Logistic Regression with Missing Covariates – Parameter Estimation, Model Selection and Prediction

Wei Jiang, Julie Josse, Marc Lavielle

Logistic Regression with Missing Covariates – Parameter Estimation, Model Selection and Prediction

Wei Jiang, Julie Josse, Marc Lavielle
Inria XPOP and CMAP, École Polytechnique, France
and
TraumaBase® Group (Tobias Gauss, Sophie Hamada)
Hôpital Beaujon, APHP, France

December 14, 2018

Abstract

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 develop a complete approach, including the estimation of parameters and variance of estimators, derivation of confidence intervals and a model selection procedure, for cases where the missing values can be anywhere in covariates. By well organizing different patterns of missingness in each observation, 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 also tackle the problem of prediction for a new individual with missing values, which is never addressed. 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 multiple imputation by chained equation can lead to biased estimates while our method is unbiased. 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, variable selection, major trauma, 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 is modifying an estimation process so that it can be applied to incomplete data. For example, one can use the EM algorithm (Dempster et al., 1977) to obtain the maximum likelihood estimate (MLE) despite missing values, and a supplemented EM algorithm (SEM) (Meng and Rubin, 1991) or Louis' formula (Louis, 1982) for the variance of the estimate. This strategy is valid under missing at random (MAR) mechanisms (Little and Rubin, 2002; Seaman et al., 2013), 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 it often happens that the expectation step of the EM algorithm involves infeasible computations. One solution in the framework of generalized linear models, suggested in Ibrahim et al. (1999) and Ibrahim et al. (2005), is to use a Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990; McLachlan and Krishnan, 2008), replacing the integral by its empirical sum using Monte Carlo sampling. Ibrahim et al. (1999) also estimate the variance using a Monte Carlo version of Louis' formula. For sampling, Ibrahim et al. (1999) used Gibbs samplers with an adaptive rejection sampling scheme (Gilks and Wild, 1992). 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 new statistical methodology for logistic regression with missing values where the missing data can be anywhere in the covariates. By well organizing different patterns of missingness in each observation, we derive a stochastic approximation version of the EM algorithm (SAEM) (Lavielle, 2014), based on Metropolis-Hasting sampling, to perform statistical inference for logistic regression with incomplete data. 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: it takes 5 minutes to estimate parameters with MCEM in a dataset of size 200×5, compared to a few seconds for SAEM as illustrated in our simulation. 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, Claeskens and Consentino (2008) and Consentino and Claeskens (2011) suggested an approximation of AIC, while Jiang et al. (2015) defined generalized information criteria and adaptive fence, and in the framework of imputation with Random Lasso, Liu et al. (2016) 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 French multicenter prospective Trauma Registry. Section 3 presents the assumptions and notations used throughout this paper. In Section 4, we derive an algorithm SAEM to obtain the maximum likelihood estimate of parameters in a logistic regression model for continuous 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 individual containing missing values in covariates. Section 6 presents a simulation study where our approach is compared to alternative methods such as multiple imputation (Rubin, 2009), which may suffer from biases and under-coverage. In Section 7, we apply our 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 vary 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 our work and provides a discussion.

Our contribution is to provide a complete methodology with theoretical foundation and computational efficiency, to perform logistic regression with missing values, available

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

to users, which have never existed, as far as we know. The methodology presented in this article is implemented as an R (R Core Team, 2017) package *misaem*, available in CRAN: `https://CRAN.R-project.org/package=misaem`. The code to reproduce all the experiment is also provided in GitHub: `https://github.com/wjiang94/miSAEM_logReg`.

2 Example

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 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 (Hay et al., 2017). 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.

Major trauma is comprised of several stages:

1. At the accident site where a patient is taken care of by paramedics and/ or doctors. A first assessment is made, and immediate emergency management is provided.

2. The patient is transferred to the resuscitation room of a trauma center, for a profound assessment and stabilization of vital functions as needed.

