Régression pénalisée

Notes de cours - ENSAI

Contents

1 Objectifs du cours et rappels

Dans ce cours nous considérons les modèles linéaires généralisés lorsque le nombre de variables dépasse le nombre d'observations, ou bien lorsque les variables sont très corrélées.

Nous verrons quelques solutions, principalement basées sur la pénalisation des coefficients. Nous les appliquerons sur des jeux de données via le logiciel R.

Les objectifs du cours sont

- Maîtriser la régression Ridge, LASSO et Elastic Net et les variantes du LASSO.
- Savoir sélectionner les meilleures variables explicatives suivant différents critères.
- Maîtriser les package R associés.

1.1 Rappel sur le modèle linéaire

On dispose d'une réalisation de $(Y_1, x_1), \cdots, (Y_n, x_n)$ avec Y_i v.a. indépendantes à valeurs dans $\mathbb{R}$, x_i vecteur de $\mathbb{R}^p$, tels que :

$$\begin{aligned} \mathbb{E}(Y_i) &= \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \\ &= (X\beta)_i, \end{aligned}$$

avec

$$\begin{aligned} X &= \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{1n} \end{pmatrix} \\ \beta &= (\beta_1, \cdots, \beta_p)'. \end{aligned}$$

Le vecteur β est le vecteur des coefficients de régression, X est la matrice de design de dimension $n \times p$.

Remarque On pourrait aussi prendre en compte une constante β_0 et ajouter une colonne à X qui serait alors de dimension $n \times (p+1)$.

On suppose en général que les Y_i suivent des loi normales de même variance, notée σ^2. Ou encore que Y est un vecteur gaussien de matrice de variance covariance $\sigma^2 \times I$, avec I matrice identité.

Quelques exemples On étudiera par exemple le score qu'un individu attribue à un produit en fonction de différentes variables.

1.2 Cadre standard

La matrice $X'X$ est inversible. Dans ce cas l'estimateur du maximum de vraisemblance (MV) est donné par

$$\widehat{\beta}_{MV} \quad = \quad (X'X)^{-1}X'Y.$$

Il maximise la vraisemblance gaussienne, ce qui revient à écrire

$$\widehat{\beta}_{MV} \quad = \quad \underset{\beta \in \mathbb{R}^p}{\arg\min} \|Y - X\beta\|_2^2.$$

C'est aussi l'estimateur des moindres carrés (MC) : $\widehat{\beta}_{MC}$.

1.3 Cadre non standard

La matrice $(X'X)$ n'est pas inversible, ou bien elle est mal conditionnée, et donc d'inverse très instable.

Plusieurs raisons possibles :

- Soit le nombre de variables p dépasse le nombre d'observations n. En effet, par contruction $X'X$ est une matrice carré $p \times p$ dont le rang est inférieur à $\max(p, n)$.
- Soit il y a une combinaison linéaire entre les covariables x_i. Dans ce cas la matrice $X'X$ n'est pas inversible, mais on peut en général facilement détecter la (ou les) combinaison(s) linéaire(s) et supprimer la (ou les) covariable(s) redondante(s).
- Enfin, les covariables peuvent être très corrélées entre elles (pas forcément deux à deux), sans qu'il y ait de liaison linéaire. Dans ce cas la matrice $X'X$ peut être mal conditionnée, avec une inverse très instable numériquement.

$\hookrightarrow$ Dans le cas où il n'y a pas d'inverse l'estimateur MV (ou MC) n'est plus unique. Le système

$$(X'X)\beta \quad = \quad X'Y$$

n'admet pas une seule solution pour β.
$\hookrightarrow$ Dans le cas où la matrice est mal conditionnée, la solution pour $\widehat{\beta}$ est trop instable.

Dans ces deux cas il faut proposer une autre solution.

1.4 Facteur d'inflation de la variance

La corrélation entre les covarables entraîne une augmentation de la variance des estimateurs des coefficients de régression. La diagonale de la matrice de variance des $\widehat{\beta}$ peut s'écrire sous la forme :

$$\begin{aligned}\widehat{\mathbb{V}}(\beta_j) &= \frac{\widehat{\sigma}^2}{(n-1)\mathbb{V}(X_j)}\frac{1}{1-R_j^2} \\ &= \frac{\widehat{\sigma}^2}{(n-1)\mathbb{V}(X_j)}F_j\end{aligned}$$

où R_j^2 est le coefficient de détermination de X_j sur les autres variables. Pour détecter la corrélation on peut donc utiliser la quantité F_j qui est le facteur d'inflation de la variance. Mais si beaucoup de facteurs sont grands (au delà de 5, 10, ...) il est difficile de retirer des covariables de l'analyse.

1.5 Indice de conditionnement

Si on note $v_1 > v_2 > \cdots > v_p$ les valeurs propres de la matrice $X'X$, alors l'indice de conditionnement est donné par

$$I = \frac{v_1}{v_p}.$$

Un indice trop grand (¿ 100, 1000) coïncide avec une matrice mal conditionnée.

