TELECOM ParisTech **MDI220**
2013/2014 **Joseph Salmon - Anne Sabourin**

Travaux pratiques n° 3 : Moindres carrés pénalisés

Quelques éléments de corrections sont disponibles à l'adresse suivante :
http://josephsalmon.eu/enseignement/TELECOM/MDI220/sources/methode_penalisees/TP_selec_var.R

Lancer d'abord cette exemple pour vous guider dans l'affichage graphique, et aussi dans la manière d'utiliser les différents packages.

On rappelle tout d'abord le formalisme des méthodes pénalisées. On se place dans le cadre d'un modèle linéaire :

$$Y = X\theta + \varepsilon$$

où le bruit est centré, et notre objectif est d'estimer θ.

Dans ce TP on considère trois types de méthodes pénalisées, le Ridge, le Lasso et l'Elastic net. La forme générale de ces estimateurs est la suivante :

$$\hat{\theta}(\lambda) = \arg\min_{\theta \in \mathbb{R}^d} \|Y - X\theta\|_2^2 + \mathrm{pen}(\theta, \lambda)$$

où l'on choisit respectivement

$$\begin{aligned} \mathrm{pen}(\theta,\lambda) &= \lambda\|\theta\|_2^2 && \text{(Ridge/Tikhonov)} \\ \mathrm{pen}(\theta,\lambda) &= \lambda\|\theta\|_1 && \text{(Lasso)} \\ \mathrm{pen}(\theta,\lambda) &= \lambda_1\|\theta\|_1 + \lambda_2\|\theta\|_2^2 && \text{(Elastic Net)} \end{aligned}$$

Le paramètre de régularisation λ est un réel positif pour le Ridge et le Lasso, mais c'est un vecteur $\lambda = (\lambda_1, \lambda_2)^\top$ ayant des coordonées positives pour l'Elastic Net.

- Types d'expériences -

On va considérer quatre types d'expériences, qui nécessitent de coder une fonction `datageneration.R`, qui génère des données Y et des variables $(X_j)_{j=1\dots,p}$ (ou de manière équivalent la matrice X). Cette fonction prend en argument n, p, K, L et θ, et construit X et Y de la façon suivante :

- $Y = X\theta + \varepsilon$
- $\varepsilon \in \mathbb{R}^p$, et $\varepsilon \sim \mathcal{N}(0, \mathrm{Id}_p)$ (penser à la fonction `rnorm`)
- $\theta \in \mathbb{R}^p$ (on supposera qu'on ne fait pas intervenir les constantes dans le modèle).
- $X \in \mathbb{R}^{n\times p}$. De plus les colonnes de X sont des réalisations d'une loi normale centrée et de covariance $\Sigma \in \mathbb{R}^{p\times p}$ où Σ est une matrice diagonale par blocs, avec K blocs de même taille ℓ (et tels que $\ell = p/K$ soit un nombre entier) et chaque bloc est de la forme :

$$\Sigma_{jj} = \frac{j-1}{K}\mathbf{1}_{\ell\times\ell} + \frac{K+1-j}{K}\mathrm{Id}_\ell \ .$$

Par exemple pour cinq blocs, on obtient une matrice de la forme :

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 & 0 & 0 & 0 \\ 0 & \Sigma_{22} & 0 & 0 & 0 \\ 0 & 0 & \Sigma_{33} & 0 & 0 \\ 0 & 0 & 0 & \Sigma_{44} & 0 \\ 0 & 0 & 0 & 0 & \Sigma_{55} \end{pmatrix}.$$

où chaque bloc de taille 10×10 vaut respectivement

$$\Sigma_{11} = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 & 0 \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix}, \quad \Sigma_{22} = \begin{pmatrix} 1 & \frac{1}{5} & \dots & \dots & \frac{1}{5} \\ \frac{1}{5} & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 & \frac{1}{5} \\ \frac{1}{5} & \dots & \dots & \frac{1}{5} & 1 \end{pmatrix}, \Sigma_{33} = \cdots$$

Remarque : pour générer des vecteurs selon une loi gaussienne multivariée on pourra utiliser `mvrnorm`. Pour plus de détails théoriques sur cette opération, voir Wikipidia si besoin. Pour définir des matrices par blocs on pourra utiliser une structure de liste et la fonction `bdiag`.

On va alors appliquer les trois méthodes mentionnées ci-dessus (Ridge, Lasso, Elastic-Net) sur les quatre expériences :

Expérience 1. (Variable explicatives non corrélées - paramètre dense)

$$\begin{aligned} &n = 100, \\ &K = 1, \\ &\ell = 50, \\ &\theta = (\underbrace{-2, \cdots, -2}_{\ell \text{ fois}}, \underbrace{2, \cdots, 2}_{(K-1)\times\ell \text{ fois}}) \end{aligned}$$

Expérience 2. (Variable explicatives non corrélées - paramètre sparse)

$$\begin{aligned} &n = 100, \\ &K = 1, \\ &\ell = 50, \\ &\theta = (\underbrace{2, \cdots, 2}_{5 \text{ fois}}, \underbrace{0, \cdots, 0}_{40 \text{ fois}}, \underbrace{2, \cdots, 2}_{5 \text{ fois}}) \end{aligned}$$

