Imputation by PLS regression for linear mixed models

Emilie Guyon, Denys Pommeret

Imputation by PLS regression for linear mixed models

Emilie Guyon[a], Denys Pommeret[a]

[a] *Institut de Mathématiques de Luminy (IML), CNRS Marseille, Case 907, Campus de Luminy, 13288 Marseille Cedex 9, France*

Abstract

The problem of handling missing data in linear mixed models with correlated covariates is considered when the missing mechanism concerns both the dependent variable and the design matrix. We propose an imputation algorithm combining multiple imputation and Partial Least Squares (PLS) regression. The method relies on two steps: removing random effects, fixed effects are first imputed and PLS components are constructed on the corresponding complete case. The dependent variable is then imputed inside the linear mixed model obtained by adding the random effects to PLS components. The method is applied on simulations and on real data.

1. Introduction

The problem of handling missing data has been extensively studied in the statistical literature. The best general reference here is Little and Rubin (2002), where three non-response mechanisms are described: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Suppose that the data consist of an incomplete variable Y_K and a set of fully observed variables $Y_1, ..., Y_{K-1}$. The data are MCAR if the probability $P(Y_K missing)$ is a constant that does not depend on the variables. The data are MAR if the probability of missingness may depend on the fully observed variables $Y_1, ..., Y_{K-1}$ but does not depend on the incomplete variable Y_K. The MNAR mechanism assumes that the probability of missingness depends on the variable that contains missing values.

This work will be concerned with the MAR mechanism. When the missing data are MAR, multiple imputation (MI) was first proposed in Rubin (1987) in the context of a non-response mechanism which concerns both the response variable and covariates. This method of imputation consists of replacing each missing value by a vector of $M \geq 2$ imputed values. It is shown that MI is efficient in the sense that " *When the M sets of imputations are repeated random draws from the predictive distribution of the missing values under a particular*

Email addresses: emilie.guyon@univmed.fr (Emilie Guyon), denys.pommeret@univmed.fr (Denys Pommeret)

model for nonresponse, the M complete-data inferences can be combined to form one inference that properly reflects uncertainty due to nonresponse under that model", (Little and Rubin, 2002).

This technique has been adapted to imputation in linear models and in generalized linear models by Schafer (1997) and Ibrahim (1990) respectively. Recently, Bastien (2008) studied the problem of missing data imputation in generalized linear models when covariates are correlated. This author combined the Partial Least Squares (PLS) regression technique (Wold, 1975) with the multiple imputation method, obtaining a successfull method, called Multiple Imputation with Partial Least Squares (MI-PLS). It consists of imputing the missing data on the variable of interest by a PLS regression after imputation of missing values on each explicative variable. The problem of missing data in linear mixed models has also been investigated by Schafer and Yucel (1998), who proposed a method denoted by MI-L2M (Multiple Imputation in Linear Mixed Models) using multiple imputation on covariates. However, this method can break down when covariates are linearly dependent, due to the singularity of the design matrix.

In this paper, we consider the problem of missing data in the case of a linear mixed model when the covariates are highly correlated. An algorithm is proposed which combines the PLS approach defined in Bastien (2008) with the multiple imputation method proposed by Schafer and Yucel (1998). This method will be denoted by MI-PLS-L2M (Multiple Imputation with Partial Least Squares in Linear Mixed Models). It consists in several steps. First, omitting the random effect, a linear model with dependent errors is considered. Missing data on covariates are imputed following the method of Little and Rubin (2002). Several complete dataset of covariates are obtained and transformed into PLS components. Second, random effects are reintroduced in the model and a linear mixed model is obtained in which the fixed PLS components are constructed by linear combinations of the observed regressors.The Henderson method (1959, see Appendix) is used to estimate the parameters of the model. Finally, the predictions of the final model based on the PLS components and the random components are used to reconstruct the dependent variables. Moreover, following Bastien *et al.* (2005), a bootstrap validation procedure allowed to test the significance of the fixed effects.

The MI-PLS-L2M method is applied on simulations and compared to two alternative procedures: the first one is the MI-PLS proposed by Bastien (2008), restricted to the linear model. The second one is the MI-L2M introduced by Schafer and Yucel (1998). This algorithm was then applied on a real dataset, namely coffee dataset from Vivien and Sabatier (2001): 17 samples of coffee were evaluated by 7 judges. In Vivien and Sabatier (2001), for each judge, characteristics of the coffee were explained by physico-chemical properties. Here, we have chosen to model the perception of the coffee by a score associated to a linear mixed model where the fixed effects are physico-chemical variables which are correlated and with a 7-level random effect corresponding to the 7 judges.

The paper is organized as follows: in Section 2 we review some standard facts on PLS and multiple imputation methods. In Section 3 we derive the

imputation algorithm by combining these two former methods. Section 4 is devoted to the simulation study and Section 5 presents the study of the coffee dataset. Section 6 contains a short discussion.

2. Multiple imputation and PLS background

Let $Y = (Y_1, \ldots, Y_n)$ be a vector of n observations. We denote by X the $(n \times p)$-matrix of the p covariate vectors X_j, $j = 1, ..., p$, referred to as the fixed effects. The random effect is a vector $u \in \mathbb{R}^q$. Given u, we consider the linear mixed model

$$Y_i \quad = \quad x_i'\beta + z_i'u + \epsilon_i, \forall i \in \{1, ..., n\}, \tag{1}$$

where x_i' and z_i' are row vectors of X and Z respectively, $\beta \in \mathbb{R}^p$ is an unknown vector of regression coefficients and Z is a known matrix associated to the random effects. We will denote by $\mathcal{N}_d(a, b)$ the d-dimensional normal distribution with mean a and variance b. It is assumed that

$$u \sim \mathcal{N}_q(0, D),\ \epsilon_i \sim \mathcal{N}_1(0, R_i),\ Cov(u, \epsilon_i) = 0, \forall i = \{1, ..., n\},$$

where $D = \sigma_q^2 A$, and A is a positive-definite matrix.