2 Des solutions basiques

2.1 L'inverse généralisée

Une première solution au problème de non unicité du MV est l'utilisation de l'inverse généralisée pour $(X'X)$. On note alors

$$\widehat{\beta}_G = (X'X)^+X'Y$$

où est $(X'X)^+$ est l'inverse généralisée (ou pseudo-inverse) définie par

$$A^+AA^+ = A^+ \qquad AA^+A = A.$$

On peut choisir $A^+ = \lim_{\lambda\to 0}(A'A + \lambda A)^{-1}A$

2.2 Sélection automatique

On peut faire une sélection automatique parmi les modèles qui admettent un MV. On considère dans ce cas la norme 0 de β :

$$\|\beta\|_0 = \sum_{j=1}^{p}\mathbb{I}_{\beta_j\neq 0},$$

qui est la dimension du modèle et que l'on utilise pour pénaliser les critère AIC ou BIC. Mais cette approche est trop longue en temps de calcul et souvent trop limitée.

2.3 Moindres carrés restreints

On peut restreindre le nombre de variables à $p' < p$, avec $p' < n$. On sélectionne ensuite le meilleur modèle au sens du MSE (Mean Square Error) avec p' covariables.

2.4 Approche bayésienne

Une approche de plus en plus répandue, surtout pour les modèles complexes, est l'utilisation de loi a priori sur les paramètres. Elle permet de prendre en compte la dépendance des observations.

2.5 Régression sur composantes principales

On peut réaliser une régression sur les composantes principales issues de l'ACP sur le nuage de points des covariables X centrées réduites. En notant $C_1, \cdots, C_K$ les K premières composantes on pose :

$$\mathbb{E}(Y) = \alpha_1 C_1 + \cdots + \alpha_K C_K.$$

Très souvent on n'incorpore pas la constante. Les variables sont centrées réduites pour l'ACP et on peut aussi centrer Y. On peut retrouver les β de la régression en développant linérairement en x toutes les composantes principales. En écrivant

$$C_i = \sum_{j=1}^{p} w_{ij} x_j$$

on a alors

$$\mathbb{E}(Y) = \beta_1 x_1 + \cdots + \beta_p x_p.$$

avec

$$\beta_j = \sum_{i=1}^{K} w_{ij} \alpha_i$$

- Avantage : permet de prendre toutes les variables, quel que soit leur nombre et leur corrélation.
- Inconvénient : ne permet pas de sélectionner les covariables. Ne prend pas en compte Y lors de la construction des composantes.

Avec R:
LIBRARY(MASS)
RESACP=PCR(Y ∼ X,DATA=)
PLOT(RESACP,NCOMP=) # *on choisit au départ un nombre de composantes K arbitraire (assez grand)*
ABLINE(0,1)
RESCV=CROSSVAL(RESACP) # *pour sélectionner le meilleur K par validation croisée*
PLOT(MSEP(RESCV))
SUMMARY(RESCV)
RESACP2=PCR(Y ∼ X,DATA=,NCOMP=) # *avec le nombre de composantes sélectionné*
COEFACP=COEF(RESACP2)
PLOT(COEFACP)
PREDACP=PREDICT(RESACP2,NEWDATA=TEST,NCOMP=)

2.6 Régression PLS (Partial Least Square)

La régression PLS reprend l'idée précédente, en construisant des composantes linéaires en x, mais qui dépendent de Y. La première composante, notée C_1, est construite de la manière suivante :

- Elle est combinaison linéaire des x :

$$C_1 = \sum_{j=1}^{p} w_{1j} x_j$$

- Elle maximise (en valeur absolue) la covariance avec Y (parmi toutes les combinaisons)
- Elle est de norme 1.

On récupère ensuite le résidu de la régression de Y sur C_1, noté $\epsilon_1 = Y - \widehat{\beta}_1 C_1$. Ce résidu est donc la partie de Y non expliquée par la première composante C_1. On cherche alors une autre composante, C_2, combinaison linéaire des x, normalisée et orthogonale à C_1, et qui maximise la covariance avec ce résidu. On récupère le nouveau rédidu, ϵ_2, et on construit ainsi toutes les composantes PLS. On peut retrouver les β de la régression en développant linérairement en x toutes les composantes PLS. En écrivant

$$C_i = \sum_{j=1}^{p} w_{ij} x_j$$

on a alors

$$\mathbb{E}(Y) = \beta_1 x_1 + \cdots + \beta_p x_p.$$

avec

$$\beta_j = \sum_{i=1}^{K} w_{ij} \alpha_i$$

Le choix de K peut se faire en fonction du pourcentage de variance de Y expliquée ou bien par validation croisée.

- Avantage : celui de l'ACP avec prise en compte de Y.
- Inconvénient : ne sélectionne pas les covariables.

Avec R :

```
LIBRARY(PLS)
RESPLS=PLSR(Y ~ X,DATA=)
PLOT(RESPLS,NCOMP=)
ABLINE(0,1)
RESCV=CROSSVAL(RESPLS) # pour sélectionner le nombre de composantes PLS par validation croisée
PLOT(MSEP(RESCV))
SUMMARY(RESCV)
RESPLS2=PLSR(Y ~ X,DATA=OZONE,NCOMP=)
COEFPLS=COEF(RESPLS2)
COEFPLS
PLOT(COEFPLS) # on pourrait faire du seuillage ici
PREDPLS=PREDICT(RESPLS2,NEWDATA=TEST,NCOMP=)
```

3 Les méthodes pénalisées

Une autre famille de solutions consiste à considérer une maximisation contrainte de la vraisemblance.

