Imputation by PLS regression for generalized linear mixed models

Emilie Guyon, Denys Pommeret

Imputation by PLS regression for generalized linear mixed models

E. Guyon[a,b,*], D. Pommeret[a]

[a] *Institut de Mathématiques de Luminy, CNRS Marseille, case 907, Campus de Luminy, 13288 Marseille Cedex 9, France*
[b] *ORS, INSERM, Marseille , France*

Abstract

The problem of handling missing data in generalized linear mixed models with correlated covariates is considered when the missing mechanism concerns both the response variable and the covariates. An imputation algorithm combining multiple imputation and Partial Least Squares (PLS) regression is proposed. The method relies on two steps. In a first step, using a linearization technique, the generalized linear mixed model is approximated by a linear mixed model. A latent variable is introduced and its associated PLS components are constructed. In a second step these PLS components are used in the generalized linear mixed model to impute the response variable. The method is applied on simulations and on a real data.

Keywords: Missing data, Multiple imputation, PLS regression, Schall linearization, Generalized linear mixed models

1. Introduction

Missing data are frequently encountered in dataset leading to various challenging problems. When the non-response is independent of the response variable the missing mechanism is said to be Missing At Random (MAR), according to Rubin (1987). In this case different authors proposed multiple imputation to recover the non-response value. This method of imputation is the most reliable method both from accuracy and efficiency point of view.

*Corresponding author
Email addresses: emilie.guyon@univmed.fr (E. Guyon), pommeret@univmed.fr (D. Pommeret)

It consists of replacing each missing value by a vector of imputed values. Multiple imputations have been developed in linear models and in generalized linear models by Schafer (1997) and Ibrahim (1990), respectively.

In the context of generalized linear models with correlated covariates, Bastien (2008) combined the Partial Least Squares (PLS) regression technique 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. Another method appropriate to multicollinear data is the use of principal components as done in Aguilera et al. (2006) for estimating logistic regression with high-dimensional data. In linear mixed models the problem of missing data has been investigated by Schafer and Yucel (1998). In generalized linear mixed models, Wu and Lang (2006) proposed an extension of multiple imputation combining Monte-Carlo EM algorithms with Gibbs sampler. Multiple imputation in generalized linear mixed models have also been studied by Little (1995), Ten Have et al. (1998), and Wu (2004) among others. However these methods can break down when covariates are linearly dependent.

Recently Guyon and Pommeret (2011) considered the case of collinearity and proposed a solution for the problem of handling missing data in linear mixed models with correlated covariates. Their method can be decomposed into two steps: a first one consists in deleting the random effect to apply a PLS regression on standard linear model. Random effect is then reintroduced in a second step taking into account the PLS components. This method gave satisfactory results on simulations as well as on real dataset. The aim of this paper is to extend this work to the case of generalized linear mixed models.

We assume that a MAR missing mechanism acts both on the response variable and on the covariates. Our aim is to recover a complete dataset using the multiple imputation technique. The main idea is to apply the method of Schall (1991) to construct a latent variable by linearization. In this way we obtain a linear mixed model with unobserved responses. Although the latent variable is unobserved, we can estimate its correlation with the covariates which permits to determine the PLS components and to adapt the method of Guyon and Pommeret (2011). Finally, by reconstructing the original response variable from the complete latent ones, we get a complete dataset .

This proposed method, that we shall call multiple imputation by PLS regression for generalized linear mixed models (MI-PLS-GLMMs), is applied

on simulations for Poisson and logistic mixed models. It is also applied to the plant vegetation dataset (Zuur et al., 2009).

The paper is organized as follow: in Section 2 the model is introduced, the linearization procedure of Schall (1991) and the PLS regression are used in combination with multiple imputation procedure to predict values for missing data. In Section 3, two simulations experiment designed to analyze the efficiency of the method are implemented. Section 4 presents the analysis of the plant vegetation dataset. Finally, Section 5 contains a brief conclusion.

