Submitted to the Annals of Applied Statistics
arXiv: arXiv:0000.0000

STOCHASTIC APPROXIMATION EM FOR LOGISTIC REGRESSION WITH MISSING VALUES

BY WEI JIANG* , JULIE JOSSE* , MARC LAVIELLE*

Inria Saclay and École Polytechnique, Paris *

Logistic regression is a common classification method in supervised learning. Surprisingly, there are very few solutions for performing it and selecting variables in the presence of missing values. We propose a stochastic approximation version of the EM algorithm based on Metropolis-Hasting sampling, to perform statistical inference for logistic regression with incomplete data. We propose a complete approach, including the estimation of parameters and their variance, derivation of confidence intervals, a model selection procedure, and a method for prediction on test sets with missing values. The method is computationally efficient, and its good coverage and variable selection properties are demonstrated in a simulation study. We then illustrate the method on a dataset of polytraumatized 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*.

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 is modifying an estimation process so that it can be applied to incomplete data. For example, one can use the EM algorithm [DLR77] to obtain the maximum likelihood estimate (MLE) despite missing values, and a supplemented EM algorithm (SEM) [MR91] or Louis' formula [Lou82] for the variance of the estimate. This strategy is valid under missing at random (MAR) mechanisms [Rub76, LR02, SGJC13], in which data missingness 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 it is often the case that the expectation step of the EM algorithm involves infeasible computations. One solution to this, sug-

Keywords and phrases: incomplete data, observed likelihood, variable selection, major trauma, public health

gested in [CC08, GW92, ICL99, ICLH05] in the framework of generalized linear models, is to use a Monte Carlo EM (MCEM) algorithm [WT90, MK08], replacing the integral by its empirical sum using Monte Carlo sampling. These authors also estimate the variance using a Monte Carlo version of Louis' formula. For sampling, Ibrahim et al. [ICL99] used Gibbs samplers with an adaptive rejection sampling scheme. However, their approach is computationally expensive and they considered an implementation only for monotone patterns of missing values, or for missing values in only two variables in a dataset.

In this paper, we develop a stochastic approximation EM (SAEM) [Lav14], an alternative to MCEM. 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. SAEM has an undeniable computational advantage over MCEM. In addition, it allows for model selection using criterion based on penalized observed likelihood. This latter characteristic is very useful in practice as only few methods are available to select a model when there are missing values. For example, [CC08, CC11] suggested approximation of AIC while [JNR15] defined generalized information criteria and adaptive fence and [LWFW16] in the framework of imputation with Random Lasso proposed to combine penalized regression techniques with multiple imputation and stability selection.

This paper proceeds as follows: In Section 2 we describe the motivation for our work, the Traumabase[1] project, a study related to the management of polytraumatized patients. Section 3 presents the assumptions and notations used throughout this paper. In Section 4, we propose the algorithm SAEM for estimating the parameters in a logistic regression model for continuous data, under the MAR mechanism. 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. Section 6 presents a simulation study where our approach is compared to alternative methods such as multiple imputation [Rub78]. In Section 7, we apply our approach to predict the hemorrhagic shock for the Traumabase study. Finally Section 8 concludes our work and provides a discussion.

The methodology presented in this article is available as an R package *misaem* provided in `https://github.com/wjiang94/misaem`.

2. Example. Our work is motivated by a collaboration with the Traumabase group at APHP (Public Assistance - Hospitals of Paris), which is

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

dedicated to the management of polytraumatized patients. Major trauma is defined as any injury that endangers the life or the functional integrity of a person. The global burden of disease working group of the WHO has recently shown that major trauma in its various forms, including traffic accidents, interpersonal violence, self-harm, and falls, remains a public health challenge and a major source of mortality and handicap around the world [DC17]. Effective and timely management of trauma is critical to improving outcomes. Delay, or errors in treatment have a direct impact on survival, especially for the two main causes of death in major trauma: hemorrhage and traumatic brain injury.

Management of a polytraumatized patient has several stages:

1. At the accident site where a patient is taken under charge by the ambulance, emergency doctors make a first assessment on the gravity of the patient's state, and start the first stage of emergency management.
2. The patient is transferred to a trauma center (intensive care unit) and put in a recovery room where new measurements are taken, and immediate interventions are made, if needed.
3. The patient is either directed to an operating room or to a radiology room, followed by comprehensive care at the hospital.

Using a patient's records in stage 1, 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. Such models intend to give support to clinicians and professionals. 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 [HGD+14]. 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 [HGP+15]. 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, now has data from more than 7000 trauma cases, and is continually updated. The granularity of collected data makes this dataset unique in Europe. However, the data is highly heterogeneous, as it comes from multiple sources, and furthermore, is often missing, which makes modeling challenging. More precisely, 250 quantitative and categorical variables contain numerous missing values, coded in various different ways: NA for Not Applicable, Imp for

Impossible, NR for Not Recorded, NM for Not Made, etc.

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. We first introduce the basic notation and assumptions that we use throughout the paper.

