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 for a linear mixed model in presence of correlation between covariates is considered. The missing mechanism concerns both dependent variable and design matrix. We propose an imputation algorithm combining multiple imputation and Partial Least Squares (PLS) analysis methods. Our 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 built by adding the random effects to PLS components. The method is applied on simulations and on real data.

Key words: Multiple Imputation, Missing Data, Linear Mixed Regression Model.

1. Introduction

The problem of handling missing data has been extensively studied in the statistical literature. In Little and Rubin (2002), three non-response mechanisms are described: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). The MCAR mechanism assumes that the missingness on a variable does not depend on the variable nor on the others variables. The MNAR mechanism assumes that missingness depends only on the variable that contains missing values. The MAR mechanism assumes that missingness only depends on the observed components and not on the components that are missing. This work will be concerned with the MAR mechanism. A method to deal with MAR data in statistical analyses is the multiple imputation introduced by Rubin (1987). Rubin proposes this method when the nonresponse mechanism concerns both response variable and covariates. It consists in replacing each missing value by a vector of $M \geq 2$ imputed values. It 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*

*Corresponding author

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

a particular 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).

So far, this technique has been widely used for imputation in linear models (Schafer, 1997), as well as in generalized linear models (Ibrahim, 1990). Recently, Bastien (2008) studied the problem of missing data imputation in generalized linear models when covariates are correlated. The author combines the technique of Partial Least Squares (PLS) regression (Wold, 1975) to the multiple imputation method, obtaining a successfull method, called Multiple Imputation with Partial Least Squares (MI-PLS). It consists, after imputation of missing values on each explicative variable, to impute the missing data on the variable of interest by a PLS regression instead of a linear regression. The problem of missing data in linear mixed models (McCullagh and Nelder, 1989) has also been investigated by Schafer and Yucel (1998), using multiple imputation on covariates. However this method breaks down when covariates are linearly dependent, due to the singularity of the design matrix.

In this paper, the problem of handling missing data in linear mixed models with correlated covariates is considered. An algorithm is proposed combining PLS method, defined by Bastien (2008) and multiple imputation, defined by Schafer and Yucel (1998). This algorithm is denoted by Multiple Imputation with Partial Least Squares in Linear Mixed Models (MI-PLS-L2M). The method relies into two 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). So several completed datasets are obtained, on which are computed the PLS components. Random effects are then reintroduced in the model and a linear mixed model is obtained where the fixed PLS latent variables are constructed by linear combinations of the observed regressors. The Henderson method (1959) is used to estimate the parameters of the model. Afterwards, we use the prediction of the final model based on the latent variables and the random components to reconstruct the dependent variable. Following Bastien *et al.* (2005), a bootstrap validation procedure allows to test the significance of the fixed effects.

Our method is applied on simulations and is 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 that introduced by Schafer and Yucel (1998) and denoted Multiple Imputation in Linear Mixed Models (MI-L2M) method. Based on the Mean Square Ratio, our algorithm gives better approximations of the data than the two others methods. Moreover, the results of simulations show good performance in situation of high colinearity. Our algorithm is applied on a real data set, the coffee example used in Vivien and Sabatier (2001): 17 samples of coffee are evaluated by 7 judges. In Vivien and Sabatier (2001), for each judge, characteristics of the coffee are 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 levels random effect corresponding to the 7 judges . The results are similar to those of Vivien

and Sabatier (2001). In addition we obtain significant random effects.

The paper is organized as follows: in Section 2 we review 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 to the study of the coffee data set. Section 6 takes part of a short discussion.

2. Multiple imputation and PLS background

Let $Y = (Y_1, \ldots, Y_n)$ be a vector of n observations. Let $X = (X_1', \cdots, X_p')$ be a matrix of p covariates referred to the fixed effects, where x' stands for the transpose of x. Let $U = (U_1', \cdots, U_K')$, $U_j \in \mathbb{R}^{q_j}$, be a matrix of K random vectors referred to the random effects. Given the random effects, we assume that Y, X and U satisfy a linear mixed model, that is

$$Y_i \quad = \quad X_i'\beta + Z_i'U + \epsilon_i, \forall i \in \{1, ..., n\}, \tag{1}$$

where $\beta \in \mathbb{R}^p$ is an unknown vector of regression coefficients, and Z is a known matrix associated to the random effects. Writing $\mathcal{N}_d(a, b)$ for the d-dimensional normal distribution with mean a and variance b, we assume that

$$U \sim \mathcal{N}_q(0, D),\ \epsilon_i \sim \mathcal{N}_1(0, R_i),\ Cov(U, \epsilon_i) = 0,$$