3. The patient is oriented to further care either to the operating theatre/ interventional radiology, the Intensive Care Unit or ward, 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 (Hamada et al., 2014). 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 (Hamada et al., 2015). 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, a multicenter prospective Trauma registry, now has data from more than 7000 trauma cases, and is continually updated. The granularity of collected data (with more than 250 variables) 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.

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 $r_i = (r_{ij}, 1 \leq j \leq p)$, 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 (Little and Rubin, 2002) 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 by 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

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

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

Since the expectation (2) in the E-step for the logistic regression model has no explicit expression, MCEM (Wei and Tanner, 1990; Ibrahim et al., 1999) 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 (Lavielle, 2014) which replaces the E-step (2) by a stochastic approximation. Note that, SAEM often deals with data (x, z), s.t., x is fully observed while z is an unknown variable. Here we assume missing data everywhere, and as a result, each observation may have a different pattern of missingness.

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

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

- **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 (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) is important for ensuring the almost sure convergence of SAEM to a maximum of the observed likelihood (Delyon et al., 1999). 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 (3), 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 (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 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 (Louis, 1982), 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 (3), 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 and prediction

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 $\beta_{\mathcal{M}}$. Note that, contrary to our approach, the existing method Claeskens and Consentino (2008) and Consentino and Claeskens (2011) use an approximation of AIC without estimating the observed likelihood.

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 (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):

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

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 $\mathcal{LL}(\theta; x_{\text{obs}}, y)$.

5.3 Prediction on test set with missing values

In supervised learning, after applying a model on the training set, a natural step is to evaluate the prediction performance, which can be done with a test set. Assume $x = (x_{\text{obs}}, x_{\text{mis}})$ an observation in the test set, we want to predict the binary response y. One important point is that test set has the same distribution as the training set and consequently also contains missing values. Therefore, we can't directly apply the fitted model to predict y for the observation 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 M Monte Carlo samples

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

we estimate directly the response by maximum a posteriori

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

Note that in the literature there are not many solutions to deal with the missing values in the test set. In the Subsection 7.2, we compare our 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 σ is the vector of standard deviations $\sigma = (1, 2, 3, 4, 5)$, and C the correlation matrix

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

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, provided in supplementary material.

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 450 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.