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:

$$\mathbb{P}\left(y_i = 1 | x_i; \beta\right) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}, \quad i = 1, \ldots, n, \tag{3.1}$$

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 $r_i = (r_{ij}, 1 \leq j \leq p)$, with $r_{ij} = 1$ if x_{ij} is missing and $r_{ij} = 0$ otherwise. The matrix $r = (r_i, 1 \leq i \leq n)$ then defines the missing data pattern. The missing data mechanism is characterized by the conditional distribution of r given x and y, with parameter ϕ, i.e., $\mathtt{p}(r_i | x_i, y_i, \phi)$. Throughout this paper, we assume the Missing at Random (MAR) mechanism which implies that the

missing values mechanism can therefore be ignored [LR02] and the maximum likelihood estimate of θ can be obtained by maximizing $\mathcal{LL}(\theta; y, x_{\text{obs}})$. A reminder of these concepts is given in the Appendix A.1.

4. Parameter estimation using SAEM.

4.1. *The EM and MCEM algorithms.* We aim to estimate the parameter θ 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 θ_0, iteration k updates θ_{k-1} to θ_k with the following two steps:

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

$$\begin{aligned} Q_k(\theta) &= \mathbb{E}[\mathcal{LL}(\theta; x, y)|x_{\text{obs}}, y; \theta_{k-1}] \\ &= \int \mathcal{LL}(\theta; x, y)\mathtt{p}(x_{\text{mis}}|x_{\text{obs}}, y; \theta_{k-1})dx_{\text{mis}}. \end{aligned} \tag{4.1}$$

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

Since the expectation (4.1) in the E-step for the logistic regression model has no explicit expression, MCEM [WT90, ICL99] can be used. The E-step of MCEMf 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.

4.2. *The SAEM algorithm.* To achieve improved computational efficiency, the SAEM algorithm [Lav14] replaces the E-step (4.1) by a stochastic approximation based on a single simulation of x_{mis}. Starting from an initial guess θ_0, the kth iteration consists of three steps:

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

$$\mathtt{p}(x_{i,\text{mis}}|x_{i,\text{obs}}, y_i; \theta_{k-1}). \tag{4.2}$$

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

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left(\mathcal{LL}(\theta; x_{\text{obs}}, x_{\text{mis}}^{(k)}, y) - Q_{k-1}(\theta) \right), \tag{4.3}$$

  where γ_k is a decreasing sequence of positive numbers.
- **Maximization:** update the estimation of θ:

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

The choice of the sequence (γ_k) in (4.3) is important for ensuring the almost sure convergence of SAEM to a maximum of the observed likelihood [DLM99]. 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 be drawn exactly from its conditional distribution (4.2), which has no explicit 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 M iterations are then used as a sample from the target distribution. To define a proposal distribution for our MH algorithm, observe that the target distribution (4.2) 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:

$$x_{i,\text{mis}}|x_{i,\text{obs}} \sim \mathcal{N}_p(\mu_i, \Sigma_i), \tag{4.4}$$

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 μ for individual i. The covariance matrix Σ 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 [Lou82], 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 M samples $(x_{i,\text{mis}}^{(m)}, 1 \leq i \leq n, 1 \leq m \leq M)$ of the missing data drawn from the conditional distribution (4.2), the observed FIM can be estimated as $\hat{\mathcal{I}}_M(\hat{\theta}) = \sum_{i=1}^n -(D_i + G_i - \Delta_i \Delta_i^T)$, where

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

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

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

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.

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

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

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 $\beta_{\mathcal{M}}$.

5.2. *Observed log-likelihood.* For a given model and parameter θ, the observed log-likelihood is, by definition,

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

For any i, the density $\mathtt{p}(y_i, x_{i,\text{obs}}; \theta)$ cannot 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 (4.4). 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 (4.4):

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

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

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

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

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 (3.1). 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 σ is the vector of standard deviations $\sigma = (1, 2, 3, 4, 5)$, and C the correlation matrix

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

Then we randomly introduced 10% missing values in the covariates first with the completely at random (MCAR) mechanism where each entry has the same probability to be observed. The code to reproduce these experiments is available on github – see Supplement A.

6.2. *The behavior of SAEM.* The algorithm was initialized with the parameters obtained after mean imputation, i.e., imputing missing entries of each variable with the mean of the variable over its observed values.

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 (γ_k), we fixed the value of $k_1 = 50$ and used $\tau = (0.6,\ 0.8,\ 1)$ during the next 250 iterations. Representative plots of the convergence of SAEM for the coefficient β_1, obtained from four simulated data sets, are shown in Figure 1. For larger τ, 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. The behavior of SAEM in estimating the other components of β was quite similar, as shown in Appendix A.4. We therefore use $\tau = 1$ in the following.

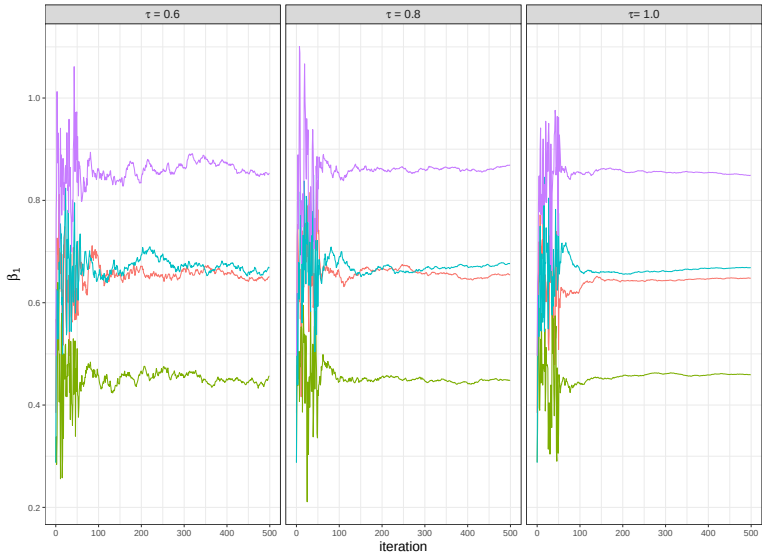

FIG 1. *Convergence plots for β_1 obtained with three different values of τ (0.6, 0.8, 1.0). Each color represents one simulation.*

