ibr: un paquet R pour la réduction itérative de biais

P.A. Cornillon[a] and **E. Matzner-Løber**[a]

[a]IRMAR 6625
Université de Rennes 2
CS 24307 - 35043 RENNES CEDEX
pac@uhb.fr, eml@uhb.fr

Mots clefs : Statistique, Régression, Lissage, Splines, Noyaux, L2 boosting.

Introduction

Si nous souhaitons expliquer une variable Y par un ensemble de d variables explicatives $X_1, \cdots, X_d$, la régression constitue un outil classique de la statistique. Cette famille de modélisation comprend la régression paramétrique linéaire ou non-linéaire, la régression non-paramétrique utilisant des lisseurs construits à partir d'ondelettes, de noyaux, de splines [1].
Dès que le nombre d'observations est modéré (de l'ordre de plusieurs centaines) et que le nombre de variables d est plus grand que 3 ou 4, les approches non-paramétriques classiques rencontrent le problème dit du fléau de la dimension. Dans ce cas, un modèle structurel est souvent utilisé: par exemple un modèle additif, des directions révélatrices ou MARS.
Le boosting est aussi une réponse possible au problème de régression [2] et cette méthode possède maintenant de nombreux développements comme adaboost, logitboost pour la discrimination ou le L_2 boosting pour la régression. Cette dernière peut être utilisée avec de nombreux lisseurs et donne lieu à une modélisation additive par composante [3].
A contrario, la réduction itérative de biais permet d'estimer une fonction de régression multivariée directement via un lisseur multivarié sans hypothèse paramétrique ou sans contraintes structurelles.

Réduction itérative de biais

Soit le modèle de régression suivant

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

où $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $m(\cdot)$ est une fonction lisse inconnue et les erreurs ε_i sont des variables aléatoires indépendantes des covariables, indépendantes entre elles, de moyenne nulle et de variance constante σ^2. Sous forme vectorielle en rassemblant les observations en vecteurs colonnes:

$$Y = m + \varepsilon. \tag{2}$$

Considérons une matrice de lissage S_λ. L'estimateur classique via le lisseur S_λ est

$$\hat{m}_1 = S_\lambda Y.$$

Le biais est alors

$$B(\hat{m}_1) = \mathbb{E}[\hat{m}_1 | X] - m = (S_\lambda - I)m$$

et une manière de l'estimer est de remplacer m inconnue par un estimateur, par exemple $\hat{m}_1$. L'estimateur corrigé de l'estimation du biais est alors

$$\hat{m}_2 \;=\; \hat{m}_1 - \hat{B}(\hat{m}_1) = (S_\lambda + S_\lambda(I - S_\lambda))Y = (I - (I - S_\lambda)^2)Y,$$

et on peut recommencer, donnant ainsi à l'itération k :

$$\hat{m}_k \;=\; S_\lambda[I + (I - S_\lambda) + (I - S_\lambda)^2 + \ldots + (I - S_\lambda)^{k-1}]Y = [I - (I - S_\lambda)^k]Y.$$

Le lissage revient alors à choisir le nombre d'itérations k, ce qui peut être fait par des critères de choix de modèles classiques comme GCV, AIC ou AICc. Des résultats théoriques sont envisagés dans [4].

Le paquet ibr

Le paquet ibr, disponible sur CRAN, implémente simplement la procédure ci-dessus avec quelques lisseurs multivariés classiques comme les splines plaques minces, un estimateur à noyaux gaussiens ou des lisseurs moins classiques comme les splines de Duchon [5].
La procédure utilise la méthode S3 de manière classique et permet un enchaînement standard:

```
R> data("ozone",package="ibr")
R> res.ibr <- ibr(x=ozone[,-1],y=ozone[,1],df=1.1,smoother="k",crit="gcv")
R> summary(res.ibr)
Residuals:
     Min       1Q   Median       3Q      Max
-13.5581  -2.0566  -0.3481   1.9816  12.6049
Residual standard error: 71.69 on 309.6 degrees of freedom

Initial df: 2.06 ; Final df: 20.42
  gcv
2.809

Number of iterations: 64 chosen by gcv
Base smoother: gaussian kernel (with 2.06 df)
```

Le choix des lisseurs S_λ et du critère d'arrêt est contrôlé par les arguments `smoother` et `crit`. Une prédiction pour la classe `ibr` est aussi disponible via la fonction `predict`. L'utilisation de cette méthode est donc en tout point semblable aux méthodes classiques comme `lm`.

Références

[1] Hastie, T.J., Tibshirani, R.J., Friedman, J.H. (2001). *The elements of statistical learning: data mining, inference and prediction.* Springer, New-York.
[2] Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, **28**, 337–407.
[3] Bühlmann, P., Hothorn, T. (2007). Boosting algorithms: regularization, prediction and model fitting (with discussion). *Statistical Science*, **22**, 477–505.
[4] Cornillon, P.A., Hengartner, N., Matzner-Løber, E. (2014). Recursive bias estimation for multivariate regression smoothers. *ESAIM / probability and statistics*, **18**, 483–502.
[5] Duchon, J., (1977). Splines minimizing rotation-invariant semi-norms in Solobev spaces. In W. Shemp, K. Zeller (eds.), *Construction theory of functions of several variables,* pp. 85–100. Springer, Berlin.