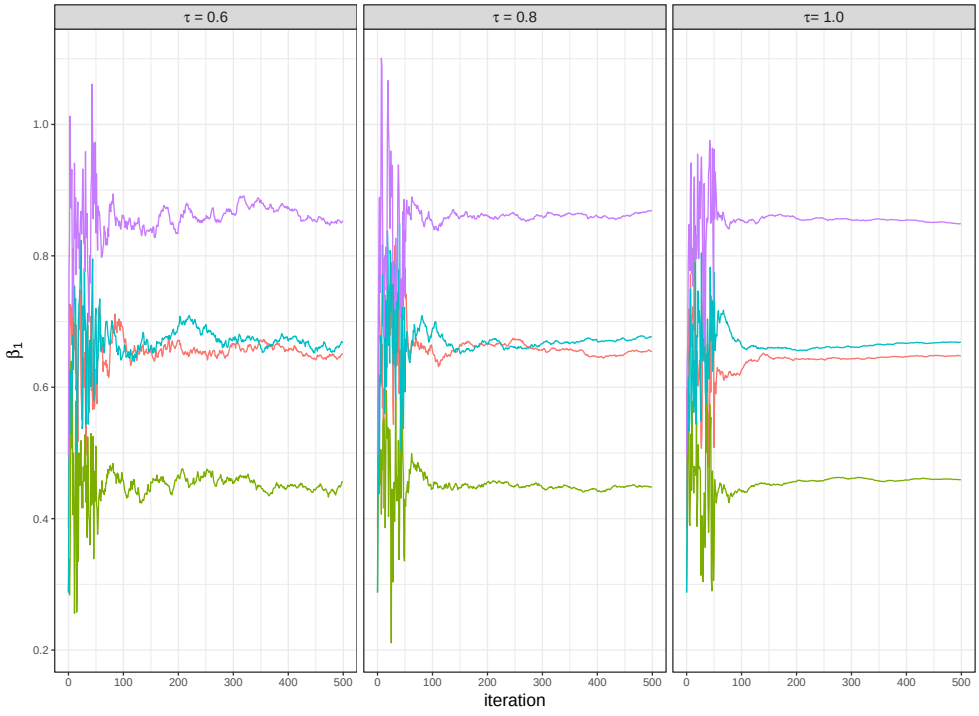

Figure 1: Convergence plots for β_1 obtained with three different values of τ (0.6, 0.8, 1.0). Each color represents one simulation. The true value of $\beta_1 = 0.5$.

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 (van Buuren and Groothuis-Oudshoorn, 2011). 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., regression models are used for quantitative variables; logistic regression models are used for categorical variables and uncertainty of the parameters is reflected within the Bayesian framework. More details are in van Buuren and Groothuis-Oudshoorn (2011). 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 β_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 Figure 13 in Appendix A.5. This plot is representative of the results obtained with the other components of β. As expected, larger samples yielded smaller bias. 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 shows the coverage of the confidence interval for all parameters and inside the

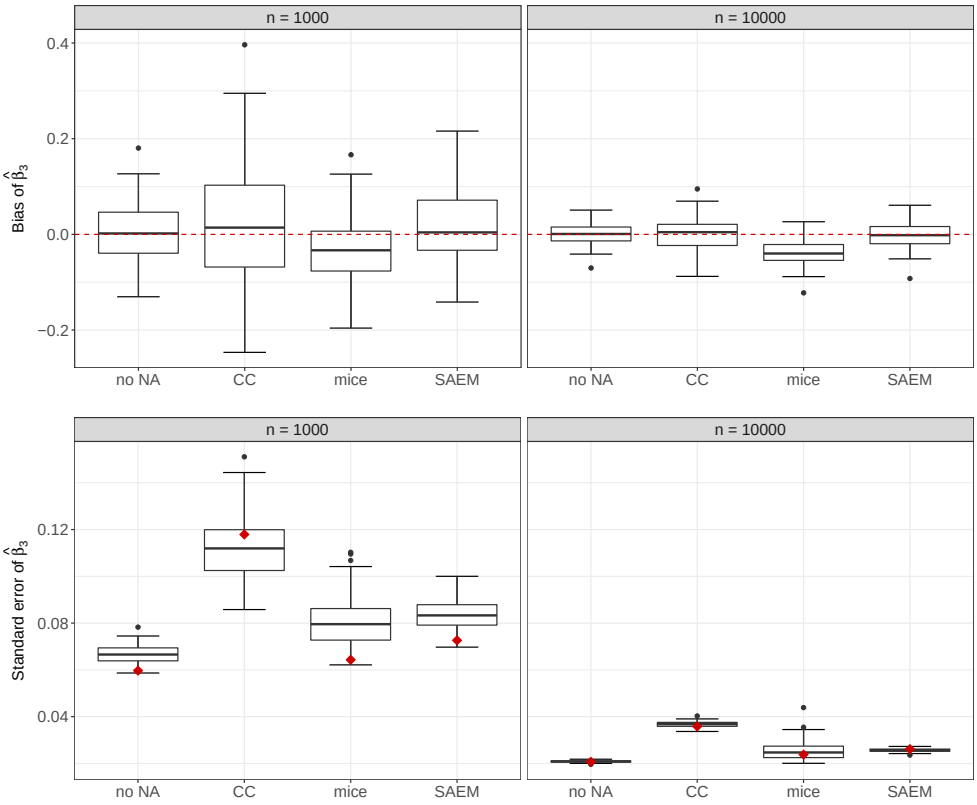

Figure 2: Top: Empirical distribution of bias of $\hat{\beta}_3$. Bottom: Distribution of the estimated standard errors of $\hat{\beta}_3$; for each method, the red point corresponds to the empirical standard deviation of $\hat{\beta}_3$ calculated over the 1000 simulations. Results for 10% MCAR and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