6.3. *Comparison with other methods.* We ran 1000 simulations and compared SAEM to several other existing methods, intially in terms of estimation errors of the parameters. We considered *i)* the complete case (CC) method (all rows containing at least one unobserved data value were removed), *ii)* multiple imputation based on conditional modeling as implemented in the R package *mice* [vGO11] (with its default settings and Rubin's combining rules), and *iii)* the MCEM algorithm [ICL99] that we implemented using adaptive rejection sampling (MCEM-AR). We used the dataset without missing values (no NA) as a reference, with parameters estimated with the Newton-Raphson algorithm as implemented in the *glm* function in R.

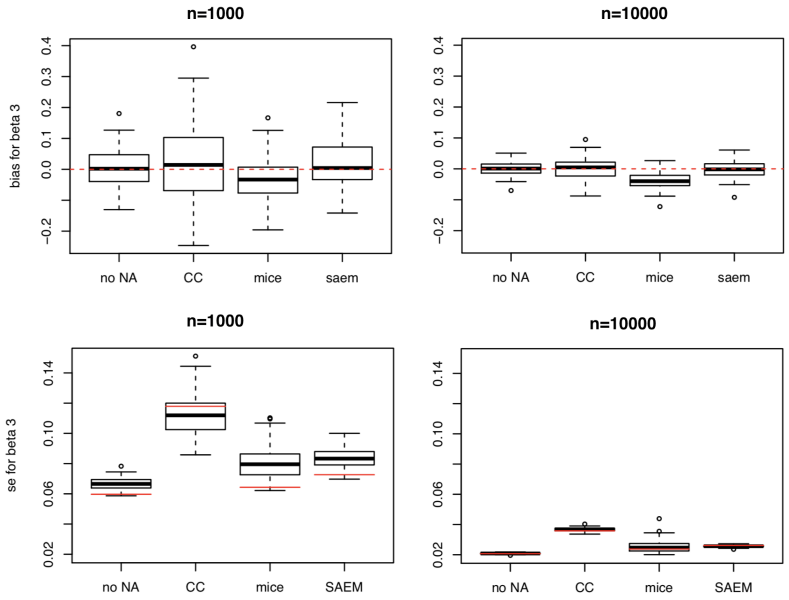

FIG 2. *Top: Empirical distribution of the estimates of β_3 obtained under MCAR. Bottom: Distribution of the estimated standard errors of $\hat{\beta}_3$ obtained under MCAR. For each method, the red line corresponds to the empirical standard deviation of $\hat{\beta}_3$ calculated over the 1000 simulations.*

Figure 2 (top) displays the distribution of the estimates of β_3, for $n = 1000$ and $n = 10\,000$ under MCAR mechanism and a strong correlation structure between covariates. This plot is representative of the results obtained with the other components of β. Note that here we do not show the results of MCEM since it spent more than one hour to converge with $n = 10\,000$ for one simulation. In fact, we show in Table 2 that even in the $n = 1000$ case, MCEM was computationally costly, and thus not recommended in

this situation. As expected, larger samples yielded smaller bias. Moreover, we observe that the estimation obtained by multiple imputation 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 the same simulations as in Figure 2 (top). 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 their variance, 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 $n = 10\,000$, calculated over 1000 simulations.

parameter	no NA	CC	mice	SAEM
β_0	95.2	94.4	94.4	94.5
β_1	94.8	94.6	86.3	95.3
β_2	94.8	94.3	85.7	94.1
β_3	94.5	94.2	77.0	94.7
β_4	94.1	92.0	86.9	92.9
β_5	95.2	92.3	78.8	93.0

Table 1 shows the coverage of the confidence interval for all parameters. We had expected coverage at the nominal 95% level. SAEM reached around 95% coverage, while *mice* struggled for certain parameters. Even though CC showed reasonable results in terms of coverage, the range of its confidence interval was still too large.

TABLE 2
Comparison of execution time between no NA, MCEM, mice, and SAEM with $n = 1000$ calculated over 1000 simulations.

Execution time (seconds)	no NA	MCEM	mice	SAEM
min	2.87×10^{-3}	492	0.64	9.96
mean	4.65×10^{-3}	773	0.70	13.50
max	43.50×10^{-3}	1077	0.76	16.79

Lastly, Table 2 highlights large differences between the methods in terms of execution time. MCEM was computationally intensive because in each iteration, it needed to generate a huge quantity of samples, while multiple imputation took less than 1 second per simulation, and SAEM around 13 seconds, which remains reasonable.

The results obtained when the covariates were not correlated are presented in Figure 9 in Appendix A.5. In this setting, SAEM could result in a non-zero estimation of the covariance, so its precision could be affected, but it could still out-perform CC and *mice*. The results obtained under a MAR mechanism are presented in Figure 10 in Appendix A.5. They were very similar to those presented here, except, of course, for the complete case method, which would be much more biased, especially in the case where the missingness in x is related to the outcome y.

