# Intégrer les données manquantes dans la sélection de variables pour données longitudinales

Julia Geronimi [1,2] & Gilbert Saporta [2]

[1] *IRIS, 50 rue Carnot, Suresnes, France geronimi.julia@gmail.com*
[2] *Cedric, CNAM, 292 rue Saint-Martin, 75003 Paris, France gilbert.saporta@cnam.fr*

**Résumé.** Les Generalized estimating equations (GEE) sont une méthode de régression utile pour l'analyse marginale en présence de mesures répétées. Dans le contexte longitudinale, il est fréquent de faire face aux données manquantes ainsi qu'à de nombreuses variables mesurées au cours du temps. L'imputation multiple, outil populaire pour le traitement des données manquantes et plus particulièrement les MI-GEE peuvent être utilisés pour l'inférence. Bien que les méthodes pour traiter les données manquantes telles que les MI-GEE aient été mises place, la sélection de variables pour GEE n'a pas été systématiquement développée pour intégrer les données manquantes. Le multiple imputation-least absolute shrinkage and selection operator (MI-LASSO) propose une sélection consistante au sein des jeux de données imputés, mais ne permet pas de prendre en compte les corrélations intra-patient. Nous présentons le MI-PGEE, multiple imputation-penalized generalized estimating equations, extension du MI-LASSO pour les données longitudinales. Cette méthode utilise les GEE pénalisés par une pénalité ridge et des poids adaptatifs qui sont communs à l'ensemble des coefficients de régression estimés de la même variable sur les échantillons multi-imputés. Nous présentons un critère de type BIC pour le choix du paramètre de régularisation. Le MI-PGEE fournit une sélection consistante sur l'ensemble des imputations, ce qui en fait une méthode de sélection pour données longitudinales capable d'intégrer les données manquantes et les corrélations intra-sujet. Une application sur le sous groupe placebo de la base de données Strontium ranelate Efficacy in Knee OsteoarthrItis triAl (SEKOIA) est présentée.

**Mots-clés.** Données longitudinales, données manquantes, imputation multiple, sélection de variables, ...

**Abstract.** Generalized estimating equations (GEE) are a useful tool for marginal regression analysis with repeated measurements. Missing data as well as a large number of variables combined with small sample size are usual issues faced with longitudinal data. Multiple imputation is a popular tool for handling missing data and in particular, the MI-GEE can be used for inference. The multiple imputation-least absolute shrinkage and selection operator (MI-LASSO) proposes a consistent selection through the multiply-imputed datasets but cannot handle correlation among individual observations.

We present MI-PGEE, a new multiple imputation-penalized generalized estimating equations as an extension of the MI-LASSO to be applied on longitudinal data. MI-PGEE applies the penalized GEE with ridge penalty and adaptive weights that are common to the group of estimated regression coefficients of the same variable across multiply-imputed datasets. In order to select the tuning parameter, a new BIC-like criterion is presented. MI-PGEE yields a consistent variable selection across multiply-imputed datasets, making this a selection method for longitudinal data able to manage missing data and within subject correlation. The usefulness of the new method is illustrated by an application on the placebo arm of the Strontium ranelate Efficacy in Knee OsteoarthrItis triAl (SEKOIA) study.

**Keywords.** Longitudinal data, missing data, multiple imputation, variable selection, ...

# 1 Introduction

Les études longitudinales donnent l'opportunité d'étudier le lien entre un critère clinique dintérêt et des covariables à l'aide de données collectées dans le temps. Les Generalized Estimating Equations (GEE) de Liang et Zeger (1986) permettent de prendre en compte les corrélations intra-sujet dans la régression marginale. La plus part des critères de sélection de modèles ont été étendus pour l'application aux GEE. Cependant quand le nombre de variables est grand l'utilisation de régressions pénalisées est recommandée. Fu (2003) propose les Penalized GEE qui utilisent une pénalité bridge afin de réduire les coefficients ou de sélectionner des variables.
Une limite importante de ces méthodes est qu'elles ne prennent pas en compte les données manquantes. La méthode des cas complets, qui supprime chaque ligne qui présente une donnée manquante entraîne une forte perte de données mais peut aussi introduire du biais. L'Imputation Multiple (MI) est une méthode populaire et utile pour résoudre le problème des données manquantes. L'imputation par régressions séquentielles utilise des modèles conditionnels séparés pour chaque variable sachant les autres. Les régressions séquentielles, aussi connue sous le nom d'imputation multivariée par équations en chaîne, sont flexibles et implémentées en `R` dans le package `mice` de Van Buuren et Groothuis-Oudshoorn (2011).
On en sait peu sur la façon de mettre en place une sélection de variables efficace et fiable avec des ensembles de données multi-imputés. Shen et Chen (2013) proposent le MI-QIC et le MI-MLIC pour la sélection de modèles dans les analyses sur MI-GEE. Le calcul de cet estimateur peut être coûteux en termes de calculs quand le nombre de variables est grand. Pour une réponse univarié, la sélection stepwise sur jeux de données imputés à été développée par Wodd, White et Royston (2009). Le MI-LASSO introduit par Chen et Wang (2013) pour les données indépendantes propose une sélection consistante à travers les jeux muli-imputés. Aucune de ces deux méthodes ne permet d'intégrer les corrélations