Expérience 3. (Variable explicatives corrélées - paramètre dense)

$$\begin{aligned} &n = 100, \\ &K = 3, \\ &\ell = 50, \\ &\theta = (\underbrace{-2, \cdots, -2}_{\ell \text{ fois}}, \underbrace{3, \cdots, 3}_{(K-1)\times\ell \text{ fois}}) \end{aligned}$$

Expérience 4. (Variable explicatives corrélées - paramètre sparse)

$$\begin{aligned} &n = 100, \\ &K = 3, \\ &\ell = 50, \\ &\theta = (\underbrace{2, \cdots, 2}_{5 \text{ fois}}, \underbrace{0, \cdots, 0}_{40 \text{ fois}}, \underbrace{2, \cdots, 2}_{5 \text{ fois}}) \end{aligned}$$

- Ridge/Tikhonov -

Pour cet estimateur on utilise la fonction `ridge` du package `MASS`.

- Utiliser `lm.ridge` pour calculer l'estimateur Ridge $\hat{\theta}^{\text{Ridge}}(\lambda)$ pour 30 valeurs différentes du paramètre de régularisation λ (que l'on prendra selon une grille géométrique, par exemple : `lambda[i]=exp(i-(k/2))`).
- Afficher sur un même graphique les p fonctions $\lambda \to \hat{\theta}_j^{\text{Ridge}}(\lambda)$, pour $j = 1, \cdots, p$. On affichera cette fonction avec des couleurs différentes pour chaque groupe de variables.

On pourra consulter les sites suivants pour plus de détails sur l'estimateur Ridge lui-même :

⋆⋆⋆ https://onlinecourses.science.psu.edu/stat857/node/155 (cours en ligne)

⋆⋆⋆ http://stat.genopole.cnrs.fr/_media/members/jchiquet/teachings/ridge.pdf (en français)

⋆⋆ http://lear.inrialpes.fr/people/harchaoui/teaching/2013-2014/ensl/m2/lecture3.pdf (page 8, pour les détails sur la décomposition biais variance)

- Lasso -

Pour cet estimateur on utilise la fonction `lars` du package `lars`.

- Noter que pour cet estimateur le chemin de régularisation est affine par morceaux et les changements de pentes sont données par `object_lasso$lambda` si `object_lasso=lars(X,Y,type="lasso")`. On peut ainsi afficher sur un même graphique les p fonctions $\lambda \to \hat{\theta}_j^{\text{Lasso}}(\lambda)$, pour $j = 1, \cdots, p$. De nouveau on affichera cette fonction avec des couleurs différentes pour chaque groupe de variables.
- Il est classique de considérer une variante non biaisée du Lasso (nommée Gauss-Lasso, debiased-Lasso, etc.) : au lieu de prendre l'estimateur du Lasso, on garde en fait uniquement les coordonnées non nulles produites par la procédure et l'on utilise un estimateur des moindre carrés sur cet ensemble. Ecrivez une fonction qui produit un tel résultat.

On pourra consulter les sites suivants pour plus de détails sur l'estimateur Lasso :

⋆⋆⋆ http://statweb.stanford.edu/~tibs/lasso.html (papier original)

⋆⋆⋆ http://arxiv.org/pdf/1112.3450.pdf (extension)

⋆⋆ http://jmlr.org/proceedings/papers/v9/lorbert10b/lorbert10b.pdf (extension)

- Elastic Net -

Pour cet estimateur on utilise la fonction `enet` du package `enet` (remarquer que le package `glmnet` est également envisageable).

- Il est bon de voir que même si l'estimateur Elastic Net est présenté comme ayant une pénalité de la forme $\lambda_1\|\theta\|_1 + \lambda_2\|\theta\|_2^2$, dans la pratique l'algorithme fournit à λ_2 fixé toutes les solutions obtenues en faisant varier λ_1.
- Afficher sur un même graphique les p fonctions $\lambda_1 \to \hat{\theta}_j^{\text{E-net}}(\lambda_1, \lambda_2)$, pour $j = 1, \cdots, p$, avec λ_2 fixé. De nouveau on affichera cette fonction avec des couleurs différentes pour chaque groupe de variables.
- Faite varier λ_2 dans la question ci-dessus.

On pourra consulter les sites suivants pour plus de détails sur l'estimateur l'Elastic Net :

⋆⋆⋆ http://www.stanford.edu/~hastie/TALKS/glmnet.pdf

⋆⋆⋆ http://www.stanford.edu/~hastie/Papers/B67.2%20%282005%29%20301-320%20Zou%20&%20Hastie.pdf (papier original)

⋆⋆⋆ http://arxiv.org/pdf/1112.3450.pdf

⋆⋆ http://jmlr.org/proceedings/papers/v9/lorbert10b/lorbert10b.pdf (extension)

- Bilan (Bonus) -

Pour chacune des expériences déterminer laquelle des trois méthodes est la plus performante. On pourra comparer les erreurs de prédiction et d'estimation, en faisant par exemple de la validation croisée.