2. The model and the proposed method

2.1. The model

We consider a generalized linear mixed model with Y a n-vector of responses, given a set of p potential covariates, through the classical regression:

$$g(\mathbb{E}(Y_i|\xi)) = X_i^T\beta + U_i^T\xi,\ \forall i \in \{1, ..., n\}, \tag{1}$$

where g denotes the link function, X_i and U_i are fixed covariates and random effect, $\beta \in \mathbb{R}^p$ is the fixed effects coefficient and $\xi = (\xi_1', ..., \xi_K')$ is the random effect coefficient with ξ_j a random vector of dimension q_j. The design matrices X and U are known. We will restrict our attention to the case where the covariates are correlated or collinear. Write $\eta_\xi = X\beta + U\xi$ and $\mu_\xi = \mathbb{E}(Y|\xi)$. It is assumed, that given ξ, the Y components are independent and $\forall i \in \{1, ..., n\}$, $Y_i|\xi$ follows a distribution from an exponential family. It is also required that ξ follows a centered normal distribution with variance matrix D, written as $\xi \sim \mathcal{N}(0, D)$.

2.2. The proposed method

The proposed method is a multiple imputation method and can be decomposed into three parts: *imputation, analysis* and *pooling*. The *imputation* results in m complete datasets. The *analysis* consists in analyzing each of the m completed datasets. The *pooling* integrates the m analysis results into a final result. We describe these three consecutive parts.

Imputation. For each missing data entry, m imputations ($m \geq 2$) are generated by bootstrap with replacement of the covariates $X_1, ..., X_p$.

Analysis. Following Schall (1991), we consider the first order Taylor expansion of the link function,

$$Z = X\beta + U\xi + (Y - \mu_\xi)g'(\mu_\xi),$$

where g' stands for the first derivative of g. Note that we construct Z for each of the m dataset obtained by imputation, with

$$\mathbb{E}(Z) = X\beta, \quad \text{and} \quad Var(Z) = UDU' + W_\xi,$$

where $W_\xi = Var(Z|\xi)$. To estimated β, D and W_ξ, we solve the following Henderson's system (Henderson et al., 1959):

$$\begin{bmatrix} X'W_\xi^{-1}X & X'W_\xi^{-1}U \\ U'W_\xi^{-1}X & U'W_\xi^{-1}U + D^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ \xi \end{bmatrix} = \begin{bmatrix} X'W_\xi^{-1}z \\ U'W_\xi^{-1}z \end{bmatrix}.$$

To overcome the random effect, imitating Guyon and Pommeret (2011) we put

$$\widetilde{Z} = Z - U\xi,$$

where Z is the m previous latent variables obtained by Schall linearization. The associated PLS model with k components fits the following expansion:

$$\begin{aligned} f_k(x, c, w) &= \sum_{h=1}^{k} c_h t_h, \\ &= \sum_{h=1}^{k} c_h \sum_{j=1}^{p} x_{h-1,j} w_{hj}, \\ &= \sum_{h=1}^{k} c_h \sum_{j=1}^{p} x_j w_{hj}^*, \end{aligned}$$

where $w_h = \arg\max_{w_h} cov(X_{h-1}w_h, \widetilde{Z})$, c_h is the regression coefficient of t_h in the regression of $\widetilde{Z}$ on $t_1, ..., t_k$ and $w_h^* = w_h/||w_h||$. It is clear that the variable $\widetilde{Z}$ is not observed since it depends on the latent variable Z. However, it is possible to calculate the PLS components through the covariances between Z and the covariates X. For instance,

$$\begin{aligned} cov(\widetilde{Z}, X) &= cov(Z - U\xi, X), \\ &= cov(X\beta + (Y - \mu)g'(\mu), X), \\ &= \beta Var(X) + g'(\mu)cov(Y, X), \end{aligned}$$

which can be determined as soon as β and ξ are estimated.

The selection of the appropriate number of components h is based on the cross-validation criterion Q_h^2 (Tenenhaus, 1998),

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

where RSS and PRESS denote the Residual Sum of Squares and the PRedictor Error Sum of Squares, respectively (see Appendix).

In order to take into account the random effect, let us consider the new linear mixed model

$$g(\mathbb{E}(Y_i|\xi, C)) = TC + U_i^T\xi. \tag{2}$$

The fixed parameters C are estimated and the random parameter ξ is predicted using the Henderson method. We denote by $\hat{C}$ and $\hat{\xi}$ their respective estimators. We reformulate $\hat{\beta}$ according to the weights of the PLS components and $\hat{C}$ is recovered using the formulas given in Bastien et al. (2005) (see Appendix).

As in Guyon and Pommeret (2011), we use a bootstrap validation procedure to assess the statistical significance of the explanatory variables. This selection method is inspired from the work of Bastien et al. (2005). More precisely, the bootstrap procedure consists in sampling with replacement from Y with their associated components T and U. Applying (2) to the B (B fixed) bootstrap samples, we obtain a vector β^* of B estimators of β. It allows us to calculate a bootstrapped confidence interval of the regressors and empirical distributions are used to retain the significant variables at a prescribed level (arbitrarily 5%).

