Méthodes prédictives pour données longitudinales en grande dimension : Etat de l'art des packages R

Perrine SORET

3^{me} Rencontre R - Montpellier

26 Juin 2014

Sous la tutelle de :
Marta AVALOS (INSERM, Bordeaux)

- Données longitudinales
  - Observations différentes dans le temps = Mesures répétées dans le temps
  - Domaines d'application : Médecine et Biologie
  - Méthodes statistiques adaptées $\Rightarrow$ Observations d'un même sujet corrélées dans le temps $\Rightarrow$ **Modèles à effets mixtes**

- Données longitudinales
  - Observations différentes dans le temps = Mesures répétées dans le temps
  - Domaines d'application : Médecine et Biologie
  - Méthodes statistiques adaptées $\Rightarrow$ Observations d'un même sujet corrélées dans le temps $\Rightarrow$ **Modèles à effets mixtes**

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \varepsilon_i \qquad i = 1, \cdots, N \tag{1}$$

Où

- $\boldsymbol{Y}_i$ est la variable réponse pour l'individu i de taille n_i
- $\boldsymbol{X}_i \in \mathbb{M}_{(n_i,p)}$ est la matrice des covariables explicatives pour l'individu i
- β le vecteur des coefficients à effets fixes de taille p
- $\boldsymbol{Z}_i \subset \boldsymbol{X}_i$ et $\boldsymbol{Z}_i \in \mathbb{M}_{(n_i,q)}$est la matrice des effets aléatoires pour l'individu i
- $\boldsymbol{b}_i$ le vecteur des coefficients à effets aléatoires pour l'individu i de taille q
- ε_i le vecteur des résidus de taille n_i pour l'individu i supposés iid

Avec :

$$\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \Psi) \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_{n_i})$$

$$\tilde{\phi} = (\beta^T, \Psi, \sigma^2)^T$$

- Données de grandes dimensions
  - Nombre de variables supérieur au nombre d'individus ($N << p$)
  - Référence aux données "-omics"
  - Méthode d'Apprentissage Statistique (Machine Learning)
    - Modélisation et Prédiction
    - Sélectionner des modèles en équilibrant biais (erreur d'approximation) et variance (erreur d'estimation)
    - Méthodes pour des données indépendantes

Objectif : Extension des modèles de machine learning aux données longitudinales

- Données de grandes dimensions
  - Nombre de variables supérieur au nombre d'individus ($N << p$)
  - Référence aux données "-omics"
  - Méthode d'Apprentissage Statistique (Machine Learning)
    - Modélisation et Prédiction
    - Sélectionner des modèles en équilibrant biais (erreur d'approximation) et variance (erreur d'estimation)
    - Méthodes pour des données indépendantes

Objectif : Extension des modèles de machine learning aux données longitudinales

Etat de l'art

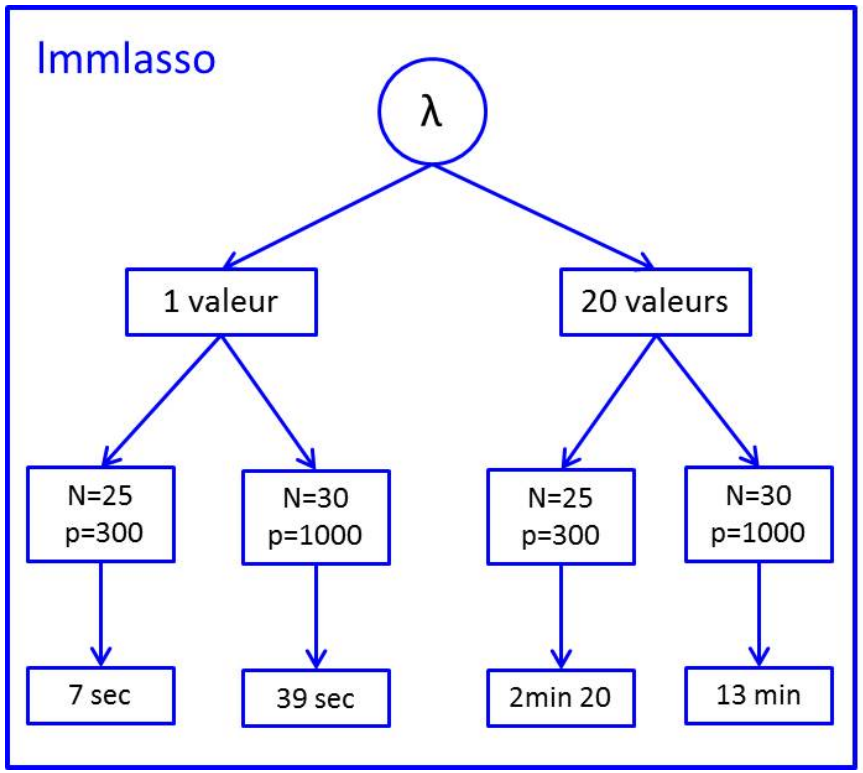
ACP (2)	Fonctional modelling and Classification of Longitudinal Data, **HG. Müller, 2005** Properties of principal component methods for functional and longitudinal data anlysis, **P. Hall, 2006**	
PLS (2)	Imputation by PLS regression for linear mixed models, **E. Guyon & al., 2011**	
Méthode pénalisée L_1 (22)	Estimation for high-dimensional linear mixed-effects models using l_1 penalization, **J. Schelldorfer & al., 2010**	`lmmlasso`
	Selection of fixed effects in high dimensional linear mixed models using a multicycle ECM algorithm, **F. Rohart & al., 2012**	`MMS`
	Variable selection for generalized linear mixed models by l_1 penalized estimation, **A. Groll & G. Tutz, 2011 et J. Schelldorfer & al., 2013**	`glmmLasso` `glmmixedlasso`
	LMM-lasso : A Lasso Multi-Marker mixed model for association maping with population structure, **B. Rakitsh & al., 2013**	Code Python
Arbre de régression et Forêt aléatoire (3)	RE-EM trees : a data mining approach for longitudinal and clustered data, **RJ. Sela & JS. Simonoff, 2010 et A. Hajjem & al.,2011**	`REEMtree` et `MERT`
	Mixed-effects random forest for clustered data, **A. Hajjem & al.,2012**	`MERF`
	Generalized Mixed Effects Regression Trees, **A. Hajjem & al.,2013**	
SVM et Méthodes à noyaux (3)	A mixed effects LSSVM model for classification of longitudinal data, **J. Luts & al., 2011**	Code Matlab
	Semiparametric regression of multidimensional genetic pathway : LSKN et LMM, **D. Liu & al., 2007,**	
	Explicit connections between longitudinal data analysis and kernel machines, **N.D. Pearce & al., 2009**	
Bayésien (2)	Bayesien mixed-effects inference on classification performance in hierarchical data sets, **Cheng Soon Ong& al., 2012**	
	Bayesian machine learning approaches for longitudinal latent class, **D. Belgrave & al., 2012**	

