CHOIX DU PARAMETRE DE LISSAGE DANS LES MODELES DE STATISTIQUE NON PARAMETRIQUE EN GRANDE DIMENSION

Soutenance de Stage du Master 1 Statistique et Sciences de Données (SSD)

Kodjo Mawuena AMEKOE

28 août 2020

sous la direction de :

Mme Sana LOUHICHI
M. Didier GIRARD
M. Karim BENHENNI

Plan de l'exposé

Contexte

Considérons le jeu de données $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, où $x_i \in \mathbb{R}^p$ et $y_i \in \mathbb{R}$ avec les observations (iid), issues de la distribution conjointe $q(y_i | x_i^\top \beta^*) p(x_i)$.

Contexte

Considérons le jeu de données $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, où $x_i \in \mathbb{R}^p$ et $y_i \in \mathbb{R}$ avec les observations (iid), issues de la distribution conjointe $q(y_i|x_i^\top \beta^*)p(x_i)$.

On cherche à estimer β^* pour faire la prédiction pour une nouvelle observation (x_{new}, y_{new}) issue de la distribution $q(y|x^\top \beta^*)p(x)$, indépendante de $\mathcal{D}$. Une méthode classique consiste à choisir la fonction de pénalité r, la fonction de perte l, le paramètre de lissage λ et à résoudre le problème d'optimisation :

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left[\sum_{i=1}^{n} l(y_i|x_i^\top \beta) + \lambda r(\beta) \right] \tag{1}$$

Contexte

Considérons le jeu de données $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, où $x_i \in \mathbb{R}^p$ et $y_i \in \mathbb{R}$ avec les observations (iid), issues de la distribution conjointe $q(y_i|x_i^\top \beta^*)p(x_i)$.

On cherche à estimer β^* pour faire la prédiction pour une nouvelle observation (x_{new}, y_{new}) issue de la distribution $q(y|x^\top \beta^*)p(x)$, indépendante de $\mathcal{D}$. Une méthode classique consiste à choisir la fonction de pénalité r, la fonction de perte l, le paramètre de lissage λ et à résoudre le problème d'optimisation :

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left[\sum_{i=1}^{n} l(y_i|x_i^\top \beta) + \lambda r(\beta) \right] \quad (1)$$

De même on s'interesse à l'estimation à partir de $\mathcal{D}$ de l'erreur de prédiction définie par :

$$Err_{extra} = E[\phi(y_{new}, x_{new}^\top \hat{\beta})|\mathcal{D}] \quad (2)$$

.

Contexte

Considérons le jeu de données $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, où $x_i \in \mathbb{R}^p$ et $y_i \in \mathbb{R}$ avec les observations (iid), issues de la distribution conjointe $q(y_i|x_i^\top \beta^*)p(x_i)$.

On cherche à estimer β^* pour faire la prédiction pour une nouvelle observation (x_{new}, y_{new}) issue de la distribution $q(y|x^\top \beta^*)p(x)$, indépendante de $\mathcal{D}$. Une méthode classique consiste à choisir la fonction de pénalité r, la fonction de perte l, le paramètre de lissage λ et à résoudre le problème d'optimisation :

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left[\sum_{i=1}^{n} l(y_i|x_i^\top \beta) + \lambda r(\beta) \right] \tag{1}$$

De même on s'interesse à l'estimation à partir de $\mathcal{D}$ de l'erreur de prédiction définie par :

$$Err_{extra} = E[\phi(y_{new}, x_{new}^\top \hat{\beta})|\mathcal{D}] \tag{2}$$

.

Une technique simple et intuitive est **la validation croisée**.

Motivation

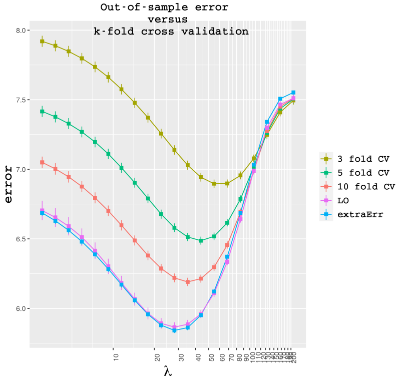

FIGURE: Comparaison de la validation croisée à 3, 5, 10 parties (folds) et du LO avec l'erreur de prédiction.