3.1 Le compromis biais-variance

Lorsque l'on dispose d'un estimateur de β, un critère de qualité du modèle est le Mean Squared Error (MSE) qui vaut, pour des valeurs de y et du vecteur x données :

$$\begin{aligned} MSE &= \mathbb{E}\left((y - \sum_{i=1}^{p}(\widehat{\beta}_i x_i))^2\right) \\ &= \mathbb{E}\left((y - \widehat{\beta}'x)^2\right) \\ &= \mathbb{E}\left(y - \beta'x + \beta'x - \mathbb{E}(\widehat{\beta}'x) + \mathbb{E}(\widehat{\beta}'x) - \widehat{\beta}'x)\right) \\ &= \sigma^2 + \mathbb{V}\left(\widehat{\beta}'x\right) + (biais\left(\widehat{\beta}'x\right))^2. \end{aligned}$$

Ainsi on veut réduire la variance de $\widehat{\beta}$ en lui appliquant un coefficient de rétrécissement (shrinkage), mais au risque d'augmenter le biais. Une solution est de diminuer $\widehat{\beta}$, c'est ce que propose les méthodes suivantes en imposant une contrainte sur l'espace des solutions. Certaines méthodes vont même imposer la nullité de certains coefficients.

3.2 Sparsité

Pour justifier les approches pénalisées on considère qu'il y a peu de coefficients non nuls, ou bien peu de coefficients grands.

Definition 1. *On définit l'ensemble de sparsité associé à β par $\mathcal{A} = \{i \in \{1, \cdots, p\}; \beta_j \neq 0\}$ et on appelle indice de sparsité son cardinal $|\mathcal{A}|$.*

L'hypothèse de sparsité est que $|\mathcal{A}| < p$. On peut aussi affaiblir l'hypothèse de sparité en proposant un seuil $s > 0$ tel que $\mathcal{A}_s = \{i \in \{1, \cdots, p\}; |\beta_j| > s\}$. On pratique alors une technique de seuillage pour sélectionner les coefficients. Ce seuil peut être fixé a posteriori.

4 Régression Ridge

(Haerl et Kennard, 1970). Une méthode naïve pour obtenir un estimateur dans le cas où $X'X$ est mal conditionnée ou non inversible est d'ajouter un coefficient $\lambda > 0$ (appelé coefficient ridge) qui biaisera un peu le modèle mais qui permettra de résoudre le système. Considérons

$$\widehat{\beta}_R \quad = \quad (X'X + \lambda I)^{-1} X'Y,$$

avec $\lambda > 0$. On obtient une solution unique. On voit que λ biaise le modèle, mais stabilise la solution. On peut réécrire la régression Ridge comme une régression pénalisée. Par la suite les covariables sont normalisées.

Definition 2. *On définit l'estimateur ridge par*

$$\widehat{\beta}_R = \underset{\beta \in \mathbb{R}^p}{\arg\min} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \; = \underset{\beta \in \mathbb{R}^p}{\arg\min} \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$$

où $\lambda > 00$ est un paramètre de shrinkage qui contrôle la force de la pénalité.

Propriétés

- On retrouve l'estimateur précédent.
- L'estimateur ridge existe toujours.
- Si $\lambda \to 0$ on retombe sur le modèle linéaire (avec une possible inverse généralisée).
- Si $\lambda \to \infty$ alors $\widehat{\beta}_R$ tend vers 0.
- L'estimateur ridge est biaisé, avec un biais égal à $-(X'X + \lambda I)^{-1}\beta$.
- On peut exprimer facilement la variance de l'estimateur ridge. On montre que sa trace tend vers zéro quand $\lambda \to \infty$. Mais alors le biais augmente.
- La régression ridge peut être vue comme une régression ordinaire augmentée. Il suffit d'ajouter des $y_i = 0$ et des $x_i = \sqrt{\lambda}$.

Vision bayésienne On peut montrer que l'approche ridge revient à mettre un a priori gaussien sur le coefficient de régression β. On peut voir alors l'estimateur comme le maximum (ou le mode) a posteriori.

Choix du paramètre On choisit λ par validation croisée (k-fold). On découpe la population en K groupes d'indices $G_1, \cdots, G_K$. Souvent K est fixé proche de 10. On fixe λ et on itère les étapes suivantes :

- On choisit un premier indice $i = 1$
- Sur les groupes $G_j, j \neq i$, on estime β par ridge $\rightarrow \widehat{\beta}(i)$.
- On calcule le MSE sur G_i :

$$MSE(i) \quad = \quad \frac{1}{|G_i|} \sum_{j \in G_i} \left(y_j - (X\widehat{\beta}(i))_j \right)^2 .$$

- On recommence avec $i \leftarrow i + 1$ jusqu'au dernier indice $i = K$.
- On calcule l'erreur moyenne :

$$MSE(\lambda) \quad = \quad \frac{1}{K} \sum_{j=1}^{K} MSE(j).$$

  Finalement le meilleur λ sera celui qui minimise les $MSE(\lambda)$ lorsque λ parcourt une grille (assez fine). On l'appelera $\lambda_{\min}$.