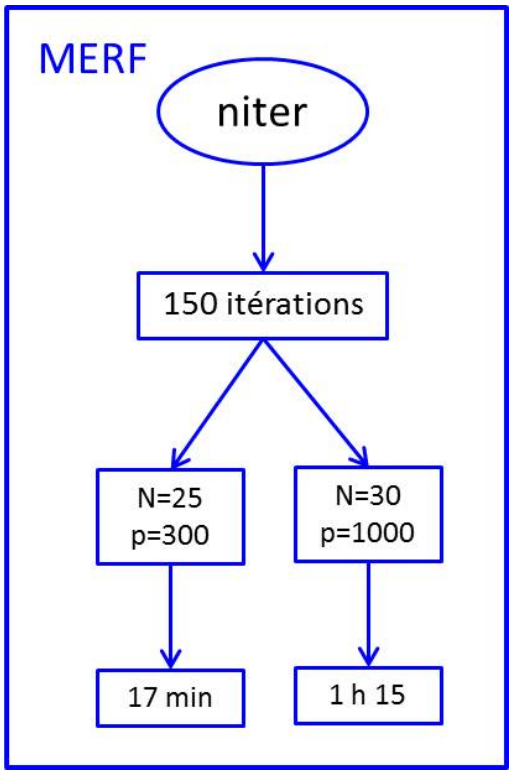
Etat de l'art

ACP (2)	Fonctional modelling and Classification of Longitudinal Data, **HG. Müller, 2005** Properties of principal component methods for functional and longitudinal data anlysis, **P. Hall, 2006**	
PLS (2)	Imputation by PLS regression for linear mixed models, **E. Guyon & al., 2011**	
Méthode pénalisée L_1 (22)	Estimation for high-dimensional linear mixed-effects models using l_1 penalization, **J. Schelldorfer & al., 2010**	`lmmlasso`
	Selection of fixed effects in high dimensional linear mixed models using a multicycle ECM algorithm, **F. Rohart & al., 2012**	`MMS`
	Variable selection for generalized linear mixed models by l_1 penalized estimation, **A. Groll & G. Tutz, 2011 et J. Schelldorfer & al., 2013**	`glmmLasso` `glmmixedlasso`
	LMM-lasso : A Lasso Multi-Marker mixed model for association maping with population structure, **B. Rakitsh & al., 2013**	Code Python
Arbre de régression et Forêt aléatoire (3)	RE-EM trees : a data mining approach for longitudinal and clustered data, **RJ. Sela & JS. Simonoff, 2010 et A. Hajjem & al.,2011**	`REEMtree` et `MERT`
	Mixed-effects random forest for clustered data, **A. Hajjem & al.,2012**	`MERF`
	Generalized Mixed Effects Regression Trees, **A. Hajjem & al.,2013**	
SVM et Méthodes à noyaux (3)	A mixed effects LSSVM model for classification of longitudinal data, **J. Luts & al., 2011**	Code Matlab
	Semiparametric regression of multidimensional genetic pathway : LSKN et LMM, **D. Liu & al., 2007,**	
	Explicit connections between longitudinal data analysis and kernel machines, **N.D. Pearce & al., 2009**	
Bayésien (2)	Bayesien mixed-effects inference on classification performance in hierarchical data sets, **Cheng Soon Ong& al., 2012**	
	Bayesian machine learning approaches for longitudinal latent class, **D. Belgrave & al., 2012**	