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

2.1. Multiple Imputation

When the data are Missing At Random, Rubin (1987) and Schafer (1997) introduce the multiple imputation as the most reliable method both on accuracy and efficiency. The principle can be divided into three steps. First, $m > 1$ samples of possible values for the missing data are created *(Imputation)*. Each of the m complete data set is analyzed using the standard statistical method that would be used in the absence of nonresponse *(Analysis)*. Finally, the results of the m analysis are combined *(Pooling)*.

Honacker *et al.* (2009) proposed to impute missing data on the fixed effects by multiple imputation and bootstrap Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin, 1977, Wu, 1983). The main idea is the following. For the first step *Imputation*, the m samples of value are created by bootstraping a dataset with the same dimension as the original data. Then, for the *Analysis* step, the model is estimated by EM algorithm, and leads to replace the missing values of the sample. Finally, the results of the m analysis are combined *(Pooling)*, as usually.

The EM part of the algorithm can be described as follows. Assume that the missing value are only on the response variable $Y = (Y_{obs}, Y_{mis})$. We note θ the vector of parameters of the model on the complete dataset. The 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) Iteration until convergence of θ.

2.2. PLS Regression

From now, we consider that the vectors Y, X_1, .., X_p are centered.
When the covariates are correlated, PLS regression can be used in place of the classical linear regression model. The idea is to replace 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 $j \in \{1, ..., p\}$, with

  $$Y_i = a_{1j} X_{ij} + \epsilon_i, \forall i \in \{1, ..., n\}.$$

  Normalize the parameter a_{1j}

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

  and write

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

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

  $$Y_i = a_{hj} X_{ij} + c_1 t_{1i} + ... + c_{h-1} t_{(h-1)i} + \epsilon_{Y_{(h-1)i}}, \forall i \in \{1, ..., n\},$$

  where a_{hj} is the parameter associated to X_{ij}, and c_h is the parameter associated to t_h, both being estimated by the model. The significance of each parameter t_h permits to choose the number of components h.
  Normalizing the parameter a_{hj}

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

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

  $$X_{ij} = p_1 t_{1i} + ... + p_{(h-1)} t_{(h-1)i} + X_{(h-1)ij}, \forall j \in \{1, ..., p\} \text{ and } \forall i \in \{1, ..., n\},$$

  with p_h the parameter associated to t_h, estimated by the model and $X_{(h-1)j}$ the residual term of the model, we write

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

3. Adapting PLS to the linear mixed model

3.1. Description of the method

We consider the model (1) with Y and X centered. We assume that both Y and X observations are affected by missing values, according to a Missing At Random mechanism. Let Y_{obs} and Y_{mis} denote the vectors of observed and missing values of Y. The algorithm can be decomposed into different steps:

Multiple imputation on fixed effects. We first impute X by the multiple imputation method described in Subsection 2.1. We obtain m complete datasets.

PLS regression. Temporarily eliminating the random effects and according to Subsection 2.2, we compute the PLS components T of the heteroscedatic linear model:

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

where $\widetilde{\epsilon} = \epsilon + ZU$. Clearly, $\widetilde{\epsilon}_i$ is a vector of dependent random variables such that $\widetilde{\epsilon} \sim \mathcal{N}(0; \sigma^2)$, with $\sigma^2 = ZDZ' + R \neq 1$. But this transformation of ϵ does not modify the PLS components, since they are completely specified by the correlation between X and Y.

Estimation of the parameters. Reintroducing the random effects, let us consider the new linear mixed model

$$Y_{obs} = TC + ZU + \epsilon. \tag{2}$$

The fixed parameters C and the random parameters U are estimated using Henderson's method (1959) and we denote by $\hat{C}$ and $\hat{U}$ their respective estimators. Then, we reformulate $\hat{\beta}$ according to W and $\hat{C}$, as defined in Bastien (2005).