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
β_0	95.2 (21.36)	94.4 (27.82)	95.2 (22.70)	94.9 (22.48)
β_1	96.0 (18.92)	94.7 (24.65)	93.9 (21.77)	95.1 (21.51)
β_2	95.5 (9.53)	94.6 (12.41)	94.0 (10.97)	94.3 (10.83)
β_3	94.9 (8.17)	94.3 (10.66)	**86.5** (9.03)	94.7 (9.03)
β_4	94.6 (4.00)	94.2 (5.21)	96.2 (4.49)	95.4 (4.42)
β_5	95.9 (5.52)	94.4 (7.19)	**89.6** (6.20)	94.7 (6.17)

parentheses is the average length of corresponding confidence interval. 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 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 Table 6 in Appendix A.5.

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

Execution time (seconds) for one simulation	no NA	MCEM	mice	SAEM
$\boldsymbol{n = 1000}$				
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
$\boldsymbol{n = 200}$				
min	1.26×10^{-3}	67.91	0.24	2.64
mean	2.32×10^{-3}	291.47	0.28	3.91
max	21.53×10^{-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 (Ibrahim et al., 1999), available in GitHub provided in supplementary material, using adaptive rejection sampling. MCEM was computationally intensive because in each iteration, it needed to generate a huge quantity of samples, and thus not recommended in this situation. 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 Figure 13 in the Appendix A.5. Due to this computational difficulty, we didn't perform MCEM to compare with others in the experiments with larger sample sizes.

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.

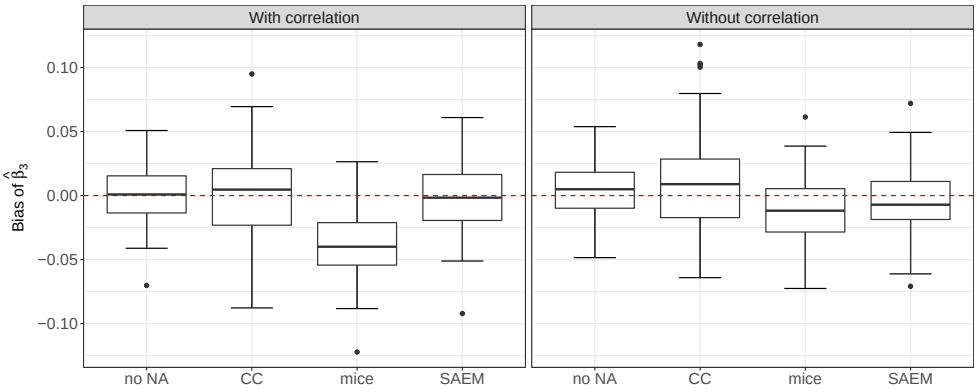

Figure 3: Empirical distribution of the estimates of β_3 obtained under MCAR, with $n = 10\,000$ and 10% of missing values; left: the covariates are correlated; right: no correlation between the covariates.

Meanwhile, We considered MAR mechanism. We introduced 10% of missing values in the covariates according to different MAR mechanisms: *i)* Missing values are introduced in some covariates according to a logistic regression model on other covariates; *ii)* missing

values are introduced in some covariates and missingness depends both on other covariates and on the response variable. Details are given in the implementation in GitHub.