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

6.4. *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 β are zero except $\beta_0 = -0.2$, $\beta_1 = 0.5$, $\beta_3 = 1$ and $\beta_5 = -0.6$. We compared the AIC_{obs} and BIC_{obs} based on the observed log-likelihood, as described in Section 5, to those based on the complete cases (AIC_{cc}, BIC_{cc}) and those obtained from the the original complete data (AIC_{orig}, 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
AIC_{obs}	60	40	0	65	32	3
AIC_{orig}	73	27	0	75	20	5
AIC_{cc}	67	32	1	77	16	7
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.

The results show that with AIC, there was a large possibility of selecting an overfitted model, while the BIC results were better. Therefore, in the following experiment with the Traumabase data set, we chose BIC to perform

model selection. These results are representative of those obtained with other simulation schemes.

7. The Traumabase application. One of the aims is to help emergency medics know whether patients will suffer from hemorrhagic shock when they arrives at the trauma center. Hemorrhagic shock remains the leading cause of death in severe trauma, and could be avoided if there were no delay in its detection and management. An optimized organization is essential to control blood loss as quickly as possible and to reduce mortality.

7.1. *Details on the dataset.* There were 7495 individuals in the trauma data we investigated, collected from May 2011 to March 2016. As suggested by doctors, patients were excluded if they had suffered from penetrating trauma (where an object such as a weapon or burns have pierced the skin and entered body tissue, creating an open wound), had been in pre-hospital traumatic cardiac arrest, or when no pre-hospital data was available. 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 are shown in Appendix A.6. Another reason to choose these variables is that they were all available in the ambulance, and therefore could be used in real situations.

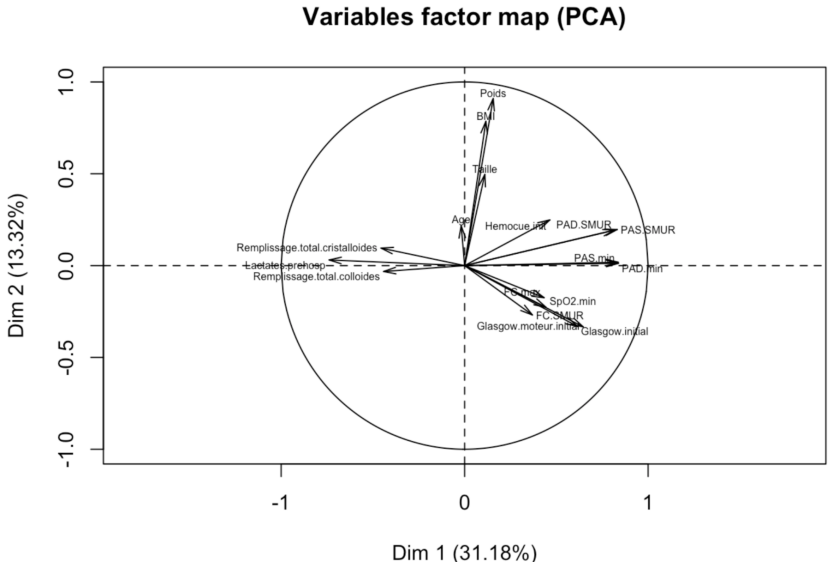

FIG 3. *The factor map of the variables from PCA.*

There was strong collinearity between variables, as can be seen in the variables PCA factor map (obtained by running an EM-PCA algorithm which performs PCA with missing values [JH16]) in Figure 3, in particular between the minimum systolic (PAS.min) and diastolic blood pressure (PAD.min).

Doctors preferred using the recoded variables, SD.min and SD.SMUR (SD.min = PAS.min − PAD.min; SD.SMUR = PAS.SMUR − PAD.SMUR) which have more clinical significance, and at the same time, allowed us to have fewer highly correlated variables, even though it was not necessarily an issue when performing the variable selection procedure. Thus, we had 14 variables to predict hemorrhagic shock.

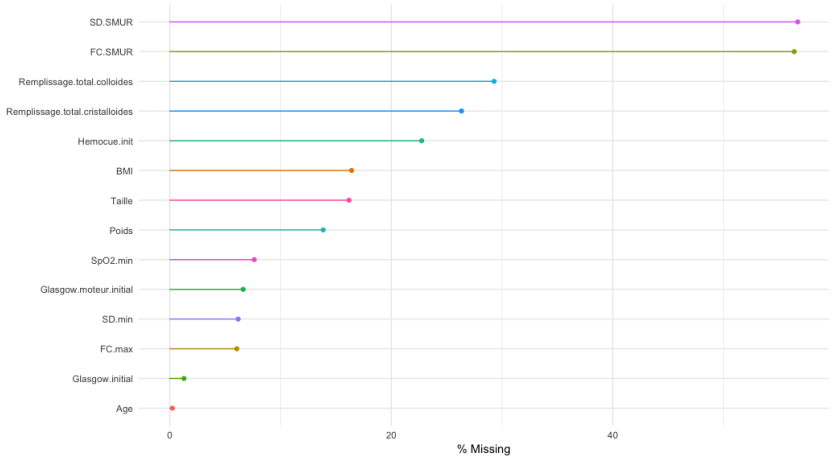

FIG 4. *Percentage of missing values in each variable.*

Figure 4 (visaulized by *naniar* package [TCMF18]) 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, as mentioned in Section 1, there are many types of missing data and also many reasons why missingness occurs, in the end, considering them all to be MAR remains a plausible assumption. FC.SMUR and SD.SMUR (heart beat and the difference between blood pressure measured when the ambulance arrives at the accident site) contain many missing values because doctors collected heart beat and blood pressure data only during transportation, but many other medical institutes and scientific publications performed analyses based on these measurement when the ambulance arrived. Consequently, it was 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, as implemented in the *mice* package. Next, we used the model selection procedure described in Section 5 based on the penalized observed log-likelihood. 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 5, patient number 3302 (circled in blue) is one of such outliers.