Bootstrap. We use a bootstrap procedure consisting in sampling with replacement in Y_{obs} and the associated components T and Z. The third step is then applied on B (B fixed) bootstrap samples, giving a vector β^* of B estimators of β. In view to predict Y_{mis}, we only keep significative coefficients, that is with confidence intervals that do not contain zero. We then simulate $N_1 \sim \mathcal{N}(Z\hat{U}, \hat{V}_U)$ and $N_2 \sim \mathcal{N}(0, \hat{V}_\epsilon)$. The prediction of Y_{mis} is given by

$$\widehat{Y_{mis}} = X\hat{\beta} + N_1 + N_2,$$

with β selected as above.

Pooling. The last step consists in pooling the m estimates of Y and X into one estimate by calculation of the mean of the estimates parameters, as done in Little and Rubin (2004). Let $\hat{\theta_1}, ..., \hat{\theta_m}$ denote the m estimates of the parameter θ and $\hat{V_1}, ..., \hat{V_m}$ the m estimates of the variance-covariance matrix V, obtained by the m complete datasets. Thus, the global estimator of θ is given by

$$\overline{\theta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}_i.$$

The variance-covariance matrix of $\overline{\theta}$ has two components: one given by the variability between the m datasets, and one obtained from the variability within every dataset. The variability between the m datasets is the mean of the estimated variance

$$\overline{V} = \frac{1}{m}\sum_{i=1}^{m}\hat{V}_i,$$

and the variability within every dataset is given by

$$\overline{W} = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i - \overline{\theta})(\hat{\theta}_i - \overline{\theta})'.$$

Then the global variance of $\overline{\theta}$ is given by

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

3.2. Description of the algorithm

Hereafter, we summarize the steps of our algorithm.

Algorithm steps.

1. Step 1. *Imputation of covariates.* Missing data on the matrix X are imputed and m complete datasets are obtained.

2. Step 2. *PLS procedure.* On each dataset, assuming that Y is centered and X are centered and normalized, the PLS components T are calculated on the linear model $Y = X\beta + \widetilde{\epsilon}$, with $\widetilde{\epsilon} = \epsilon + ZU$.

3. Step 3. *Considering random effects.* The estimations of C and U, denoted by $\hat{C}$ and $\hat{U}$, are obtained from the mixed model $Y = TC + ZU + \epsilon$, and $\hat{\beta}$ is reconstructed using $\hat{C}$ and the PLS components.

4. Step 4. *Testing fixed effects.* Bootstrap testing procedures are used to select significant fixed effects.

5. Step 5. *Imputation of Y.* When Y is missing, Y is simulated from $\mathcal{N}(X\hat{\beta} + Z\hat{U}; ZDZ' + R)$.

6. Step 6. *Pooling.* Pool the m estimates of Y and X into one estimate.

4. Simulation study

We compare the performance of our algorithm with that of several competitors on simulated data arising from a linear mixed model with correlated fixed effects. The measure of performance is the Mean Square Error for missing values

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

4.1. The method

In our simulations, for $N = 500$ observations, the design matrix X consists of a 15-dimensional vector such that the last 10 components are deeply 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

$$\begin{aligned}
x_6 &= x_1 + 4x_2 - x_3 + 2x_4 + 3x_5, & x_7 &= 2x_1 - x_2 + 5x_3 - 3x_4 - 2x_5,\\
x_8 &= x_1 + 4x_2 - x_3 + 4x_4 + 0.5x_5, & x_9 &= 2x_1 - x_2 + 5x_3 + x_4 - 3x_5,\\
x_{10} &= 3x_1 + 3x_2 + 4x_3 + 5x_4 + 0.5x_5, & x_{11} &= 3x_1 - 2x_2 - 5x_3 + 0.5x_4 + x_5,\\
x_{12} &= -x_1 + x_2 - x_3 + x_4 - x_5, & x_{13} &= x_1 - 4x_2 + 4x_3 + 0.5x_4 + 0.5x_5,\\
x_{14} &= 2x_1 + 2x_2 - 2x_3 - 2x_4 - 2x_5, & x_{15} &= 0.5x_1 - 0.5x_2 + x_3 + x_4 + 2x_5.
\end{aligned}$$

The random effect is a 3 levels vector $U \sim \mathcal{N}_3(0, 2I)$, where I denotes the identity matrix.

