Estimateur de type Lasso pour modèle mixte non-paramétrique

Perrine Soret, Cristian Meza, Marta Avalos, Karine Bertin

Estimateur de type Lasso pour modèle mixte non-paramétrique

Perrine SORET [1,2,3,4] & Cristian Meza [5] & Marta Avalos [1,2,3] & Karine Bertin [5]

[1] *Univ. Bordeaux, ISPED, F–33000 Bordeaux, France*
[2] *INRIA SISTM Bordeaux – Sud–Ouest, F–*33405 *Talence, France*
[3] *INSERM, Centre INSERM U1219–Epidémiologie–Biostatistique, F–33000 Bordeaux, France*
[4] *Vaccine Research Institute (VRI), F–94000 Créteil, France*
[5] *CIMFAV–Facultad de Ingeniería, Univ. de Valparaíso, Valparaíso, Chile*
[*] *perrine.soret@isped.fr*

Résumé. La vraisemblance pénalisée par une norme L1 est devenue relativement standard en grande dimension quand le modèle est supposé basé sur n observations indépendantes et identiquement distribuées. Ces techniques peuvent améliorer la capacité de prédiction (la régularisation implique une réduction de la variance) tout en restant interprétable (la sparsité identifie un sous ensemble de variable avec des effets forts). D'un point de vue computationnel, ces pénalités sont attractives et leurs propriétés théoriques ont été largement étudiées ces dernières années.
Plusieurs auteurs ont récemment suggérer des méthodes pour analyser les données longitudinales ou groupées de grandes dimensions utilisant une pénalisation L_1 dans des modèles mixtes. Ces approches ont été développées pour la sélection de variables dans le cas modèle linéaire mixte et modèle linéaire mixte généralisé mais moins dans le cas de modèle non linéaire mixte.
Peu de travaux ont considéré le problème de sélection de fonctions non linéaire utilisant une méthode de pénalisation de type L_1 dans un modèle mixte non paramétrique avec ou non des covariables. Dans ce cas, les fonctions non linéaire sont approximées par une combinaison linéaire de fonction de lissage (spline, wavelet ou bases de Fourier) possiblement combinées à des fonctions irrégulières (bases de Spiky).

Mots-clés. Données longitudinales, Données complexes, Apprentissage

Abstract. The penalization of likelihoods by L1-norms has become a relatively standard technique for high-dimensional data when the assumed models are based on n independent and identically distributed observations. These techniques may improve prediction accuracy (since regularization leads to variance reduction) together with interpretability (since sparsity identifies a subset of variables with strong effects). Computationally, these penalties are attractive and their theoretical properties have been intensively studied during the last years.
Several authors have recently developed suggestions to analyze high-dimensional clustered

or longitudinal data using L1-penalization methods in mixed effects models. These approaches are mostly developed for variable selection purposes in linear and generalized linear mixed effects models and also, but less extensive, in parametric nonlinear mixed effects models.
Only a few works have considered the problem of selecting nonlinear functions using L1-penalization methods in nonparametric mixed effects models, with additive or nonadditive predictors. Nonlinear functions are approximated by a linear combination of smooth functions (spline, wavelet or Fourier basis functions) possibly combined with more irregular functions (spiky basis functions).

Keywords. Longitudinal data, Complex data, Machine learning

1 Contexte

La vraisemblance pénalisée par une norme L_1 fait est devenue une méthode relativement standard en grande dimension quand le modèle est supposé basé sur n observations indépendantes et identiquement distribuées. Les méthodes de pénalisations ont pour objectif premier de prédire au mieux la réponse tout en équilibrant le biais et la variance.
Les données de grandes dimensions sont de plus en plus courantes dans les études cliniques qui sont le plus souvent longitudinales. Plusieurs auteurs ont récemment suggérer des méthodes pour analyser les données longitudinales ou groupées de grandes dimensions utilisant une pénalisation L_1 dans des modèles mixtes. Ces approches ont été développées pour la sélection de variables dans le cas modèle linéaire mixte [3] et modèle linéaire mixte généralisé [2]. Cependant, on peut utiliser la pénalisation L_1 dans le cas de modèle non linéaire mixte. Dans ce cas, on utilise la pénalisation pour sélectionner des fonctions non linéaire [1].
Les fonctions non linéaire sont approximées par une combinaison linéaire de fonction de lissage (spline, wavelet ou bases de Fourier) possiblement combinées à des fonctions irrégulières (bases de Spiky). Cependant ces méthodes ont été développées dans le cas où le nombre de covariable est faible.

2 Modèle

On considère un modèle non linéaire mixte semiparamétrique.
Soit $i = 1, \cdots, N$ et $j = 1, \cdots, n_i$

$$y_{ij} = g(x_{ij}, \phi_i, f) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.} \tag{1}$$

où $y_{ij} \in \mathbb{R}$ est la $j$ième observation du $i$ème individu, $x_{ij} \in \mathbb{R}^p$ représente les variables de régression connues, g est une fonction commune connue et f est une fonction inconnue

nonparamétrique qui doit être estimée. Les effets aléatoires $\phi_i \in \mathbb{R}^d$ satisfont:

$$\phi_i = A_i\beta + \eta_i, \quad \eta_i \sim (0, \Gamma) \text{ i.i.d.} \tag{2}$$

avec $A_i \in \mathcal{M}_{d,q}$ des matrices de design connues, $\beta \in \mathbb{R}^q$ le vecteur des effets fixes à estimer.

Les paramètres du modèle sont (θ, f) où $\theta = (\beta, \Gamma, \sigma^2)$.

$$g(x_{ij}, \phi_i, f) = a(\phi_i; x_{ij}) + b(\phi_i; x_{ij}) f(c(\phi_i; x_{ij})) \tag{3}$$

a, b et c sont des fonctions connues qui doivent dépendre de i.