2.1. Multiple Imputation

When the data are Missing At Random, Rubin (1987) (see also Schafer, 1997) introduced the multiple imputation (MI) as the most reliable method both from accuracy and efficiency point of view. The standard scheme of the MI-algorithm consists of three steps:

- *Imputation*: $m > 1$ samples of possible values for the missing data are created.
- *Analysis*: each of the m complete dataset is analyzed using the standard statistical method that would be used in the absence of non-response.
- *Pooling*: the results of the m analysis are combined to get a single complete dataset.

In the *Analysis* step, we follow Honacker *et al.* (2010) who proposed an Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin, 1977, Wu, 1983) to estimate and replace the missing values. The EM part of the algorithm can be described as follows. Assume that the missing values only concern the response variable $Y = (Y_{obs}, Y_{mis})$. We denote by θ the vector of parameters of the model for the complete dataset. The EM algorithm can be summarized in the following way:

- (1) Complete the missing data Y_{mis} using the first estimation of θ.
- (2) Using Y_{obs} and Y_{mis} completed, re-estimate θ.
- (3) Using θ estimated, re-estimate Y_{mis}.
- (4) Iterate until convergence of θ.

2.2. PLS Regression

From now on, we assume that the vectors Y, X_1, .., X_p are centered.
In presence of correlation between covariates, PLS regression can replace the classical linear regression model. The idea is to take over the matrix X by a matrix $T = (t_1, \ldots, t_h)$, for $h < p$, iteratively obtained by linear transformation of the columns of X (Tenenhaus, 1998) according to the following algorithm:

- *Determination of the first component* t_1

  Compute the linear regression of Y on each X_j, for $j \in \{1, ..., p\}$,

  $$Y = a_{1j} X_j + \epsilon.$$

  Normalize the estimator $\hat{a}_{1j}$

  $$w_{1j} = \frac{\hat{a}_{1j}}{||\hat{a}_1||}, \forall j \in \{1, ..., p\},$$

  and so

  $$t_1 = \sum_{j=1}^{p} w_{1j} X_j.$$

  Compute the linear regression of Y on t_1

  $$Y = c_1 t_1 + \epsilon_{Y_1},$$

  where the residuals ϵ_{Y_1} express the deviation between the observed and the first PLS component. Then, for $j \in \{1, ..., p\}$, compute the linear regression of each covariate X_j on t_1,

  $$X_j = p_{1j} t_1 + X_{1j},$$

  where p_{1j} is the loading vector associated to t_1 and the residuals X_{1j} express the deviation between the covariates and the first PLS component.

- *Determination of the components* t_h

  For $h = 2, ..., H$, $H < p$, and for $j \in \{1, ..., p\}$, compute the linear regression of Y on each X_j as well as on the other PLS components t_{h-1},

  $$Y = a_{hj} X_{(h-1)j} + c_1 t_1 + ... + c_{h-1} t_{(h-1)} + \epsilon_{Y_{(h-1)}},$$

  where a_{hj} and c_h are the parameters associated to $X_{(h-1)j}$ and t_h respectively, both of them being estimated by the model.

Normalizing the estimator $\hat{a}_{hj}$

$$w_{hj} = \frac{\hat{a}_{hj}}{||\hat{a}_h||}, \forall j \in \{1, ..., p\},$$

and using a linear regression of the matrix X on the h PLS components

$$X_j = p_{1j}t_1 + ... + p_{(h-1)j}t_{(h-1)} + X_{(h)j}, \forall j \in \{1, ..., p\}$$

where p_{hj} is the loading vector associated to t_h and $X_{(h)j}$ is the residual term of the model,

$$t_h = \sum_{j=1}^{p} w_{hj} X_{(h)j}.$$

- *Number of PLS components*

The number h of PLS components to be retained is estimated by cross-validation. For that task, consider the regression model of y on the h PLS components:

$$y = \underbrace{c_1 t_1 + ... + c_h t_h}_{\hat{y}_h} + y_h. \tag{2}$$

At each step h, a criterion is calculated for each new component t_h:

$$Q_h^2 = 1 - \frac{PRESS_h}{RSS_{h-1}},$$

where RSS_h (Residual Sum of Squares) and $PRESS_h$ (PRediction Error Sum of Squares) are defined as:

$$RSS_h = \sum_{i=1}^{n} (y_i - \hat{y}_{hi})^2 \quad \text{and} \quad PRESS_h = \sum_{i=1}^{n} (y_i - \hat{y}_{h(-i)})^2$$

where $\hat{y}_{h(-i)}$ is the prediction of y_i obtained by (2) without the observation i.

For $h = 1$,

$$RSS_0 = \sum_{i=1}^{n} (y_i - \overline{y_i})^2$$

Referring to Tenenhaus (1998), a new component is considered as significant as soon as $Q_h^2 \geq 0.0975$.

3. Adapting imputation by PLS regression to the linear mixed model

3.1. Description of the method

Let Y_{obs} and Y_{mis} denote the vectors of observed and missing values of Y, respectively. The proposed method can be decomposed into six consecutive steps that we describe below.

Multiple imputation on fixed effects. We first impute $X_1, ..., X_p$ by the multiple imputation method described in 2.1. We obtain m complete dataset.

PLS regression without random effects. Temporarily eliminating the random effect and according to 2.2, we compute the PLS components T of the heteroscedastic linear model:

$$Y_{obs} = \beta X + \widetilde{\epsilon},$$

where $\widetilde{\epsilon} = \epsilon + Zu$. Clearly, $\widetilde{\epsilon}$ is a vector of dependent random variable such that $\widetilde{\epsilon} \sim \mathcal{N}_n(0; \Sigma)$, with $\Sigma = ZDZ' + R$. But this transformation of ϵ does not modify the PLS components, since they are completely specified by the correlation between X and Y. The selection of the appropriate number of components h is based on the cross-validation criterion Q_h^2.