We consider an independent error term $\epsilon \sim \mathcal{N}(0,1)$. The output variable Y belongs to $\mathbb{R}$. Algorithm MI-PLS-L2M as well as competitors MI-PLS and MI-L2M are run on a set of size $N = 500$ observations, with prescribed percentage p of missing value on Y and X. We choose $p \in \{8\%, 10\%, 15\%, 20\%, 25\%\}$.

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 is fixed to $m = 5$ following Rubin (1987) who shown only three to five imputations are enough to have excellent results.

4.2. Competitor 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, we confront the result of our method MI-PLS-L2M with MI-PLS and MI-L2M. By construction the method MI-PLS proposed by Bastien (2008) will not take into account the random effect. Concerning the method MI-L2M proposed by Schafer and Yucel (1998), it is not adapted to the case of colinearity. The regression of Y_{obs} is made on each vector of the design matrix and each component of β is obtained from a one-dimensional model. On each model, a parameter of the random effect is estimated and for the final model, the mean of the parameters estimates is calculated. We use the Mean Square Error to quantify the difference between the imputation of Y and its true value.

The first step *imputation of covariates* is the same for the three algorithms. We use 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.

4.3. Results

Table 2 presents ratio of the Mean Square Error obtained for the three methods of imputation according to the proportion of missing values.

Table 2: Mean Square Error Ratio for MI-PLS and MI-L2M versus MI-PLS-L2M, N_{mis} denoting the number of missing values

Missing values (%)	N_{mis}	$\frac{\text{MSE(MI}-\text{PLS)}}{\text{MSE(MI}-\text{PLS}-\text{L2M)}}$	$\frac{\text{MSE(MI}-\text{L2M)}}{\text{MSE(MI}-\text{PLS}-\text{L2M)}}$
8	40	3.544	2.735
10	50	3.493	4.716
15	75	3.178	5.155
20	100	3.186	5.18
25	125	2.222	3.625

For these situations our method performs better than the multiple imputation with PLS regression (MI-PLS). In fact, as the random effect has an important variance, the error term is more important. MI-PLS overestimates the error term since this term is made up of the residuals and the random effects parameters of the model. It is then not appropriated in this situation. For the method of multiple imputation on a linear mixed model (MI-L2M), it can be implemented only on univariate regressions. Thus, this method has a term of error more important than our method.

Figures 1 shows the associated Mean square errors. The three methods have a term of error increasing with the number of missing values, but MI-PLS-L2M has a term of error smaller than the others.

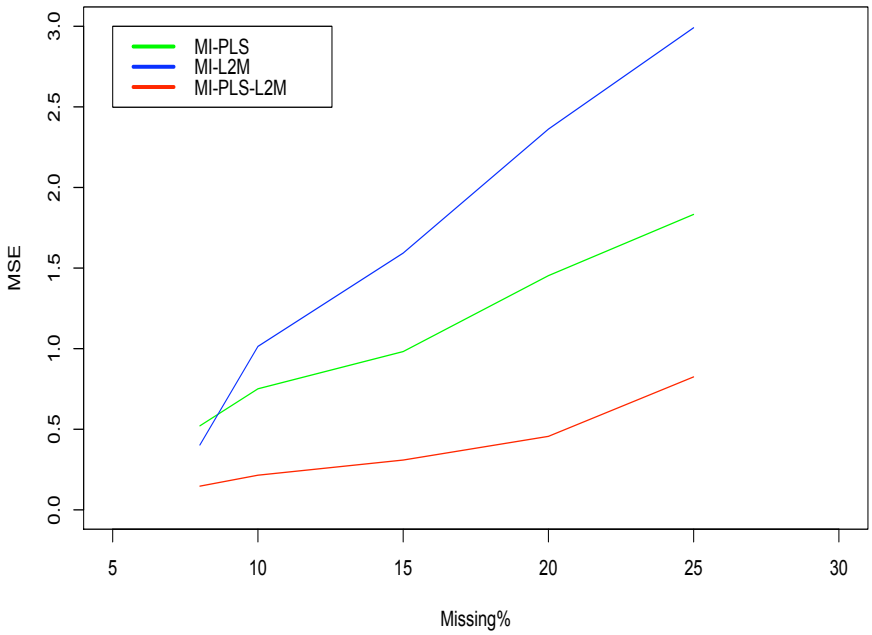

Figure 1: MSE associated to the three methods with respect to the % of missing data

.