Motivation

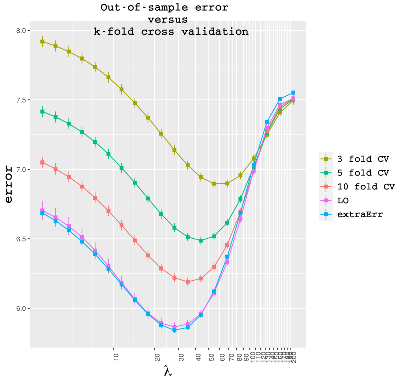

FIGURE: Comparaison de la validation croisée à 3, 5, 10 parties (folds) et du LO avec l'erreur de prédiction.

20 différentes valeurs λ pour une régression linéaire avec la pénalité LASSO. $(p,n,k)=(1000,250,50)$. $y \sim \mathcal{N}(X\beta^\star, \sigma^2 I)$ et $Err_{extra} = \sigma^2 + \|\beta^\star - \hat{\beta}\|_2^2$.

Motivation

- La technique de validation croisée souffre d'un biais large sauf si le nombre de folds (parties) est grand.

Motivation

- La technique de validation croisée souffre d'un biais large sauf si le nombre de folds (parties) est grand.
- Le Leave-One-Out (LO) est très coûteux en terme de ressources de calcul en grande dimension.

Challenge

Trouver une technique aussi précise que le LO mais moins coûteuse en ressource

Alternative

Approximate Leave-One-Out (ALO) pour une large classe des estimateurs comme LASSO [Tib96], Bridge [IEFF93] Elastic-net [ZH05] et méthodes de Régression linéaire, logistique, Poisson, robuste proposée par [RM18].

Temps de calcul

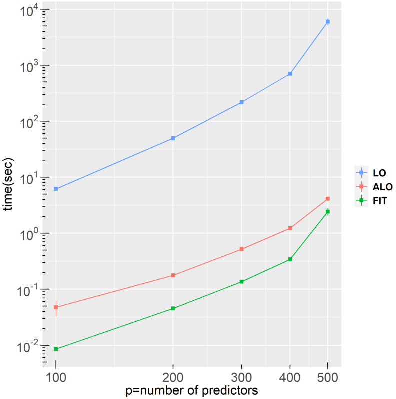

FIGURE: Régression linéaire avec pénalisation Elastic-net pour $\frac{n}{p} = 5$.

Objectifs

- Comprendre le contexte du travail de [RM18] sur l'approximation du leave-one-out.

Objectifs

- Comprendre le contexte du travail de [RM18] sur l'approximation du leave-one-out.
- Discuter les résultats.

Objectifs

- Comprendre le contexte du travail de [RM18] sur l'approximation du leave-one-out.
- Discuter les résultats.
- Reproduire et faire d'autres simulations pour les configurations $n < p$, $n > p$ et $n = p$.

Objectifs

- Comprendre le contexte du travail de [RM18] sur l'approximation du leave-one-out.
- Discuter les résultats.
- Reproduire et faire d'autres simulations pour les configurations $n < p$, $n > p$ et $n = p$.
- Prendre connaissance des problèmes de statistique en grande dimension.

Objectifs

- Comprendre le contexte du travail de [RM18] sur l'approximation du leave-one-out.
- Discuter les résultats.
- Reproduire et faire d'autres simulations pour les configurations $n < p$, $n > p$ et $n = p$.
- Prendre connaissance des problèmes de statistique en grande dimension.
- Développer les compétences en programmation.

Plan de l'exposé

LO

La formule de l'estimation Leave-One-Out (LO) est donnée par :

$$\mathrm{LO} = \frac{1}{n}\sum_{i=1}^{n} \phi(y_i, x_i^\top \hat{\beta}_{/i}) \tag{3}$$

où

LO

La formule de l'estimation Leave-One-Out (LO) est donnée par :

$$\text{LO} = \frac{1}{n}\sum_{i=1}^{n} \phi(y_i, x_i^\top \hat{\beta}_{/i}) \tag{3}$$

où

$$\hat{\beta}_{/i} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left[\sum_{j \neq i} l(y_j | x_j^\top \beta) + \lambda r(\beta) \right] \tag{4}$$

LO

La formule de l'estimation Leave-One-Out (LO) est donnée par :

$$\text{LO} = \frac{1}{n}\sum_{i=1}^{n} \phi(y_i, x_i^\top \hat{\beta}_{/i}) \tag{3}$$

où

$$\hat{\beta}_{/i} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left[\sum_{j \neq i} l(y_j | x_j^\top \beta) + \lambda r(\beta) \right] \tag{4}$$

Challenge

LO est très coûteux en grande dimension : n calculs de $\hat{\beta}_{/i}$.