Estimation of fixed and random parameters. In order to take into account the random effects, let us consider the new linear mixed model

$$Y_{obs} = TC + Zu + \epsilon. \tag{3}$$

The fixed parameters C and the random parameter u are estimated using Henderson's method (1959, see Appendix) and we denote by $\hat{C}$ and $\hat{u}$ their respective estimators. Then, we reformulate $\hat{\beta}$ according to W and $\hat{C}$, using the recovery formulas given in Bastien *et al.* (2005, see Appendix).

Selection of the fixed effects. We use a bootstrap validation procedure to assess the statistical significance of explanatory variables. This selection method is inspired from the jacknife technique introduced by Westad and Martens (1999). More precisely, the bootstrap procedure consists in sampling with replacement from Y_{obs} with their associated components T and Z. Applying (3) to the B (B fixed) bootstrap samples, we finally obtain a vector β^* of B estimators of β. It allows us to calculate a bootstrapped confidence interval of the regressors and Student tests are used to retain the significant variables at a prescribed level (arbitrarily 5%).

Pooling. The last step consists of pooling the m estimates into a single one which equals the estimated parameters mean, as it is done in Little and Rubin (2004). If $\hat{\theta}_1, ..., \hat{\theta}_m$ denote the m estimators of the parameter $\theta = (\beta, u)$, the global estimator of θ is given by

$$\overline{\theta} = (\overline{\beta}, \overline{u}) = \frac{1}{m} \sum_{k=1}^{m} \hat{\theta_k}.$$

Moreover the pooling method takes into account the variability of the m estimates as follows: Let $\hat{D}_1, ..., \hat{D}_m$ denote the m estimators of $Var(u)$. The variability between the m dataset is the empirical mean of the estimated variance

$$\overline{V} = \frac{1}{m} \sum_{k=1}^{m} \hat{D}_k,$$

while the variability within every dataset is given by

$$\overline{W} = \frac{1}{m-1} \sum_{k=1}^{m} (\hat{u}_k - \overline{u})(\hat{u}_k - \overline{u})'.$$

Therefore, the global variance of $\overline{u}$ is given by

$$\hat{F} = \overline{V} + (1 + \frac{1}{m})\overline{W}.$$

Finally, we simulate $N_1 \sim \mathcal{N}_n(Z\overline{u}, \hat{F})$ and $N_2 \sim \mathcal{N}_n(0, \hat{V}_\epsilon)$. We replace the ith missing value of Y by the corresponding ith value of

$$\widehat{Y_{mis}} = X\overline{\beta} + N_1 + N_2.$$

3.2. Summary of the algorithm

Hereafter, we summarize the steps of the MI-PLS-L2M algorithm.

- Step 1. *Imputation of covariates to get m complete dataset.*
- Step 2. *PLS procedure without random effect.*
- Step 3. *Consideration of random effect to estimate C and u.*
- Step 4. *Bootstrap selection.*
- Step 5. *Pooling.*

4. Simulation study

In order to evaluate the performance of the proposed algorithm, simulations were performed on several sample sizes ($N = 100$ and $N = 500$) and for different variances of the random effects ($Var(u) = 0.5$ and $Var(u) = 2$). For each simulation, we have computed the algorithm and the results presented were the mean of 10 simulations. The performance criterion used is the Mean Square Error for missing values defined by

$$MSE = \frac{\sum_{i=1}^{N_{mis}} (\hat{Y}_i - Y_i)^2}{Var(Y)},$$

where N_{mis} denotes the number of missing values.

4.1. The model

In our simulations, the design matrix X consists of a N-sample of a 15-dimensional covariate vector such that the last 10 components of the covariate vector are highly correlated with the first 5 ones, as it is shown in Table 4.1. The covariates are constructed as follows: five independent normal variables $X_1 \sim \mathcal{N}(2,1)$, $X_2 \sim \mathcal{N}(0,1)$, $X_3 \sim \mathcal{N}(0,1)$, $X_4 \sim \mathcal{N}(0,1)$, $X_5 \sim \mathcal{N}(0,1)$, and ten linearly dependent variables

$$\begin{aligned} X_6 &= X_1 + 4X_2 - X_3 + 2X_4 + 3X_5, & X_7 &= 2X_1 - X_2 + 5X_3 - 3X_4 - 2X_5, \\ X_8 &= X_1 + 4X_2 - X_3 + 4X_4 + 0.5X_5, & X_9 &= 2X_1 - X_2 + 5X_3 + X_4 - 3X_5, \\ X_{10} &= 3X_1 + 3X_2 + 4X_3 + 5X_4 + 0.5X_5, & X_{11} &= 3X_1 - 2X_2 - 5X_3 + 0.5X_4 + X_5, \\ X_{12} &= -X_1 + X_2 - X_3 + X_4 - X_5, & X_{13} &= X_1 - 4X_2 + 4X_3 + 0.5X_4 + 0.5X_5, \\ X_{14} &= X_1 + X_2 - 2X_3 - 2X_4 - 2X_5, & X_{15} &= 0.5X_1 - 0.5X_2 + X_3 + X_4 + 2X_5. \end{aligned}$$

We consider an independent error term $\epsilon \sim \mathcal{N}(0, I)$, where I denotes the identity matrix. In a first case, for $N = 100$ observations, the random effect was a 3-level vector $u \sim \mathcal{N}_3(0, 2I)$. In other cases, for $N = 500$ observations, the random effect was first a 3-level vector $u \sim \mathcal{N}_3(0, 2I)$, and second a 3-level vector $u \sim \mathcal{N}_3(0, 0.5I)$.

The output variable Y belongs to $\mathbb{R}^N$ and the MI-PLS-L2M algorithm (as well as competitors MI-PLS and MI-L2M) was run with various percentage p of missing value on Y and X. We have chosen $p \in \{8\%, 10\%, 15\%, 20\%, 25\%\}$.