Pooling. The last part consists of pooling the m estimates into a single one equal to the estimated parameters mean, as done in Little and Rubin (1987). From the previous step, we obtained m estimators of D, the variance of ξ, $\hat{D}_1, ..., \hat{D}_m$ and m predictors of ξ: $\hat{\xi}_1, ..., \hat{\xi}_m$. We simulate $S \sim \mathcal{N}_n(U\overline{\xi}, \hat{F})$, where

$$\overline{\xi} = \frac{1}{m}\sum_{k=1}^{m}\hat{\xi}_k, \qquad \hat{F} = \overline{V} + (1 + \frac{1}{m})\overline{W},$$

with

$$\overline{V} = \frac{1}{m}\sum_{k=1}^{m}\hat{D}_k, \quad \text{and} \quad \overline{W} = \frac{1}{m-1}\sum_{k=1}^{m}(\hat{\xi}_k - \overline{\xi})(\hat{\xi}_k - \overline{\xi})'. \tag{3}$$

Finally, if Y_i is missing, we replaced it by

$$g(\widehat{\mathbb{E}(Y_i|\beta,\xi)}) = X_i\hat{\beta} + S_i.$$

2.3. Proposed algorithm

The MI-PLS-GLMM algorithm can be decomposed into six consecutive steps.

- Step 1. *Imputation of covariates to get m complete dataset.*
- Step 2. *Schall linearization.*
- Step 3. *PLS procedure based on the latent variable without random effect.*
- Step 4. *Taking into account the random effect to estimate C and ξ.*
- Step 5. *Bootstrap selection.*
- Step 6. *Pooling.*

3. Simulation study

In order to evaluate the performance of the proposed algorithm, simulations were performed with two sample sizes ($N = 100$ and $N = 500$) and for two variances of the random effects ($Var(\xi) = 0.5$ and $Var(\xi) = 2$). For each simulation, we have computed the algorithm and the results presented were the mean of 30 simulations. The performance criterion used is the Mean Squared Error (MSE) 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.

In our simulations, the design matrix X consists of a N-sample of a 5-dimensional covariate vector such that the last 2 components of the covariate vector are collinear with the first 3 ones, as it is shown in Table 1. The covariates are constructed as follows: three independent uniform variables $X_1 \sim \mathcal{U}([0;10])$, $X_2 \sim \mathcal{U}([-5;5])$, $X_3 \sim \mathcal{U}([5;15])$, and two linearly dependent variables

$$X_4 = X_1 + X_2 + X_3, \qquad X_5 = 2X_1 + 3X_2 + X_3.$$

We add a random effect as follows: For $N = 100$ observations, the random effect was a 3-level vector $\xi \sim \mathcal{N}_3(0, 2I)$. For $N = 500$ observations, the random effect was first a 3-level vector $\xi \sim \mathcal{N}_3(0, 2I)$, and second a 3-level vector $\xi \sim \mathcal{N}_3(0, 0.5I)$.

3.1. The Poisson model

Assuming that $Y|X, \xi$ is Poisson distributed with

$$\mathbb{E}(Y|\xi, X) = \exp(\beta X + \xi U),$$

where $\beta = (0.5, 0.5, -0.5, 0.5, 0.5)$. The MI-PLS-GLMM algorithm was run with various percentage p of missing value on Y and X. We chose $p \in \{8\%, 10\%, 15\%, 20\%\}$.

Table 1: Correlation matrix between the variables (Pearson correlation coefficients)

	Y	X1	X2	X3	X4	X5
X1	0.193	1				
X2	0.170	0.057	1			
X3	0.037	0.081	-0.046	1		
X4	0.227	0.641	0.584	0.559	1	
X5	0.243	0.580	0.809	0.252	0.930	1

Following Rubin (1987) the number of multiple imputation was fixed to $m = 5$. The different estimations of β and of the standard errors of ξ were obtained with $p = \{0\%, 8\%, 10\%, 15\%, 20\%\}$. As expected, only three PLS components were retained for all models and they were used to reconstruct the β parameters. Tables 2-4 contain estimations. The risk level was fixed at $\alpha = 0.05$ and the test of significance was based on $B = 200$ bootstrap samples. All coefficients were significant.