Finally, we compare the different estimations obtained with our method when $p = 0, 8\%, 10\%, 15\%, 20\%, 25\%$. We estimate C, the coefficient parameter associated to the PLS components, and the standard errors of U and ϵ. Table 3 contains estimations for the significant PLS components at the risk level $\alpha = 0.05$. The test of significance is based on $B = 200$ bootstrap samples. As expected only five components are retained for all models.

Table 3: Estimations associated to significant PLS component t_1, t_2, t_3, t_4, t_5. Brackets give the standard errors obtained by bootstrap.

Missing (%)	0	8	10	15	20	25
C (sd)	7.466 (0.037)	7.296 (0.034)	7.478 (0.036)	7.598 (0.038)	7.364 (0.049)	7.246 (0.044)
	2.254 (0.036)	2.462 (0.04)	2.262 (0.04)	2.170 (0.045)	2.440 (0.059)	2.480 (0.052)
	1.051 (0.046)	1.121 (0.05)	1.028 (0.048)	0.855 (0.048)	1.098 (0.072)	1.12 (0.067)
	0.191 (0.042)	0.209 (0.047)	0.164 (0.046)	0.226 (0.052)	0.166 (0.068)	0.144 (0.062)
	0.104 (0.047)	0.109 (0.048)	0.110 (0.054)	0.120 (0.055)	0.134 (0.08)	0.092 (0.065)
sd(U)	1.709	1.732	1.706	1.793	1.902	1.938
sd(ϵ)	1.017	1.007	1.025	1.028	1.048	1.121
MSE	0	0.147	0.215	0.309	0.456	0.825
N	500	460	450	425	400	375

Figure 2 shows the boxplot associated to the percentages p of missing values equal to 0%, 8%, 10%; 15%, 20%, 25%. The distribution of the predicted data seems to be relatively closed to the initial dataset with no missing values.

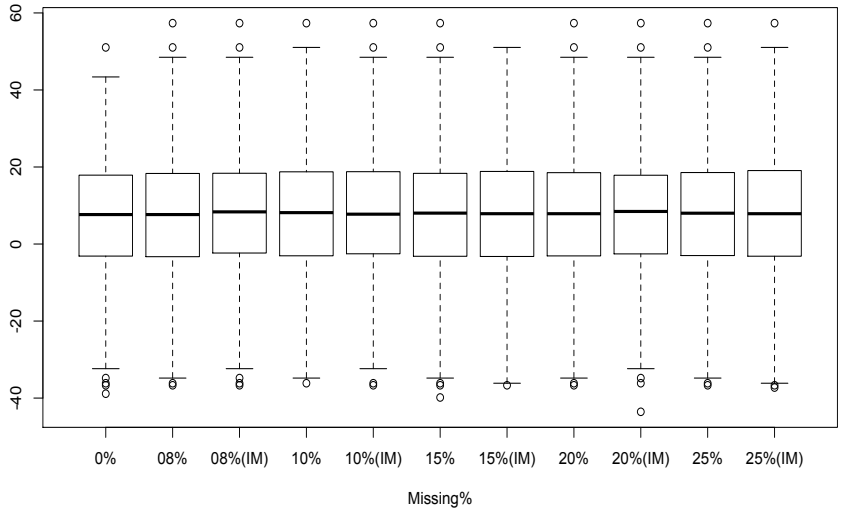

Figure 2: From the left to the right : Boxplot with 0%, 8% (before and after imputation), 10% (before and after imputation), 15% (before and after imputation), 20% (before and after imputation), 25% (before and after imputation) of missing values

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: temperature, proportion grinding-water and size, with 3, 3 and 2 modalities, respectively. 4 samples, with similar physico-chemicals properties were chosen as witness for a comparison with the others 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 witness 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 witness is strictly greater than the sample.
- 2: The witness is greater than the sample.
- 3: The judge has perceived no difference.
- 4: The sample is greater than the witness.
- 5: The sample is strictly greater than the witness.

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 4 presents the fixed effects and Table 5 provides an extract of the data.

Table 4: 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 5: 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

To justify the use of the PLS method, Table 6 provides the correlation values of the fixed effects and the response variable.

Table 6: 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 previously on the simulations, we confront the result of our method MI-PLS-L2M with MI-PLS and MI-L2M. We use the Mean Square Error to quantify the difference between the imputation of Y and its true value.