Si nous n'avons pas assez d'observations pour découper en K sous groupes alors on fait du leave-one-out (c'est-à-dire que l'on prend $K = n$).

Package R Dans R on peut utiliser le package *lm.ridge* :

```
RES <- LM.RIDGE(Y ~ X,LAMBDA=SEQ(0,100,0.1))
SELECT(RES) # sélectionne le meilleur λ par validation croisée
RESRIDGE <- LM.RIDGE(Y ~ X,LAMBDA=S)
PLOT(COEF(RIDGE)) # on peut réaliser un seuillage
```

En pratique On peut faire une régression ridge pour retenir les meilleures variables par seuillage. L'avantage est que l'on peut sélectionner des variables très corrélées et en grand nombre. L'inconvénient est que la régression ridge ne sélectionne pas automatiquement les variables.

5 Régression LASSO

Least Absolute Shrinkage Selection (Tibshirani, 1996). Les covariables sont normalisées. On peut également centrer Y.

Definition 3. *On définit l'estimateur LASSO par*

$$\widehat{\beta}_L \;=\; \underset{\beta\in\mathbb{R}^p}{\arg\min}\, \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 \;= \underset{\beta\in\mathbb{R}^p}{\arg\min}\, \|Y - X\beta\|_2^2 + \lambda\sum_{j=1}^{p}|\beta_j|,$$

où $\lambda \geq 0$ est un paramètre de shrinkage qui contrôle la force de la pénalité.

Propriétés

- L'estimateur LASSO existe toujours.
- Si $\lambda \to 0$ on retombe sur le modèle linéaire (avec une possible inverse généralisée).
- Si $\lambda \to \infty$ alors $\widehat{\beta}_L$ tend vers 0 et atteint zéro pour une valeur de λ finie.

Path LASSO Les coefficients de β vont s'annuler au fur et à mesure que la contrainte va augmenter. Ainsi il existe un chemin pour lequel des valeurs de λ correspondent à l'annulation de certains coefficients de régression. On peut noter $\lambda_0 = 0$ lorsqu'il n'y a pas de contrainte, ce qui revient à prendre le MC. Puis on augmente λ jusqu'au $\lambda_{\max}$ tel que $\widehat{\beta}_R = 0$.

Vision bayésienne On peut montrer que l'approche LASSO revient à mettre comme a priori une loi double exponentielle sur le coefficient de régression β et à étudier le maximum a posteriori.

Choix du paramètre Comme précédemment on choisit λ par validation croisée. Parfois le $\lambda_{\min}$ est en pratique trop faible et dans ce cas on garde trop de variables. Il est courant d'utiliser le λ_{1se} qui est la plus grande valeur (estimée) de λ telle que le MSE associé ne dépasse pas de plus d'un écart-type celui du $\lambda_{\min}$. Cette pratique est classique en LASSO et elle permet souvent de retenir beaucoup moins de variables. Par contre, c'est toujours le $\lambda_{\min}$ que l'on garde pour la régression ridge puisque dans tous les cas aucun coefficient n'est annulé.

Package R Dans R on peut utiliser le package lars :

```
library(lars)
reslasso=lars( x,y, type="lasso")
plot(reslasso)
rescv=cv.lars(x,y)
min(rescv$cv) # donne le plus petit cv
```

I = WHICH.MIN(RES$CV) # *indique la fraction donnant le plus petit cv*
F = RES$CV[I] # *nous donne la valeur de cette fraction*
PREDICT(RESLASSO,S=F,TYPE="COEFFICIENTS", MODE="STEP") # *nous donne les coeff associés à la fraction*

En pratique On peut faire une régression LASSO pour retenir automatiquemnt les meilleures variables. L'un des inconvénients est que le nombre de variables sélectionnées ne peut dépasser le nombre d'observations. D'autre part, la régression LASSO aura tendance à moins sélectionner une variables si elle est corrélée à d'autres variables déjà sélectionnées. Cela fait que la convergence vers l'ensemble de sparsité peut être longue dans ce cas.

Approximation de la solution LASSO On peut approcher les coefficients LASSO avec le programme suivant :

- Fixer $\delta > 0$ petit.
- Démarrer par $\widehat{\beta} = 0$ et $\epsilon = Y$.
- Calculer le résidu $\epsilon = Y - \bar{Y}$
- Trouver le x_i le plus corrélé (en valeur absolue) à ϵ et poser

$$\widehat{\beta}_i \quad = \quad \widehat{\beta}_i + \delta \; s(x_i, \epsilon),$$

  où $s(x, y)$ désigne le signe de la corrélation entre x et y.
- Calculer le nouveau résidu $\epsilon = \epsilon - \widehat{\beta}_i x_i$.
- Recommencer jusqu'à ce que $\min(n, p)$ composantes de $\widehat{\beta}$ soient non nulles.

6 Régression ELASTICNET

Les covariables sont normalisées comme pour ridge et LASSO. On va associer les qualités du LASSO à celles de ridge pour pouvoir sélectionner davantage de variables, notamment corrélées (voir Zou et Hastie, 2005).

