# Logistic Regression with Missing Covariates – Parameter Estimation, Model Selection and Prediction within a Joint-Modeling Framework

Wei Jiang[a,*], Julie Josse[a], Marc Lavielle[a], TraumaBase Group[b]

[a]*Inria XPOP and CMAP, École Polytechnique, France*
[b]*Hôpital Beaujon, APHP, France*



**Abstract**

Logistic regression is a common classification method in supervised learning. Surprisingly, there are very few solutions for performing logistic regression with missing values in the covariates. We suggest a complete approach based on a stochastic approximation version of the EM algorithm to do statistical inference with missing values including the estimation of the parameters and their variance, derivation of confidence intervals and a model selection procedure. We also tackle the problem of prediction for new observations (on a test set) with missing covariate data. The methodology is computationally efficient, and its good coverage and variable selection properties are demonstrated in a simulation study where we contrast its performances to other methods. For instance, the popular approach of multiple imputation by chained equations can lead to estimates that exhibit meaningfully greater biases than the proposed approach. We then illustrate the method on a dataset of severely traumatized patients from Paris hospitals to predict the occurrence of hemorrhagic shock, a leading cause of early preventable death in severe trauma cases. The aim is to consolidate the current red flag procedure, a binary alert identifying patients with a high risk of severe hemorrhage. The methodology is implemented in the R package misaem.

*Keywords:* incomplete data, observed likelihood, Metropolis-Hastings, public health

## 1. Introduction

Missing data exist in almost all areas of empirical research. There are various reasons why missing data may occur, including survey non-response, unavailability of measurements, and lost data. One popular approach to handle missing values consists in modifying an estimation process so that it can be applied to incomplete data. For example, one can use the EM algorithm [1] to obtain the maximum likelihood estimate (MLE) despite missing values, accompanied by a supplemented EM algorithm (SEM) [2] or Louis' formula [3] for

*Corresponding author
*Email address:* wei.jiang@polytechnique.edu (Wei Jiang)

their variance. This strategy is valid under missing at random (MAR) mechanisms [4, 5], in which the missingness of data is independent of the missing values, given the observed data. Even though this approach is perfectly suited to specific inference problems with missing values, there are few solutions or implementations available, even for simple models such as logistic regression, the focus of this paper.

One explanation is that the expectation step of the EM algorithm often involves unfeasible computations. In the framework of generalized linear models Ibrahim et al. [6, 7], suggested to use a Monte Carlo EM (MCEM) algorithm [8, 9], replacing the integral by its empirical sum using Monte Carlo sampling. Ibrahim et al. [6] also estimated the variance using a Monte Carlo version of Louis' formula by Gibbs sampling with an adaptive rejection sampling scheme [10]. However, their approach is computationally expensive and they considered an implementation only for monotone patterns of missing values, or for missing values only in two variables in a dataset.

In this paper, we develop a stochastic approximation version of the EM algorithm (SAEM) [11], based on Metropolis-Hastings sampling, to perform statistical inference for logistic regression with incomplete data, where the missing data can be anywhere in the covariates. SAEM uses a stochastic approximation procedure to estimate the conditional expectation of the complete-data likelihood, instead of generating a large number of Monte Carlo samples which lead to an undeniable computational advantage over MCEM as illustrated in the simulation studies. In addition, it allows for model selection using criterion based on a penalized version of the observed-data likelihood. This latter characteristic is very useful in practice, as few methods are available to select a model when there are missing values. For example, Claeskens and Consentino [12], Consentino and Claeskens [13] suggested an approximation of AIC, while Jiang et al. [14] defined generalized information criteria and in the framework of imputation Liu et al. [15] proposed to combine penalized regression techniques with multiple imputation and stability selection. Besides aiming at maximizing the MLE for observed data, Chow [16], Yuen Fung and A. Wrobel [17] studied the linear discriminant function for logistic regression, using pairs of observed values in columns to calculate the covariance matrix. Note that another solution is to use Laplace approximation to compute integrals, however, this approximation linearizes the likelihood function by differentiation whereas SAEM performs exactly the inference.

This paper proceeds as follows: In Section 2 we describe the motivation for this work, the TraumaBase[1] project based on a French multicenter prospective Trauma Registry. Section 3 presents the assumptions and notation used throughout this paper. In Section 4, we derive an algorithm SAEM to obtain the maximum likelihood estimate of parameters in an logistic regression model for continuous covariate data, under the MAR mechanism and a general pattern of missing data. Following the estimation of parameters, we present how to estimate the Fisher information matrix using a Monte Carlo version of Louis' formula. Section 5 describes the model selection scheme based on a Bayesian information criterion (BIC) with missing values. In addition, we propose an approach to perform prediction for a new iobservation with missing values. Section 6 presents a simulation study where the

[1] http://www.traumabase.eu/

proposed approach is compared to alternative methods such as multiple imputation [18], which may suffer from greater biases than the proposed approach and under-coverage. In Section 7, we apply the newly developed approach to predict the occurrence of hemorrhagic shock in patients with blunt trauma to the TraumaBase dataset, where it is crucial to efficiently manage missing data because the percentage of missing data varies from 0 to 60% depending on the variables. Compared to the predictions made by emergency doctors, the results are improved with SAEM. Finally, Section 8 concludes this work and provides a discussion.

Our contribution is to provide users the ability to perform logistic regression with missing values within a joint-modeling methodological framework that combines computational efficiency and a sound theoretical foundation. The methodology presented in this article is implemented as an R [19] package *misaem* [20], available in CRAN. The code to reproduce all the experiment is also provided in GitHub [21].

## 2. Medical emergency

Our work is motivated by a collaboration with the TraumaBase group at APHP (Public Assistance - Hospitals of Paris), which is dedicated to the management of severely traumatized patients.

Major trauma refers to injuries that endanger a person's life or functional integrity. The WHO has recently shown that major trauma - road accidents, interpersonal violence, falls, etc. - are a worldwide public health challenge and a major source of mortality (first cause in the age group 16-45) and disability (2nd cause) in the world [22]. The two leading causes of death are hemorrhagic shock and traumatic brain injury.

The path of a traumatized patient takes place in several stages: from the accident site where he is taken care of by the ambulance to the transfer to intensive care unit for immediate interventions and finally comprehensive care at the hospital. Using a pre hospital patient's records, we aim to establish models to predict the risk of severe hemorrhage to prepare an appropriate response upon arrival at the trauma center; e.g., massive transfusion protocol and/or immediate haemostatic procedures.

Due to the highly stressful and multi-player environments involved, evidence suggests that patient management – even in mature trauma systems – often exceeds acceptable time frames [23]. In addition, discrepancies may be observed between the diagnoses made by emergency doctors in the ambulance, and those made when the patient arrives at the trauma center [24]. These discrepancies can result in poor outcomes such as inadequate hemorrhage control or delayed transfusion.

To improve decision-making and patient care, 15 French trauma centers have collaborated to collect detailed high-quality clinical data from the accident scene, to the hospital. The resulting database, TraumaBase, is a multicenter prospective trauma registry that is continually updated and now has data from more than 7,000 trauma cases. The granularity of collected data (with more than 250 variables) makes this dataset unique in Europe. However, the data from multiple sources, are highly heterogeneous, and are often missing, which makes modeling challenging.

In this paper, we focus on performing logistic regression with missing values to help propose an innovative response to the public health challenge of major trauma.

## 3. Assumptions and notation

Let $(y, x)$ be the observed data with $y = (y_i, 1 \leq i \leq n)$ an $n$-vector of binary responses coded with $\{0, 1\}$ and $x = (x_{ij}, 1 \leq i \leq n, 1 \leq j \leq p)$ a $n \times p$ matrix of covariates, where $x_{ij}$ takes its values in $\mathbb{R}$. The logistic regression model for binary classification can be written as:

$$\mathbb{P}\left(y_i = 1 | x_i; \beta\right) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}, \quad i = 1, \ldots, n, \tag{1}$$

where $x_{i1}, \ldots, x_{ip}$ are the covariates for individual $i$ and $\beta_0, \beta_1, \ldots, \beta_p$ unknown parameters. We adopt a probabilistic framework by assuming that $x_i = (x_{i1}, \ldots, x_{ip})$ is normally distributed:

$$x_i \underset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Sigma), \quad i = 1, \cdots, n.$$

Let $\theta = (\mu, \Sigma, \beta)$ be the set of parameters of the model. Then, the log-likelihood for the complete data can be written as:

$$\begin{aligned} \mathcal{LL}(\theta; x, y) &= \sum_{i=1}^{n} \mathcal{LL}(\theta; x_i, y_i) \\ &= \sum_{i=1}^{n} \Big( \log(\mathtt{p}(y_i | x_i; \beta)) + \log(\mathtt{p}(x_i; \mu, \Sigma)) \Big). \end{aligned}$$

Our main goal is to estimate the vector of parameters $\beta = (\beta_j, 0 \leq j \leq p)$ when missing values exist in the design matrix, i.e., in the matrix $x$. For each individual $i$, we note $x_{i,\text{obs}}$ the elements of $x_i$ that are observed and $x_{i,\text{mis}}$ those that are missing. We also decompose the matrix of covariates as $x = (x_{\text{obs}}, x_{\text{mis}})$, keeping in mind that the missing elements may differ from one individual to another.

For each individual $i$, we define the missing data indicator vector $M_i = (M_{ij}, 1 \leq j \leq p)$, with $M_{ij} = 1$ if $x_{ij}$ is missing and $M_{ij} = 0$ otherwise. The matrix $M = (M_i, 1 \leq i \leq n)$ then defines the missing data pattern. The missing data mechanism is characterized by the conditional distribution of $M$ given $x$ and $y$, with parameter $\phi$, i.e., $\mathtt{p}(M_i | x_i, y_i, \phi)$. Throughout this paper, we assume a missing at random (MAR) mechanism which implies that the missing values mechanism can therefore be ignored [4] and the maximum likelihood estimate of $\theta$ can be obtained by maximizing $\mathcal{LL}(\theta; y, x_{\text{obs}})$. A reminder of these concepts is given in *Appendix A.1*.

## 4. Parameter estimation by SAEM

### *4.1. The EM and MCEM algorithms*

We aim to estimate the parameter $\theta$ of the logistic regression model by maximizing the observed log-likelihood $\mathcal{LL}(\theta; x_{\text{obs}}, y)$. Let us start with the classical EM formulation

for obtaining the maximum likelihood estimator from incomplete data. Given some initial value $\theta_0$, iteration $k$ updates $\theta_{k-1}$ to $\theta_k$ with the following two steps:

- **E-step:** Evaluate the quantity

$$\begin{aligned} Q_k(\theta) &= \mathbb{E}[\mathcal{LL}(\theta; x, y)|x_{\text{obs}}, y; \theta_{k-1}] \\ &= \int \mathcal{LL}(\theta; x, y)\mathtt{p}(x_{\text{mis}}|x_{\text{obs}}, y; \theta_{k-1})dx_{\text{mis}}. \end{aligned} \tag{2}$$

- **M-step:** Update the estimation of $\theta$: $\theta_k = \arg\max_\theta Q_k(\theta)$.

Since the expectation (2) in the E-step for the logistic regression model has no explicit expression, MCEM [8, 6] can be used. The E-step of MCEM generates several samples of missing data from the target distribution $\mathtt{p}(x_{\text{mis}}|x_{\text{obs}}, y; \theta_{k-1})$ and replaces the expectation of the complete log-likelihood by an empirical mean. However, an accurate Monte Carlo approximation of the E-step may require a significant computational effort, as illustrated in the Section 6.