Table 2: Estimations for the model with $N = 100$ and $Var(\xi) = 2$ (se = standard error)

Missing (%)	0%	8%	10%	15%	20%
$\hat{\beta}1$ (se)	0.477 (0.042)	0.425 (0.039)	0.409 (0.044)	0.355 (0.051)	0.359 (0.047)
$\hat{\beta}2$ (se)	0.329 (0.061)	0.300 (0.068)	0.191 (0.068)	0.155 (0.066)	0.124 (0.067)
$\hat{\beta}3$ (se)	-0.560 (0.007)	-0.440 (0.020)	-0.474 (0.021)	-0.658 (0.021)	-0.424 (0.026)
$\hat{\beta}4$ (se)	0.440 (0.012)	0.366 (0.023)	0.528 (0.023)	0.517 (0.031)	0.529 (0.027)
$\hat{\beta}5$ (se)	0.220 (0.008)	0.176 (0.014)	0.257 (0.013)	0.257 (0.022)	0.244 (0.026)
$se(\xi)$	1.455	1.408	1.402	1.402	1.402
N observed	100	92	90	85	80

Table 3: Estimations for the Poisson model with $N = 500$ and $Var(\xi) = 0.5$ (se = standard error)

Missing (%)	0%	8%	10%	15%	20%
$\hat{\beta}1$ (se)	0.530 (0.018)	0.528 (0.019)	0.525 (0.019)	0.530 (0.021)	0.537 (0.021)
$\hat{\beta}2$ (se)	0.613 (0.023)	0.612 (0.026)	0.624 (0.028)	0.624 (0.027)	0.659 (0.031)
$\hat{\beta}3$ (se)	-0.517 (0.006)	-0.529 (0.008)	-0.547 (0.009)	-0.529 (0.008)	-0.534 (0.010)
$\hat{\beta}4$ (se)	0.583 (0.007)	0.580 (0.009)	0.598 (0.008)	0.593 (0.009)	0.598 (0.009)
$\hat{\beta}5$ (se)	0.535 (0.006)	0.579 (0.006)	0.596 (0.005)	0.597 (0.006)	0.597 (0.006)
$se(\xi)$	0.552	0.564	0.498	0.599	0.494
N observed	500	460	450	425	400

Table 4: Estimations for the Poisson model with $N = 500$ and $Var(\xi) = 2$ (se = standard error)

Missing (%)	0%	8%	10%	15%	20%
$\hat{\beta}1$ (se)	0.510 (0.009)	0.508 (0.007)	0.511 (0.008)	0.513 (0.008)	0.514 (0.012)
$\hat{\beta}2$ (se)	0.389 (0.015)	0.384 (0.010)	0.385 (0.009)	0.340 (0.010)	0.386 (0.010)
$\hat{\beta}3$ (se)	-0.482 (0.003)	-0.493 (0.003)	-0.495 (0.003)	-0.500 (0.003)	-0.515 (0.003)
$\hat{\beta}4$ (se)	0.511 (0.004)	0.519 (0.003)	0.507 (0.003)	0.534 (0.003)	0.503 (0.003)
$\hat{\beta}5$ (se)	0.511 (0.002)	0.515 (0.002)	0.522 (0.002)	0.523 (0.002)	0.506 (0.002)
$se(\xi)$	1.248	1.232	1.288	1.253	1.398
N observed	500	460	450	425	400

For $N = 100$ and $Var(\xi) = 2$, the estimations of β were less accurate when the number of missing values increased. However, the estimation of $Var(\xi)$ was relatively close to the initial value. For $N = 500$ and $Var(\xi) = 0.5$, the estimations of β were more stable with respect to the proportion p of missing values. Further, $Var(\xi)$ was overpredicted, but this bias seems to be due to the Schall linearization. Finally, for $N = 500$ and $Var(\xi) = 2$, there was clearly a gain of precision for the estimations of β and $Var(\xi)$.

In order to evaluate the efficiency of the algorithm, we considered the case $p = 0$ in Tables 2-4. We observed a bias due to the hight correlation between covariates. This bias was balanced among the components of β. Finally a certain stability through the sample size and the variance of the random effect ξ was observed.

Concerning the MSE, it increased with the number of missing values and figure 1 shows the three MSE associated to the three simulations. It can be observed that the increasing is slow with the percentage of missing value.