intra-sujet. Nous proposons donc le MI-PGEE, extension du MI-LASO aux données longitudinales.

# 2 Generalized Estimating Equations

Soit $Y_i = (Y_{i1}, \ldots, Y_{iT_i})$ et $X_i = (X_{i1}, \ldots, X_{iT_i})$ les informations collectées pour l'individu $i \in \{1, \ldots, K\}$. $Y_{it}$ et $X_{it}$ sont la réponse est le vecteur de $p$ covariables observées au temps $t \in \{1, ..., T_i\}$. L'espérance marginale $\mu_{it} = \mathbf{E}(Y_{it}|X_{it})$ est reliée aux covariables à l'aide d'une fonction $g$ de lien $g(\mu_{it}) = X'_{it}\beta$ oú $\beta$ est le vecteur $p \times 1$ de paramètres de régression. La variance marginale dépend de l'espérance marginale à travers $Var(Y_{it}) = \phi\nu(\mu_{it})$ oú $\phi$ est le paramètre de dispersion et $\nu$ est une fonction de variance qui définie la relation espérance-variance. Les estimateurs donnés par les GEE de Liang et Zeger (1986) sont solution de :

$$S(\beta) = \sum_{i=1}^{K} S_i(\beta) = \sum_{i=1}^{K} D'_i V_i^{-1}(Y_i - \mu_i) = 0 \tag{1}$$

Oú $D_i = \partial\mu_i/\partial\beta'$ et $V_i = A_i^{1/2} R_i(\alpha) A^{1/2}$. $A_i$ est une matrice $T_i \times T_i$ diagonale composée des variance marginales $Var(Y_{it})$. $R_i(\alpha)$ est la matrice de corrélations de *travail* choisie par l'utilisateur qui dépend d'un paramètre de corrélation $\alpha$. Liang et Zeger (1986) proposent de calculer cet estimateur en alternant entre une un algorithme de type Newton-Raphson et une méthode des moments jusqu'à convergence. L'estimateur robuste, de type sandwich, de la matrice de variance covariance de l'estimateur des GEE est donnée par :

$$W = I^{-1} \left\{ \sum_{i=1}^{K} D'_i V_i^{-1} Cov(Y_i) V_i^{-1} D_i \right\} I^{-1}$$

Oú $Cov(Y_i)$ peut être estimée par $(Y_i - \mu_i)(Y_i - \mu_i)'$, $I = \sum_{i=1}^{K} D'_i V_i^{-1} D_i$ et $\beta$, $\alpha$ et $\phi$ peuvent être remplacés par leur estimation. Les données manquantes peuvent impacter les covariables ainsi que la variable réponse à chaque temps d'observation. Nous avons donc utilisé la méthode d'imputation multivariée par équations en chaîne implémentée dans le package `mice` de `R` sur $D$ imputations. Les analyses effectuées sur chaque jeu de données imputées peuvent être combinées à l'aide des règles de Rubin (1987) :

$$\overline{\beta} = \frac{1}{D} \sum_{d=1}^{D} \hat{\beta}_d \tag{2}$$

Oú $\hat{\beta}_d$ est l'estimateur de $\beta$ dans le $d$ème jeu de données imputé, $d \in \{1, \ldots, D\}$. La variance $\Sigma$ de l'estimateur combiné $\overline{\beta}$ est composée d'un terme de variance intra-imputation $\overline{W} = \frac{1}{D} \sum_{d=1}^{D} \hat{W}_d$ et d'un terme de variance inter-imputation $B = \frac{1}{D-1} \sum_{d=1}^{D} (\hat{\beta}_d - \overline{\beta})'(\hat{\beta}_d - \overline{\beta})$.

$$\Sigma = \overline{W} + \frac{D+1}{D} B \tag{3}$$

Comme noté dans Chen et Wang (2013), ces règles ne peuvent être utilisées que si les même variables sont utilisées sur toutes les imputations. Il est donc nécessaire d'obtenir une méthode de sélection qui aboutisse à un modèle consistant. Les règles de Rubin seront appliquées une fois le modèle optimal choisi.

