\lambda, \nu > 0$ par

$$\hat{\beta}_{\lambda}^{a-lasso} \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|Y - \mathbb{X}\beta\|_2^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{\left|\left[\hat{\beta}_{\nu}^{Gauss-lasso}\right]_j\right|},$$

où $\hat{\beta}_{\nu}^{Gauss-lasso} = \Pi_{vect(X^j,\ j \in S(\hat{\beta}_{\lambda}^{lasso}))} Y$.

L'estimateur adaptive LASSO - le plus populaire sans doute - permet de retrouver le support du vrai modèle sous des conditions plus faibles que celles de l'estimateur LASSO.

Calcul par les algorithmes LARS et descente de coordonnées possible.

Le coin du UseR : packages `lqa`, `parcor`, `lassogrp`

Group LASSO

Sélection de groupes de variables / Group sparsity

Groupes de variables pré-définis par des sous-matrices $\mathbb{X}_l$ de $\mathbb{X}$, pour $l = 1 \dots L$, dont les colonnes correspondent à un groupe donné de colonnes de $\mathbb{X}$, et les vecteurs β_l de longueur n_l associés.

L'estimateur Group LASSO (Yuan, Lin 2007) est défini pour $\lambda > 0$ et $K_1, \dots, K_L$ matrices définies positives $n_l \times n_l$, par

$$\hat{\beta}_\lambda^{grp-lasso} \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|Y - \mathbb{X}\beta\|_2^2 + \lambda \sum_{l=1}^{L} \sqrt{\beta_l' K_l \beta_l}.$$

Cas particulier : $K_l = n_l I_{n_l}$, $\sqrt{\beta_l' K_l \beta_l} = \sqrt{n_l}\|\beta_l\|_2$.

Pour appliquer l'algorithme de descente de coordonnées, les $\mathbb{X}_l$ doivent être orthonormales : condition restrictive.

Le coin du UseR : algorithme LARS implémenté dans `lassogrp`, algorithme Blockwise-Majorization-Descent implémenté dans `gglasso`.

Sparse-Group LASSO

Sparse-Group sparsity

Groupes de variables pré-définis par des sous-matrices $\mathbb{X}_l$ de $\mathbb{X}$, pour $l = 1 \dots L$, dont les colonnes correspondent à un groupe donné de colonnes de $\mathbb{X}$, et les vecteurs β_l de longueur n_l associés.

L'estimateur Sparse-Group LASSO (Simon et al. 2013) est défini pour $\lambda, \mu > 0$ et $K_1, \dots, K_L$ matrices définies positives $n_l \times n_l$, par

$$\hat{\beta}^{SGL}_{\lambda} \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \| Y - \mathbb{X}\beta \|_2^2 + \lambda \sum_{l=1}^{L} \sqrt{\beta_l' K_l \beta_l} + \mu \|\beta\|_1.$$

Le coin du UseR : package `SGL`.

Fused LASSO

Variation sparsity

Situation où peu d'incréments $\beta_{j+1} - \beta_j$ sont non nuls.

L'estimateur Fused LASSO (Tibshirani et al. 2005) est défini pour $\lambda > 0$ par

$$\hat{\beta}_\lambda^{f-lasso} \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|Y - \mathbb{X}\beta\|_2^2 + \lambda \sum_{j=1}^{p-1} |\beta_{j+1} - \beta_j|.$$

$\hookrightarrow$ Idem au LASSO après un changement de variables : $Z^j = \sum_{k=j}^{p} X^k$.

Régression logistique (Group) LASSO

Rappel des notations

Données observées de type entrée-sortie : $d_1^n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ avec $x_i \in \mathbb{R}^p$), $y_i \in \{-1, 1\}$ pour $i = 1 \ldots n$.

Objectif : prédire la sortie y associée à une nouvelle entrée x, sur la base de d_1^n (problème de discrimination binaire).

Modèle non paramétrique :
On suppose que d_1^n est l'observation d'un n-échantillon $D_1^n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ d'une loi conjointe P sur $\mathbb{R}^p \times \{-1, 1\}$, totalement inconnue.
On suppose que x est une observation de la variable X, (X, Y) étant un couple aléatoire de loi conjointe P indépendant de D_1^n.