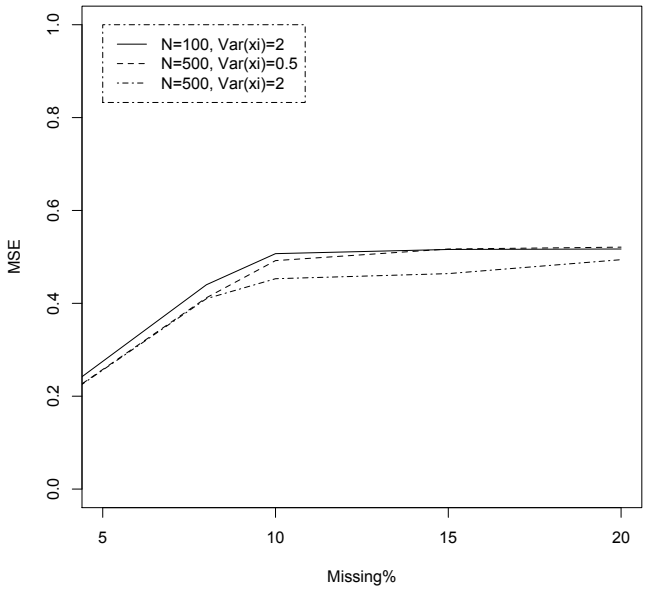

Figure 1: MSE associated to the three simulations with respect to the % of missing data

3.2. The logistic model

Assume that $Y|\xi, X$ follows a logistic distribution

$$\mathbb{E}(Y|\xi, X) = \frac{\exp(\beta X + \xi U)}{1 + \exp(\beta X + \xi U)},$$

with $\beta = (0.5, 0.5, -0.5, 0.5, 0.5)$. The MI-PLS-GLMM algorithm was run with $p \in \{8\%, 10\%, 15\%, 20\%\}$. Only three PLS components were retained for all models and were used to reconstruct the β coefficient. Tables 5-7 contain estimations for the estimators at the risk level $\alpha = 0.05$. The test of significance was based on $B = 200$ bootstrap samples.

Table 5: Estimations for the logistic model with $N = 100$ and $Var(\xi) = 2$ (se = standard error)

Missing (%)	0%	8%	10%	15%	20%
$\hat{\beta}1$ (se)	0.867 (0.233)	0.880 (0.299)	0.821 (0.357)	0.785 (0.635)	0.799 (0.931)
$\hat{\beta}2$ (se)	0.467 (0.330)	0.450 (0.365)	0.490 (0.302)	0.445 (0.325)	0.449 (0.320)
$\hat{\beta}3$ (se)	-0.525 (0.048)	-0.560 (0.060)	-0.549 (0.056)	-0.483 (0.061)	-0.471 (0.058)
$\hat{\beta}4$ (se)	0.185 (0.048)	0.170 (0.061)	0.126 (0.053)	0.200 (0.076)	0.122 (0.080)
$\hat{\beta}5$ (se)	0.206 (0.295)	0.207 (0.187)	0.270 (0.239)	0.294 (0.214)	0.272 (0.219)
$se(\xi)$	1.166	1.234	1.326	1.124	1.121
N observed	100	92	90	85	80

Table 6: Estimations for the logistic model with $N = 500$ and $Var(\xi) = 0.5$ (se = standard error)

Missing (%)	0%	8%	10%	15%	20%
$\hat{\beta}1$ (se)	0.757 (0.096)	0.640 (0.103)	0.702 (0.091)	0.598 (0.108)	0.708 (0.135)
$\hat{\beta}2$ (se)	0.268 (0.162)	0.170 (0.157)	0.198 (0.175)	0.165 (0.210)	0.199 (0.190)
$\hat{\beta}3$ (se)	-0.416 (0.020)	-0.489 (0.024)	-0.419 (0.022)	-0.458 (0.028)	-0.468 (0.048)
$\hat{\beta}4$ (se)	0.293 (0.120)	0.268 (0.144)	0.233 (0.119)	0.205 (0.166)	0.184 (0.187)
$\hat{\beta}5$ (se)	0.774 (0.293)	0.739 (0.468)	0.721 (0.301)	0.608 (0.316)	0.593 (0.310)
$se(\xi)$	0.747	0.702	0.702	0.753	0.764
N observed	500	460	450	425	400

Table 7: Estimations for the logistic model with $N = 500$ and $Var(\xi) = 2$ (se = standard error)