3 Estimation

Basé sur le modèle de Ke et Wang, Arribas et al. propose d'estimer (θ, f) par itération suivant les deux étapes suivantes.
A l'étape (k):

i) **Etape paramétrique**
Soit $\hat{f}^{(k-1)}$ l'estimation de f obtenu à l'itération précédente, θ et ϕ sont estimés par NLME à l'aide d'un algorithme SAEM.
La vraisemblance complète s'écrit de la façon suivante:

$$\begin{aligned} p(y, \phi, \theta) &= p(y|\phi, \theta)p(\phi, \theta) \\ &= \frac{1}{(2\pi)^{\frac{n+Np}{2}}(\sigma^2)^{\frac{n}{2}}|\Gamma|^{\frac{N}{2}}} \exp\{\frac{-1}{2}\left(\frac{1}{\sigma^2}\|y - g(\phi, \hat{f}^{(k-1)})\|^2 + \|\tilde{\Gamma}^{-1/2}(\phi - A\beta)\|^2\right)\} \end{aligned}$$

où $n = \sum_{i=1} nn_i$. La log-vraisemblance peut donc s'écrire de la façon suivante:

$$\begin{aligned} \log p(y, \phi; \theta) = -\frac{1}{2}\{C + n\log\sigma^2 + N\log|\Gamma| \\ + \frac{1}{\sigma^2}\|y - g(\phi, \hat{f}^{(k-1)})\|^2 + \sum_{i=1}^{N}(\phi_i - A_i\beta)'\Gamma^{-1}(\phi_i - A_i\beta)\} \end{aligned}$$

où C est une constante qui ne dépend pas de θ.
L'algorithme SAEM remplace l'étape E d'un algorithme EM par une étape (S) de simulation des données manquantes (ϕ) et d'une étape (A) d'approximation. Soit l l'itération, l'algorithme SAEM s'écrit de la façon suivante:

- Etape S: Simulation de m valeurs des effets aléatoires, $\phi^{(l+1,1)}, \cdots, \phi^{(l+1,m)}$ par une loi conditionnelle $p(.|, \theta^{(l)})$

- Etape A: mise à jour de s_{l+1} de la façon suivante:

$$s_{l+1} = s_l + \chi_l\{\frac{1}{m}\sum_{q=1}^{m} S(y, \phi^{(l+1,q)}) - s_k\}$$

- Etape M: mise à jour de θ

$$\theta^{l+1} = \text{argmax}\{-\Psi(\theta) + \langle s_{l+1}, \Phi(\theta)\rangle\}$$

où $(s_l)_l$ est initialisé à s_0 et $(\chi_l)_l$ est une suite décroissantes de nombres qui accélère la convergence.

ii) **Etape non paramétrique**
Soit $\theta^{(k)}$ et $\phi^{(k)}$ estimés à l'étape précédente, nous estimons f est par régression non paramétrique à l'aide d'une méthode de type Lasso.
Le but est de construire une approximation sparse de f à l'aide d'une combinaison linéaire de fonctions. Pour cela, nous construisons un ensemble de fonctions $\{\psi_1, \cdots, \psi_M\}$ appelé le *dictionnaire*. Il peut contenir tout type de base de fonctions tels que des Splines, Wavelets, Fourier ou encore Spiky.
Pour $\gamma \in \mathbb{R}^M$, on note

$$f_\lambda = \sum_{k=1}^{M} \gamma_k \psi_k \tag{4}$$

L'objectif est de trouver un bon candidat pour estimer f qui est une combinaison linéaire de fonctions comprises dans le dictionnaire.

$$\text{crit}(\gamma) = \frac{1}{n}\sum_{i=1}^{N} \|\tilde{Y}_i - b_i f_\gamma(x_i)\|^2 + 2\sum_{k=1}^{M} r_{n,k}|\gamma_k| \tag{5}$$

avec $n = \sum_{i=1}^{N} n_i$ et $r_{n,k} = \sigma\|\psi_k\|_n\sqrt{\frac{\lambda log M}{n}}$ où $\lambda > 0$ et pour une fonction h, $\|h\|_n^2 = \frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{n_i} b_i^2 h^2(x_i)$. On note $\hat{\gamma}$ le paramètre qui minimise $crit(\gamma)$ pour tout $\gamma \in \mathbb{R}^M$ et on note $\hat{f} = f_{\hat{\gamma}}$

4 Implémentation

Plusieurs cas peuvent être considérées:

- Si f dépend d'une fonction c connue comme décrit dans l'équation (3), on applique l'algorithme décrit dans la Section (3). Pour cela on utilise le package `saemix` pour l'étape i) et le package `glmnet` pour l'étape ii).

- Si f ne dépend pas de c, l'utilisation d'un algorithme SAEM n'est pas utile. Dans ce cas la on combine entre un simple algorithme EM et une estimation Lasso. Pour cela, l'étape i) serait l'estimation des paramètres par un algorithme EM et le package `glmnet` pour l'étape ii).
- Si f ne dépend pas de c, on peut également construire un algorithme itératif combinant une estimation NLME par le package `nlme` ou `lme4` et une estimation Lasso par le package `glmnet`.

References

[1] Arribas-Gil, A., Bertin, K., Meza, C., and Rivoirard, V. Lasso-type estimators for semiparametric nonlinar mixed-effects models estimation. *Statistics and Computing 24* (2012), 443–460.

[2] Groll, A., and Tuts, G. Variable selection for generalized linear mixed models by l1-penalized estimation. *Statistics and computing* (2011).

[3] Schelldorfer, J., Bühlman, P., and der Geer, S. V. Estimation for high dimensional linear mixed effects models using l1-penalization. *The scandinavian Journal of Statistics* (2010).