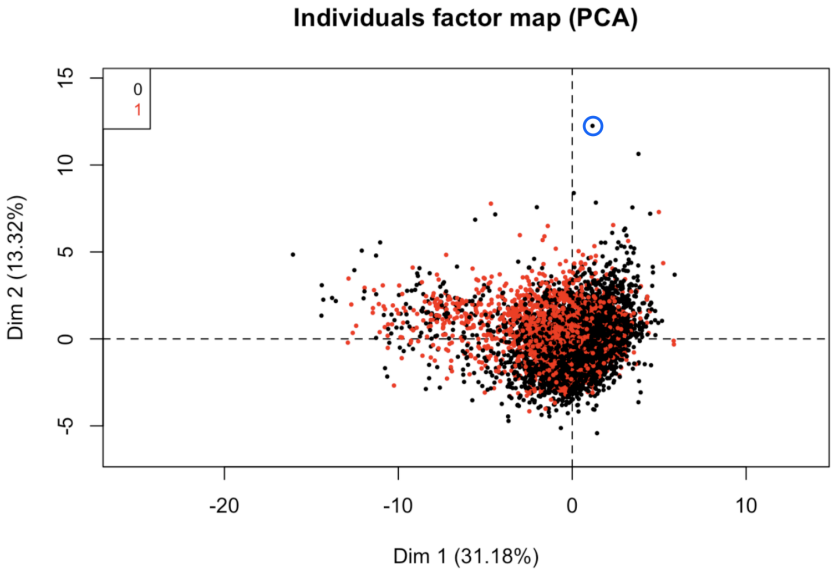

FIG 5. *Observation's factor map of PCA. The blue circle shows the outlier. Red points are hemorrhagic shock patients, black points are patients who did have hemorrhagic shock.*

7.2. *Predictive performance.* We divided the dataset into training and test sets. The training set contained a random selection of 80% of observations, and the test set contained the remaining 20%. In the training set, we selected a model with the suggested BIC with missing values, and used forward selection. Using the BIC (Figure 6), we selected a model with 8 variables. The estimation of parameters and their standard errors is shown in Table 4.

Variables	Estimate (se)
$(Intercept)$	-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)

TABLE 4
Estimation of β and its standard errors obtained by SAEM, using BIC as the model selection criterion.

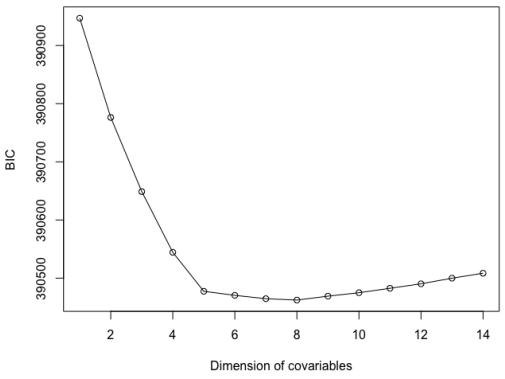

FIG 6
BIC as a function of the number of covariates involved in the regression model.

The Traumabase medical team indicated us that the signs of the coefficients are in agreement with their a priori ideas: Older people are more likely to have a hemorrhagic shock; and a low Glasgow score implies little or no motor response, which often is the case for hemorrhagic shock patients; One typical sign of hemorrhagic shock is rapid heart rate. 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. Therapy involving two types of volume expander: 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. Therefore, we chose a threshold of 0.1 in our analyses; i.e., individuals with a prediction probability higher than 0.1 were predicted as hemorrhagic shock patients. Note that the test set is also incomplete, and as far as we know, there is no standard solution to deal with this. We suggest a strategy based on the maximum a posteriori (MAP) – see Appendix A.7. The confusion matrix of the predictive performance on the test set is shown in Table 5. The associated ROC curve is shown in Figure 7, and has an AUC (area under the curve) of 0.88.

		Predicted outcome 1	0
Observed value	**1**	True Positive (84)	False Negative (24)
	0	False Positive (166)	True Negative (1003)

TABLE 5
Confusion matrix for prediction on test set.

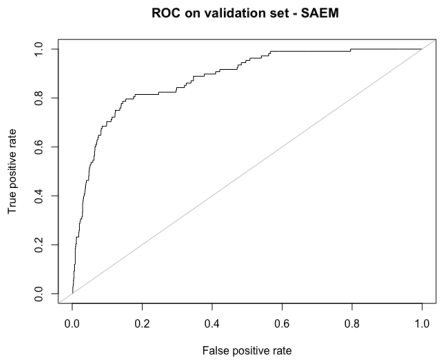

FIG 7
ROC curve of the test set predictions.