Missing (%)	0%	8%	10%	15%	20%
$\hat{\beta}1$ (se)	0.521 (0.109)	0.610 (0.115)	0.501 (0.109)	0.597 (0.134)	0.476 (0.141)
$\hat{\beta}2$ (se)	0.374 (0.152)	0.391 (0.158)	0.365 (0.157)	0.239 (0.191)	0.239 (0.165)
$\hat{\beta}3$ (se)	-0.456 (0.055)	-0.432 (0.052)	-0.523 (0.031)	-0.547 (0.038)	-0.440 (0.073)
$\hat{\beta}4$ (se)	0.456 (0.092)	0.409 (0.105)	0.549 (0.096)	0.437 (0.111)	0.534 (0.115)
$\hat{\beta}5$ (se)	0.509 (0.179)	0.579 (0.187)	0.434 (0.214)	0.595 (0.303)	0.552 (0.223)
$se(\xi)$	1.349	1.045	1.303	1.544	1.578
N observed	500	460	450	425	400

The same conclusions than with the Poisson model can be drawn and we omit it for brievity. We just illustrate the increasing of the MSE in Figure 2.

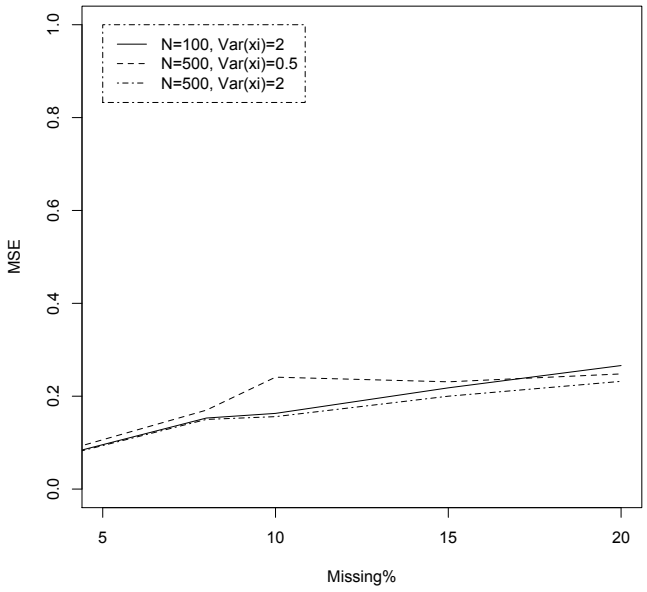

Figure 2: MSE associated to the three simulations with respect to the % of missing data

4. Real dataset

4.1. The plant vegetation example

We considered the data presented in Zuur et al. (2009): they consist of grassland data from a monitoring program from communities in Montana, USA. The data were measured in eight different transects and each transect was measured repeatdly over time with time intervals of about four to ten years. The aim of the study was to determinate whether the biodiversity of the bunchgrass communities changes over time and if they did, whether the changes in biodiversity relate to specific environmental factors. In order to quantify biodiversity, species richness were used. Richness is defined as the different number of species per site, and it is assumed that this response variable follows a Poisson distribution. Since the data are longitudinal, we take into account the time and we consider that the site is a random effect with 8 levels. Note that in the original analysis of Zuur et al. (2009) the longitudinal aspect was ignored. The explanatory variables are rock content,

litter content, bare soil, rainfall in the fall, maximum temperature in the spring. They are correlated and Table 8 shows the correlation values of the fixed effects and the response variable.

Table 8: Correlation between the response variable and the fixed effects

	Richness	Rock	Litter	Baresoil	Rainfall	Temperature
Rock	-0.3	1				
Litter	0.09	-0.7	1			
Baresoil	-0.6	0.04	-0.2	1		
Rainfall	0.2	0.07	-0.2	0.1	1	
Temperature	-0.6	-0.02	0.2	0.4	-0.2	1

The response variable and some of the explanatory variables have 9.7% of missing values. The missing mechanism is considered as MAR. We apply the algorithm MI-PLS-GLMM on the dataset.

4.2. Results

We retained two PLS components based on the Q_h^2 criterion. In Zuur et al. (2009), two components were also obtained in their PLS approach with a Poisson model. The parameters β were reconstructed from the PLS regression. Table 9 presents their estimates for the fixed effects, before and after imputation. Based on the bootstrap procedure, they are all significant at the risk level $\alpha = 0.05$.

Table 9: Estimates associated to the fixed effects and to the random effect (se = standard error).

