Projet n°8 : stability selection

M1 parcours SSD - UE Apprentissage Statistique I

La sélection de variables par pénalisation Lasso est connue pour être sensible au degré de corrélation entre les prédicteurs, ce qui se traduit par une instabilité et une tendance à sélectionner plus de variables que nécessaire. Plusieurs techniques de ré-échantillonnage ont été proposées pour palier ces limitations, comme par exemple le "bootstrap lasso" [1] et la "stability selection" [2].

L'approche dite de "stability selection" est particulièrement simple. Elle consiste :

1. à construire plusieurs chemins de régularisation (de l'ordre de la centaine) à partir de sous-ensembles aléatoires du jeu de données disponible (en général de taille égale à la moitié du jeu de données total).
2. à calculer pour chaque variable un "chemin de stabilité" (stability path), correspondant à la fréquence à laquelle elle est sélectionnée le long du chemin de régularisation dans les différentes répétitions.
3. à sélectionner les variables dont les chemins de stabilité dépassent un certain seuil (fixé a priori ou optimisé par validation croisée).
4. à construire le modèle final sur la base de ces variables, de façon non pénalisée.

La Figure 1 illustre la notion de chemin de stabilité.

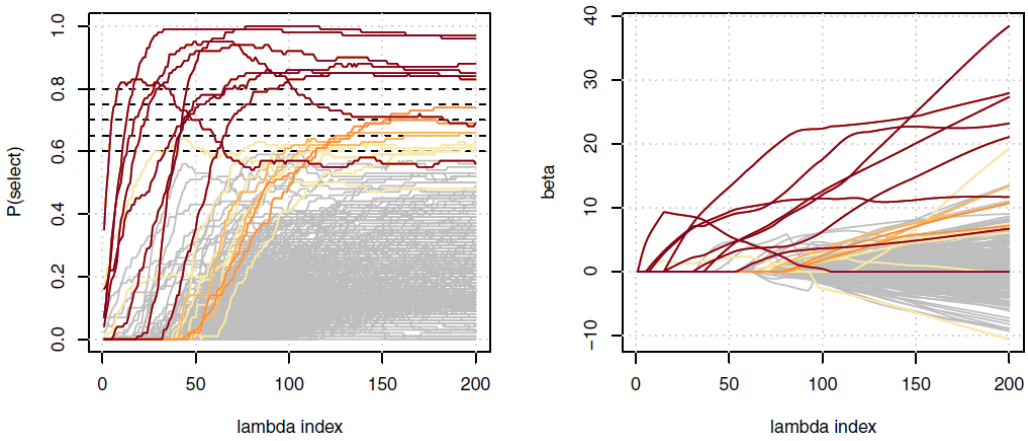

FIGURE 1 – Illustration des chemins de stabilité. Droite : chemin du régularisation obtenue par pénalisation Lasso "classique". Gauche : chemin de stabilité obtenue en agrégeant 100 chemins de régularisation, obtenus à partir de sous-ensembles aléatoires du jeu de données global. L'axe des abscisses correspond aux indices du paramètre de régularisation (λ) dans la grille proposée par `glmnet`. Chaque courbe correspond à une variable et représente la fréquence à laquelle elle a été choisie, à un niveau de pénalisation donnée, lors des 100 répétitions du Lasso. Le code couleur illustre les variables qui sont sélectionnées pour des seuils de sélection allant de 0.8 (en rouge) à 0.6 (en jaune).

L'objectif de ce projet est d'évaluer l'intérêt de cette approche sur une problématique de génomique bactérienne : prédire la résistance à un antibiotique (la streptomycine) d'une souche bactérienne (de l'espèce *Mycobacterium tuberculosis*) à partir de son génome. Le jeu de données considéré provient d'une publication récente [3]. Il est stocké dans le fichier `TB.Rdata` contenant :

- un jeu d'apprentissage, constitué (1) d'une matrice `X.train` de taille 966×53677, encodant la présence de 53677 motifs génomiques dans le génome de 966 souches et (2) d'un vecteur `y.train` contenant le phénotype de résistance de ces 966 souches (+1 : résistant ; -1 : sensible).
- un jeu de test, constitué d'une matrice `X.test` et d'un vecteur `y.test`, contenant les mêmes informations pour 200 nouvelles souches.

Notez que les matrices `X.train` et `X.test` sont stockées au format "sparse".

Objectifs

L'objectif est donc d'implémenter la procédure de stability selection et d'évaluer son intérêt en terme (1) de performance de prédiction et (2) de support (i.e., de nombre de variables sélectionnées). Pour cela on se comparera à l'approche Lasso classique en mettant en oeuvre la procédure suivante :

1. Optimiser un modèle de régression logistique avec pénalisation Lasso par validation croisée, en utilisant le package `glmnet`.
2. Implémenter une procédure de "stability selection" et construire les modèles obtenus en faisant varier le seuil de sélection de 0.6 à 1 par pas de 0.05.
3. Comparer les différents modèles obtenus selon (1) leurs performances sur le jeu de test, et (2) la taille de leur supports.

Pour aller plus loin, et si les ressources informatiques le permettent, on pourra mettre en oeuvre une procédure de validation croisée pour optimiser le seuil de sélection de la procédure de stability selection.

Quelques points sont importants à noter :

- On considèrera 100 répétitions du Lasso pour la procédure de stability selection, en tirant à chaque fois un jeu de données de taille $n/2 = 483$ parmi les $n = 966$ observations disponibles.
- En pratique, il est préférable d'utiliser la même grille pour le paramètre de régularisation pour les différentes répétitions de la procédure de stability selection. On utilisera celle qui est obtenue lorsque l'on réalise la procédure Lasso "classique".
- La parcimonie du support étant un critère important, nous considèrerons la stratégie `lambda.1se` proposée par `glmnet` pour choisir le paramètre de régularisation de l'approche Lasso "classique".
- Pour mesurer la performance du modèle, on considèrera le taux de bonne classification global.
- Enfin, on note qu'on peut se contenter de ne faire qu'une fois la répétition des 100 Lassos pour en déduire les modèles à différents seuils de sélection.

Le minimum

Voici les consignes qui empêcheraient d'avoir la moyenne si elles ne sont pas respectées :

- Le compte-rendu ne doit pas faire plus de 6 pages (sans compter les éventuelles annexes).
- Il doit contenir le nom des auteurs, un titre explicite, une introduction et une conclusion.
- Un graphique similaire à la Figure 1 comparant les chemins de régularisation et de stabilité doit être présenté et commenté.

— Un graphique comparant les performances obtenues sur le jeu de test et la taille des supports doit être présenté et commenté.
— Le taux d'agrément entre les différents supports et l'amplitude des coefficients des différents modèles doit être analysée.
— Le code mis en oeuvre doit apparaître, commenté un minimum, en annexe (et uniquement en annexe).

Références

[1] Francis R. Bach. Bolasso : model consistent lasso estimation ,through the bootstrap. In William W. Cohen, Andrew Mccallum, and Sam T. Roweis, editors, *International Conference on Machine Learning*, pages 33–40, 2008.

[2] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society, Series B*, 72 :417–473, 2010.

[3] James J. Davis, Sébastien Boisvert, Thomas Brettin, Ronald W. Kenyon, Chunhong Mao, Robert Olson, Ross Overbeek, John Santerre, Maulik Shukla, Alice R. Wattam, Rebecca Will, Fangfang Xia, and Rick Stevens. Antimicrobial resistance prediction in PATRIC and RAST. *Scientific Reports*, 6 :27930, 2016.