Limites des méthodes (Algorithme et codes)

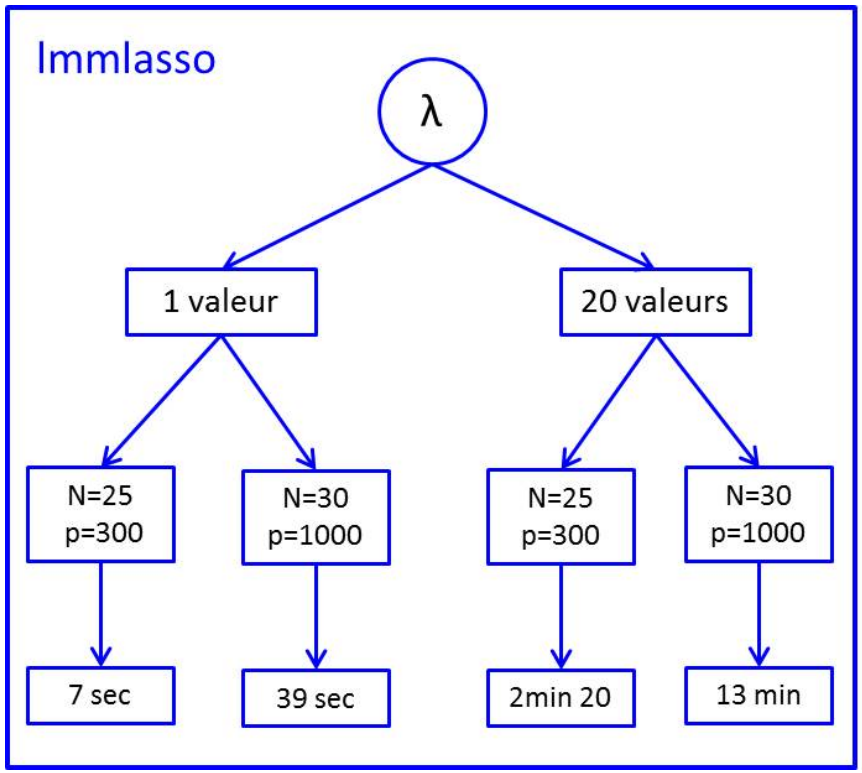
- Temps de calcul conséquent

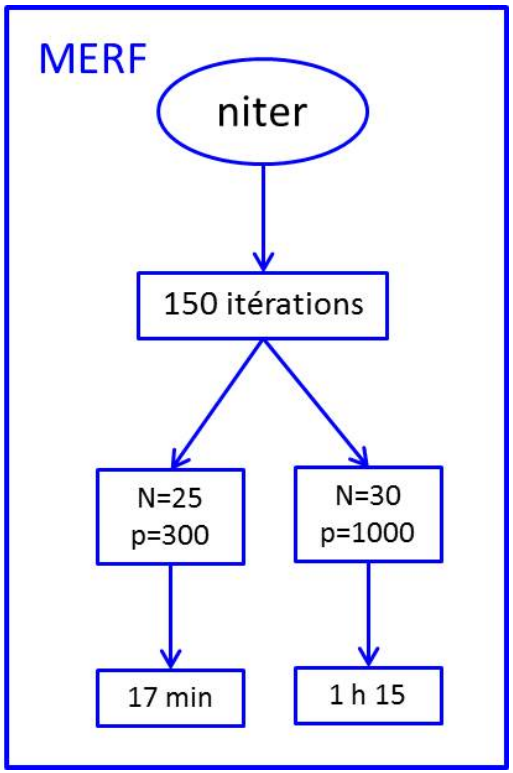




Limites des méthodes (Algorithme et codes)

- Temps de calcul conséquent





- Désintérêt des estimations des composantes de la variances
  - Point important dans les modèles à effets mixtes
  - Plus adaptées à des modèles tels que les GEE
- Structure de corrélation simple

Conclusion

- Intérêt et développement croissant pour résoudre les problèmes engendrés par ce type de données
- Peu de codes disponibles
- Algorithmes non optimaux

Conclusion

- Intérêt et développement croissant pour résoudre les problèmes engendrés par ce type de données
- Peu de codes disponibles
- Algorithmes non optimaux

Perspective

- Méthode utilisant la pénalisation Lasso
- Proposer des estimations pertinentes et interprétables de la covariance
- Optimiser l'algorithme
- Application à des données issues de la vaccination contre le VIH

MERCI DE VOTRE ATTENTION