Finally we compared our method (SAEM with BIC on the training set and the use of MAP on the test set) to other approaches. We considered single imputation methods followed by variable selection on the imputed training dataset, such as single imputation by PCA (impPCA) implemented in the *missMDA* package [JH16], as well as mean imputation (impMean) as a naive baseline method. For multiple imputation (MI), 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 [WWR], where they kept the variables selected in each imputed dataset to define the final model. In the same way, there is no common solution to deal with missing values in the test set. For MI, impPCA and impMean, we imputed the test set and then applied the model that had been selected on the training set.

Table 6 compare the methods in terms of their predictive performance. An initial observation is that there is no large difference between the approaches. Nevertheless, our method has a great advantage over mean imputation in terms of true negative performance, and also outperforms PCA imputation in terms of true positives. We also see that multiple imputation performed by *mice* gives prediction performance, and that there is no obvious advantage for SAEM. However, since we are focused more on reducing the false negative rate, the SAEM result is more relevant to clinical needs of emergency doctors. In addition, one of the advantages of our methodology is that, from estimation to selection and prediction on a test sample with missing data, it is theoretically well-founded.

TABLE 6
Comparison of the predictive performance of different methods dealing with missing data. The sensitivity is defined as the true positive rate; (1-specificity) as the false positive rate; and the accuracy as the number of true positive and true negative over the total number of observations.

Variables	SAEM	impMean	impPCA	*mice*
$True\ Positive$	**84**	**84**	81	81
$True\ Negative$	1003	987	1004	**1010**
$Accuracy$	**0.852**	0.838	**0.850**	**0.854**
$Sensitivity$	**0.778**	**0.778**	0.750	0.750
$1 - Specificity$	0.858	0.844	0.859	**0.863**
$Precision$	**0.336**	0.316	0.329	**0.338**
AUC	0.882	0.882	0.882	0.882

8. Discussion. The EM algorithm can be considered a natural solution for dealing with missing values, as it provides maximum likelihood estimates despite missing values in the MAR setting. However, it is often impossible to directly compute the expectation in the E-step; SAEM offers a good solution to this. In this paper, we have developed a comprehensive framework for logistic regression with missing values. Our experiments indicate that our method is computationally efficient, and can be easily implemented. In addition, compared with multiple imputation implemented in the *mice* package – especially in the case with correlation between variables – estimation using SAEM is unbiased and leads to accurate coverage of the confidence interval. Based on our algorithm, model selection by BIC with missing data can be performed in a natural way.

The approach we suggest assumes that the covariates follow a normal distribution, and the performance of the method could be improved by applying certain variable transformations. Paths for possible future research include further developing the method to handle quantitative and categorical data. In addition, in the trauma data set, 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 our 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

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:

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

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,

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

MAR assumption implies that, the observed likelihood can be maximize and the distribution of r can be ignored [LR02]. Indeed,

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

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

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.

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

A.3. Calculation of observed information matrix. Procedure 2 shows how we calculate the observed information matrix by using MH sampling.

Procedure 2 Calculation of observed information matrix.

Input: MH samples $(x_{i,\text{mis}}^{(m)}, 1 \le i \le n, 1 \le m \le M)$ for unobserved data $(x_{i,\text{mis}}, 1 \le i \le n)$, $\Delta_i = \mathbf{0}$, $D_i = \mathbf{0}$, $G_i = \mathbf{0}$, $\hat{\mathcal{I}}_M(\hat{\beta}) = \mathbf{0}$;
for $n = 1, 2, \cdots, n$ **do**
  for $m = 1, 2, \cdots, M$ **do**
    Calculate the gradient $\nabla f_{im} = \frac{\partial \mathcal{LL}(\theta|x_{i,\text{obs}}, z_{im}, y_i)}{\partial \beta}$;
    Calculate the Hessian matrix $H_{im} = \frac{\partial^2 \mathcal{LL}(\theta|x_{i,\text{obs}}, z_{im}, y_i)}{\partial \beta \partial \beta^T}$;
    $\Delta_i \leftarrow \frac{1}{m}[(m-1)\Delta_i + \nabla f_{im}]$;
    $D_i \leftarrow \frac{1}{m}[(m-1)D_i + H_{im}]$;
    $G_i \leftarrow \frac{1}{m}[(m-1)G_i + \nabla f_{im} \nabla f_{im}^T]$;
  end for
  $\hat{\mathcal{I}}_M(\hat{\beta}) \leftarrow \hat{\mathcal{I}}_M(\hat{\beta}) - (D_i + G_i - \Delta_i \Delta_i^T)$;
end for
Output: $\hat{\mathcal{I}}_M(\hat{\beta})$.

A.4. Behavior of SAEM: convergence plots for all betas. Figure 8 shows the convergence plot for all the β in one simulation.

