Elastic Net avec gestion des interactions et débiaisage

Florent Bascou*, Sophie Lèbre**
Joseph Salmon***

*IMAG, Univ. Montpellier, CNRS Montpellier, France
florent.bascou@umontpellier.fr,
**IMAG, Univ. Montpellier, CNRS Montpellier, France
Univ. Paul-Valéry-Montpellier 3, Montpellier, France
sophie.lebre@umontpellier.fr
***IMAG, Univ. Montpellier, CNRS Montpellier, France
joseph.salmon@umontpellier.fr

Résumé. Nous présentons quelques résultats statistiques pour un modèle de régression pénalisée et dé-biaisée pour ajuster, en grande dimension, un modèle linéaire parcimonieux avec interactions. L'analyse statistique s'intéresse à l'approche proposée par Bascou et al. (2020), notamment afin d'illustrer le fonctionnement du débiaisage. De plus, on montre qu'il permet de sélectionner moins de variables que le modèle sans débiaisage sur données simulées et réelles.

1 Introduction

Grâce à leur interprétabilité, les modèles linéaires sont populaires, néanmoins, le nombre de variables explicatives est souvent supérieur au nombre d'observations, de sorte qu'une régularisation est nécessaire. Des techniques de régularisation exploitant la norme ℓ_1 ont conduit à la création de nombreux estimateurs, dont les plus connus sont le Lasso (Tibshirani, 1996) et l'Elastic Net (Zou et Hastie, 2005). Pour traiter les interactions entre variables, la parcimonie est cruciale, car même limité aux d'ordre deux, le nombre de variables grandit déjà de manière quadratique. Nous estimons les coefficients à l'aide de l'estimateur Elastic Net qui permet de réduire le nombre de variables grâce à la pénalité ℓ_1, tout en tenant compte des fortes corrélations entre variables grâce à la pénalité ℓ_2 (Tikhonov, 1943; Hoerl et Kennard, 1970). Dans Bascou et al. (2020), nous avons adapté un algorithme de descente de coordonnées (popularisé par `glmnet` (Friedman et al., 2007, 2010)) pour que la matrice d'interaction n'ait pas besoin d'être stockée en mémoire. Par ailleurs, sachant que l'Elastic Net contracte les grands coefficients vers zéro, dans Bascou et al. (2020) nous suggérions de calculer une version non dé-biaisée des coefficients (Deledalle et al., 2017) afin de proposer un algorithme approchant le LS Elastic Net (Elastic Net suivi des moindres carrés sur le support). Dans ce travail, nous comparons l'Elastic Net et sa version débiaisée sur données simulées avec différents scénarios d'hérédité pour montrer que le débiaisage améliore la sélection de variables. Nous terminons par une étude sur données réelles, pour confirmer ce comportement.