ALO

Solution

En supposant que $\hat{\beta}_{/i}$ est proche de $\hat{\beta}$, on utilise la méthode de Newton qui permet d'obtenir une approximation de $\hat{\beta}_{/i}$.

Solution

Avec quelques manipulations algébriques , on obtient ALO :

$$\text{ALO} = \frac{1}{n}\sum_{i=1}^{n}\phi(y_i, x_i^\top \tilde{\beta}_{/i}) = \frac{1}{n}\sum_{i=1}^{n}\phi\left(y_i, x_i^\top \hat{\beta} + \left(\frac{\dot{l}_i(\hat{\beta})}{\ddot{l}_i(\hat{\beta})}\right)\left(\frac{H_{ii}}{1-H_{ii}}\right)\right) \tag{5}$$

avec

$$H = X(\lambda \text{diag}[\ddot{r}(\hat{\beta})] + X^\top \text{diag}[\ddot{l}(\hat{\beta})]X)^{-1}X^\top \text{diag}[\ddot{l}(\hat{\beta})]. \tag{6}$$

ALO

Challenge

L'approche utilisée ci-dessus nécessite des fonctions de perte et pénalité lisses, donc ne peut pas s'appliquer directement à :

- LASSO :

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left[\sum_{i=1}^{n} l(y_i | x_i^\top \beta) + \lambda \|\beta\|_1 \right] \tag{7}$$

ALO

Challenge

L'approche utilisée ci-dessus nécessite des fonctions de perte et pénalité lisses, donc ne peut pas s'appliquer directement à :

- LASSO :

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left[\sum_{i=1}^{n} l(y_i | x_i^\top \beta) + \lambda \|\beta\|_1 \right] \tag{7}$$

- Elastic-net :

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left[\sum_{i=1}^{n} l(y_i | x_i^\top \beta) + \lambda_1 \|\beta\|_2^2 + \lambda_2 \|\beta\|_1 \right]. \tag{8}$$

Solution

Faire une approximation lisse de la fonction de pénalité r : approximation de Schmidt [MGR07] pour LASSO.

Plan de l'exposé

Ecart ALO et LO

Théorème important

[RM18] a montré avec une probabilité qui tend vers 1 avec n et p que $|ALO - LO| \leq O_p(\frac{PolyLog(n)}{\sqrt{n}})$ pour $r(\beta) = \gamma\beta^2 + (1-\gamma)r^{\alpha}(\beta)$ avec $0 < \gamma < 1$ (Elastic-net).

En supposant que :

- $\frac{n}{p} = \delta_0$, avec δ_0 un nombre fini ,borné loin de zéro(0) ;

Ecart ALO et LO

Théorème important

[RM18] a montré avec une probabilité qui tend vers 1 avec n et p que $|ALO - LO| \leq O_p(\frac{PolyLog(n)}{\sqrt{n}})$ pour $r(\beta) = \gamma\beta^2 + (1-\gamma)r^\alpha(\beta)$ avec $0 < \gamma < 1$ (Elastic-net).

En supposant que :

- $\frac{n}{p} = \delta_0$, avec δ_0 un nombre fini ,borné loin de zéro(0) ;
- les lignes de $X \in \mathbb{R}^{n \times p}$ sont indépendantes et suivent une loi Gaussienne de moyenne 0 , de matrice de covariance Σ dont la valeur propre maximale est $\rho_{\max} = \frac{c}{n}$, c une constante ;

Ecart ALO et LO

Théorème important

[RM18] a montré avec une probabilité qui tend vers 1 avec n et p que $|ALO - LO| \leq O_p(\frac{PolyLog(n)}{\sqrt{n}})$ pour $r(\beta) = \gamma\beta^2 + (1-\gamma)r^\alpha(\beta)$ avec $0 < \gamma < 1$ (Elastic-net).

