Données longitudinales en grande dimension : état des lieux des packages R

P. Soret[1,2,3] et M. Avalos[1,2,4]

[1] INSERM U897–Epidémiologie–Biostatistique, Univ. Bordeaux, ISPED
[2] INRIA-SISTM, Bordeaux
[3] Univ. de Montpellier
[4] Univ. Bordeaux, ISPED, INSERM U897–Epidémiologie–Biostatistique
Perrine.Soret@isped.u-bordeaux2.fr, Marta.Avalos@isped.u-bordeaux2.fr

Mots clefs : longitudinal data, high-dimensionality, Statistical Machine Learning

Les données longitudinales constituent un domaine important de la statistique. On entend par données longitudinales des données telles que, pour chaque individu considéré, on dispose d'observations à différents instants, autrement dit répétées dans le temps. Les principaux domaines d'application de ce type de données sont la médecine ou la biologie. On peut prendre comme exemple des données de séquençage pour rechercher l'efficacité d'un vaccin sur une maladie, des données d'imagerie au cours du temps pour rechercher la localisation d'un tumeur dans le cerveau, mais également des données sportives pour étudier la performance suivant les entrainements des athlètes.
L'analyse de ces données longitudinales requiert des méthodes statistiques adaptées. En effet, les séries des données de chaque sujet sont supposées indépendantes les unes des autres, mais les données d'un même sujet sont vraisemblablement corrélées dans le temps. Les modèles à effets mixtes permettent de tenir compte de cette corrélation (Verbeke et Molenberghs, 2000). Ces modèles permettent d'expliquer la variabilité d'une suite d'observations par deux types d'effets : les effets fixes de population et les effets individuels, considérés comme aléatoires puisqu'ils varient d'un individu à l'autre. Quand le nombre d'observations est faible par rapport au nombre d'effets fixes, les modèles mixtes classiques présentent des limites.
Nous présentons ici une revue des méthodes prédictives issues du champ de l'apprentissage statistique (ou *machine learning*) qui ont été proposées dans la littérature permettant de tenir compte de la nature dépendante des données longitudinales par des adaptations des modèles à effets mixtes. Nous effectuons également une revue et une évaluation des différents packages R implémentant ces méthodes. Nous étudions leurs capacités et leurs limites[1].

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*, 2012.
[2] Groll A and Tuts G. Variable selection for generalized linear mixed models by l1-penalized estimation. *Statistics and computing*, 2011.
[3] Hajjem A, Bellavance F, and Larocque D. Mixed effects random forest for clustered data. *JSCS*, 2012.
[4] Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, and Heckerman D. Fast linear mixed models for genome-wide association studies. *Nature methods*, 8:833–835, 2011.

[1]Ce travail a été effectué dans le cadre du stage de M2 Biostatistique, Univ de Montpellier, de Perrine SORET au sein de l'équipe Biostatistique de l'INSERM U897 et SISTM-INRIA, Bordeaux.

Table 1: Packages R : méthodes prédictives adaptées aux modèles à effets mixtes

Méthodes	**Packages**
PLS regression for linear mixed models	`nlme`, `pls`
Linear Mixed-Effects models using l_1-penalisation	`lmmlasso`, `LLMM`, `MMS`
Generalized linear mixed models using l_1-penalisation	`glmmLasso`, `GLMMLasso`
Random-Effect and EM-algorithm	`nlme`, `rpart`, `REEMtree`
Mixed-effects random forest	`randomForest`, `rpart`, `lme4`, `nlme`

[5] Liu D, Xin X, and Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed-models. *Biometrics*, 63:1079–1088, 2007.

[6] Guyon E and Pommeret D. Imputation by pls regression for linear mixed models. *Journal de la société Française de Statistique*, pages 30–16, 2011.

[7] Rohart F, San-Cristobal M, and Laurent B. Fixed effects selection in high-dimensional linear mixed models. *Technical report INSA Toulouse*, 2012.

[8] Verbeke G and Molenberghs G. Linear mixed models for longitudinal data. *Springer series in Statistics*, 2000.

[9] Bondell HD, Krishna A, and Gosh SK. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66:1069–1077, 2010.

[10] Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, and Heckerman D. Improved linear mixed models for genome-wide association studies. *Nature methods*, 9:525–526, 2012.

[11] Luts J, Molenberghs G, Verbeke G, Van Huffel S, and Suykens JAK. A mixed effects least squares support vector machine model for classification of longitudinal data. *Comput Stat Data Anal*, 56:611–628, 2012.

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

[13] Schellforfer J, Meier L, and Bühlmann P. Glmmlasso: An algorithm for high-dimensional generalized linear mixed models using l1-penalization. *Comp Graph Stat*, 2013.

[14] Ibrahim JG, Fhu H, Garcia RI, and Guo R. Fixed and random effects selection in mixed effects models. *Biometrics*, 67:495–503, 2011.

[15] Pearce ND and Wand MP. Explicit connections between longitudinal data analysis and kernel machines. *Electron. J. statist*, 3:797–823, 2009.

[16] Sela RJ and Simonoff JS. Re-em trees: a data mining approach for longitudinal and clustered data. *Mach Learn*, 2012.

[17] Fieuws S and Verbeke G. Joint models for high-dimensional longitudinal data. *Chapman and Hal/CRC*, 2009.

[18] Foster SD, Verbyla AP, and Pitchford WS. Estimation, prediction and inference for the lasso random effects model. *Australian and New Zealand Journal of Statistics*, 51:43–61, 2009.