*4.2. The SAEM algorithm*

To achieve improved computational efficiency, we suggest deriving a SAEM algorithm [11] which replaces the E-step (2) by a stochastic approximation. Starting from an initial guess $\theta_0$, the $k$th iteration consists of three steps:

- **Simulation:** For $i = 1, 2, \cdots, n$, draw $x_{i,\text{mis}}^{(k)}$ from

$$\mathtt{p}(x_{i,\text{mis}}|x_{i,\text{obs}}, y_i; \theta_{k-1}). \tag{3}$$

- **Stochastic approximation:** Update the function $Q$ according to

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left( \mathcal{LL}(\theta; x_{\text{obs}}, x_{\text{mis}}^{(k)}, y) - Q_{k-1}(\theta) \right), \tag{4}$$

  where $(\gamma_k)$ is a non-increasing sequence of positive number.

- **Maximization:** Update the estimation of $\theta$:

$$\theta_k = \arg\max_\theta Q_k(\theta).$$

The choice of the sequence $(\gamma_k)$ in (4) is important for ensuring the almost sure convergence of SAEM to a maximum of the observed likelihood [25]. We will see in Section 6 that, in our case, very good convergence is obtained using $\gamma_k = 1$ during the first iterations, followed by a sequence that decreases as $1/k$.

*4.3. Metropolis-Hastings sampling*

In the logistic regression case, the unobserved data cannot in general be drawn exactly from the conditional distribution (3), which depends on an integral that is not calculable in closed form. One solution is to use a Metropolis-Hastings (MH) algorithm, which consists of constructing a Markov chain that has the target distribution as its stationary distribution. The states of the chain after $S$ iterations are then used as a sample from the target distribution. To define a proposal distribution for MH algorithm, we observe that the target distribution (3) can be factorized as follows:

$$\mathtt{p}(x_{i,\text{mis}}|x_{i,\text{obs}}, y_i; \theta) \propto \mathtt{p}(y_i|x_i; \beta)\mathtt{p}(x_{i,\text{mis}}|x_{i,\text{obs}}; \mu, \Sigma).$$

We select the proposal distribution as the second term $\mathtt{p}(x_{i,\text{mis}}|x_{i,\text{obs}}, \mu, \Sigma)$, which is normally distributed:

$$x_{i,\text{mis}}|x_{i,\text{obs}} \sim \mathcal{N}_p(\mu_i, \Sigma_i), \tag{5}$$

where

$$\begin{aligned}\mu_i &= \mu_{i,\text{mis}} + \Sigma_{i,\text{mis,obs}}\Sigma^{-1}_{i,\text{obs,obs}}(x_{i,\text{obs}} - \mu_{i,\text{obs}}),\\ \Sigma_i &= \Sigma_{i,\text{mis,mis}} - \Sigma_{i,\text{mis,obs}}\Sigma^{-1}_{i,\text{obs,obs}}\Sigma_{i,\text{obs,mis}},\end{aligned}$$

with $\mu_{i,\text{mis}}$ (resp. $\mu_{i,\text{obs}}$) the missing (resp. observed) elements of $\mu$ for individual $i$. The covariance matrix $\Sigma$ is decomposed in the same way. The MH algorithm is described further in Appendix A.2.

*4.4. Observed Fisher information*

After computing the MLE $\hat{\theta}_{\text{ML}}$ with SAEM, we estimate its variance. To do so, we can use the observed Fisher information matrix (FIM): $\mathcal{I}(\theta) = -\frac{\partial^2 \mathcal{LL}(\theta; x_{\text{obs}}, y)}{\partial\theta\partial\theta^T}$. According to Louis' formula [3], we have:

$$\begin{aligned}\mathcal{I}(\theta) = &- \mathbb{E}\left(\frac{\partial^2 \mathcal{LL}(\theta; x, y)}{\partial\theta\partial\theta^T}\Big|x_{\text{obs}}, y; \theta\right)\\ &- \mathbb{E}\left(\frac{\partial \mathcal{LL}(\theta; x, y)}{\partial\theta}\frac{\partial \mathcal{LL}(\theta; x, y)^T}{\partial\theta}\Big|x_{\text{obs}}, y; \theta\right)\\ &+ \mathbb{E}\left(\frac{\partial \mathcal{LL}(\theta; x, y)}{\partial\theta}|x_{\text{obs}}, y; \theta\right)\mathbb{E}\left(\frac{\partial \mathcal{LL}(\theta; x, y)}{\partial\theta}|x_{\text{obs}}, y; \theta\right)^T.\end{aligned}$$

The observed FIM can therefore be expressed in terms of conditional expectations, which can also be approximated using a Monte Carlo procedure. More precisely, given $S$ samples $(x^{(s)}_{i,\text{mis}}, 1 \leq i \leq n, 1 \leq s \leq S)$ of the missing data drawn from the conditional distribution

(3), the observed FIM can be estimated as $\hat{\mathcal{I}}_S(\hat{\theta}) = \sum_{i=1}^{n} -(D_i + G_i - \Delta_i\Delta_i^T)$, where

$$\Delta_i = \frac{1}{S}\sum_{s=1}^{S} \frac{\partial \mathcal{LL}(\hat{\theta}; x_{i,\text{mis}}^{(s)}, x_{i,\text{obs}}, y_i)}{\partial \theta},$$

$$D_i = \frac{1}{S}\sum_{s=1}^{S} \frac{\partial^2 \mathcal{LL}(\hat{\theta}; x_{i,\text{mis}}^{(s)}, x_{i,\text{obs}}, y_i)}{\partial \theta \partial \theta^T},$$

$$G_i = \frac{1}{S}\sum_{s=1}^{S} \left(\frac{\partial \mathcal{LL}(\hat{\theta}; x_{i,\text{mis}}^{(s)}, x_{i,\text{obs}}, y_i)}{\partial \theta}\right)\left(\frac{\partial \mathcal{LL}(\hat{\theta}; x_{i,\text{mis}}^{(s)}, x_{i,\text{obs}}, y_i)}{\partial \theta}\right)^T.$$

Here, the gradient and the Hessian matrix can be computed in closed form. The procedure for calculating the observed information matrix is described in Appendix A.3.

## 5. Model selection and prediction

### *5.1. Information criteria*

In order to compare different possible covariate models, we can consider penalized likelihood criteria such as the Bayesian information criterion (BIC). For a given model $\mathcal{M}$ and an estimated parameter $\hat{\theta}_{\mathcal{M}}$, BIC is defined as:

$$\text{BIC}(\mathcal{M}) = -2\mathcal{LL}(\hat{\theta}_{\mathcal{M}}; x_{\text{obs}}, y) + \log(n) d(\mathcal{M}),$$

where $d(\mathcal{M})$ is the number of estimated parameters in a model $\mathcal{M}$. The distribution of the complete set of covariates $(x_{ij}, 1 \leq i \leq n, 1 \leq j \leq p)$ does not depend on the regression model used for modeling the binary outcomes $(y_i, 1 \leq i \leq n)$: we assume the same normal distribution $\mathcal{N}_p(\mu, \Sigma)$ for all regression models. Thus, the difference between models between the number $d(\mathcal{M})$ of estimated parameters is equivalent to the difference between the number of non-zero coefficients in $\beta_{\mathcal{M}}$. Note that, contrary to the suggested approach, the existing methods Claeskens and Consentino [12], Consentino and Claeskens [13] use an approximation of the Akaike information criterion (AIC) without estimating the observed likelihood.

### *5.2. Observed log-likelihood*

For a given model and parameter $\theta$, the observed log-likelihood is, by definition:

$$\mathcal{LL}(\theta; x_{\text{obs}}, y) = \sum_{i=1}^{n} \log\left(\mathtt{p}(y_i, x_{i,\text{obs}}; \theta)\right).$$

With missing data, the density $\mathtt{p}(y_i, x_{i,\text{obs}}; \theta)$ cannot in general be computed in closed form. We suggest to approximate it using an importance sampling Monte Carlo approach. Let $g_i$

be the density function of the normal distribution defined in (5). Then,

$$\begin{aligned}
\mathtt{p}(y_i, x_{i,\text{obs}}; \theta) &= \int \mathtt{p}(y_i, x_{i,\text{obs}} | x_{i,\text{mis}}; \theta) \mathtt{p}(x_{i,\text{mis}}; \theta) dx_{i,\text{mis}} \\
&= \int \mathtt{p}(y_i, x_{i,\text{obs}} | x_{i,\text{mis}}; \theta) \frac{\mathtt{p}(x_{i,\text{mis}}; \theta)}{g_i(x_{i,\text{mis}})} g_i(x_{i,\text{mis}}) dx_{i,\text{mis}} \\
&= \mathbb{E}_{g_i} \left( \mathtt{p}(y_i, x_{i,\text{obs}} | x_{i,\text{mis}}; \theta) \frac{\mathtt{p}(x_{i,\text{mis}}; \theta)}{g_i(x_{i,\text{mis}})} \right).
\end{aligned}$$

Consequently, if we draw $M$ samples from the proposal distribution (5):

$$x_{i,\text{mis}}^{(s)} \underset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_i, \Sigma_i), \quad m = 1, 2, \cdots, S,$$

we can estimate $\mathtt{p}(y_i, x_{i,\text{obs}}; \theta)$ by:

$$\hat{\mathtt{p}}(y_i, x_{i,\text{obs}}; \theta) = \frac{1}{S} \sum_{m=1}^{S} \mathtt{p}(y_i, x_{i,\text{obs}} | x_{i,\text{mis}}^{(s)}; \theta) \frac{\mathtt{p}(x_{i,\text{mis}}^{(s)}; \theta)}{g_i(x_{i,\text{mis}}^{(s)})},$$

and derive an estimate of the observed log-likelihood $\mathcal{LL}(\theta; x_{\text{obs}}, y)$.

### *5.3. Prediction on test set with missing values*

In supervised learning, after fitting a model using a training set, a natural step is to evaluate the prediction performance, which can be done with a test set. Assuming $x = (x_{\text{obs}}, x_{\text{mis}})$ an observation in the test set, we want to predict the binary response $y$. One important point is that test set also contains missing values, since the training set and the test set have the same distribution (*i.e.*, the distribution of covariates and the distribution of missingness). Therefore, we can't directly apply the fitted model (which uses $p$ coefficients) to predict $y$ from an incomplete observation of the test $x$.

Our framework offers a natural way to tackle this issue by marginalizing over the distribution of missing data given the observed ones. More precisely, with $S$ Monte Carlo samples

$$(x_{\text{mis}}^{(s)}, 1 \leq s \leq S) \sim \mathtt{p}(x_{\text{mis}} | x_{\text{obs}}),$$

we estimate directly the response by maximum a posteriori

$$\begin{aligned}
\hat{y} = \arg\max_y \mathtt{p}(y | x_{\text{obs}}) &= \arg\max_y \int \mathtt{p}(y|x) \mathtt{p}(x_{\text{mis}} | x_{\text{obs}}) dx_{\text{mis}} \\
&= \arg\max_y \mathbb{E}_{\mathtt{p}_{x_{\text{mis}}|x_{\text{obs}}}} \mathtt{p}(y|x) \\
&= \arg\max_y \sum_{s=1}^{S} \mathtt{p}\left(y | x_{\text{obs}}, x_{\text{mis}}^{(s)}\right).
\end{aligned}$$

Note that in the literature there are not many solutions to deal with the missing values in the test set. In Subsection 7.2, we compare the suggested approach to some methods used in practice based on imputation of the test set.