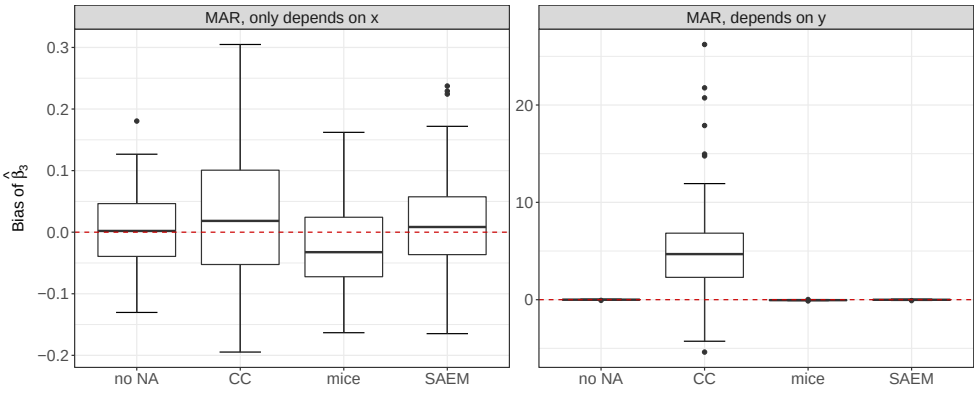

Figure 4: Empirical distribution of the bias of $\hat{\beta}_3$ obtained under MAR mechanism, with $n = 1000$ and 10% of missing values; left: missingness only depends on covariates x; right: missingness depends both on covariates x and on response y.

Figure 4 (left) shows that the biases were very similar to the ones obtained under a MCAR mechanism, expect for the CC method, which would be much more biased, especially in the case where the missingness in x was related to the outcome y, as shown in Figure 4 (right).

With 30% of missing values, the results (not shown here) were similar to the case with 10% missing data.

In summary, not only did these simulations allow us to verify that SAEM lead to unbiased estimators, but also they ensured that we made 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 illustrate 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 dataset, we chose BIC to perform model selection. These results are representative of those obtained with other simulation schemes.

6.5 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 1000×5. We also generated a test set of size 100×5.

We compared our 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 of column (impMean) *ii)* imputation by PCA (impPCA) (Josse and

Husson, 2016) which is based on low-rank assumption of the data matrix to impute. 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 our method does not encounter this issue.

In the framework of logistic regression, another method to perform imputation (impSAEM) could be considered, where the missing values of the test set are imputed with the conditional expectation of the missing entries given the observed values and the parameters estimated on the training set by SAEM. Due to the normal assumption of the covariates, it boils down to imputing the missing values with: $\hat{x}_{i,\text{mis}} = \hat{\mu}_{i,\text{mis}} + \hat{\Sigma}_{i,\text{mis,obs}}\hat{\Sigma}^{-1}_{i,\text{obs,obs}}(x_{i,\text{obs}} - \hat{\mu}_{i,\text{obs}})$, then to predict the probabilities with: $\widehat{\mathtt{p}(y_i = 1)} = \frac{\exp[(x_{i,\text{obs}},\hat{x}_{i,\text{mis}})^T\hat{\beta}]}{1+\exp[(x_{i,\text{obs}},\hat{x}_{i,\text{mis}})^T\hat{\beta}]}$.

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 (Brier, 1950) and Logarithmic score (Good, 1952). Figure 5 shows that on average, marginalizing over