Table 1: Correlation matrix between the covariates (Pearson correlation coefficients, $N = 500$)

	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
X1	0.336	1													
X2	0.797	0.051	1												
X3	-0.014	-0.076	-0.018	1											
X4	0.354	0.032	-0.024	-0.021	1										
X5	0.383	-0.002	0.013	-0.076	0.028	1									
X6	0.940	0.234	0.706	-0.253	0.336	0.574	1								
X7	-0.298	0.224	-0.142	0.784	-0.436	-0.385	-0.553	1							
X8	0.880	0.245	0.691	-0.225	0.650	0.130	0.848	-0.520	1						
X9	-0.160	0.247	-0.161	0.797	0.128	-0.535	-0.460	0.827	-0.175	1					
X10	0.701	0.406	0.399	0.495	0.627	0.050	0.493	0.164	0.674	0.520	1				
X11	0.010	0.507	-0.265	-0.824	0.115	0.217	0.207	-0.558	0.147	-0.536	-0.268	1			
X12	0.191	-0.399	0.438	-0.429	0.414	-0.434	-0.211	-0.558	0.542	-0.260	0.009	-0.021	1		
X13	-0.437	0.087	-0.694	0.700	0.091	0.029	-0.547	0.623	-0.529	0.682	0.203	-0.285	-0.681	1	
X14	0.183	0.489	0.475	-0.451	-0.414	-0.431	0.127	0.040	0.185	-0.130	-0.151	0.336	0.235	-0.634	1
X15	0.350	0.168	-0.192	0.323	0.396	0.795	0.400	-0.088	0.170	0.012	0.476	0.038	-0.540	0.494	-0.685

The number of multiple imputation was fixed to $m = 5$ following Rubin (1987).

4.2. Study of the MI-PLS-L2M performance

Using the algorithm without missing data. In order to evaluate the efficiency of the algorithm, we have first calculated estimators based on the original simulated data without missing value for the three considered scenarios. Table 2 shows the estimations of the vector β based on five retained PLS components. We can observe a bias due to the hight correlation between covariates. This bias is compensated on all the components of β. Finally a certain stability through the sample size and the variance of the random effect u is observed.

Table 2: Values of β and its estimators for $N = 100, 500$ and $Var(u) = 0.5, 2$

β	$\hat{\beta}$ $(N = 100; Var(u) = 2)$	$\hat{\beta}$ $(N = 500; Var(u) = 0.5)$	$\hat{\beta}$ $(N = 500; Var(u) = 2)$
1	1.05	1.06	1.05
2	1.28	1.13	1.28
-1	-0.97	-0.85	-0.97
-2	-1.64	-1.15	-1.64
3	3.35	-2.68	3.35
1	1.15	0.49	1.15
2	2.29	1.7	2.29
-1	-0.53	-0.4	-0.53
-2	-2.27	-1.99	-2.27
3	2.64	2.43	2.64
1	1.05	1.07	1.05
2	2.27	2.17	2.27
-1	-1	-0.88	-1
-2	-1.47	-1.87	-1.47
-3	-2.82	-2.56	-2.82

Analysis of MI-PLS-L2M performance. We have studied estimations of C, the coefficient vector associated to the PLS components, and estimations of the standard errors of u and ϵ when using MI-PLS-L2M. The different estimations were obtained with $p = \{0\%, 8\%, 10\%, 15\%, 20\%, 25\%\}$. Tables 3-5 contain estimations for the significant PLS components at the risk level $\alpha = 0.05$. The test of significance was based on $B = 200$ bootstrap samples. Only five components were retained for all models. It can be seen that the estimation of $Var(\epsilon)$ was increasing with the percentage of missing data while the estimations of $Var(u)$ was more stable. The MSE clearly increased with the number of missing values. Tables 3 and 5 illustrate the gain of precision due to the sample size.

Table 3: Estimations associated to significant PLS component t_1, t_2, t_3, t_4, t_5. Brackets give the standard deviations obtained by bootstrap, $N = 100$ and $Var(u) = 2$

Missing (%)	0%	8%	10%	15%	20%	25%
C (sd)	7.16 (0.05)	7.37 (0.06)	7.54 (0.06)	7.48 (0.06)	7.45 (0.07)	7.35 (0.09)
	2.29 (0.06)	2.13 (0.07)	1.96 (0.07)	2.06 (0.08)	2.13 (0.09)	2.02 (0.01)
	1.12 (0.07)	0.82 (0.07)	0.79 (0.07)	0.82 (0.07)	0.83 (0.09)	0.88 (0.1)
	0.30 (0.06)	0.26 (0.07)	0.21 (0.07)	0.25 (0.08)	0.22 (0.09)	0.28 (0.12)
	0.14 (0.08)	0.09 (0.12)	0.08 (0.13)	0.11 (0.14)	0.89 (0.47)	0.15 (0.37)
sd(u)	1.57	1.23	1.22	1.36	1.29	1.71
sd(ϵ)	1	1.07	1.04	1.11	1.3	1.48
MSE	0	0.21	0.19	0.36	0.4	0.51
N observed	100	92	90	85	80	75

Table 4: Estimations associated to significant PLS component t_1, t_2, t_3, t_4, t_5. Brackets give the standard deviations obtained by bootstrap, $N = 500$ and $Var(u) = 0.5$

Missing (%)	0%	8%	10%	15%	20%	25%
C (sd)	7.76 (0.02)	7.66 (0.02)	7.62 (0.02)	7.63 (0.03)	7.58 (0.03)	7.61 (0.04)
	2.25 (0.03)	2.04 (0.03)	2.04 (0.03)	2.07 (0.03)	2.11 (0.03)	2.02 (0.04)
	0.93 (0.03)	0.88 (0.03)	0.91 (0.04)	0.9 (0.03)	0.93 (0.03)	0.96 (0.04)
	0.26 (0.04)	0.22 (0.03)	0.22 (0.03)	0.21 (0.04)	0.22 (0.04)	0.23 (0.05)
	0.04 (0.03)	0.04 (0.03)	0.03 (0.03)	0.03 (0.04)	0.04 (0.04)	0.08 (0.07)
sd(u)	0.60	0.58	0.53	0.56	0.54	0.69
sd(ϵ)	0.98	0.99	0.99	1.07	1.11	1.34
MSE	0	0.18	0.25	0.33	0.47	0.57
N observed	500	460	450	425	400	375