## 6. Simulation study

### *6.1. Simulation settings*

We first generated a design matrix $x$ of size $n = 1000 \times p = 5$ by drawing each observation from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$. Then, we generated the response according to the logistic regression model (1). We considered as the true parameter values: $\beta = (-0.2, 0.5, -0.3, 1, 0, -0.6)$, $\mu = (1, 2, 3, 4, 5)$, $\Sigma = \text{diag}(\sigma)C\text{diag}(\sigma)$, where the $\sigma$ is the vector of standard deviations $\sigma = (1, 2, 3, 4, 5)$, and $C$ the correlation matrix

$$C = \begin{bmatrix} 1 & 0.8 & 0 & 0 & 0 \\ 0.8 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.3 & 0.6 \\ 0 & 0 & 0.3 & 1 & 0.7 \\ 0 & 0 & 0.6 & 0.7 & 1 \end{bmatrix}. \tag{6}$$

Before generating missing values, we performed classical logistic regression on the complete dataset, the results (ROC curve) are provided in Appendix A.4. Then we randomly introduced 10% missing values in the covariates first with a missing completely at random (MCAR) mechanism where each entry has the same probability to be observed.

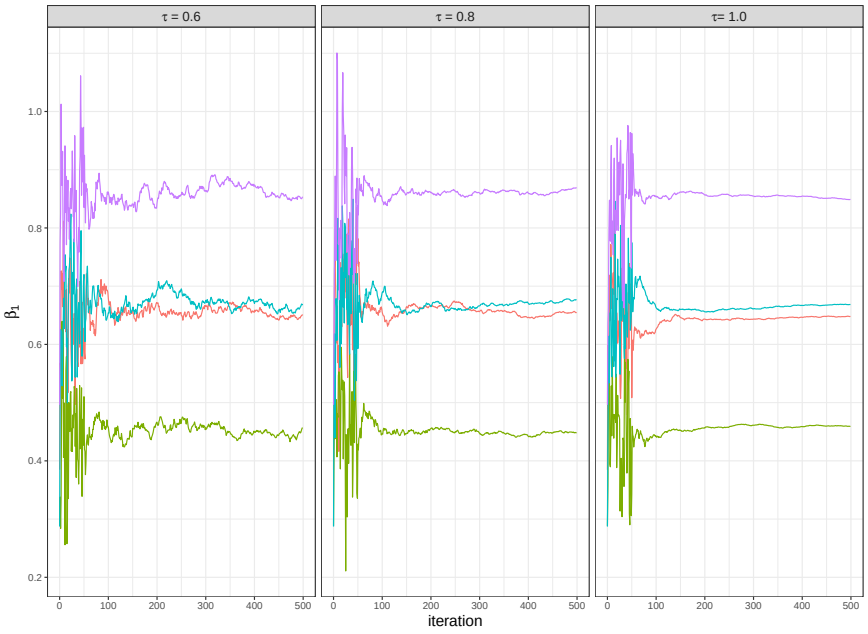



Figure 1: Convergence plots for $\beta_1$ obtained with three different values of $\tau$ (0.6, 0.8, 1.0). Each color represents one simulation. The true value of $\beta_1 = 0.5$.

### *6.2. The behavior of SAEM*

The algorithm was initialized with the parameters obtained after mean imputation, i.e., where missing value in a variable are replaced by the unconditional mean calculated from

the the available cases and the logistic regression is applied on the completed data. For the non-increasing sequence $(\gamma_k)$ in the Stochastic Approximation step of SAEM, we chose $\gamma_k = 1$ during the first $k_1$ iterations in order to converge quickly to a neighborhood of the MLE, and from $k_1$ iterations on, we set $\gamma_k = (k - k_1)^{-\tau}$ to assist the almost sure convergence of SAEM. In order to study the effect of the sequence of stepsizes $(\gamma_k)$, we fixed the value of $k_1 = 50$ and used $\tau = (0.6,\ 0.8,\ 1)$ during the next 450 iterations. Representative plots of the convergence of SAEM for the coefficient $\beta_1$, obtained from four simulated data sets, are shown in Figure 1. For larger $\tau$, SAEM converged faster, and with less fluctuation. For a given simulation, the three sequences of estimates converged to the same solution, but using $\tau = 1$ yielded the fastest convergence, and showed less fluctuation. We therefore use $\tau = 1$ in the following.

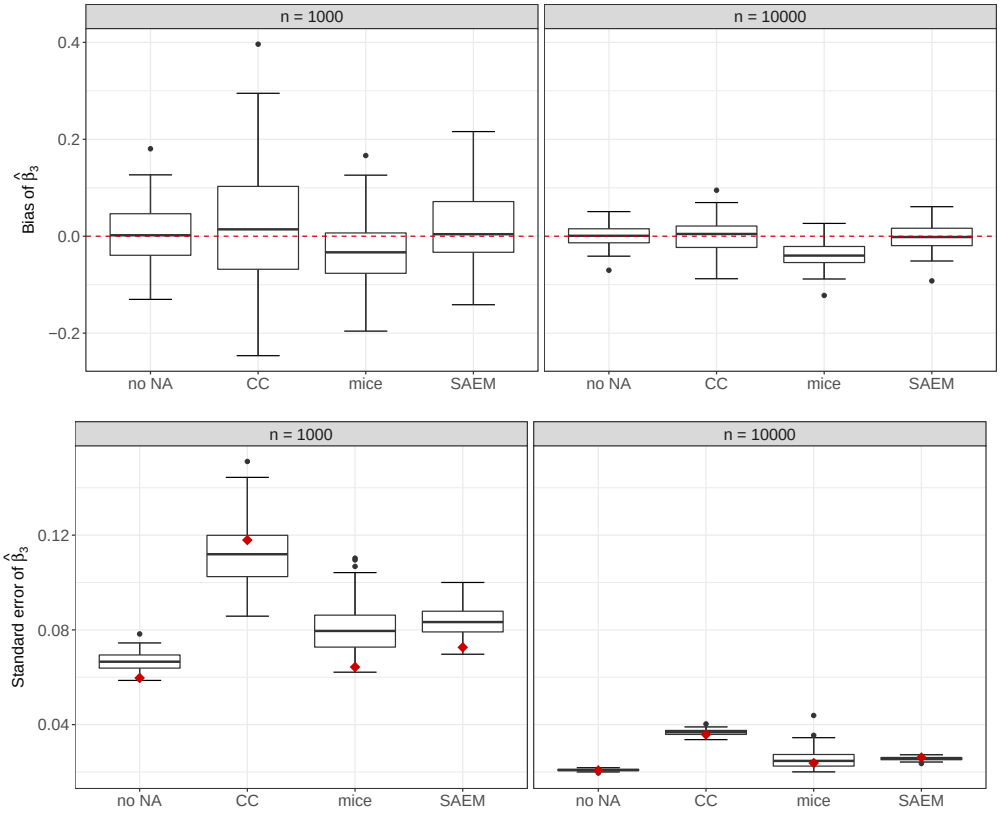



Figure 2: Top: Empirical distribution of bias of $\hat{\beta}_3$. Bottom: Distribution of the estimated standard errors of $\hat{\beta}_3$; for each method, the red point corresponds to the empirical standard deviation of $\hat{\beta}_3$ calculated over the 1000 simulations. Results for 10% MCAR and correlation $C$.

### *6.3. Comparison with other methods*

We ran 1000 simulations and compared SAEM to several other existing methods, initially in terms of estimation errors of the parameters. We mainly focused on *i)* the complete case (CC) method, i.e., all rows containing at least one unobserved data value were removed, *ii)*

multiple imputation by chained equations (mice) with Rubin's combining rules [26]. More precisely, missing values are imputed successively by drawing from conditional distribution. We use the default arguments of the function implemented in R, i.e., conditional models based on regression models are used for quantitative variables and on logistic regression models are used for binary variables and uncertainty of the parameters is reflected within a Bayesian framework. More details are in van Buuren and Groothuis-Oudshoorn [26]. Finally, we used the dataset without missing values (no NA) as a reference, with parameters estimated with the Newton-Raphson algorithm. We varied the number of observations $n = 200, 1000$ and $10\,000$, the missing value mechanism MCAR and MAR, the percentage of missing values 10% and 30%, as well as the correlation structure either using $C$ given by (6) or an orthogonal design.

Figure 2 (top) displays the distribution of the estimates of $\beta_3$, for $n = 1000$ and $n = 10\,000$ under MCAR mechanism and the correlation between covariates is given by (6). Results of simulation with $n = 200$ are presented in supplementary materials [27]. This plot is representative of the results obtained with the other components of $\beta$. As expected, larger samples yielded less variability. Moreover, we observe that in both cases, the estimation obtained by mice could be biased, whereas SAEM provided unbiased estimates with small variances. Figure 2 (bottom) represents the empirical distribution of the estimated standard error of $\hat{\beta}_3$. For SAEM it was calculated using the observed Fisher information as described in Section 4.4. With a larger $n$, not only the estimated standard errors, but also variance of estimation, clearly decreased for all of the methods. In the case where $n = 1000$, SAEM and mice slightly overestimated the standard error, while CC underestimated it, on average. Globally, SAEM led to the best result, since compared with its competitor mice, it had a similar estimation of the standard error on average, but with much less variance.

Table 1: Coverage (%) for $n = 10\,000$, correlation $C$ and 10% MCAR, calculated over 1000 simulations. Bold indicates under coverage. Inside the parentheses is the average length of corresponding confidence interval over 1000 simulations (multiplied by 100).

| parameter | no NA | CC | mice | SAEM |
|---|---|---|---|---|
| $\beta_0$ | 95.2 (21.36) | 94.4 (27.82) | 95.2 (22.70) | 94.9 (22.48) |
| $\beta_1$ | 96.0 (18.92) | 94.7 (24.65) | 93.9 (21.77) | 95.1 (21.51) |
| $\beta_2$ | 95.5 (9.53) | 94.6 (12.41) | 94.0 (10.97) | 94.3 (10.83) |
| $\beta_3$ | 94.9 (8.17) | 94.3 (10.66) | **86.5** (9.03) | 94.7 (9.03) |
| $\beta_4$ | 94.6 (4.00) | 94.2 (5.21) | 96.2 (4.49) | 95.4 (4.42) |
| $\beta_5$ | 95.9 (5.52) | 94.4 (7.19) | **89.6** (6.20) | 94.7 (6.17) |

Table 1 shows the coverage of the confidence interval for all parameters and inside the parentheses is the average length of corresponding confidence interval. We had expected coverage at the nominal 95% level. The simulation margin of error corresponding to coverage results is 1.35%. SAEM reached from 94.3% to 95.4% coverage, while mice struggled for certain parameters: the coverage rates for a few estimates are 89.6% or 86.5%, which are significantly below the nominal level. Even though CC showed reasonable results in terms of coverage, the width of its confidence interval was still too large. Simulation with smaller

sample size had the same results, for example, coverages for $n = 200$ are presented in supplementary materials [27].

Table 2: Comparison of execution time between no NA, MCEM, mice, and SAEM with correlation $C$ and 10% MCAR, for $n = 200$ or $n = 1000$, calculated over 1000 simulations.

| Execution time (seconds) for one simulation | no NA | MCEM | mice | SAEM |
|---|---|---|---|---|
| $\boldsymbol{n = 1000}$ | | | | |
| min | $2.87 \times 10^{-3}$ | 492 | 0.64 | 9.96 |
| mean | $4.65 \times 10^{-3}$ | 773 | 0.70 | 13.50 |
| max | $43.50 \times 10^{-3}$ | 1077 | 0.76 | 16.79 |
| $\boldsymbol{n = 200}$ | | | | |
| min | $1.26 \times 10^{-3}$ | 67.91 | 0.24 | 2.64 |
| mean | $2.32 \times 10^{-3}$ | 291.47 | 0.28 | 3.91 |
| max | $21.53 \times 10^{-3}$ | 1003 | 0.48 | 6.04 |

Lastly, Table 2 highlights large differences between the methods in terms of execution time. In fact we also implemented MCEM algorithm [6] using adaptive rejection sampling. Even with a very small sample size $n = 200$, MCEM took on average 5 minutes for one simulation; while multiple imputation took less than 1 second per simulation, and SAEM less than 10 seconds, which remains reasonable. However, the bias and standard error for the estimation of SAEM and MCEM were quite similar, as presented in supplementary materials [27]. Due to this computational difficulty, we didn't perform MCEM to compare with others in the experiments with larger sample sizes.