# 3 MI-PGEE

## 3.1 MI-LASSO

Dans un contexte d'étude transversale, Shen et Wang (2013) proposent la méthode du MI-LASSO pour intégrer les imputations dans la sélection de variables. Considérons une étude pour $K$ sujets oú $Y$ et $X$ sont le vecteur réponse et la matrice de covariables. Nous notons alors $Y_d$ et $X_d = (X_{d,1}, \ldots, X_{d,p})$ la réponse et les covariable du $d$-ème jeu de données imputé, $d \in \{1, \ldots, D\}$. L'idée est d'utiliser une pénalité group LASSO en considérant les coefficients de régression d'une même variable comme un groupe. Le problème d'optimisation est alors :

$$\min_{\beta_d, j} \left\{ \sum_{d=1}^{D} ||Y_d - X_d' \beta_d||_2^2 + \lambda \sum_{j=1}^{p} \sqrt{\sum_{d=1}^{D} \beta_{d,j}^2} \right\} \tag{4}$$

oú $\sum_{j=1}^{p} \sqrt{\sum_{d=1}^{D} \beta_{d,j}^2}$ est la pénalité groupe LASSO. La méthode d'approximation quadratique locale est utilisée pour parer la singularité de la pénalité. Supposons $\hat{\beta}_{d,j}^l, d \in \{1, \ldots, D\}$ connue à la $l$ème itération de l'algorithme, quand $\sqrt{\sum_{d=1}^{D} \beta_{d,j}^2} > 0$, le LQA utilise l'approximation suivante :

$$\sqrt{\sum_{d=1}^{D} \beta_{d,j}^2} \approx \frac{\sum_{d=1}^{D} \beta_{d,j}^2}{\sqrt{\sum_{d=1}^{D} (\beta_{d,j}^l)^2}}$$

L'équation 4 peut être réécrite comme $D$ régressions ridge :

$$\min_{\beta_d, j} \sum_{d=1}^{D} \left\{ ||Y_d - X_d' \beta_d||_2^2 + \lambda \sum_{j=1}^{p} c_j \beta_{d,j}^2 \right\} \tag{5}$$

oú $c_j = 1/\sqrt{\sum_{d=1}^{D} (\beta_{d,j}^l)^2}$ et $\lambda$ est choisi minisant un critère de type BIC.

## 3.2 MI-PGEE

Les GEE pénalisés développés par Fu (2003) sont solutions de

$$S^p(\beta) = S(\beta) - \dot{P}(\beta) = 0$$

oú $\dot{P}(\beta) = \partial P(\beta)/\partial\beta$ est le vecteur dérivé de la fonction de pénalité qui peut être une pénalité bridge ou une combinaison de pénalités. La specification de la vraisemblance jointe est evite par l'utilisation des GEE ce qui transforme le problème d'optimisation de l'équation 4 en un système d'équation. Par conséquent, nous proposons le MI-PGEE, solution de $D$ GEE pénalisés par une pénalité ridge et des poids adaptatifs :

$$S^p(\beta_d) = \sum_{i=1}^{K} D'_{d,i} V_{d,i}^{-1}(Y_{d,i} - \mu_{d,i}) - \lambda \mathbf{C}\beta_d = 0 \tag{6}$$
$$d \in \{1, ..., D\}$$

Oú $\mathbf{C} = diag(c_j)$ est la seule quantité commune aux $D$ équations. A mesure que $\lambda$ croit, les coefficients vont être réduit à zéro sans être *exactement* nuls. Pour permettre à la régression ridge de réduire et sélectionner les coefficients nous fixons $\hat{\beta}^l_{d,j} = 0$ pour $d \in \{1, \ldots, D\}$ dès que $\sum_{d=1}^{D} \hat{\beta}^2_{d,j} \leq 5^{-10}$. La solution à ces équations pénalisées peut être approchée par une méthode des moindres carrés pondérés iteritivement. Notre méthode permet à l'utilisateur d'ajuster $D$ modèles conjointement ce qui permet d'obtenir une sélection consistante qui intègre les corrélations intra patients.