Table 5: Estimations associated to significant PLS component t_1, t_2, t_3, t_4, t_5. Brackets give the standard deviations obtained by bootstrap, $N = 500$ and $Var(u) = 2$

Missing (%)	0%	8%	10%	15%	20%	25%
C (sd)	7.47 (0.04)	7.30 (0.03)	7.48 (0.04)	7.60 (0.04)	7.36 (0.05)	7.25 (0.04)
	2.25 (0.04)	2.46 (0.04)	2.26 (0.04)	2.17 (0.05)	2.44 (0.06)	2.48 (0.05)
	1.05 (0.05)	1.12 (0.05)	1.03 (0.05)	0.86 (0.05)	1.10 (0.07)	1.12 (0.07)
	0.19 (0.04)	0.21 (0.05)	0.16 (0.05)	0.23 (0.05)	0.17 (0.07)	0.14 (0.06)
	0.10 (0.05)	0.11 (0.05)	0.11 (0.05)	0.12 (0.06)	0.13 (0.08)	0.09 (0.07)
sd(u)	1.71	1.73	1.71	1.79	1.90	1.94
sd(ϵ)	1.02	1.01	1.03	1.03	1.05	1.12
MSE	0	0.15	0.22	0.31	0.46	0.83
N observed	500	460	450	425	400	375

Figure 1 illustrates the stability of the distribution of Y after multiple imputation as showed by the boxplot of Y after imputation using MI-PLS-L2M, with respect to the percentages of missing values. The sample size was $N = 500$ and the variance of the random effect was $Var(u) = 2$. A general remark is that the distribution of the predicted data seems to be relatively close to the initial dataset with no missing values.

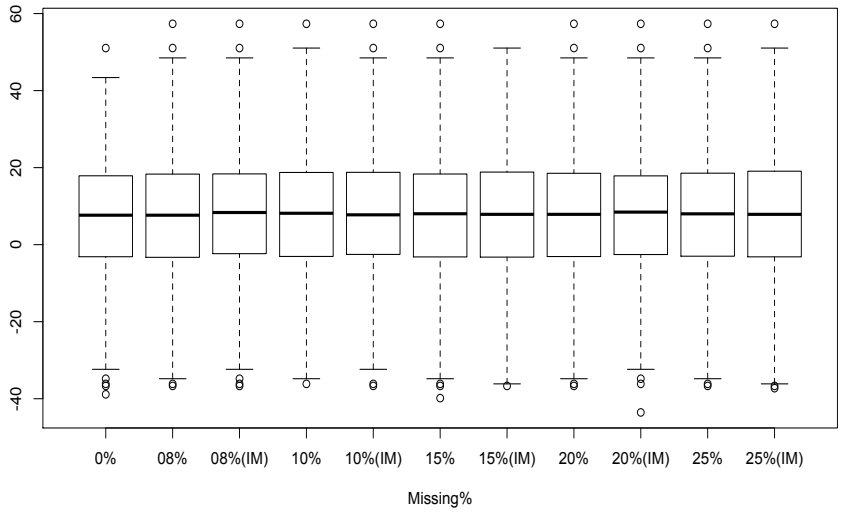


Figure 1: From the left to the right : Boxplot with 0%, 8% (before and after imputation), 10% (before and after imputation), 15% (before and after imputation), 20% (before and after imputation), 25% (before and after imputation) of missing values

4.3. Comparing with MI-PLS and MI-L2M methods

There is no known method to deal with the problem of missing data for a linear mixed model in presence of correlation between covariates. However, to assess the contribution of the method we have confronted the result of MI-PLS-L2M with the results obtained with MI-PLS and MI-L2M. By construction, the method MI-PLS proposed by Bastien (2008) will not take into account the random effect and then for assessing the contribution of our method we have considered different variances of the random effects: it was expected that with large variance of u ($Var(u) = 2$) the MI-PLS-L2M will be better, although with a small variance of u ($Var(u) = 0.5$) the results obtained with the two algorithms will be similar. Concerning the MI-L2M method proposed by Schafer and Yucel (1998), it was not adapted to the case of collinearity because of singularity of the design matrix, but we proposed to make the regression of Y_{obs} on each vector of the design matrix and each component of β was obtained from a one-dimensional model. On each model, a parameter of the random effect was estimated and for the final model, the mean of the estimated parameters was calculated. The MI-PLS-L2M method took into account all variables simultaneously and we have therefore expected to obtain better results. We have used the ratio of the MSE to quantify the difference between the three methods.

The first step *imputation of covariates* was the same for the three algorithms. We used the method of Honacker and King (2010), that is a multiple imputation by EM algorithm of X, as defined in 2.1 and using the R package *Amelia*.

Table 6 presents ratio between MSE obtained for the three methods of imputation according to the proportion of missing values for $N = 100$ and for $Var(u) = 2$. Clearly, the variance of the random effects was large but the sample size was too small to detect this effect. Then, the variability of random effect was spread within the random errors and the estimated model was close to a linear one. Hence, MI-PLS and MI-PLS-L2M gave very close results in terms of MSE. For all cases MI-L2M gave larger MSE since it was used as consecutive univariate methods.