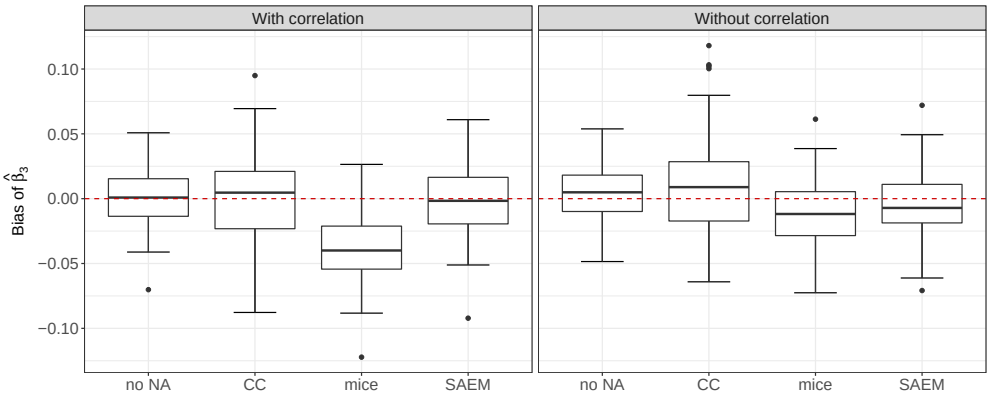



Figure 3: Empirical distribution of the estimates of $\beta_3$ obtained under MCAR, with $n = 10\,000$ and 10% of missing values; left: the covariates are correlated; right: no correlation between the covariates.

The results obtained, when the covariates were independent, are also presented. Figure 3 (right) shows the results of estimation in the case with orthogonal design. SAEM was a little biased since it estimated non-zero terms for the covariance, but it stills outperformed CC and mice.

*6.4. Extended simulations*

*Missing at Random mechanisms..* We first simulated a binary vector $\eta = (\eta_1, \eta_2, \cdots, \eta_p)$ of dimension $n \times p$ from Bernoulli distribution, where $\eta_{ij} = 0$ indicates that the corresponding $x_{ij}$ will be missing while 1 indicates observed. Then the probability of having missing data on one variable is calculated by a logistic regression function. For example in our case $p = 5$ and the realizations of $\eta$ (the pattern) $(1, 0, 1, 0, 0)$, the probability that covariates $(x_2, x_4, x_5)$ can be missing, depends only on $x_1$ and $x_3$ with a logistic regression model. The weights in the linear combination impact the proportion of missingness. We introduced 10% of missing values in the covariates according to the MAR mechanisms. The results presented in Appendix A.5 highlight that as expected they are similar to the ones obtained under MCAR and the parameters are estimated without bias.

*Robustness to the Gaussian assumption for covariates..* First we generated a design matrix of size $n = 1000 \times p = 5$ by drawing each observation from a multivariate Student distribution $t_v(\mu, \Sigma)$ with degree of freedom $v = 5$ or $v = 20$, and $(\mu, \Sigma)$ the same as those in Normal distribution in Subsection 6.1. Then, we considered the Gaussian mixture model case by generating half of the samples from $\mathcal{N}(\mu_1, \Sigma)$ and the other half from $\mathcal{N}(\mu_2, \Sigma)$, where $\mu_1 = (1, 2, 3, 4, 5)$ and $\mu_2 = (1, 1, 1, 1, 1)$, and the same $\Sigma$ as previously. Then, we generated the response according to the same logistic regression model as described in Subsection 6.1 and considered either MCAR or MAR mechanisms.

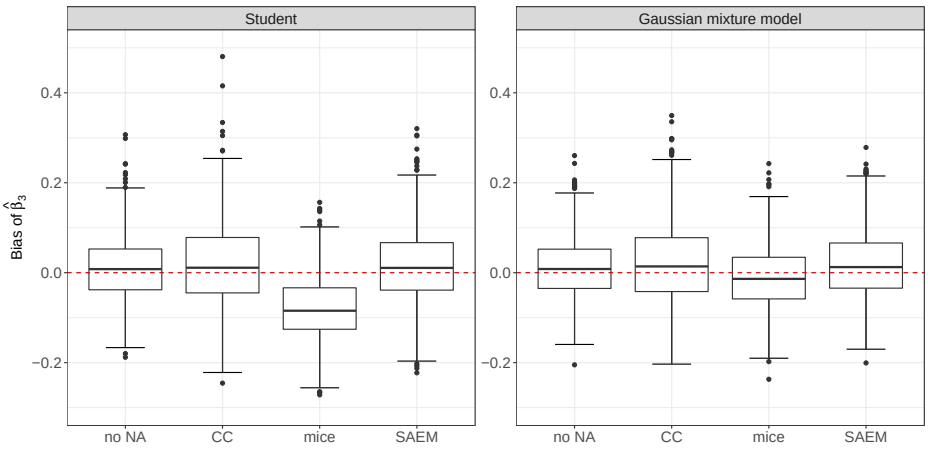



Figure 4: Empirical distribution of the bias of $\hat{\beta}_3$ obtained for misspecificated models under MCAR, with $n = 1000$; left: Student distribution with degree of freedom $v = 5$; right: Gaussian mixture model.

Figure 4 illustrates the estimation bias of the parameter $\beta_3$ and Appendix A.6 shows the coverage for all parameters and inside the parentheses is the average length of corresponding confidence interval. This experiment shows that the estimation bias for regression coefficient with the proposed method even based on normal assumption, is robust to such a model misspecification. Indeed, the bias may increase when covariates don't follow exactly a normal distribution, but the increase is negligible compared to the bias of imputation based methods.

We also observe only a small undercoverage compared to mice, and a more reasonable length of confidence interval compared to CC.

*Varying the percentage of missing values.* When the percentage of missing values increases, the variability of the results increases but the methods still provide satisfactory results as illustrated in supplementary materials [27].

*Varying the separability of the classes.* When the classes are very separated SAEM can exibit a biais and large variance as illustrated in supplementary materials [27]. However, the logistic regression without missing values also encounters difficulties.

In summary, not only did these simulations allow us to verify that SAEM leads to estimators with limited bias, but also they ensured that we made correct inferences by taking into account the additional variance due to missing data.

### *6.5. Model selection*

To look at the capabilities of the method in terms of model selection, we considered the same simulation scenarios as in Section 6.1, with some parameters set to zero. We now describe the results for the case where all parameters in $\beta$ are zero except $\beta_0 = -0.2$, $\beta_1 = 0.5$, $\beta_3 = 1$ and $\beta_5 = -0.6$. We compared the $BIC_{obs}$ based on the observed log-likelihood, as described in Section 5, to that based on the complete cases $BIC_{cc}$ and that obtained from the the original complete data $BIC_{orig}$.

Table 3: For data with or without correlations, the percentage of times that each criterion selects the correct true model (C), overfits (O), and underfits (U).

| | Non-Correlated | | | Correlated | | |
|---|---|---|---|---|---|---|
| Criterion | C | O | U | C | O | U |
| $BIC_{obs}$ | 92 | 3 | 5 | 94 | 2 | 4 |
| $BIC_{orig}$ | 96 | 2 | 2 | 93 | 0 | 7 |
| $BIC_{cc}$ | 79 | 1 | 20 | 91 | 0 | 9 |

Table 3 shows, with or without correlation between covariates, the percentage of cases where each criterion selects the true model (C), overfits (O) – i.e., selects more variables than there were – or underfits (U) – i.e., selects less variables than there were. In the case where the variables were correlated, the correlation matrix was the same as in Section 6.1. These results are representative of those obtained with other simulation schemes.

### *6.6. Prediction on a test set with missing values*

To evaluate the prediction performance on a test set with missing values, we considered the the same simulation scenarios for the training set as in Subsection 6.1 with sample size $1000 \times 5$. We also generated a test set of size $100 \times 5$. We compared the suggested approach described in Subsection 5.3, with imputation methods. More precisely, we considered single imputation methods on the training set followed by classical logistic regression and variable selection by BIC on the imputed dataset such as *i)* imputation by the mean (impMean) *ii)*

imputation by PCA (impPCA) [28] which is based on low-rank assumption of the data matrix to impute. *iii)* imputation by mice. Note that Hentges and Dunsmore [29] highlighted from a simulation study that, imputation methods can have good performance when the aim is to predict in logistic regression for MCAR data. For all the imputation methods, we also imputed the test set independently and then applied the model that had been selected on the training set. Note that this can be a limitation if there is only one individual in the test set to predict whereas the suggested method does not encounter this issue.
We compared all these approaches with classical measures to evaluate predicted probability of logistic regression, such as AUC (area under the ROC curve), Brier score [30] and Logarithmic score [31]. Figure 5 shows that on average, marginalizing over distribution of missing values has the best performances: it gave the largest AUC and Logarithmic score, and the smallest Brier scores.

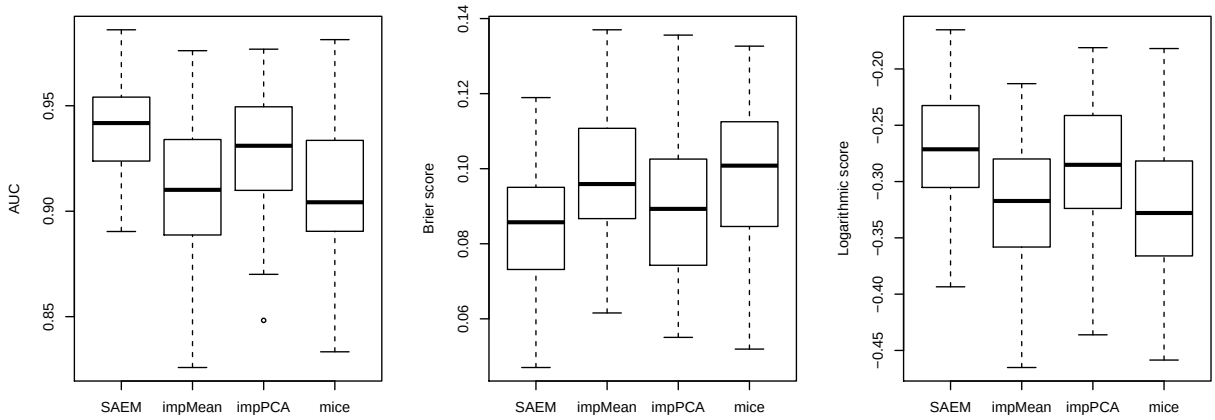



Figure 5: Comparison of empirical distribution of AUC, Brier score and Logarithmic score obtained on the test set, for the proposed approach SAEM without imputation, impMean, impPCA and mice, over 100 simulations.

## 7. Risk of severe hemorrhage for TraumaBase

The aim of our work is to accelerate and simplify the detection of patients presenting in hemorrhagic shock due to blunt trauma to speed up the management of this most preventable cause of death in major trauma. An optimized organization is essential to control blood loss as quickly as possible and to reduce mortality.

### *7.1. Details on the dataset*

This study has used the data collected from a trauma registry (TraumaBase®) shared between six trauma centers within the Ile de France region (Paris area) in France. These

centers have joined TraumaBase progressively between January 2011 and June 2015. Since then, data collection is exhaustive and covers the whole administrative area around Paris. The structure of the database integrates algorithm for consistency and coherence, and the data monitoring is performed by a central administrator. Sociodemographic, clinical, biological and therapeutic data (from the prehospital phase to the discharge if hospital) are systematically recorded for all trauma patients, and all patients transported in the trauma rooms of the participating centers are included in the registry. As a result, there were 7495 individuals in the trauma data that we investigated, collected from January 2011 to March 2016, with age ranged from 12 to 96. The study group decided to focus on patients with blunt trauma to be able to compare to the existing prediction rules. Patients with prehospital cardiac arrest and missing pre-hospital data were excluded. After this selection, 6384 patients remained in the data set. Based on clinical experience, 16 influential quantitative measurements were included. Detailed descriptions of these measurements and their histograms are shown in Appendix A.7. These variables were chosen because they were all available to the pre-hospital team, and therefore could be used in real situations.