Missing	Before Imputation	After Imputation
Rock (se)	-0.004 (0.004)	-0.003 (0.004)
Litter (se)	0.005 (0.007)	0.006 (0.007)
Baresoil (se)	-0.018 (0.006)	-0.010 (0.006)
Rainfall (se)	0.019 (0.003)	0.006 (0.004)
Temperature (se)	0.104 (0.030)	0.107 (0.038)
Time (se)	0.002 (0.002)	0.028 (0.017)
$se(\xi)$	0.606	0.829

4.3. Interpretation

From Table 9, we distinguished two groups of covariates. A first group with negative coefficients including, for instance, bare soil that reduces the mean of the number of species, by 1.78% per unit more. In contrast, the second group increases the number of species. For instance for an additional unit of temperature, the mean of the number of species increases by 11.29%.

The prediction of the random effect leads to detect two other groups of sites, as shown in Table 10.

Table 10: Prediction of the random effect

	Site1	Site2	Site3	Site4	Site5	Site6	Site7	Site8
$\hat{\xi}$	0.047	-0.077	0.050	0.004	0.051	-0.008	-0.047	0.096

A first type of site, formed by sites 1, 3, 4, 5 and 8, seems to be more propitious to the biodiversity a priori. In contrast, the second type formed the other sites seems to be less propitious to the diversity of species.

Finally, Figure 4.3 shows the boxplots before and after imputation. They were very close, indicating a stability of the original distribution shape after imputation.

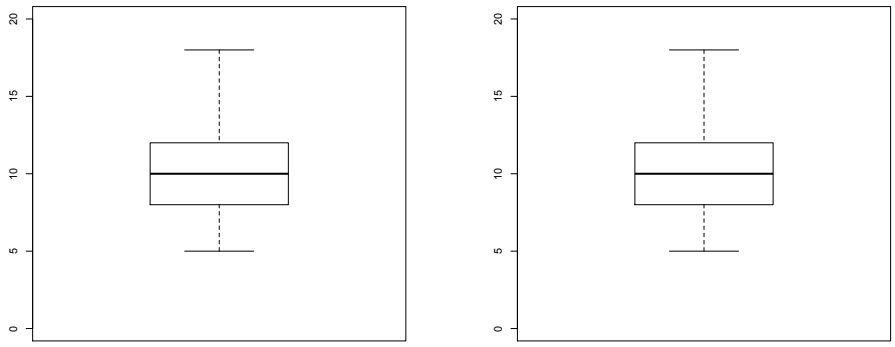

Figure 3: Boxplot of the distribution of Richness: before imputation (left) and after imputation (right)

5. Conclusion

The algorithm MI-PLS-GLMM was proposed to deal with the problem of missing data in a generalized linear mixed model when covariates are correlated. It combines the multiple imputation theory adapted to the generalized linear mixed models with the PLS method. It is based on the Schall method which allows to reduce the problem to a linear mixed model. It is also an adaptation of the MI-PLS initiated by Bastien (2008) and the MI-PLS-LMM algorithm proposed by Guyon and Pommeret (2011) since it is dedicated to the problem of missing data in the presence of both collinearity and random effect.

Simulation studies were carried out which suggest that the proposed method works well for mixed Poisson and mixed logistic models. Thus, MI-PLS-GLMM provided good estimates of the parameters and kept the distribution shape of the original data before imputation. It was also shown that the MSE increase slowly with the percentage of missing values. The limit of the method seemed to be inherent to the Schall linearization which may lead to overestimations of the variance of the random effect.

The application of this method to a real dataset shown it is easily relevant in the case of longitudinal data.

References

Aguilera, A. M., Escabias, M., Valderrama, M. J., 2006. Using principal components for estimating logistic regression with high-dimensional multicollinear data. Computational Statistics & Data Analysis 50 (8), 1905–1924.

Bastien, P., 2008. Régression pls et données censurées. Ph.D. thesis, Conservatoire National des Arts et Métiers, Paris.

Bastien, P., Vinzi, V. E., Tenenhaus, M., 2005. PLS generalised linear regression. Computational Statistics & Data Analysis 48 (1), 17–46.

Guyon, E., Pommeret, D., 2011. Imputation by PLS regression for linear mixed models. Journal de la Société Française de Statistique 152 (4), 30–46.

Henderson, C., Kempthorne, O., Searle, S., von Krosigk, C., 1959. The estimation of environmental and genetic trends from records subject to culling. Biometrics 15, 192–218.