Definition 4. *On définit l'estimateur Elastic Net par*

$$\widehat{\beta}_{EN} \quad = \quad \underset{\beta \in \mathbb{R}^p}{\arg\min} \, \|Y - X\beta\|_2^2 + \lambda(\alpha\|\beta\|_1^2 + (1-\alpha)\|\beta\|_2^2),$$

où $\lambda \geq 0$ *et* $\alpha \in [0, 1]$.

Propriétés

- On retrouve Ridge ou LASSO lorsque $\alpha = 0$ ou 1.
- On a encore un chemin qui permet d'annuler les coefficients de régression.

Vision bayésienne L'approche Elastic Net peut être vue comme une approche bayésienne avec une loi a priori non classique sur le paramètre β.

Choix du paramètre En général on fixe $\alpha \in [0, 1]$. C'est la proportion de LASSO dans le modèle. Puis on choisit λ par validation croisée. On peut aussi exprimer l'estimateur Elastic Net de manière naïve comme

$$\widehat{\beta}_{EN} = \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1^2 + \lambda_2 \|\beta\|_2^2,$$

où $\lambda_1, \lambda_2 \geq 0$. Dans ce cas, on peut d'abord fixer un paramètre λ en faisant la régression associée (LASSO ou Ridge), puis trouver le second paramètre par validation croisée. Cette approche avantage la méthode associée au premier choix du paramètre.

Package R Dans R on peut utiliser

- Le package enet :
  LIBRARY(ELASTICNET)
  RESENET=ENET(X,Y,LAMBDA=) # *le lambda est celui de Ridge qui peut être déterminé par lm.ridge*

  PLOT(RESENET)
  CV.ENET(X,Y,LAMBDA=,S=SEQ(0,1,LENGTH=),MODE='FRACTION')
  MIN(RESCV$CV) # *donne le plus petit cv*
  I = WHICH.MIN(RES$CV) # *indique la fraction donnant le plus petit cv*
  F = RES$CV[I] # *nous donne la valeur de cette fraction*
  PREDICT(RESLASSO,S=F,TYPE="COEFFICIENTS", MODE="STEP") # *nous donne les coeff associés à la fraction*
- Le package glmnet :
  library(glmnet)
  res=glmnet(x,y,familiy="gaussian", alpha=0)
  rescv=cv.glmnet(x,y,familiy="gaussian", alpha=) # alpha est la proportion de LASSO
  predict(res,type= "coefficients",s=rescv$lambda.1se)

  Ou bien :
  resglmnet=glmnet(x,y,familiy="gaussian", lambda=rescv$lamda.1se,alpha=0)
  coef(resglmnet)

  Son avantage est qu'il permet d'obtenir directement le $\lambda_{\min}$ et le λ_{1se}. On peut faire du LASSO ou du ridge en fixant alpha=1 ou 0.

En pratique On peut faire varier le paramètre α pour renforcer la partie LASSO et ainsi sélectionner moins de variables corrélées. On peut inversement diminuer α pour avantager ridge et sélectionner davantage de variables. On peut aussi faire une boucle sur une grille du paramètre α et retenir celui qui minimise les MSE obtenus pour chaque lambda optimal.

7 Premières conclusions

En résumé,

- En comparant les résultats obtenus par Ridge et LASSO on peut deviner si certaines variables significatives sont très corrélées entre elles.
- La combinaison Elastic Net permet de sélectionner automatiquement un modèle qui répond à certains critères (plus ou moins de variable corrélées).
- Les paramètres retenus dépendent fortement de la validation croisée et il est conseillé de faire une boucle pour la rendre plus robuste au choix des partitions.
- On peut également faire une boucle pour trouver un alpha optimal.
- On peut réaliser un modèle classique une fois les variables retenues par Elastic Net. L'avantage est de pouvoir alors utiliser les test et les méthodes de sélection classiques, ainsi que de construire des intervalles de confiance. Les interprétations sont également plus simples, sans normalisation des variables.

8 GLM pénalisés

La régression pénalisée se généralise aux GLM. On travaille toujours sur les données centrées réduites. On considère la log vraisemblance pénalisée comme suit :

$$\widehat{\beta} \;=\; \underset{\beta\in\mathbb{R}^p}{\arg\min}\left(-\log(L(Y_, \cdots, Y_n;\beta)) + \lambda_1\|\beta\|_1^2 + \lambda_2\|\beta\|_2^2\right).$$

On peut également introduire un coefficient $\alpha \in [0,1]$ qui détermine les pourcentages de LASSO et de Ridge dans la régression :

$$\widehat{\beta} \;=\; \underset{\beta\in\mathbb{R}^p}{\arg\min}\left(-\log(L(Y_, \cdots, Y_n;\beta)) + \lambda(\alpha\|\beta\|_1^2 + (1-\alpha)\|\beta\|_2^2)\right),$$

où $\lambda \geq 0$ et $\alpha \in [0,1]$.