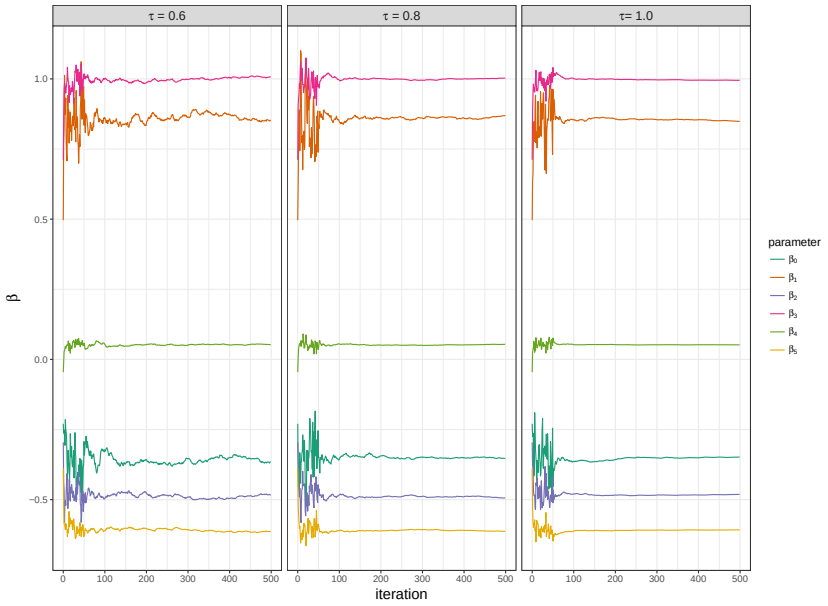

FIG 8. *Convergence plots for all β in SAEM. Each color represents one parameter.*

A.5. Simulation results.

Existence of correlation. Figure 9 (left) shows the results of estimation in the case without correlation. SAEM is a little biased since it estimates non-zero terms for the covariance, but it stills outperforms CC and *mice*.

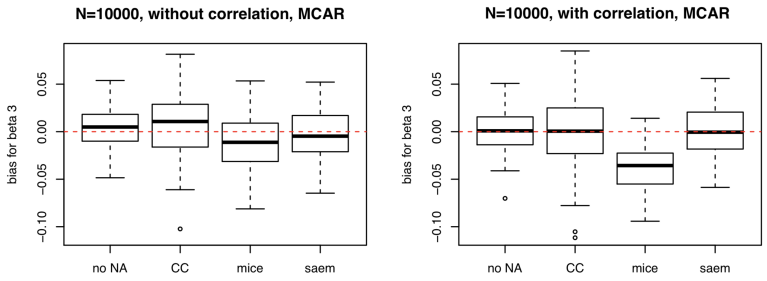

FIG 9. *Empirical distribution of the estimates of β_3 obtained under MCAR, with $n = 10\,000$ and 10% of missing values; left: no correlation between the covariates; right: the covariates are correlated.*

MAR mechanism. We introduce 10% of missing values in the covariates according to different MAR mechanisms: first missing values are introduced

in some covariates according to a logistic regression model on other covariates; second missing values are introduced in some covariates and depend both on other covariates and on the response variable.

Figure 10 (left) shows that the results are very similar to the ones obtained under a MCAR mechanism. However, the CC method is inaccurate when the missingness depends on the outcome y (right).

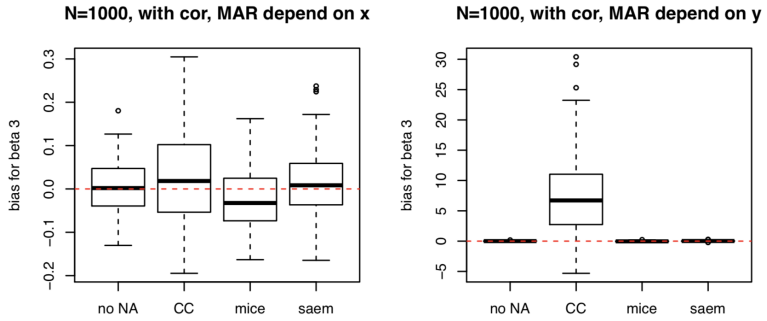

FIG 10. *Empirical distribution of the estimates of β_3 obtained under MAR mechanism, with N=1000 and 10% of missing values; left : missingness only depends on covariates x; right: missingness also depends on y.*

A.6. 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*: A neurological scale which aims to give a reliable and objective way of recording the conscious state of a person.
- *Glasgow.moteur*: Initial Glasgow for motor response.
- *PAS.min*: The minimum systolic blood pressure. The systolic value corresponds to the minimum pressure encountered during contraction of the chambers of the heart (or systole), in centimeters of mercury (mmHg).
- *PAD.min*: The minimum diastolic blood pressure. Diastolic blood pressure number or the bottom number indicates the pressure in the arteries when the heart rests between beats.
- *FC.max*: The maximum number of heartbeat (or pulse) per unit time

(usually a minute).

- $PAS.SMUR$: Systolic blood pressure at one time point.
- $PAD.SMUR$: Diastolic blood pressure at one time point.
- $FC.SMUR$: Heart rate at one time point.
- $Hemocue.init$: Initial Hemoglobin measured by blood sample on finger top.
- $SpO2.min$: An estimate of the amount of oxygen in the blood. It represents the percentage of oxygenated hemoglobin relative to the total amount of hemoglobin in the blood.
- $RT.cristalloides$ and $RT.colloides$: Prehospital volume expander. There are two main types: crystalloids and colloids.
- $SD.min$ $(= PAS.min - PAD.min)$: Pulse pressure for the minimum value of diastolic and systolic blood pressure.
- $SD.SMUR$ $(= PAS.SMUR - PAD.SMUR)$: Pulse pressure on arrival in the ambulance.