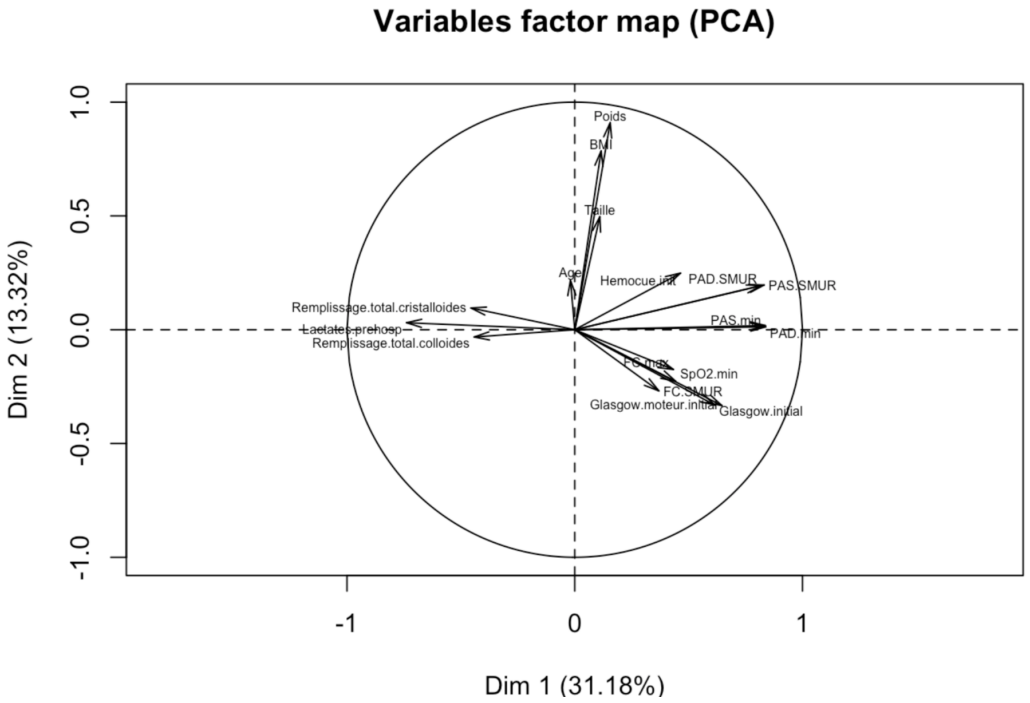



Figure 6: The factor map of the variables from PCA.

There was strong collinearity between variables, as can be seen in the variables PCA factor map (obtained by running an EM-PCA algorithm [28] which performs PCA with missing values) in Figure 6, in particular between the minimum systolic (PAS.min) and diastolic blood pressure (PAD.min). Based on expert advice, the recoded variables, SD.min and SD.SMUR (SD.min = PAS.min − PAD.min; SD.SMUR = PAS.SMUR − PAD.SMUR) were used since they have more clinical significance [32]. Thus, we had 14 variables to predict hemorrhagic shock.

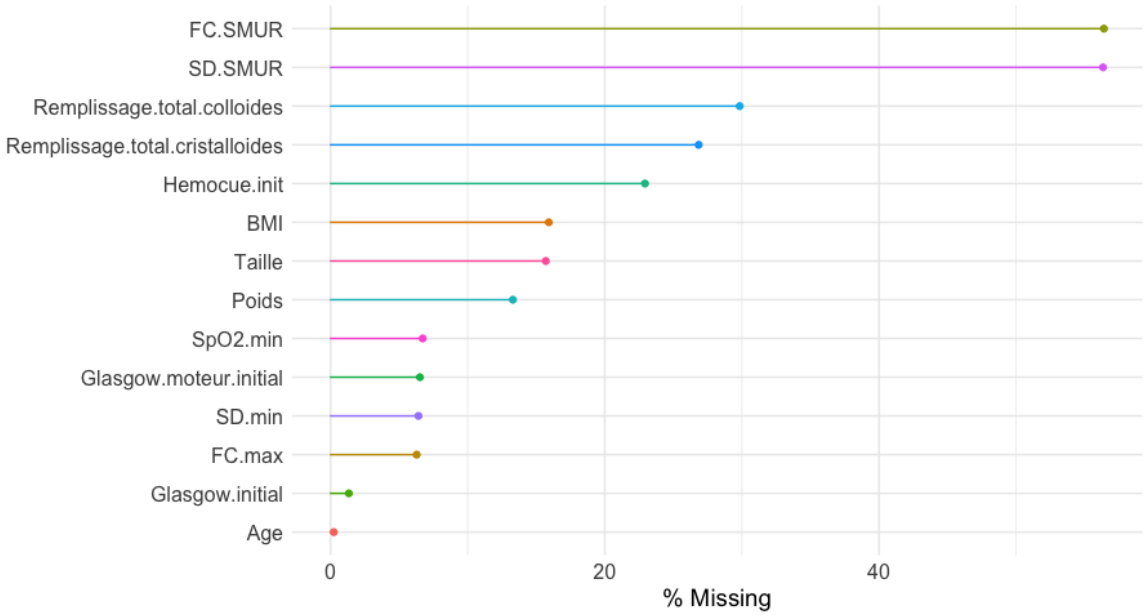



Figure 7: Percentage of missing values in each variable.

Figure 7 shows the percentage of missingness per variable, varying from 0 to 60%, which demonstrates the importance of taking appropriate account of missing data. Even though, there may be many reasons why missingness occurred, in the end, considering them all to be MAR remains a plausible assumption. For instance, FC.SMUR (heart rate) and SD.SMUR (the difference between blood pressure measured when the ambulance arrives at the accident site) contain many missing values because doctors collected these data during transportation. However, many other medical institutes and scientific publications used measurement on arrival at the accident scene. Consequently, doctors decided to record these measures as well but after the TraumaBase was set up.

We first applied SAEM for logistic regression with all 14 predictors and for the whole dataset. The estimation obtained by SAEM was of the same order of magnitude as that obtained by multiple imputation. Next, we used the model selection procedure described in Section 5. There were two observations leading to a very small value of the log-likelihood. Upon closer inspection, we found that for patient number 3302, the BMI was obtained using an incorrect calculation, and for patient number 1144, the weight (200 kg) and height (100 cm) values were likely to be incorrect. Hence, the observed log-likelihood allowed us to discover undetected outliers. On the observations' map of PCA, as shown in Figure 8, patient number 3302 (circled in blue) is one of such outliers.

### *7.2. Predictive performances*

We divided the dataset into training and test sets. The training set contained a random selection of 70% of observations, and the test set contained the remaining 30%. In the training set, we selected a model with the suggested BIC with missing values, and used

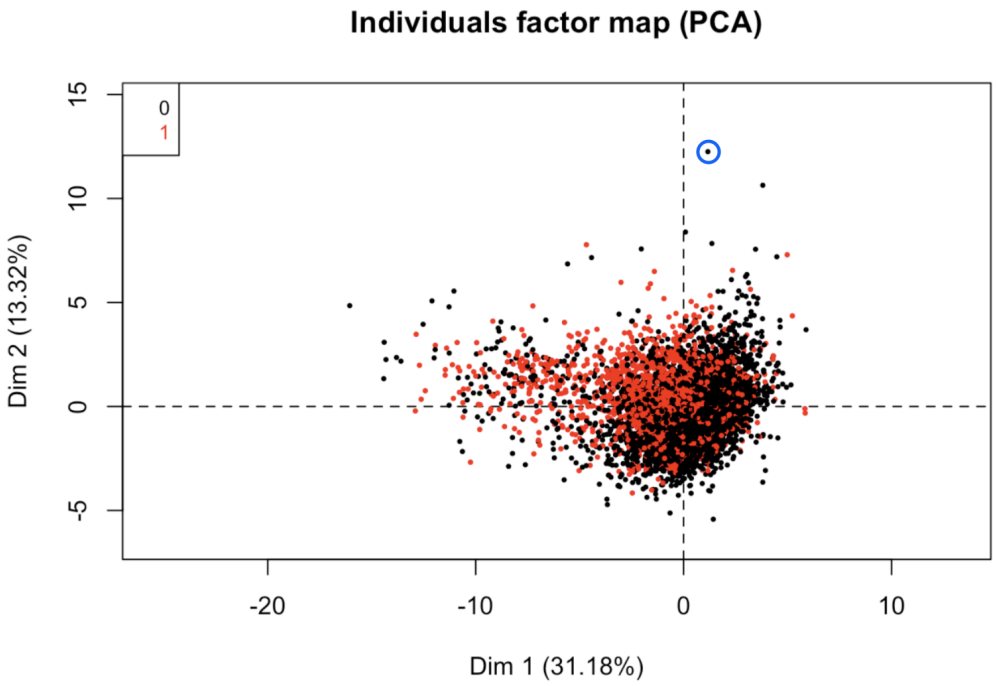



Figure 8: Observation's factor map of PCA. Red points are hemorrhagic shock patients, and black points are patients who did not have hemorrhagic shock. Patient number 3302 (circled in blue) has wrong calculation of BMI.

forward selection resulting in a model with 8 variables. The estimates of parameters and their standard errors are shown in Table 4.

Table 4: Estimation of $\beta$ and its standard errors obtained by SAEM, using BIC for model selection.

| Variables | Estimate (standard errors) |
|---|---|
| $(Intercept)$ | -0.52 (0.59) |
| $Age$ | 0.011 (0.0033) |
| $Glasgow.moteur$ | -0.16 (0.036) |
| $FC.max$ | 0.026 (0.0025) |
| $Hemocue.init$ | -0.23 (0.031) |
| $RT.cristalloides$ | 0.00090 (0.00010) |
| $RT.colloides$ | 0.0019 (0.00021) |
| $SD.min$ | -0.025 (0.0050) |
| $SD.SMUR$ | -0.021 (0.0056) |

The TraumaBase medical team indicated to us that the signs of the coefficients were in agreement with their a priori ideas: all the others things being equal *a)* Older people are more likely to have a hemorrhagic shock; *b)* A low Glasgow score implies little or no motor response, which often is the case for hemorrhagic shock patients; *c)* One typical sign of hemorrhagic shock is rapid heart rate; *d)* The more a patient bleeds, the lower their

Hemocue is, and the more blood must be transfused. Eventually, it is more likely they will end up in hemorrhagic shock; *e)* Therapy involving two types of volume expanders, cristalloides and colloides, can be conducted to treat hemorrhagic shock. If extremely low difference between blood pressure is observed, its cause may be low stroke volume, as is usually the case in hemorrhagic shock.

Next, we assessed the prediction quality on the test set with usual metrics based on the confusion matrix (false positive rate, false negative rate, etc.). We need to ensure that the cost of a false negative is much more than that of a false positive, as non-recognition of a potential hemorrhagic shock leads to a higher risk of patient mortality. We define the validation error on test set as:

$$l(\hat{y}, y) = \frac{1}{n} \sum_{i=1}^{n} w_0 \mathbb{1}_{\{y_i=1, \hat{y}_i=0\}} + w_1 \mathbb{1}_{\{y_i=0, \hat{y}_i=1\}} \tag{7}$$

where $w_0$ and $w_1$ are user defined weight for the cost of false negative and false positive respectively, s.t.,$w_0 + w_1 = 1$. Therefore, we can choose a threshold for logistic regression by given the value for $w_0$ and $w_1$. For instance, we chose $\frac{w_0}{w_1} = 5$, i.e., the false negative was 5 times costly than the false positive. The cost function was chosen in agreement with the experts. Note that the test set was also incomplete, so we used the strategy described in Subsection 5.3. The confusion matrix of the predictive performance on the test set is shown in Table 5. The associated ROC curve is shown in Figure 9, and the AUC is 0.8487.

| | | **Predicted outcome** | |
|---|---|---|---|
| | | 1 | 0 |
| **Observed value** | **1** | True Positive (98) | False Negative (29) |
| | **0** | False Positive (146) | True Negative (808) |

Table 5: Confusion matrix for prediction on test set.