Table 6: Ratio between MSE for MI-PLS and MI-L2M versus MI-PLS-L2M, N_{mis} denoting the number of missing values, for $N = 100$ and $var(u) = 2$

Missing values (%)	N_{mis}	$\frac{\text{MSE(MI - PLS)}}{\text{MSE(MI - PLS - L2M)}}$	$\frac{\text{MSE(MI - L2M)}}{\text{MSE(MI - PLS - L2M)}}$
8	8	1.1	4.1
10	10	1.11	5.47
15	15	1.14	4.61
20	20	1.08	5.05
25	25	1.16	5.75

In Table 7 ratio for MSE are presented for $N = 500$ and $Var(u) = 2$. This larger sample size allowed to well estimate the random effect and then the MI-PLS-L2M method took into account this variable. Moreover, it gave better results than MI-PLS where the random effect was omitted. Concerning MI-L2M, the conclusion was the same than in the first situation because of the collinearity.

Table 7: Ratio between MSE for MI-PLS and MI-L2M versus MI-PLS-L2M, N_{mis} denoting the number of missing values, for $N = 500$ and $var(u) = 0.5$

Missing values (%)	N_{mis}	$\frac{\text{MSE(MI - PLS)}}{\text{MSE(MI - PLS - L2M)}}$	$\frac{\text{MSE(MI - L2M)}}{\text{MSE(MI - PLS - L2M)}}$
8	40	1	4.06
10	50	1	4.82
15	75	1.03	4.84
20	100	1	4.60
25	125	1.04	5.09

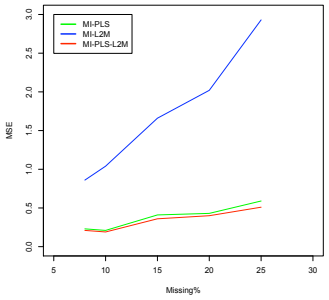
In Table 8 the ratio of MSE were obtained for $N = 100$ and $Var(u) = 0.5$. The variance of the random effect was too small to get significant estimations with MI-PLS-L2M. Then, this situation concords with the first one where the random effect was not detected and MI-PLS-L2M gave the same results than MI-PLS. Concerning MI-L2M, the conclusion is the same as previously.

Table 8: Ratio between MSE for MI-PLS and MI-L2M versus MI-PLS-L2M, N_{mis} denoting the number of missing values, for $N = 500$ and $var(u) = 2$

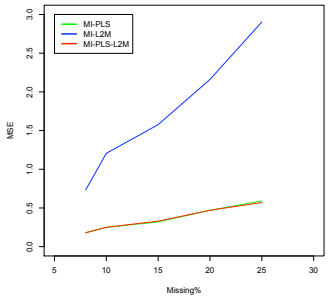
Missing values (%)	N_{mis}	$\frac{\text{MSE(MI - PLS)}}{\text{MSE(MI - PLS - L2M)}}$	$\frac{\text{MSE(MI - L2M)}}{\text{MSE(MI - PLS - L2M)}}$
8	40	3.54	2.74
10	50	3.49	4.72
15	75	3.18	5.16
20	100	3.19	5.18
25	125	2.22	3.63

In conclusion, MI-PLS-L2M performed better than the two other methods when the random effect is significantly estimated; that is, when the sample size

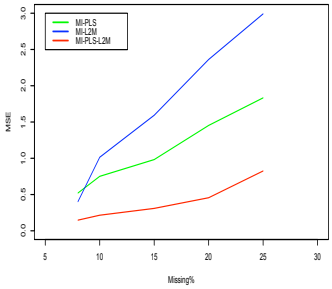
allows to estimate its variability. When random effect was not significant (too small variance, or too small sample size), MI-PLS-L2M and MI-PLS gave similar results and they were better than MI-L2M as soon as there was a collinearity between covariates. Figure 2 summarizes this analysis by showing the associated MSE, respectively for $N = 100$ and $Var(u) = 2$, $N = 500$ and $Var(u) = 0.5$, and $N = 500$ and $Var(u) = 2$. The three methods have a term of error increasing with the number of missing values.



(a) $N = 100$ and $Var(u) = 2$



(b) $N = 500$ and $Var(u) = 0.5$



(c) $N = 500$ and $Var(u) = 2$

Figure 2: MSE associated to the three methods with respect to the % of missing data

5. The coffee example

5.1. The data

We used the data presented in Vivien and Sabatier (2001): 17 coffee samples were prepared with 3 parameters which are temperature, proportion grinding-water and size, with 3, 3 and 2 modalities, respectively. 4 samples, with similar physico-chemical properties were chosen as reference sample for a comparison with the other 13 samples. In order to evaluate the different samples, 7 judges from the ENSBANA (Dijon, France) have compared 3 times each of the 13

samples to a reference sample and have answered 6 questions: *Which both is the more characteristic in flavor, the more intense in flavor, the most bitter, the more acid, the most characteristic in aroma ans the most intense in aroma ?* by specifying if the difference was *very easy, easy, complicated, hard,* or *almost impossible* to evaluate. In order to summary these judgements, a scale, from 1 to 5 was created where:

- 1: The reference sample is strictly greater than the sample.
- 2: The reference sample is greater than the sample.
- 3: The judge has perceived no difference.
- 4: The sample is greater than the reference sample.
- 5: The sample is strictly greater than the reference sample.

Here, we have chosen to create a mean score of these marks, called Y, which constitutes our variable of interest. The judges were considered as a random effect of 7 levels and X is a design matrix associated to 10 fixed effects corresponding to the 10 physico-chemical properties and we simulate 15% of missing values on Y and X, by random resampling without replacement. Table 9 presents the fixed effects and Table 10 provides an extract of the data.