A.7. Prediction on test set with SAEM. Suppose an observation in test set is $x = (x_{\text{obs}}, x_{\text{mis}})$, we want to predict the binary response y. We consider two methods to obtain a predictive value for SAEM.

1. First impute the missing value by maximum a posteriori

$$\hat{x}_{mis} = \arg\max_{x_{\text{mis}}} \mathtt{p}(x_{\text{mis}}|x_{\text{obs}}),$$

then predict the value of hemorrhagic shock according to

$$\mathtt{p}(\hat{y} = 1) = \frac{\exp[(x_{\text{obs}}, \hat{x}_{mis})^T \beta]}{1 + \exp[(x_{\text{obs}}, \hat{x}_{mis})^T \beta]}.$$

2. With M Monte Carlo samples

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

we estimate directly the response by maximum a posteriori

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

ACKNOWLEDGEMENTS

The authors thank Sophie HAMADA and Tobias GAUSS for their help with the interpretation of Traumabase data.

SUPPLEMENTARY MATERIAL

Supplement A: Code R to reproduce the experiments (https://github.com/wjiang94/miSAEM_logReg).

REFERENCES

[CC08] Gerda Claeskens and Fabrizio Consentino. Variable selection with incomplete covariate data. *Biometrics*, 64:1062–9, 04 2008.

[CC11] Fabrizio Consentino and Gerda Claeskens. Missing covariates in logistic regression, estimation and distribution selection. *Statistical Modelling*, 11(2):159–183, 2011.

[DC17] GBD 2016 DALYs and HALE Collaborators. 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(10100):1260 – 1344, 2017.

[DLM99] Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a stochastic approximation version of the em algorithm. *The Annals of Statistics*, 27(1):94–128, 1999.

[DLR77] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[GW92] Wally R. Gilks and Pascal P. Wild. Adaptive rejection sampling for gibbs sampling. *Appl. Statist*, 41(2):337–348, 1992.

[HGD+14] Sophie Rym Hamada, Tobias Gauss, François-Xavier Duchateau, Jennifer Truchot, Anatole Harrois, Mathieu Raux, Jacques Duranteau, Jean Mantz, and Catherine 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(6):1476–1483, 2014.

[HGP+15] Sophie Rym Hamada, Tobias Gauss, Jakob Pann, Martin W. Dünser, Marc Léone, and Jacques Duranteau. European trauma guideline compliance assessment: the etrauss study. *Critical care*, 19:423, 2015.

[ICL99] Joseph G. Ibrahim, Ming-Hui Chen, and Stuart R. Lipsitz. Monte carlo em for missing covariates in parametric regression models. *BIOMETRICS*, 55:591–596, 1999.

[ICLH05] Joseph G. Ibrahim, Ming-Hui Chen, Stuart R. Lipsitz, and Amy H. Herring. Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469):332–346, 2005.

[JH16] Julie Josse and François Husson. missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31, 2016.

[JNR15] Jiming Jiang, Thuan Nguyen, and J. Sunil Rao. The e-ms algorithm: Model selection with incomplete data. *Journal of the American Statistical Association*, 110(511):1136–1147, 2015.

[Lav14] Marc Lavielle. *Mixed Effects Models for the Population Approach: Models, Tasks, Methods and Tools*. Chapman and Hall/CRC, 2014.

[Lou82] Thomas A. Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233, 1982.

[LR02] Roderick J.A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., 2002.

[LWFW16] Ying Liu, Yuanjia Wang, Yang Feng, and Melanie M. Wall. Variable selection and prediction with incomplete high-dimensional data. *Ann. Appl. Stat.*, 10(1):418–450, 03 2016.

[MK08] Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. Wiley series in probability and statistics. Wiley, Hoboken, NJ, 2. ed edition, 2008.

[MR91] Xiao-Li Meng and Donald B. Rubin. Using em to obtain asymptotic variance-covariance matrices: The sem algorithm. *Journal of the American Statistical Association*, 86(416):899–909, 1991.

[Rub76] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[Rub78] Donald B Rubin. Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association*, volume 1, pages 20–34. American Statistical Association, 1978.

[SGJC13] Shaun Seaman, John Galati, Dan Jackson, and John Carlin. What is meant by "missing at random"? *Statist. Sci.*, 28(2):257–268, 05 2013.

[TCMF18] Nicholas Tierney, Di Cook, Miles McBain, and Colin Fay. *naniar: Data Structures, Summaries, and Visualisations for Missing Data*, 2018. R package version 0.2.0.

[vGO11] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011.

[WT90] Greg C. G. Wei and Martin 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(411):699–704, 1990.

[WWR] Angela M. Wood, Ian R. White, and Patrick Royston. How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, 27(17):3227–3246.

Wei Jiang,
CMAP, Ecole Polytechnique,
Route de Saclay, 91128 Palaiseau, France
E-mail: wei.jiang@polytechnique.edu
URL: https://sites.google.com/site/weijiangstat/

Julie Josse,
CMAP, Ecole Polytechnique,
Route de Saclay, 91128 Palaiseau, France
E-mail: julie.josse@polytechnique.edu
URL: http://juliejosse.com/

Marc Lavielle,
CMAP, Ecole Polytechnique,
Route de Saclay, 91128 Palaiseau, France
E-mail: marc.lavielle@inria.fr
URL: http://www.cmap.polytechnique.fr/l̃avielle/