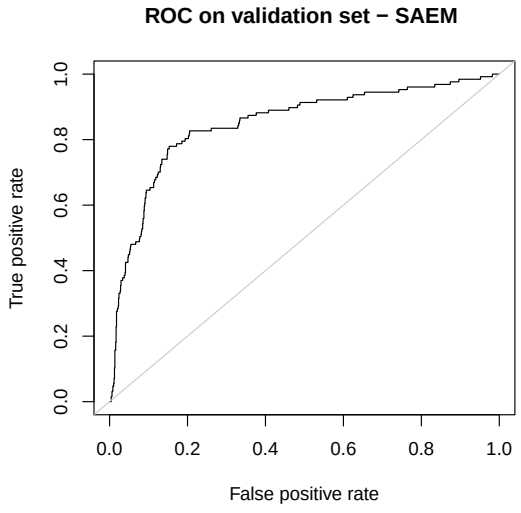



Figure 9: ROC curve of the test set predictions.

*7.3. Comparison with other approaches*

Finally we compared the proposed method to other approaches. Similar to the Subsection 7.2, we considered single imputation methods followed by classical logistic regression

and variable selection on the imputed training dataset, such as single imputation by PCA (impPCA) [28], imputation by Random Forest (missForest) [33], as well as mean imputation (impMean). Meanwhile, we compared logistic regression model with other prediction models, such as Random Forest (predRF) and SVM (predSVM), both applied on the imputed dataset by Random Forest [33]. We also considered multiple imputation by chained equation (mice): we applied logistic regression with a classical forward selection method, with BIC on each imputed data set. However, note that there is no straightforward solution for combining multiple imputation and variable selection; we followed the empirical approach suggested in Wood et al. [34], where they kept the variables selected in each imputed dataset to define the final model.

We also considered three rules used by the doctors to predict the hemorrhagic shock *i)* Doctors' prediction (doctor): the decision was recorded in the TraumaBase. It determines whether the doctor considered the patient to be at risk of hemorrhagic shock. *ii)* Assessment of Blood Consumption score (ABC): it is an examination usually performed when the patient arrives at the trauma center. As such, the score is not exactly prehospital but can be computed very early once the patient is hospitalized. *iii)* Trauma Associated Severe Hemorrhage score (TASH): this score was also designed for hemorrhage detection, but at a later stage since it uses some values that are only available after laboratory tests or radiography.

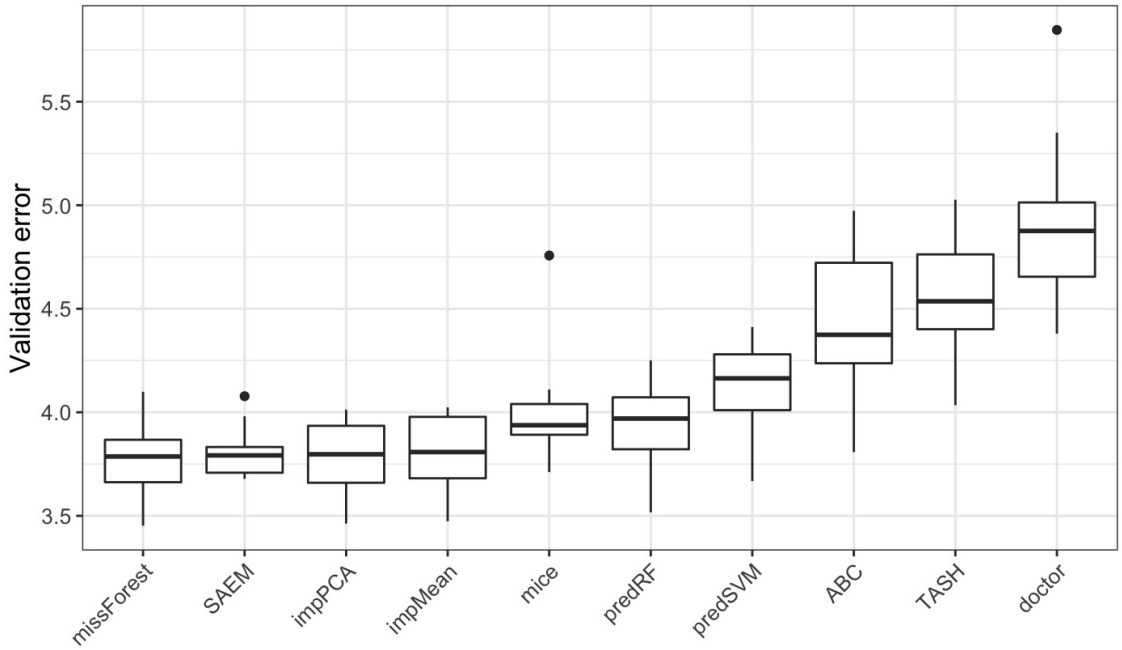



Figure 10: Empirical distribution of prediction errors of different methods over 15 replications for the TraumaBase data.

Figure 10 compares the methods in terms of their validation error (7). The splitting of data (into training and test sets) was repeated 15 times and we fixed the threshold such that the cost of false negative is 5 times that of false positive, i.e, $\frac{w_0}{w_1} = 5$. On average, SAEM

had good performance with small variability, while all the imputation methods performed similarly even the naive mean imputation. In addition, other prediction methods (Random Forest and SVM) did not result in a smaller error on the test sets than the logistic regression models. Lastly the rules used by the doctors, even the ones using more information than prehospital data, were not as competitive as SAEM. Appendix A.8 gives the details with classical measures (AUC, sensitivity, specificity, accuracy and precision) to compare the predictive performance of the methods. The suggested approach resulted in good performance on average, and in particular, had an advantage in terms of the sensitivity, i.e., it rarely misdiagnosed the hemorrhagic shock patients, which is relevant to clinical needs of emergency doctors.

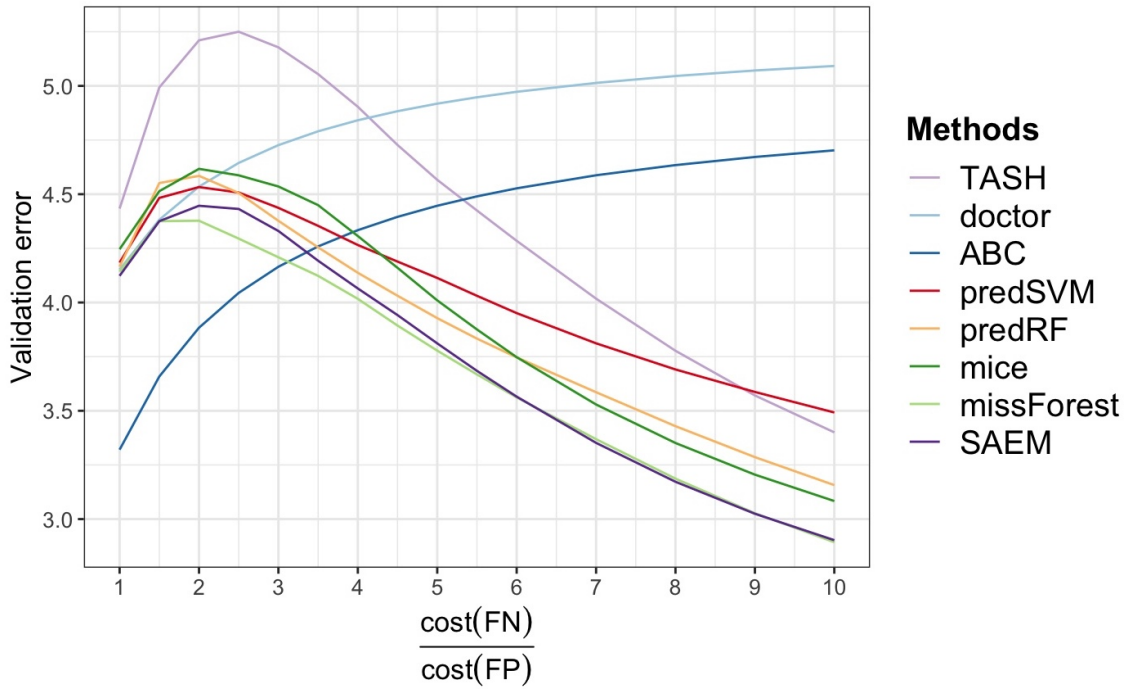



Figure 11: Average prediction errors of different methods, as function of the cost importance $\{\frac{w_0}{w_1} \mid \frac{w_0}{w_1} > 1\}$, over 15 replications for the TraumaBase data.

More generally, without defining a specific threshold, we observed in Figure 11 the average predictive loss over 15 replications as function of the cost importance $\{\frac{w_0}{w_1} \mid \frac{w_0}{w_1} > 1\}$ for all the methods. Obviously, we had the same performance evaluation as before, as SAEM had smaller error on the test sets with the respect to the choice of $\frac{w_0}{w_1}$, especially when we emphasized more on the cost of false negative. Note that the curves of doctors' rules and ABC increase as a function of the cost importance $\frac{w_0}{w_1}$, which means that, the rules of doctors are more conservative than SAEM, which can be problematic in this application.

Note that even if the proposed methodology is based on the assumption of normally distributed covariates, the performance of the proposed methodology is better than the prediction made by the widely used medical criterion, in terms of prediction error. Some

discussions on the normal assumption are provided in Appendix A.7.

In summary, the logistic regression methodology with missing values, from estimation to selection, as well as prediction on a test sample with missing data, is theoretically well founded. Based on the TraumaBase application and comparison with other methods, we have demonstrated that the proposed approach has the ability to outperform existing popular methods dealing with missing data.

## 8. Discussion

In this paper, we have developed a comprehensive joint-modeling framework for logistic regression with missing values. The experiments indicate that the proposed method is computationally efficient, and can be easily implemented. In addition, compared with multiple imputation – especially in the case with correlation between variables – estimation using SAEM is less biased than other methods and generally leads to interval-estimate coverage that is close to the nominal level. Based on the proposed algorithm, model selection by BIC with missing data can be performed in a natural way. In view of the results reported in this article, we have been invited by emergency-room doctors in one of the centers that contributes to the TraumaBase dataset to implement the missing-data methodology outlined here in a prospective study to evaluate its performance in real time in a clinical setting. Paths for possible future research include further developing the method to handle quantitative and categorical data. This paper focused on making inference with missing values but we have suggested a method to predict from a test set with missing values. More work can be done in the direction of supervised learning with missing values, especially to suggest variance of prediction. Extensions of the methods of Schafer and Schenker [35] could be studied. In addition, in the TraumaBase dataset, we can reasonably expect to have both MAR and missing not at random (MNAR) values. MNAR means that missingness is related to the missing values themselves, therefore, the correct treatment would require incorporating models for the missing data mechanisms. As a final note, the proposed method may be quite useful in the causal inference framework, especially for propensity score analysis, which estimates the effect of a treatment, policy, or other intervention. Indeed, inverse probability weighting methods (IPW) are often performed with logistic regression, and the proposed method offers a potential solution for times where there are missing values in the covariates. The method is implemented in the R package *misaem*.