Table 9: The fixed effects

Dry extract	EXS
Extraction rate	TEE
PH	PHH
Acidity	CID
Optical density to 430n. m	DO4
Optical density to 510n. m	DO5
Conductance	CDT
Caffeine	CAF
Viscosity	VIS
Retention of water in the milling	CPR

Table 10: Fixed effects for each coffee sample

Cof.	EXS	TEE	PHH	CID	DO4	DO5	CDT	CAF	VIS	CPR
1	2.5	-1.18	-0.02	1.02	0.72	0.31	3.09	0.95	0.32	0.13
2	-1.44	-4.62	0.27	-0.66	-0.42	-0.17	-2.42	-0.2	-0.11	-0.34
3	0.18	1.42	-0.10	0.10	0.09	0.04	-0.25	0.25	-0.09	-0.70
4	0.85	6.59	0.05	0.35	0.56	0.28	0.69	-0.05	-0.04	-0.42
5	0.49	4.76	0.00	0.22	0.33	0.11	0.44	0.32	-0.04	-0.07
6	-0.08	-0.15	0.05	0.00	-0.10	-0.06	-0.06	-0.20	-0.19	0.07
7	0.18	2.56	0.10	0.01	0.01	-0.01	0.32	0.00	-0.13	0.01
8	-1.16	-4.12	0.05	-0.57	-0.26	-0.10	0.32	0.00	-0.13	0.01
9	-1.07	-2.76	0.00	-0.22	-0.12	-0.04	-0.64	-0.27	-0.16	-0.09
10	-0.45	4.76	0.00	-0.50	-0.22	-0.08	-1.68	-0.50	-0.22	0.08
11	-1.06	-2.85	-0.05	-0.63	-0.17	-0.05	-2.08	-0.60	0.09	0.00
12	1.89	-5.82	0.15	0.78	0.28	0.12	2.30	1.05	0.05	-0.27
13	1.45	-6.97	0.20	0.58	0.18	0.08	2.42	.50	0.05	-0.21

Table 11 provides the correlation values of the fixed effects and the response variable. It justifies clearly the use of the PLS method.

Table 11: Correlation between the response variable and the fixed effects

Cor	Y	EXS	TEE	PHH	CID	DO4	DO5	CDT	CAF	VIS	CPR
EXS	0.624	1									
TEE	0.254	0.021	1								
PHH	-0.163	0.030	-0.492	1							
CID	0.578	0.974	-0.019	0.030	1						
DO4	0.519	0.779	0.194	0.085	0.752	1					
DO5	0.644	0.859	0.283	-0.212	0.863	0.838	1				
CDT	0.400	0.755	-0.119	0.409	0.731	0.824	0.669	1			
CAF	0.401	0.862	-0.272	0.175	0.881	0.753	0.662	0.778	1		
VIS	0.429	0.681	-0.275	-0.071	0.645	0.635	0.691	0.555	0.639	1	
CPR	0.021	-0.044	-0.016	-0.087	-0.106	-0.305	-0.128	0.016	-0.158	0.084	1

5.2. Competitor methods

As in the simulations study, we have confronted the results of the MI-PLS-L2M method with those of MI-PLS and MI-L2M, obtained after ten compilations of the algorithm. We used the MSE to quantify the difference between the imputation of Y and its true value.

5.3. Results

Table 12 presents ratio between MSE obtained for the three methods of imputation according to the 15% of missing values.

Table 12: Ratio between Mean Square Errors for MI-PLS and MI-L2M versus MI-PLS-L2M, with 15% of missing values

$\frac{\text{MSE(MI - PLS)}}{\text{MSE(MI - PLS - L2M)}}$	$\frac{\text{MSE(MI - L2M)}}{\text{MSE(MI - PLS - L2M)}}$
10.516	15.671

On this example, MI-PLS-L2M seems to perform better than the two other methods. We have compared the different estimations obtained with MI-PLS-L2M without missing values and with $p = 15\%$. We estimated C, the coefficient parameter associated to the PLS components, and the standard errors of u and ϵ. Table 13 contains estimations for the significant PLS components at the risk level $\alpha = 0.05$. The test of significance was based on $B = 200$ bootstrap samples. As expected only six components were retained for the model, as in Vivien and Sabatier (2001). The loss of accuracy obtained under the missing mechanism did not appear significant. The random effect u was estimated significantly with a variance close to that of the errors ϵ. It might explain the advantage of the use of MI-PLS-L2M on this dataset.

Table 13: Estimations associated to significant PLS component t_1, t_2, t_3, t_4, t_5, t_6. Brackets give the standard deviations obtained by bootstrap.

Missing (%)	0	15
C (sd)	0.375 (0.001)	0.366 (0.002)
	0.132 (0.097)	0.158 (0.162)
	0.066 (0.079)	0.075 (0.047)
	0.031 (0.148)	0.057 (0.158)
	0.884 (0.787)	0.629 (0.741)
	-1.283 (1.175)	-1.909 (1.968)
sd(u)	0.045	0.042
sd(ϵ)	0.086	0.096
MSE	0	0.221
N observed	91	77

Table 14 presents estimations for the random effect. It appears that the two first judges overestimate the coffee score while the effect of the third judge seems to make decrease the score. The other judges seem to estimate in the same way the coffee score.

Table 14: Estimations associated to the significant random effect.

Judge	1	2	3	4	5	6	7
$\hat{u}$	0.42	0.49	-0.72	0.30	0.33	0.33	0.36

As observed in the simulation study, the boxplots before and after imputation associated to $p = 15\%$ of missing values were very close to the initial dataset and we have omitted their representation.

6. Discussion

The algorithm MI-PLS-L2M was proposed to deal with the problem of missing data in a linear mixed model when covariates are correlated. It combines the multiple imputation theory developed by Rubin (1987) adapted to the linear mixed models with the PLS method introduced by Wold (1975) (see also Tenenhaus, 1998). It is also an adaptation of the MI-PLS algorithm proposed by Bastien (2008) and the MI-L2M initiated by Schafer and Yucel (1998), since it is dedicated to the problem of missing data in the presence of both collinearity and random effect.

Simulation studies are carried out which suggest that the proposed method is advocated as soon as a random effect is significant. When the random effect is detected, MI-PLS-L2M provides good estimations of the parameters and keeps the distribution shape of the original data before imputation. It is also shown that the MSE increases slowly with the percentage of missing values.

