

## TP: MÉTHODES DE RÉDUCTION DE DIMENSION

**Vous aurez besoin dans ce TP d'utiliser les librairies R suivantes :**

*MASS*, *glmnet* et *elasticnet*

L'objectif de ce TP est de démontrer les propriétés de procédures de sélection de modèle visant à réduire leur dimension. Ces propriétés seront illustrées à travers des exemples sur des jeux de données simulés pour lesquels le modèle théorique exact est donc connu. La régularisation est essentielle pour traiter les problèmes de prédicteurs corrélés (le LASSO en éjectera, tandis que le RIDGE les corrigera).

### Cas de la régression RIDGE

On se propose ici de simuler un jeu de données ayant les caractéristiques suivantes :

- le jeu de données comporte 500 observations, et  $z \sim \mathcal{N}(0, 1)$  ;
- on définit la réponse  $Y$  comme  $y = z + 0,2 \times \mathcal{N}(0, 1)$
- la covariable  $x_1$  est définie par  $x_1 = z + \mathcal{N}(0, 1)$
- la covariable  $x_2$  est définie par  $x_2 = z + \mathcal{N}(0, 1)$  (autre simulation que précédemment)
- la covariable  $x_3$  est définie par  $x_3 = x_1 + x_2$ .

- (1) Créer un data frame de 500 observations composé de  $y$  et du vecteur  $x = (x_1, x_2, x_3)$ .
- (2) Estimer les coefficients d'une régression linéaire multiple (sans intercept, comme dans toute la suite du TP).
- (3) Consulter les résultats de l'estimation et interpréter.
- (4) Quel problème apparaît ? Conclure.
- (5) Estimer les coefficients d'une régression Ridge en incluant les mêmes covariables, après avoir déterminé le paramètre de tuning  $\lambda$  optimal.
- (6) Consulter les résultats de l'estimation des paramètres. Conclure.

On se propose maintenant de simuler le jeu de données suivant : mêmes données que précédemment, sauf que  $x_3 = x_1 + x_2 + 0.05 \times \mathcal{N}(0, 1)$ .

Créer un échantillon d'apprentissage de 400 individus, et un échantillon de validation de 100 personnes.

- (1) Estimer les coefficients d'une régression linéaire multiple sur l'échantillon d'apprentissage.
- (2) Consulter les résultats de l'estimation et interpréter.
- (3) Effectuer des prévisions sur l'échantillon de validation. Calculer la MSE sur cet échantillon test.
- (4) Estimer les coefficients d'une régression Ridge en incluant les mêmes covariables, après avoir déterminé le paramètre de tuning  $\lambda$  optimal.
- (5) Consulter les résultats de l'estimation et interpréter.
- (6) Effectuer des prévisions sur l'échantillon de validation. Calculer la MSE sur cet échantillon test. Comparer les MSE via ces deux approches différentes.

### Cas de la régression LASSO

On se propose ici de simuler un jeu de données de 20 observations ayant les caractéristiques suivantes :

- le vecteur théorique des coefficients de régression servant à simuler l'échantillon est le suivant :  $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8) = (3, 1.5, 0, 0, 2, 0, 0, 0)$  (il y aura donc 8 covariables)
- on considère la matrice de corrélation  $C$  de taille 8x8, dont les éléments valent

$$C[i, j] = 0.3^{|i-j|}$$

- le vecteur de covariables  $\mathbf{X}$  est simulé à partir d'une loi normale multivariée de moyennes 0 et de matrice de variance-covariance la matrice  $C$ .
- la réponse  $Y$  vaut  $Y = X\beta + 3 \times \mathcal{N}(0, 1)$ .

On va maintenant étudier la robustesse de chaque méthode d'estimation, à travers plusieurs simulations. On crée ainsi 100 jeux de données tels que celui ci-dessus.

Pour chaque jeu de données :

- (1) Créer un data frame correspondant à ces données.
- (2) Estimer les coefficients d'une régression linéaire multiple.
- (3) Consulter les résultats de l'estimation et récupérer le vecteur des coefficients de régression estimés.
- (4) Calculer l'erreur quadratique entre les coefficients estimés et les coefficients théoriques dont vous disposez.

Calculer ensuite la médiane des erreurs quadratiques dans cette modélisation classique.

Répéter les mêmes étapes en estimant les paramètres via une régression Ridge. Puis une régression LASSO. Comparer les 3 médianes obtenues. Quelle est la méthode la plus adaptée ? Pourquoi et cela vous paraît-il logique ?

### Cas de la régression ELASTIC NET

On se propose ici de simuler un jeu de données de 100 observations telles que :

- 6 prédicteurs dont 3 sont dominants ( $x_1, x_2, x_3$ ) et 3 sont négligeables ( $x_4, x_5, x_6$ ).  
On aimerait annuler l'effet de ces 3 derniers via le LASSO.
- 2 facteurs de risque indépendants  $z_1$  et  $z_2$  de loi uniforme entre 0 et 20, avec une réponse  $y = z_1 + 0.1 * z_2 + \mathcal{N}(0, 1)$
- Voici les covariables corrélées groupées :  $x = (x_1, x_2, x_3, x_4, x_5, x_6)$  avec
  - $x_1 = z_1 + \epsilon_1$  ,  $x_2 = -z_1 + \epsilon_2$  ,  $x_3 = z_1 + \epsilon_3$ ,  $\epsilon_k \sim \mathcal{N}(0, 1)$
  - $x_4 = z_2 + \epsilon_4$  ,  $x_5 = -z_2 + \epsilon_5$  ,  $x_6 = z_2 + \epsilon_6$

On essaie ici de comparer le LASSO à l'ELASTIC NET dans ce cas de données corrélés.

- (1) Créer un data frame correspondant à ces données.
- (2) Estimer les coefficients d'une régression LASSO de  $Y$  par  $\mathbf{X}$ .
- (3) Consulter les résultats de l'estimation et récupérer le vecteur des coefficients de régression estimés.
- (4) Faire de même avec une procédure ELASTIC-NET. Vous devriez distinguer des groupements de coefficients qui n'apparaissent pas dans le cadre du LASSO.