## Appendix A. Appendix

*Appendix A.1. Missing mechanism*

Missing completely at random (MCAR) means that there is no relationship between the missingness of the data and any values, observed or missing. In other words, MCAR means:

$$\mathtt{p}(M_i|y, x_i, \phi) = \mathtt{p}(M_i|\phi)$$

Missing at Random (MAR), means that the probability to have missing values may depend on the observed data, but not on the missing data. We must carefully define what this means in our case by decomposing the data $x_i$ into a subset $x_i^{(\text{mis})}$ of data that "can be missing", and a subset $x_i^{(\text{obs})}$ of data that "cannot be missing", i.e. that are always observed. Then, the observed data $x_{i,\text{obs}}$ necessarily includes the data that can be observed $x_i^{(\text{obs})}$, while the data that can be missing $x_i^{(\text{mis})}$ includes the missing data $x_{i,\text{mis}}$. Thus, MAR assumption implies that, for all individual $i$,

$$\begin{aligned}\mathtt{p}(M_i|y_i, x_i; \phi) &= \mathtt{p}(M_i|y_i, x_i^{(\text{obs})}; \phi)\\ &= \mathtt{p}(M_i|y_i, x_{i,\text{obs}}; \phi)\end{aligned}$$

MAR assumption implies that, the observed likelihood can be maximize and the distribution of $M$ can be ignored [4]. Indeed,

$$\begin{aligned}\mathcal{L}(\theta, \phi; y, x_{\text{obs}}, M) &= \mathtt{p}(y, x_{\text{obs}}, M; \theta, \phi)\\ &= \prod_{i=1}^{n} \mathtt{p}(y_i, x_{i,\text{obs}}, M_i; \theta, \phi)\\ &= \prod_{i=1}^{n} \int \mathtt{p}(y_i, x_i, M_i; \theta, \phi) dx_{i,\text{mis}}\\ &= \prod_{i=1}^{n} \int \mathtt{p}(y_i, x_i; \theta)\mathtt{p}(M_i|y_i, x_i; \phi) dx_{i,\text{mis}}\\ &= \prod_{i=1}^{n} \int \mathtt{p}(y_i, x_i; \theta)\mathtt{p}(M_i|y_i, x_{i,\text{obs}}; \phi) dx_{i,\text{mis}}\\ &= \prod_{i=1}^{n} \mathtt{p}(M_i|y_i, x_{i,\text{obs}}; \phi) \times \prod_{i=1}^{n} \int \mathtt{p}(y_i, x_i; \theta) dx_{i,\text{mis}}\\ &= \mathtt{p}(M|y, x_{\text{obs}}; \phi) \times \mathtt{p}(y, x_{\text{obs}}; \theta)\\ &= \mathtt{p}(M|y, x^{(\text{obs})}; \phi) \times \mathtt{p}(y, x_{\text{obs}}; \theta)\end{aligned}$$

Therefore, to estimate $\theta$, we aim at maximizing $\mathcal{L}(\theta; y, x_{\text{obs}}) = \mathtt{p}(y, x_{\text{obs}}; \theta)$.

*Appendix A.2. Metropolis-Hastings sampling*

During the iterations of SAEM, the Metropolis-Hastings sampling is performed as Algorithm 1, with the target distribution $f(x_{i,\text{mis}}) = \mathtt{p}(x_{i,\text{mis}}|x_{i,\text{obs}}, y_i; \theta)$ and the proposal distribution $g(x_{i,\text{mis}}) = \mathtt{p}(x_{i,\text{mis}}|x_{i,\text{obs}}; \mu, \Sigma)$.

**Algorithm 1** Metropolis-Hastings sampling.

**Input:** An initial samples $x_{i,\text{mis}}^{(0)} \sim g(x_{i,\text{mis}})$;
  **for** $s = 1, 2, \cdots, S$ **do**
    Generate $x_{i,\text{mis}}^{(s)} \sim g(x_{i,\text{mis}})$;
    Generate $u \sim \mathcal{U}[0, 1]$;
    Calculate the ratio $w = \frac{f(x_{i,\text{mis}}^{(s)})/g(x_{i,\text{mis}}^{(s)})}{f(x_{i,\text{mis}}^{(s-1)})/g(x_{i,\text{mis}}^{(s-1)})}$;
    **if** $u < w$ **then**
      Accept $x_{i,\text{mis}}^{(s)}$;
    **else**
      $x_{i,\text{mis}}^{(s)} \leftarrow x_{i,\text{mis}}^{(s-1)}$;
    **end if**
  **end for**
**Output:** $(x_{i,\text{mis}}^{(s)}, 1 \le i \le n, 1 \le s \le S)$.

*Appendix A.3. Calculation of observed information matrix*

Procedure 2 shows how we calculate the observed information matrix.

**Procedure 2** Calculation of observed information matrix.

**Input:** After drawing MH samples $(x_{i,\text{mis}}^{(s)}, 1 \le i \le n, 1 \le s \le S)$ for unobserved data $(x_{i,\text{mis}}, 1 \le i \le n)$, we have imputed observations, noted as $(z_i^{(s)}, 1 \le i \le n, 1 \le s \le S)$, where $z_{ij}^{(s)} = x_{i,\text{obs}}$, if $x_{ij}$ is observed; else $z_{ij}^{(s)} = x_{i,\text{mis}}^{(s)}$.
  **for** $n = 1, 2, \cdots, n$ **do**
    **for** $s = 1, 2, \cdots, S$ **do**
      Calculate the gradient:
      $\nabla f_{is} = \frac{\partial \mathcal{LL}(\theta; x_{i,\text{obs}}, x_{i,\text{mis}}^{(s)}, y_i)}{\partial \beta} = z_i^{(s)} \left( y_i - \frac{\exp(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j z_{ij}^{(s)})}{1 + \exp(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j z_{ij}^{(s)})} \right)$;
      Calculate the Hessian matrix:
      $H_{is} = \frac{\partial^2 \mathcal{LL}(\theta; x_{i,\text{obs}}, x_{i,\text{mis}}^{(s)}, y_i)}{\partial \beta \partial \beta^T} = -z_i^{(s)} z_i^{(s)^T} \frac{\exp(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j z_{ij}^{(s)})}{\left(1 + \exp(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j z_{ij}^{(s)})\right)^2}$;
      $\Delta_i \leftarrow \frac{1}{s}[(s-1)\Delta_i + \nabla f_{is}]$;
      $D_i \leftarrow \frac{1}{s}[(s-1)D_i + H_{is}]$;
      $G_i \leftarrow \frac{1}{s}[(s-1)G_i + \nabla f_{is} \nabla f_{is}^T]$;
    **end for**
    $\hat{\mathcal{I}}_S(\hat{\beta}) \leftarrow \hat{\mathcal{I}}_S(\hat{\beta}) - (D_i + G_i - \Delta_i \Delta_i^T)$;
  **end for**
**Output:** $\hat{\mathcal{I}}_S(\hat{\beta})$.

*Appendix A.4. Logistic regression on simulated complete dataset*

Figure A.12 shows the ROC curve on a simulated complete dataset. The corresponding AUC (for training set) is 0.8976.

*Appendix A.5. Simulation results for Missing at Random data*

We consider a Missing at Random mechanism to generate data. Figure A.13 shows that the biases were very similar to the ones obtained under a MCAR mechanism and the parameters were estimated without bias.

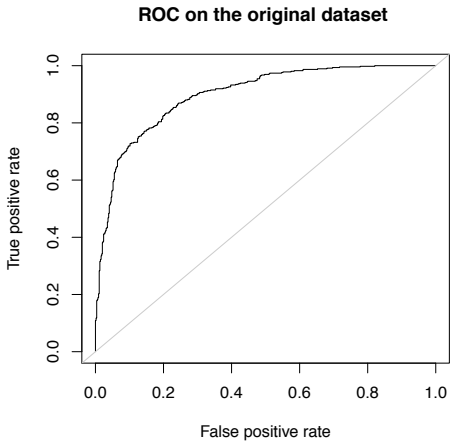



Figure A.12: ROC curve on a simulated complete dataset.

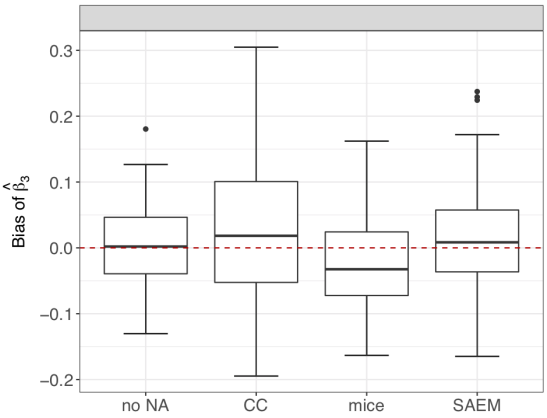



Figure A.13: Empirical distribution of the bias of $\hat{\beta}_3$ obtained under MAR mechanism, with $n = 1000$ and 10% of missing values.

*Appendix A.6. Simulation results for model misspecification: the coverage*

Table A.6 shows the coverage for all parameters and inside the parentheses is the average length of corresponding confidence interval.

Table A.6: Coverage (%) for $n = 1000$, MCAR and misspecified models, calculated over 1000 simulations. Bold indicates under coverage. Inside the parentheses is the average length of corresponding confidence interval over 1000 simulations (multiplied by 100).

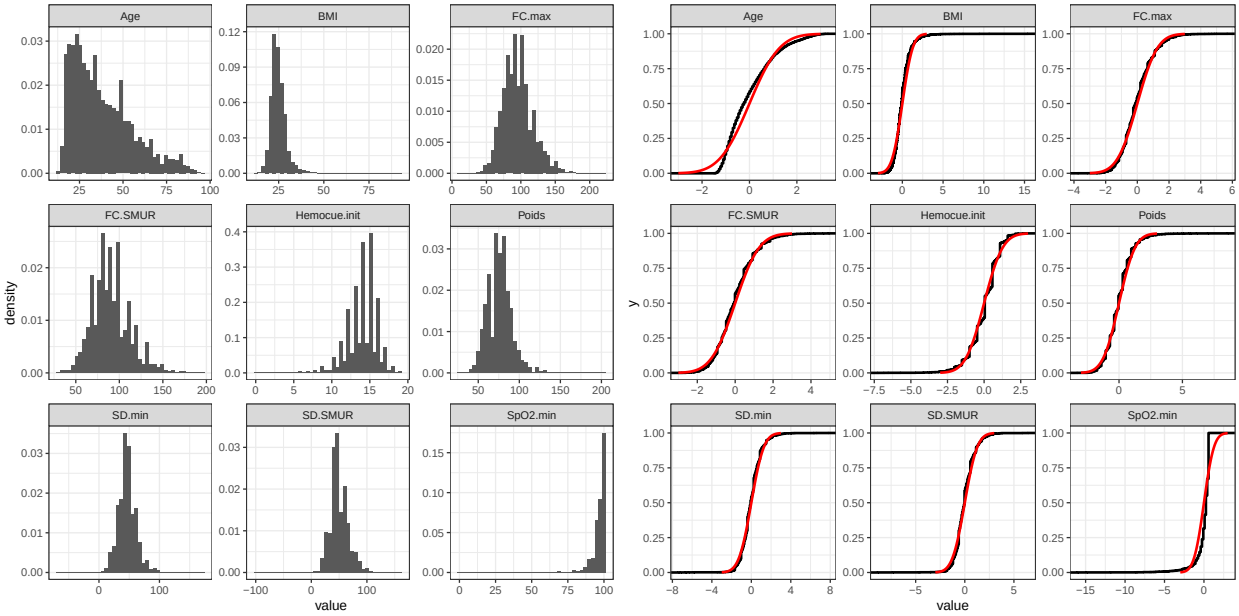
| parameter | no NA | CC | mice | SAEM |
|---|---|---|---|---|
| Student distribution: | ($v = 5$) | | | |
| $\beta_0$ | 94.7 (68.02) | 94.3 (84.14) | 94.6 (67.69) | 93.8 (68.25) |
| $\beta_1$ | 95.2 (54.78) | 94.2 (72.15) | 91.7 (61.96) | 93.5 (63.05) |
| $\beta_2$ | 94.9 (27.66) | 94.6 (36.39) | 91.4 (31.21) | 93.7 (31.84) |
| $\beta_3$ | 94.9 (26.76) | 94.3 (35.24) | **81.5** (30.46) | 94.7 (29.98) |
| $\beta_4$ | 95.2 (11.52) | 95.4 (15.16) | 95.8 (12.94) | 95.5 (12.88) |
| $\beta_5$ | 93.7 (17.63) | 94.9 (23.22) | **83.4** (20.40) | 93.3 (19.93) |
| Gaussian mixture: | | | | |
| $\beta_0$ | 94.8 (57.54) | 95.2 (75.42) | 95.4 (61.95) | 95.0 (61.33) |
| $\beta_1$ | 94.7 (58.00) | 96.2 (76.05) | 95.4 (66.66) | 95.3 (66.13) |
| $\beta_2$ | 94.3 (28.49) | 95.3 (37.35) | 95.3 (32.65) | 94.0 (32.50) |
| $\beta_3$ | 94.7 (26.16) | 94.9 (34.38) | 94.9 (28.91) | 94.5 (29.10) |
| $\beta_4$ | 94.4 (12.68) | 94.4 (16.60) | 94.4 (14.24) | 94.7 (14.09) |
| $\beta_5$ | 95.3 (17.70) | 94.7 (23.25) | 94.7 (19.86) | 95.3 (19.92) |