Moreover, the gain provided by our algorithm compared to the method used in Bastien (2008) and the method proposed in Schafer and Yucel (1998) was studied. Because of the singularity of the design matrix due to the collinearity, the MSE calculated for MI-PLS-L2M was better than that of MI-L2M. The ratio between MSE shown better performances of MI-PLS-L2M other MI-PLS when the random effects are significant. When the variance of the random effect is small, or for sample size too small to detect it, MI-PLS and MI-PLS-L2M gave similar results.

The application of our method to a real dataset showed good performance through the MSE and the estimation of the parameters. In addition, the method detected a significant random effect and then a new interpretation of the analysis of the coffe dataset.

Future research will be to adapt the MI-PLS-L2M to the generalized linear mixed models. The way is to use a step of linearization of the model, adapting for instance the algorithm of Schall (1991).

Aknowledgements We are very grateful to the reviewers for their helpful comments and suggestions. We also thank the Editor and an Associate Editor for their careful readings. This work has been supported by Region Provence Alpes Cote d'Azur.

References

[0] P. Bastien : Régression PLS et données censurées. *PhD thesis. Conservatoire National des Arts et Métiers, Paris.* 2008.

[0] P. Bastien, V. Esposito-Vinzi, and M. Tenenhaus : PLS generalised linear regression. *Computational Statistics and Data Analysis*, 48, 17-46. 2005.

[0] A.P. Dempster, N.M. Laird, and D.B. Rubin : Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39, 1-38. 1977.

[0] C.R. Henderson : Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics*, 31, 423-447. 1975.

[0] C.R. Henderson, O Kempthorne, S.R. Searle, and C.M. von Krosigk : The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, 15, 192-218. 1959.

[0] J. Honacker, and G. King : What to do about missing values in Time-Series Cross-Section Data. *American Journal of Political Science*, 54, 561-581. 2010.

[0] J.G. Ibrahim : Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85, 765-769. 1990.

[0] R.A. Little, and D.B. Rubin : *Statistical Analysis with Missing Data*, 2nd Edition. J.Wiley and Sons, New York. 2002.

[0] P. McCullagh, and J.A. Nelder : *Generalized Linear Models*, 2nd Edition. Chapman and Hall, London. 1989.

[0] D.B. Rubin : *Multiple Imputation for Non-response in Surveys*. J. Wiley and Sons, New York. 1987.

[0] D.B. Rubin : Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*, 91, 473-489. 1996.

[0] J.L. Schafer : *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London. 1997.

[0] J.L. Schafer, and R.M. Yucel : Fitting multivariate linear mixed models with incomplete data. *Proceedings of the Statistical Computing Section of the American Statistical Association.*, 177-182. 1998.

[0] R. Schall : Estimation in generalized linear models with random effects. *Biometrika*, 78, 719–727. 1991.

[0] M. Tenenhaus : *La régression PLS: théorie et pratique*. Technip, Paris. 1998.

[0] M. Tenenhaus, J.P. Gauchi, and C. Ménardo : Régression PLS et applications. *Revue de Statistique Appliquée*, 43,7-63. 1995.

[0] G. Verbeke, and G. Molenberghs : *Linear Mixed Models for Longitudinal Data*. Springer, New York. 2000.

[0] M. Vivien, and R. Sabatier : Une extension multi-tabeaux de la régression PLS. *Revue de Statistique Appliquée*, 49, 31-54. 2001.

[0] F. Westad, and H. Martens : Variable selection in NIR based on the significance testing in Partial Least Squares Regression. *Journal of Near Infrared Spectroscopy*, 8, 117-124. 1999.

[0] S. Wold : *Path models with latent variables: the non-linear iterative partial least squares (NIPALS) approach*. Academic Press. 1975.

[0] C.F.J. Wu : On the convergence properties of the EM algorithm. *Annals of Statistics*, 11, 95-103. 1983.

Appendix

The Henderson method (1959)

Given the random effects, we assume that Y, X and u satisfy (1).
The Henderson method (1959) defines an equation system

$$\hat{\beta} = (\sum_{i=1}^{n} X_i' V_i^{-1} X_i)^{-1} \sum_{i=1}^{n} X_i' V_i^{-1} y_i$$

$$\hat{u} = (\sum_{i=1}^{n} D_i' Z_i' V_i^{-1})^{-1} \sum_{i=1}^{n} y_i - X_i \hat{\beta}_i,$$

where $\tilde{\beta}$ and $\tilde{u}$ are solution of the GLS (Generalized Least Squares) and the BLUP (Best Linear Unbiased Predictor). Writing:

$$X'R^{-1}X\tilde{\beta} = X'R^{-1}(y - Z\tilde{u}) \quad (4)$$

$$(Z'R^{-1}Z + D^{-1})\tilde{u} = Z'R^{-1}(y - X\tilde{\beta}), \quad (5)$$

we obtain the following system

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}T & Z'R^{-1}Z + D^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y. \end{bmatrix}$$

Expression of PLS components in terms of the original explanatory variables (Bastien et al., 2005)

All variables $y, x_1, ..., x_j, ..., x_p$ are assumed to be centered. The PLS regression model with h components is written as

$$y = \sum_{H=1}^{h} c_H (\sum_{j=1}^{p} w_{Hj}^{*} x_j) + residual, \quad (6)$$

with the constraint that the PLS components $t_H = \sum_{j=1}^{p} w_{Hj}^* x_j$ are orthogonal and the parameters c_H and w_{Hj}^* in (6) are to be estimated.

The estimated regression equation may be then expressed in terms of the original variables x_j's:

$$\hat{y} = \sum_{H=1}^{h} c_H (\sum_{j=1}^{p} w_{Hj}^* x_j) = \sum_{j=1}^{p} (\sum_{H=1}^{h} c_H w_{Hj}^*) x_j = \sum_{j=1}^{p} \beta_j x_j.$$