Modèle de Poisson

$$\widehat{\beta} \;=\; \underset{\beta\in\mathbb{R}^p}{\arg\min}\left(-\sum_{i=1}^{n} y_i(X\beta)_i - e^{(X\beta)_i} + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2\right).$$

Modèle binomial négatif Les données complexes, avec beaucoup de corrélation ou avec trop de variables, peuvent aussi souffrir de surdispersion. En général on réalise un modèle de Poisson pénalisé pour sélectionner les variables. On regarde ensuite la surdispersion associée au modèle de Poisson classique, avec les variables retenues.

Modèle logistique L'avantage de la régression pénalisée est qu'elle permet de trouver des solutions même en présence de (quasi) séparation des données. On a

$$\widehat{\beta} \quad = \quad \underset{\beta\in\mathbb{R}^p}{\arg\min} \left(-\sum_{i=1}^{n} y_i (X\beta)_i + (1+X\beta)_i^{-1} + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right).$$

Modèle polytomique On peut également réaliser une régression polytomique ordonnée via le package glmnet de R.

Package R On peut reprendre glmnet en utilisant family =poisson, binomial, multinomial, cox, mgaussian...

9 Le fused LASSO

Le fused LASSO permet de détecter l'égalité des coefficient successifs. C'est particulièrement intéressant lorsque qu'il y a une stabilité des coefficients sur une certaine période, ce qui peut être le cas lorsque les variables sont ordonnées (temporellement).

Package R Avec les notations précédentes : FUSED.LASSO(LAMBDA=C(λ_1,λ2))

10 LASSO adaptatif

Un inconvénient de la méthode LASSO est le biais de son estimateur. Pour le diminuer on peut adapter les poids des coefficients de LASSO de la manière suivante :

- On démarre avec un estimateur ridge de β, noté $\widehat{\beta}$.
- On réalise alors un LASSO pondéré :

$$\widehat{\beta}_L \quad = \quad \underset{\beta\in\mathbb{R}^p}{\arg\min} \|Y - X\beta\|_2^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| / \|\widehat{\beta}_j|.$$

- On s'arrête, ou bien on réitère l'étape précédente jusqu' à stabilité. Mais si on continue alors on ne propose pas les coefficients déjà annulés (ils le sont définitivement).

On voit que les forts coefficients ridge vont être avantagés car moins pénalisés dans le LASSO.

Package R On construit les poids avec tout d'abord une régression ridge, et on les introduit ensuite dans LASSO :

```
cv.ridge <- cv.glmnet(x, y, family='', alpha=0)
w <- 1/abs(matrix(coef(cv.ridge, s=cv.ridge$lambda.min))
cv.lasso <- cv.glmnet(x, y, family='', alpha=1, penalty.factor=w)
plot(cv.lasso)
coef(cv.lasso, s=cv.lasso$lambda.1se)
```

11 Données longitudinales

On peut combiner la méthode GEE avec la régression pénalisée (voir par exemple Fu, 2003 ou Geronimi et Saporta, 2017).

La quasi vraisemblance Les observations $Y_i \in \mathbb{R}^p$ $(i = 1, \cdots, n)$ sont indépendantes et satisfont au modèle suivant :

- Chaque composante $Y_{ij} \in \mathbb{R}$ appartient à un GLM; i.e. il existe une fonction de lien, $g(.)$, et une fonction variance, $v(.)$, telles que

$$\begin{aligned} g(m_{ij}) &= x_{ij}^T \beta, \\ \sigma_{ij}^2 &= \phi\, v(m_{ij}), \end{aligned}$$

  où $\beta \in \mathbb{R}^q$ est le vecteur de régression, $x_{ij} \in \mathbb{R}^q$ est le vecteur des covariables associées à Y_{ij}, $m_{ij} = \mathbb{E}(Y_{i,j})$ et $\sigma_{ij}^2 = Var(Y_{ij})$.

- L'estimateur de β va satisfaire l'équation du quasi (ou pseudo) score suivante :

$$\begin{aligned} q(y, \beta) &= \sum_{i=1}^{n} \frac{\partial m_i}{\partial \beta}^T V(m_i)^{-1}(y_i - m_i) \\ &= 0, \end{aligned}$$

  où $\frac{\partial m_i}{\partial \beta}$ est la matrice $p_i \times q$ dont le (j,k)-ème élément s'écrit $\frac{\partial m_{ij}}{\partial \beta_k}$ et $V(m_i) = Var(Y_i)$ est la matrice $p \times p$ de variance covariance de Y_i dont les éléments non diagonaux sont inconnus.

Décomposition de la matrice de variance En écrivant chaque variance de la manière suivante

$$Var(Y_i) = \phi S_i R_i S_i,$$

avec

$$\begin{aligned} R_i &= Corr(Y_i), \\ S_i &= diag(\sigma_{i1} \cdots, \sigma_{ip_i})\phi^{-1/2}, \\ &= diag(\sqrt{m_{i1}}, \cdots, \sqrt{m_{ip_i}}), \end{aligned}$$

le quasi score devient :

$$q(y,\beta) = \sum_{i=1}^{n} \frac{\partial m_i}{\partial \beta}^T (S_i R_i S_i)^{-1}(y_i - m_i).$$

Mais seule S_i est définie dans le modèle. Les matrices R_i des corrélations sont inconnues.