En supposant que :

- $\frac{n}{p} = \delta_0$, avec δ_0 un nombre fini ,borné loin de zéro(0) ;
- les lignes de $X \in \mathbb{R}^{n\times p}$ sont indépendantes et suivent une loi Gaussienne de moyenne 0 , de matrice de covariance Σ dont la valeur propre maximale est $\rho_{\max} = \frac{c}{n}$, c une constante ;
- la dérivée seconde des fonctions de perte et pénalité est régulière ;

Ecart ALO et LO

Théorème important

[RM18] a montré avec une probabilité qui tend vers 1 avec n et p que $|ALO - LO| \leq O_p(\frac{PolyLog(n)}{\sqrt{n}})$ pour $r(\beta) = \gamma\beta^2 + (1-\gamma)r^{\alpha}(\beta)$ avec $0 < \gamma < 1$ (Elastic-net).

En supposant que :

- $\frac{n}{p} = \delta_0$, avec δ_0 un nombre fini ,borné loin de zéro(0) ;
- les lignes de $X \in \mathbb{R}^{n\times p}$ sont indépendantes et suivent une loi Gaussienne de moyenne 0 , de matrice de covariance Σ dont la valeur propre maximale est $\rho_{\max} = \frac{c}{n}$, c une constante ;
- la dérivée seconde des fonctions de perte et pénalité est régulière ;
- $\phi(.,.) = l(.,.)$.

Plan de l'exposé

Résultats expérimentaux de l'approximation du LO par ALO pour :

- Régression linéaire avec pénalisation Elastic-net ;

Simulations et discussions

Résultats expérimentaux de l'approximation du LO par ALO pour :

- Régression linéaire avec pénalisation Elastic-net ;
- Régression logistique avec pénalisation LASSO ;

Résultats expérimentaux de l'approximation du LO par ALO pour :

- Régression linéaire avec pénalisation Elastic-net ;
- Régression logistique avec pénalisation LASSO ;
- Régression Poisson Elastic-net ;

Régression linéaire avec pénalisation Elastic-net

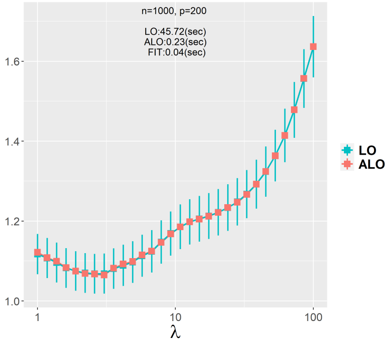

FIGURE: $n > p$

Régression linéaire avec pénalisation Elastic-net

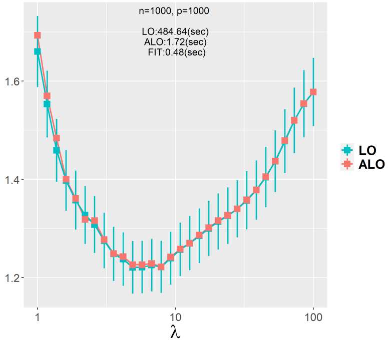

FIGURE: $n = p$

Régression linéaire avec pénalisation Elastic-net

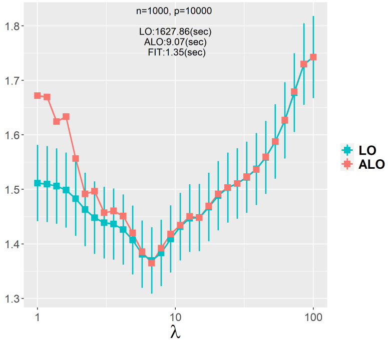

FIGURE: $n < p$

Plan de l'exposé

Contribution

Impact du paramètre de pondération de l'elastic-net mis à zéro : $\gamma = 0$.

Approximation de Schmidt :

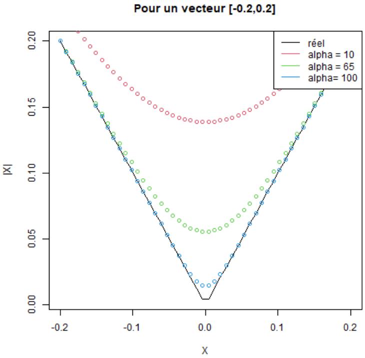

FIGURE: Séquence de 50 points dans l'intervalle $[-0.2, 0.2]$

Pénalisation de Schmidt :