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

Little, R., Rubin, D., 1987. Statistical analysis with missing data. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

Little, R. J. A., 1995. Modeling the drop-out mechanism in repeated-measures studies. Journal of the American Statistical Association 90 (431), 1112– 1121.

Rubin, D., 1987. Multiple Imputation for Non-response in Survey. J. Wiley and Sons, New York. Wiley.

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

Schafer, J., Yucel, R., 1998. Fitting multivariate linear mixed models with incomplete data. In: Proceedings of the Statistical Computing Section of the American Statistical Association. Vol. 8. pp. 177–182.

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

Ten Have, T., Pulkstenis, E., K. A., Landis, J., 1998. Mixed effects logistics regression models fro longitudinal binary response data with droupout. Biometrics 54, 367–383.

Tenenhaus, M., 1998. La régression PLS: théorie et pratique.

Wu, K., Lang, W., 2006. Generalized linear mixed models with informative dropouts and missing covariates. Metrika 66, 1–18.

Wu, L., 2004. Exact and approximate inferences for nonlinear mixed-effects models with missing covariates. Journal of the American Statistical Association 99, 700–709.

Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., Smith, G. M., 2009. Mixed effects models and extensions in ecology with r. Statistics 32 (i), 209–243.

Appendix

1. Schall linearization

The Schall algorithm can be summarized as follow: at each iteration t, we consider $\beta^{[t]}$, $\xi^{[t]}$ and $\sigma^{2[t]} = (\sigma_1^{2[t]}, ..., \sigma_K^{2[t]})$.

1 Calculation of $Z^{[t]} = X\beta^{[t]} + U\xi^{[t]} + (y - \mu_\xi^{[t]})g'(\mu_\xi^{[t]})$.

2 Calculation of $W_\xi^{[t]}$ and $\Gamma_\xi^{[t]}$.

3 Determination of $\beta^{[t+1]}$ and $\xi^{[t+1]}$ as solution of the Henderson's system.

4 Calculation of $\sigma^{2[t+1]}$.

For the fourth step, with the maximum likelihood,

$$\sigma_j^{2[t+1]} = \frac{\xi_j'^{[t+1]} A_j^{-1} \xi_j^{[t+1]}}{q_j - \frac{tr(A_j^{-1} C_{jj}^{*[t]})}{\sigma_j^{2[t]}}},$$

where C_{jj}^* is the $j \times j$ submatrix of $C^* = (U'W_\xi^{-1}U + D^{-1})^{-1}$.

2. PLS linear regression algorithm

Step1. Computation of the PLS regression components

Computation of the first PLS regression component t_1:

1. Computation of the coefficients a_{1j} for each simple linear regression of $\widetilde{Z}$ on x_j, $j = 1, ..., p$.
2. Normalization of the column vector a_1 made by a_{1j}'s: $w_1 = \frac{a_1}{||a_1||}$.
3. Computation of the PLS regression component as $t_1 = Xw_1$.

Computation of the kth PLS regression component t_k:

1. Computation of the residual $x_{k-1,1}, ..., x_{k-1,p}$ from the multiple regression of x_j, $j = 1, ..., p$ on $t_1, ..., t_{k-1}$. Let $X_{k-1} = [x_{k-1,1}, ..., x_{k-1,p}]$.
2. Computation of the coefficients a_{kj} of $x_{k-1,j}$ in the linear regression of $\widetilde{Z}$ on $t_1, ..., t_{k-1}$ and each $x_{k-1,j}$, $j = 1, ..., p$.
3. Normalization of the column vector a_k made by a_{kj}'s: $w_k = \frac{a_k}{||a_k||}$.
4. Computation of the kth PLS regression component as $t_k = X_{k-1}w_k$.
5. Expression of the component t_k in terms of X as $t_k = Xw_k^*$.

Step2. Linear regression of $\widetilde{Z}$ on the k retained PLS regression components

3. Number of PLS components

Consider the regression model of z on the h PLS components:

$$z = \underbrace{c_1 t_1 + ... + c_h t_h}_{\hat{z}_h} + z_h. \tag{4}$$

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:

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

where $\hat{z}_{h(-i)}$ is the prediction of z_i obtained by (4) without the observation i. For $h = 1$,

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

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

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

All variables $Y, X_1, ..., X_j, ..., X_p$ are assumed to be centered. The PLS regression model with h components is written as

$$Y = \sum_{H=1}^{h} c_H (\sum_{j=1}^{p} w_{Hj}^* X_j) + residual, \quad (5)$$

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 (5) are to be estimated.

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

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