Plusieurs choix sont possibles pour R_i. On propose pour R_i une matrice, appelée working correlation matrix qui peut prendre différentes formes (identité, complètement déterminée, d-dépendance, AR,...). En générale, c'est la même pour tous les Y_i et on peut donc enlever l'indice i.

GEE et pénalisation L'estimateur pénalisé de β va satisfaire l'équation du quasi score pénalisée suivante :

$$\begin{aligned} q(y,\beta) &= \sum_{i=1}^{n} \frac{\partial m_i}{\partial \beta}^T V(m_i)^{-1}(y_i - m_i) + 2\lambda\beta \\ &= 0, \end{aligned}$$

ce qui revient ici à dériver la contrainte $L2$ de ridge. On peut généraliser à une contrainte plus générale (contrainte bridge).

Estimation de la working correlation matrix par la méthode GEE On utilise donc un algorithme itératif de la forme suivante :

- Proposer une valeur initiale de R_i, notée R_{i0} (en général l'identité). En déduire l'estimation de β, notée β_0, par quasi score pénalisé (avec l'identité dans la fonction de quasi score). On estime aussi le lambda optimal dans ce cadre supposé d'indépendance.
- En utilisant l'estimation précédente de β on estime alors les versions standardisées des Y_{ij}, les paramètres de la matrice R_i et par suite R_i (notons R_{i1} cette première estimation). Et on estime également le paramètre d'échelle ϕ.
- Grâce à l'estimation précédente de R_i on peut estimer la matrice de variance covariance de β.
- On peut maintenant ré-estimer β en utilisant l'équation du score pénalisée avec la nouvelle estimation de R_i et avec comme valeur initiale l'ancienne valeur de $\widehat{\beta}$.
- On recommence jusqu'à convergence, c'est-à-dire $|\widehat{\beta}_{i+1} - \widehat{\beta}_i| < \epsilon$.

Package R
On calcule d'abord le lambda optimal pour la pénalisation en supposant l'indépendance (Working Matrice = identité)

```
library(PGEE)
```

cv <- CVfit(y ∼ x, id = identifiant, family =)
lambda = cv$lam.opt

Puis on réalise une PGEE avec le lambda précédemment retenu

res <- PGEE(y ∼ x, id = identifiant, family = , corstr =, lambda = cv$lam.opt)

On peut utiliser comme forme pour la structure de corrélation de la Working Matrix *corstr* : *"AR-1","exchangeable", "fixed", "independence", "M dépendance", et "unstructured".* L'identifiant permet de reconnaître les individus répétés. Avec l'option *family* on peut choisir *gaussian, poisson, binomial, ou gamma.*

12 Le Group LASSO

On suppose que l'on dispose a priori d'une partition de K groupes de variables $G_1, \cdots, G_K$. On veut sélectionner toutes les variables d'un groupe ou bien les éliminer toutes simultanément. Pour cela on réalise un group LASSO (voir par exemple Meier et al, 2008, ou Wang et Leng, 2008)

$$\widehat{\beta}_L \quad = \quad \underset{\beta \in \mathbb{R}^p}{\arg\min} \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{K} \sqrt{\sum_{k \in G_j} \beta_k^2}.$$

L'approche LASSO va ainsi mettre à zéro au fur et à mesure tous les coefficients d'un groupe simultanément. Cela permet par exemple de garder ou de retirer tous les coefficients associés à une variable qualitative. On peut également prendre en compte les tailles des groupes et les intégrer dans la pénalité.

Definition 5. *On dit qu'un groupe est un vrai groupe s'il contient au moins un indice dans l'ensemble de sparsité $\mathcal{A}$. Une variable dans un vrai groupe est une vraie variable si son indice est dans $\mathcal{A}$. Si son incide n'est pas dans $\mathcal{A}$ on parle de pseudo-vraie variable. Les autres groupes sont appelés des faux groupes et leurs variables associées sont des fausses variables.*

Group LASSO supervisé On peut construire les groupes à partir d'une méthode de classification (par exemple CAH). Ensuite on associe un groupe à chaque classes. On peut aussi diminuer a taille des classes en réalisant un LASSO sur chaque classe.

- Construction des classes $C_1, \cdots, C_K$ par classification.
- Eventuellement élagage dans chaque classe par un LASSO qui ne retiendra que les coefficients non nuls. On aura donc tendance à retenir les vraies variables. On élimine ainsi les pseudo-vraies variables.
- On peut faire un group LASSO sur les groupes élagués.

Package R

```
RES <- GGLASSO(X,Y,GROUP=IDENTIFIANT)
PLOT(RES)
RESCV=CV.GGLASSO(X,Y,GROUP=IDENTIFIANT)
LAMBDA1=RESCG$LAMBDA.MIN
LAMBDA2=RESCV$LAMBDA.1SE
RESGG= GGLASSO(X,Y,GROUP=IDENTIFIANT,LAMBDA=LAMBDA2)
PLOT(COEF(RESCV))
```

L'identifiant permet d'identifier les groupes. Il s'agit d'un vecteur d'indices, de taille p (nombre de variables) et avec K indices différents (nombre de groupes).