5.3. Results

Table 7 presents ratio of the Mean Square Error obtained for the three methods of imputation according to the 15% of missing values.

Table 7: Mean Square Error Ratio 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, our method seems to perform better than the two methods. Finally, we compare the different estimations obtained with our method when $p = 0\%, 15\%$. We estimate C, the coefficient parameter associated to the PLS components, and the standard errors of U and ϵ. Table 8 contains estimations for the significant PLS components at the risk level $\alpha = 0.05$. The test of significance is based on $B = 200$ bootstrap samples. As expected only six components are retained for the model, as in Vivien and Sabatier (2001).

Table 8: Estimations associated to significant PLS component t_1, t_2, t_3, t_4, t_5, t_6. Brackets give the standard errors 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	91	77

Figure 3 shows the boxplot associated to the percentages p of missing values equal to 0%, 15%. The distribution of the predicted data seems to be relatively closed to the initial dataset with no missing values.

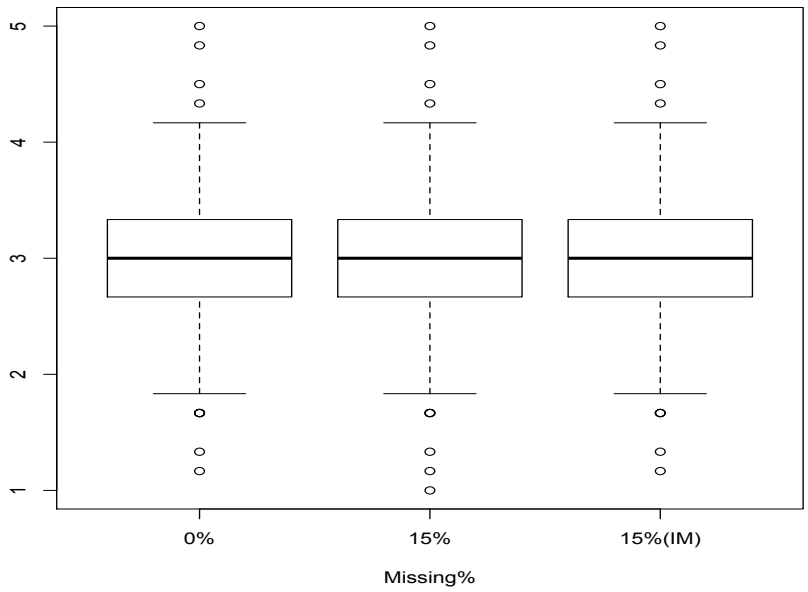

Figure 3: From the left to the right : Boxplot with 0%, 15% (before and after imputation) of missing values

6. Discussion

We have proposed an algorithm MI-PLS-L2M to deal with the problem of missing data in a linear mixed model when covariates are correlated. The algorithm combines the multiple imputation theory developed by Rubin (1987) adapted to the linear mixed models with the PLS method introduced by Wold (1975).

Simulation studies are carried out which suggest that the proposed method works well for practical situations. It is shown that the mean square error increases slowly with the percentage of missing values. Moreover it provides good estimations of the parameters and it keeps the distribution shape of the original data before imputation. We confronted our algorithm with two others, one proposed by Bastien (2008) and another proposed in Schafer and Yucel (1998). The MSE ratios shown better performances of our method.

The application of our method to a real data set also shows good performance trough the MSE and the estimation of the parameters. Moreover, the advantage of the method is to take into account the random effects.

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, using for instance the algorithm of Schall (1991).

References

P. Bastien : RÈgression PLS et donnÈes censurÈes. *PhD thesis. Conservatoire National des Arts et MÈtiers, Paris.* 2008.

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

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.

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

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.

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.

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

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

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

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

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

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

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.

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

M. Tenenhaus : *La rÈgression PLS: thÈorie et pratique.* Technip, Paris. 1998.

M. Tenenhaus, J.P. Gauchi, and C. MÈnardo : RÈgression PLS et applications. *Revue de Statistique AppliquÈe*, 43,7-63. 1995.

M. Vivien, and R. Sabatier : Une extension multi-tabeaux de la rÈgression PLS. *Revue de Statistique AppliquÈe*, 49, 31-54. 2001.

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

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