# 4 Aspect calculatoire

## 4.1 Algorithme

Les valeurs initiales de l'algorithme, $\beta^0_d = (\beta^0_{d,1}, \ldots, \beta^0_{d,p})$, proches de la solution sont les solutions de GEE non pénalisés. A chaque itération $l$, l'équation 6 peut être estimée par une série de Taylor, ce qui permet d'obtenir un algorithme itératif :

$$\beta^{l+1}_d = \beta^l_d + \left[\frac{\partial S(\beta^l_d)}{\partial\beta} + \lambda \mathbf{C}^l\right]^{-1} \left[S(\beta^l_d) - \lambda \mathbf{C}^l \beta^l_d\right]$$

Une fois qu'un groupe de coefficients est réduit à zéro il ne peut devenir non nul à nouveau. Pour surmonter cette limite, nous fixons $\hat{\beta}^l_{d,j} = \delta$ pour $d \in \{1, \ldots, D\}$ quand $\sum_{d=1}^{D} \left(\hat{\beta}^l_{d,j}\right)^2 \leq D\delta^2$ avec $\delta = 10^{-10}$. Pour un paramètre $\lambda$ donné, ce processus est répété jusqu'à convergence.

## 4.2 Choix du paramtre $\lambda$

Notre critère de type BIC est donné par :

$$BIC = DNlog(\sum_{d=1}^{D} WRSS_d/DN) + dflog(\sum_{d=1}^{D} \tilde{N}_d) \tag{7}$$

oú $WRSS_d = \sum_{i=1}^{K}(Y_{d,i} - \mu_{i,d})'\hat{R}_{d,i}^{-1}(\alpha)(Y_{d,i} - \mu_{i,d})$ est la somme des carrés résiduels pondérée par $\hat{R}_{d,i}$, la matrice de corrélation de travail estimée sur le jeu complet (i.e. quand $\lambda = 0$). Le degré de liberté *df* est estimé comme pour le groupe LASSO de Yuan et Lin (2006) :

$$df = \sum_{j=1}^{p} I\left(\sqrt{\sum_{d=1}^{D}\hat{\beta}_{d,j}^2} > 0\right) + \sum_{j=1}^{p}\frac{\sqrt{\sum_{d=1}^{D}\hat{\beta}_{d,j}^2}}{\sqrt{\sum_{d=1}^{D}\tilde{\beta}_{d,j}^2}}(D-1)$$

oú $\tilde{\beta}_{d,j}$ est le $j$ème coefficient estimé sur le $d$ème jeu de données imputé avec le modèle complet (i.e. GEE sans pénalisation). $\tilde{N}$ représente un compromis entre un BIC *lourd* qui utilise le nombre de sujet $K$ et un BIC *léger* qui utilise le nombre d'observations. Plus les corrélations sont faibles plus on s'approche du nombre dobservations, plus les corrélations sont fortes et plus on s'approche du nombre de sujets :

$$\tilde{N} = \sum_{i=1}^{K}\frac{T_i^2}{\sum_{i=1}^{K}\sum_{k=1}^{K}\hat{R}_{ik}}$$

Le paramètre de régularisation sera choisi en minimisant le critère type BIC sur une grille assez fine de valeurs possibles.

# 5 Application

Nous présentons une application sur le sous groupe placebo de la base de données Strontium ranelate Efficacy in Knee OsteoarthrItis triAl (SEKOIA). Le sous groupe étudié est composé de 166 patients suivis sur quatre visites. L'objectif est d'identifier les marqueurs ayant le plus d'impact sur le Joint Space Width (JSW), critère continu de séverité de l'arhrose du genou. Nous cherchons parmis 44 covariables dont 40 dépendantes du temps, celles qui expliquent le mieux les différences de JSW entre patients au cours du temps.

# Bibliographie

[1] Yuan, Ming et Lin, Yi (2006), Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society: Series B.
[2] Liang, K.Y. et Zeger, S.L. (1986), Longitudinal data analysis using generalized linear models, Biometrika.
[3] Rubin, Donald B (1987), Multiple Imputation for Nonresponse in Surveys, Wiley Series in Probability and Statistics.
[4] Shen, C.W. et Chen, Y.H. (2013), Model selection of generalized estimating equations with multiply imputed longitudinal data, Biometrical Journal.

[5] Chen, Q. et Wang, S. (2013),Variable selection for multiply-imputed data with application to dioxin exposure study, Statistics in medicine.
[6] Wood, Angela M et White, Ian R et Royston, Patrick (2008), How should variable selection be performed with multiply imputed data?, Statistics in medicine.
[7] Van Buuren, Stef et Groothuis-Oudshoorn, Karin, mice: Multivariate imputation by chained equations in R, Journal of statistical software.
[8] Fu, W.J., Penalized estimating equations, Biometrics.