13 Les données manquantes

En présence de données manquantes sur les p variables x_i, $i = 1, \cdots, p$, nous allons utiliser une méthode d'imputation multiple (IM) que nous allons combiner à une approche LASSO. L'inconvénient de l'imputation multiple directe est que si nous appliquons LASSO sur les D jeux de données complets obtenus par IM alors ce ne sont pas forcément les mêmes variables qui seront retenues lors des D analyses. Pour pallier ce défaut nous utilions la méthode IM-LASSO proposée par Yuan (2016). Plus précisément il y aura trois étapes

- Imputation : on crée D jeux de données complets par une méthode classique (régression, ...). En pratique D est compris entre 5 et 10.
- Group LASSO : on réalise un Group LASSO sur les p groupes. Chaque groupe est de taille D. Il est composée d'une seule variable répétée D fois.
- Les coefficients finaux sont obtenus en prenant les moyennes des coefficients de chaque groupe sélectionné.

Cette méthode est importante car les packages classiques ne fonctionnent pas en présence de données manquantes.

Package R On peut tout d'abord utiliser un package pour les données manquantes comme MICE (Multivariate Imputation by Chained Equations). Par défaut la méthode utilisée est le predictive mean matching (pmm). On construit ainsi $m = 5$ jeux de données complets.

```
DATA=MICE(X,M=5,METH='PMM')
```

On utilise ensuite un groupe LASSO sur ces m groupes obtenus pas MICE.

14 Méthodes transductives

Cette méthode s'applique lorsque l'on dispose de deux jeux de données. Le premier jeu est complet avec un vecteur de réponse Y et une matrice de design X. Le deuxième jeu ne contient que la matrice de design, $\widetilde{X}$, sans le vecteur de réponses $\widetilde{Y}$ On peut alors suivre le shéma suivant :

- Construction d'un premier estimateur (LASSO ou autre) sur le premier jeu de données, noté $\widehat{\beta}$.
- Reconstruction de la réponse du deuxième jeu de données :

$$\widetilde{Y} \quad = \widetilde{X}\widehat{\beta}$$

- Nouvelle estimation de β :

$$\widetilde{\beta} \;=\; \underset{\beta\in\mathbb{R}^p}{\arg\min} \left(\|\widetilde{Y} - \widetilde{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1^2 \right).$$

  On voit que l'on peut soit garder le premier estimateur (ce qui rend nulle la première quantité), soit l'améliorer...

Une autre approche qui permet de profiter du deuxième jeu de données en renforçant l'estimation de la variance des X est la suivante :

$$\widehat{\beta} \;=\; (\lambda \widetilde{X}'\widetilde{X} + X'X)^{-1} X'Y,$$

avec $\lambda \geq 0$.

Avec R On peut récupérer un estimateur LASSO sur le jeu de donnée complet. On peut ensuite reconstruire Y en utilisant les prédictions sur les nouveaux X. On réalise alors un LASSO sur ce jeu contenant les nouveaux X et les y prédits. On peut espérer que l'estimateur obtenu sera meilleur (en terme de MSE par exemple).

15 Références

- M.J. Azur, E.A. Stuart, C. Frangakis, P.J. Leaf (2011) Multiple Imputation by chained equations: what is it and how does it work? Psychiatric research, 20,1.
- Q. Chen, S. Wang (2013) Variable selection for multiply-imputed data with application to dioxin exposure study. Stat. Med. 32, 3646–3659.
- W.J. Fu. (2003) Penalized Estimating Equations. BIOMETRICS 59, 126-132
- J. Geronimi, G. Saporta (2017) Variable selection for multiply-imputed data with penalized generalized estimating equations. Computational Statistics and Data Analysis.

- A. E. Hoerl, A.E. R.W. Kennard (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12, 1.
- J. Huang, S. Ma, C.H. Zhang (2008) Adaptive Lasso for sparse high-dimensional regression models. Statistica Sinica 18, 1603-1618.
- K. Liang, S. Zeger (1986) Longitudinal data analysis using generalized linear models. Biometrika 13
- L. Meier, S. van de Geer, P. Buhlmann (2008), The Group Lasso for Logistic Regression, Journal of the Royal Statistical Society, 70 (1), 53 - 71
- R. J.A. Little. (1988) Missing-Data Adjustments in Large Surveys. Journal of Business and Economic Statistics, Vol. 6, No. 3, pp. 287–296
- D.B. Rubin. (1988) An overview of multiple imputation. Harvard University.
- R. Tibshirani (1996) Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B, Volume 58, Issue 1, pp. 267–288.
- S. Van Buuren, K. Groothuis-Ousdshoorn (2011) mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, Vol 45, Issue 3
- H. Wang, C. Leng (2008) A note on adaptive group lasso. Computational Statistics and Data Analysis 52, 5277–5286
- Y.C. Yuan (2016) SAS Institute Inc., Rockville, MD Multiple Imputation for Missing Data: Concepts and New Development.
- H. Zou, T. Hastie (2005) Regularization and variable selection via the elastic net. J. R. Statist. Soc. B 67, Part 2, pp. 301–320
- Yang, Y. and Zou, H. (2015), A Fast Unified Algorithm for Computing Group-Lasso Penalized Learning Problems, Statistics and Computing. 25(6), 1129-1141.