Régression logistique (Group) LASSO

Modèle de régression logistique :

$x_i = (x_i^1, \ldots, x_i^p)'$, $\pi(x_i) = P(Y_i = 1 | X_i = x_i)$

On suppose que :

$$\ln \frac{\pi(x_i)}{1 - \pi(x_i)} = \beta_0 + \beta_1 x_i^1 + \ldots + \beta_p x_i^p, \ 1 \leq i \leq n.$$

Log vraisemblance : $\mathcal{L}_n(Y_1, \ldots, Y_n, \beta_0, \beta) =$
$\sum_{i=1}^n \left(Y_i(\beta_0 + \beta_1 x_i^1 + \ldots + \beta_p x_i^p) - \ln \left(1 + \exp(\beta_0 + \beta_1 x_i^1 + \ldots + \beta_p x_i^p)\right)\right)$

Les variables qualitatives sont transformées en variables indicatrices (dummy variables).

Régression logistique (Group) LASSO

Des groupes de variables sont créés de la façon suivante : toutes les modalités d'une variable qualitative sont regroupées dans un même groupe, les variables quantitatives sont considérées individuellement.

Notations idem à la régression Group LASSO : $\mathbb{X}_l$ de taille $n \times n_l$, pour $l = 1 \dots L$, β_l de longueur n_l.

L'estimateur (Group) LASSO logistique (Lokhorst 1999, Meier, van de Geer, Bühlmann 2008) est défini pour $\lambda > 0$ par

$$\hat{\beta}_\lambda^{logGL} \in \text{argmin}_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} - \mathcal{L}_n(Y_1, \dots, Y_n, \beta_0, \beta) + \lambda \sum_{l=1}^{L} \sqrt{n_l} \|\beta_l\|_2.$$

Algorithme de descente de coordonnées par blocs.

Le coin du UseR : package `grplasso`.

Plan du cours

1. Régression linéaire : rappels
2. Régression linéaire régularisée : régressions ridge, LASSO, Elastic Net
3. Variantes du LASSO
4. Méthodes à noyaux pour la régression non linéaire
   - Apprentissage de dictionnaire
   - Kernel Ridge regression

Apprentissage de dictionnaire

On considère un dictionnaire de fonctions $\mathcal{D} = \{\eta_j,\ j = 1 \dots p\}$, $\eta_j : \mathcal{X} \to \mathbb{R}$ d'un espace de Hilbert $(\mathcal{H}, \|.\|_{\mathcal{H}})$ et on se restreint aux fonctions de régression de la forme :

$$\eta(x) = \beta_1 \eta_1(x) + \ldots + \beta_p \eta_p(x).$$

Ce dictionnaire peut être choisi pour permettre à la fonction de régression de s'adapter à une certaine structure de parcimonie de $\mathbb{X}$.

Un estimateur de minimisation de risque empirique régularisé, pour la fonction de perte l, basé sur le dictionnaire $\mathcal{D}$, est défini pour $\lambda > 0$ par

$$\hat{\beta}_\lambda \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} l(Y_i, \beta_1 \eta_1(x_i) + \ldots + \beta_p \eta_p(x_i)) + \lambda \|\beta_1 \eta_1 + \ldots + \beta_p \eta_p\|_{\mathcal{H}}^2.$$

Kernel Ridge regression

Cas particulier :

$\mathcal{D} = \{k(x_j, .),\ j = 1 \dots n\}$, où k est un noyau symétrique auto-reproduisant ou noyau de Mercer, et $\mathcal{H}$ est un espace auto-reproduisant (Reproducing Kernel Hilbert Space).

Propriété d'auto-reproduction : $\langle k(x_j, .), k(x_i, .)\rangle_{\mathcal{H}} = k(x_i, x_j)$.
Notations : K matrice des $\big(k(x_i, x_j)\big)_{i,j\in\{1,\dots,n\}}$, $\beta = (\beta_1, \dots, \beta_n)'$.

- $\sum_{i=1}^n (Y_i - \beta_1 k(x_1, x_i) - \dots - \beta_n k(x_n, x_i))^2 = \|Y - K\beta\|_2^2$
- $\|\beta_1 k(x_1, .) + \dots + \beta_n k(x_n, .)\|_{\mathcal{H}}^2 = \beta' K \beta$

L'estimateur Kernel Ridge est défini pour $\lambda > 0$ par

$$\hat{\beta}_\lambda^{kridge} \in \operatorname{argmin}_{\beta\in\mathbb{R}^n} \|Y - K\beta\|_2^2 + \lambda(\beta' K \beta).$$

Solution explicite : $\hat{\beta}_\lambda^{kridge} = (K + \lambda I_n)^{-1} Y$.