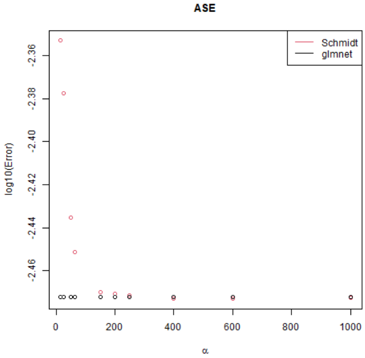

FIGURE: Comparaison ASE Estimation de Schmidt Vs LASSO glmnet

Programmation de ALO^{α}

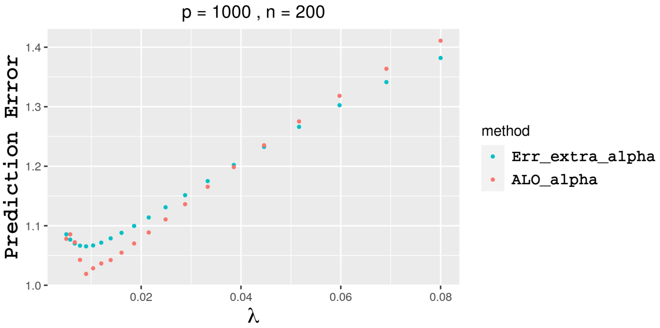

FIGURE: Comparaison ASE Estimation de Schmidt Vs LASSO glmnet

ALO$^{\alpha}$ VS ALO

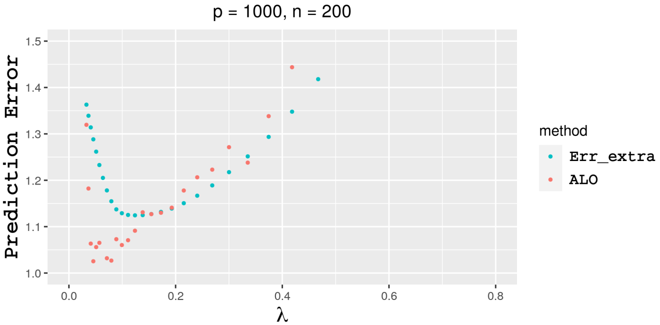

FIGURE: Comparaison de l'erreur de prédiction et ALO pour une régression linéaire LASSO

Contribution

Influence de fortes et faibles corrélations : X avec la structure de corrélation Toeplitz : $corr(X_{ij}, X_{ij'}) = \rho^{|j-j'|}$ avec $i = 1, 2, ..., n$, $j, j' = 1, 2, .., p$ et $\rho \in \{0.02, 0.7, 0.99\}$.

Plan de l'exposé

Conclusion

- Missions remplies.

Conclusion

- Missions remplies.
- Importance d'avoir les observations(iid) et $\frac{n}{p}$ fixe.

Conclusion

- Missions remplies.
- Importance d'avoir les observations(iid) et $\frac{n}{p}$ fixe.
- Visualisations avec ggplot2, optimisation des programmes, parallélisation sous R.

Difficultés et Perspectives

- Ressources de calcul, temps limité.

Difficultés et Perspectives

- Ressources de calcul, temps limité.
- Traitement des données réelles.

Difficultés et Perspectives

- Ressources de calcul, temps limité.
- Traitement des données réelles.
- Optimisation du code de l'approximation de Schmidt et calcul de ALO^{α}.

Difficultés et Perspectives

- Ressources de calcul, temps limité.
- Traitement des données réelles.
- Optimisation du code de l'approximation de Schmidt et calcul de ALO^{α}.
- Utilisation d'approximations de Monte-Carlo.

Plan de l'exposé

References I

- Ildiko E. Frank and Jerome H. Friedman, *A statistical view of some chemometrics regression tools*, Technometrics **35** (1993), no. 2, 109–135.
- Schmidt M., Fung G., and Rosales R, *Fast optimization methods for l1 regularization : A comparative study and two new approaches*, ECML **4701** (2007), 286–297.
- Kamiar Rahnama Rad and Arian Maleki, *A scalable estimate of the extra-sample prediction error via approximate leave-one-out*, arXiv : Methodology (2018).
- Robert Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological) **58** (1996), no. 1, 267–288.
- Hui Zou and Trevor Hastie, *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society. Series B (Statistical Methodology) **67** (2005), no. 2, 301–320.

Merci pour votre attention.