*Appendix A.7. Definition of the variables of the TraumaBase data set*

In this Subsection, we give the detailed explanations for the selected quantitative variables:

- *Age*: Age.
- *Poids*: Weight.
- *Taille*: Height.
- *BMI*: Body Mass index, $BMI = \frac{Weight \text{ in } kg}{(Height \text{ in } m)^2}$
- *Glasgow*: Glasgow Coma Scale .
- *Glasgow.moteur*: Glasgow Coma Scale motor component.
- *PAS.min*: The minimum systolic blood pressure.
- *PAD.min*: The minimum diastolic blood pressure.
- *FC.max*: The maximum number of heart rate (or pulse) per unit time (usually a minute).
- *PAS.SMUR*: Systolic blood pressure at arrival of ambulance.
- *PAD.SMUR*: Diastolic blood pressure at arrival of ambulance.
- *FC.SMUR*: Heart rate at arrival of ambulance.
- *Hemocue.init*: Capillary Hemoglobin concentration.
- *SpO2.min*: Oxygen saturation.
- *Remplissage.total.colloides* (or *RT.colloides*): Fluid expansion colloids.
- *Remplissage.total.cristalloides* (or *RT.cristalloides*): Fluid expansion cristalloids.
- *SD.min* ($= PAS.min - PAD.min$): Pulse pressure for the minimum value of diastolic and systolic blood pressure.
- *SD.SMUR* ($= PAS.SMUR - PAD.SMUR$): Pulse pressure at arrival of ambulance.

Figure A.14 shows the histogram and the empirical *c.d.f.* of several covariates from the TraumaBase data. Several of these distributions are not symmetrical. In practice, it is possible to consider that some suitable transformations of the covariates can be approximated by normal distributions. For example, transformations of the form $\log(c+x)$ and $\log(c-x)$, can be very appropriate for, respectively, right-skewed and left-skewed distributions. We applied the proposed methodology to the real dataset after transformation. However, the prediction result from cross-validation didn't show advantage of the transformed version.




(a) Histograms of covariates (b) Empirical cumulative distributions

Figure A.14: Empirical distribution of variables from TraumaBase. (a) Histograms of covariates (b) Black curve illustrates the empirical cumulative distributions while the red curve represents the normal distribution.

Indeed when the log transformation is used as a prepossessing step, it only operates on the observed part, which is appropriate under MCAR calues. Consequently, taking into account the simulation study, the interpretability, the choices of transformations, and the prediction results, we have decided to keep the variables without any transformation.

*Appendix A.8. Details of predictive performance for TraumaBase data*

Details of predictive performance for TraumaBase data are given by Table A.7.

Table A.7: Comparison of the mean of the predictive performances (values are multiplied by 100) of different methods dealing with missing data. AUC is the area under ROC; the accuracy is the number of true positive plus true negative divided by the total number of observations; the sensitivity is defined as the true positive rate; specificity as the true negative rate; the precision is the number of true positive over all positive predictions. The best results are in bold.

| Metrics | SAEM | missForest | impMean | impPCA | mice | predRF | predSVM |
|---|---|---|---|---|---|---|---|
| AUC | **88.5** | **88.8** | **88.9** | **89.0** | 87.7 | 88.0 | 80.4 |
| Accuracy | 86.9 | 87.0 | 87.3 | 86.7 | 85.3 | 87.2 | **88.3** |
| Precision | 41.1 | 41.6 | 42.2 | 41.0 | 37.9 | 41.6 | **44.0** |
| Sensitivity | **74.6** | 74.3 | 73.2 | **75.0** | **75.2** | 71.5 | 66.0 |
| Specificity | 88.2 | 88.4 | 88.8 | 87.9 | 86.4 | 88.9 | **90.6** |

## Supplementary material

**R-package:** R-package "misaem" containing the implementation of algorithm SAEM to fit the logistic regression model with missing data, now available in CRAN [20].

**Codes:** Code to reproduce the experiments are provided in GitHub [21].

**Additional supplementary materials:** Some supplementary simulation results are presented [27].

## References

## References

[1] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society. Series B (Methodological) 39 (1977) 1–38.
[2] X.-L. Meng, D. B. Rubin, Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm, Journal of the American Statistical Association 86 (1991) 899–909.
[3] T. A. Louis, Finding the observed information matrix when using the EM algorithm, Journal of the Royal Statistical Society. Series B (Methodological) 44 (1982) 226–233.
[4] R. J. Little, D. B. Rubin, Statistical Analysis with Missing Data, second ed., John Wiley & Sons, Inc., 2002.
[5] S. Seaman, J. Galati, D. Jackson, J. Carlin, What is meant by "missing at random"?, Statist. Sci. 28 (2013) 257–268.
[6] J. G. Ibrahim, M.-H. Chen, S. R. Lipsitz, Monte Carlo EM for missing covariates in parametric regression models, BIOMETRICS 55 (1999) 591–596.
[7] J. G. Ibrahim, M.-H. Chen, S. R. Lipsitz, A. H. Herring, Missing-data methods for generalized linear models: A comparative review, Journal of the American Statistical Association 100 (2005) 332–346.
[8] G. C. G. Wei, M. A. Tanner, A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms, Journal of the American Statistical Association 85 (1990) 699–704.
[9] G. McLachlan, T. Krishnan, The EM algorithm and extensions, Wiley series in probability and statistics, 2. ed ed., Wiley, Hoboken, NJ, 2008.
[10] W. R. Gilks, P. P. Wild, Adaptive rejection sampling for Gibbs sampling, Appl. Statist 41 (1992) 337–348.
[11] M. Lavielle, Mixed Effects Models for the Population Approach: Models, Tasks, Methods and Tools, Chapman and Hall/CRC, 2014.
[12] G. Claeskens, F. Consentino, Variable selection with incomplete covariate data, Biometrics 64 (2008) 1062–9.
[13] F. Consentino, G. Claeskens, Missing covariates in logistic regression, estimation and distribution selection, Statistical Modelling 11 (2011) 159–183.
[14] J. Jiang, T. Nguyen, J. S. Rao, The E-MS algorithm: Model selection with incomplete data, Journal of the American Statistical Association 110 (2015) 1136–1147.
[15] Y. Liu, Y. Wang, Y. Feng, M. M. Wall, Variable selection and prediction with incomplete high-dimensional data, Ann. Appl. Stat. 10 (2016) 418–450.
[16] W. K. Chow, A look at various estimators in logistic models in the presence of missing values, Technical Report, RAND CORP SANTA MONICA CA, 1979.
[17] K. Yuen Fung, B. A. Wrobel, The treatment of missing values in logistic regression, Biometrical Journal 31 (1989) 35 – 47.
[18] D. B. Rubin, Multiple Imputation for Nonresponse in Surveys, volume 307, John Wiley & Sons, 2009.
[19] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2017.

[20] W. Jiang, misaem: Logistic regression with missing covariates, 2019. R package version 0.9.1.
[21] W. Jiang, Codes and implementations for "Logistic regression with missing covariates – parameter estimation, model selection and prediction within a joint-modeling framework", https://github.com/wjiang94/miSAEM_logReg, 2019.
[22] S. I. Hay, et al., Global, regional, and national disability-adjusted life-years (dalys) for 333 diseases and injuries and healthy life expectancy (hale) for 195 countries and territories, 1990–2016: a systematic analysis for the global burden of disease study 2016, The Lancet 390 (2017) 1260 – 1344.
[23] S. R. Hamada, T. Gauss, F.-X. Duchateau, J. Truchot, A. Harrois, M. Raux, J. Duranteau, J. Mantz, C. Paugam-Burtz, Evaluation of the performance of french physician-staffed emergency medical service in the triage of major trauma patients, Journal of Trauma and Acute Care Surgery 76 (2014) 1476–1483.
[24] S. R. Hamada, T. Gauss, J. Pann, M. W. Dünser, M. Léone, J. Duranteau, European trauma guideline compliance assessment: The ETRAUSS study, Critical care 19 (2015) 423.
[25] B. Delyon, M. Lavielle, E. Moulines, Convergence of a stochastic approximation version of the EM algorithm, The Annals of Statistics 27 (1999) 94–128.
[26] S. van Buuren, K. Groothuis-Oudshoorn, mice: Multivariate imputation by chained equations in R, Journal of Statistical Software 45 (2011) 1–67.
[27] W. Jiang, Additional supplementary materials for "Logistic regression with missing covariates – parameter estimation, model selection and prediction within a joint-modeling framework", https://github.com/wjiang94/miSAEM_logReg/tree/master/Supplement, 2019.
[28] J. Josse, F. Husson, missMDA: A package for handling missing values in multivariate data analysis, Journal of Statistical Software 70 (2016) 1–31.
[29] A. L. Hentges, I. R. Dunsmore, Predictive distributions in binary models with missing data, Communications in Statistics-Simulation and Computation 27 (1998) 735–759.
[30] G. W. Brier, Verification of forecasts expressed in terms of probability, Monthly Weather Review 78 (1950) 1–3.
[31] I. J. Good, Rational decisions, Journal of the Royal Statistical Society. Series B (Methodological) (1952) 107–114.
[32] S. R. Hamada, A. Rosa, T. Gauss, J.-P. Desclefs, M. Raux, A. Harrois, A. Follin, F. Cook, M. Boutonnet, A. Attias, S. Ausset, G. Dhonneur, O. Langeron, C. Paugam-Burtz, R. Pirracchio, B. Riou, G. de St Maurice, B. Vigué, A. Rouquette, J. Duranteau, Development and validation of a pre-hospital "Red Flag" alert for activation of intra-hospital haemorrhage control response in blunt trauma, Critical Care 22 (2018) 113.
[33] D. J. Stekhoven, P. Buehlmann, MissForest – non-parametric missing value imputation for mixed-type data, Bioinformatics 28 (2012) 112–118.
[34] A. M. Wood, I. R. White, P. Royston, How should variable selection be performed with multiply imputed data?, Statistics in Medicine 27 (2008) 3227–3246.
[35] J. L. Schafer, N. Schenker, Inference with imputed conditional means, Journal of the American Statistical Association 95 (2000) 144–154.