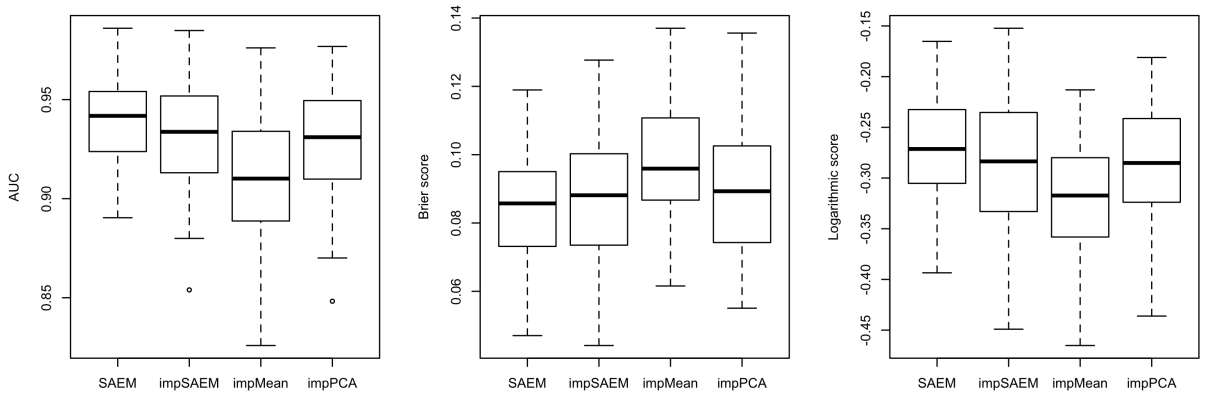

Figure 5: Comparison of empirical distribution of AUC, Brier score and Logarithmic score obtained on the test set, for our approach SAEM without imputation, impSAEM, impMean and impPCA, over 100 simulations.

distribution of missing values has the best performances: it gave the largest AUC and Logarithmic score, and the smallest Brier scores.

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

There were 7495 individuals in the trauma data we investigated, collected from May 2011 to March 2016. The study group decided to focus on patients with blunt trauma to be able to compare to the existing prediction rules. Patients with pre-hospital 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 are shown in Appendix A.6. These variables were chosen because they were all available to the pre-hospital team, and therefore could be used in real situations.

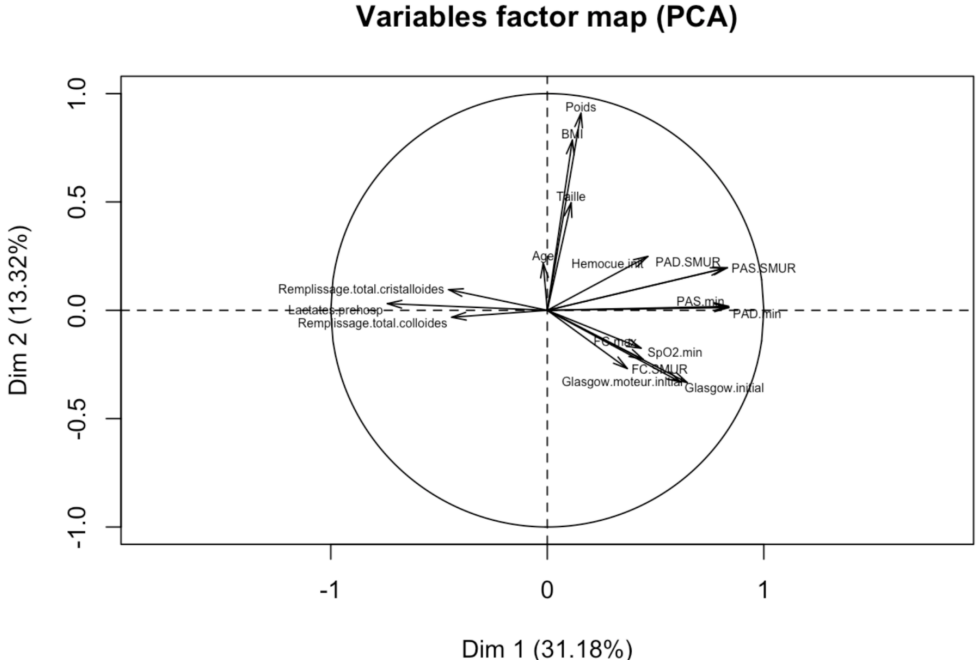

Figure 6: The factor map of the variables from PCA.

There was strong collinearity between variables, as can be seen in the variables PCA factor map (obtained by running an EM-PCA algorithm (Josse and Husson, 2016) 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 (Hamada et al., 2018). Thus, we had 14 variables to predict hemorrhagic shock.

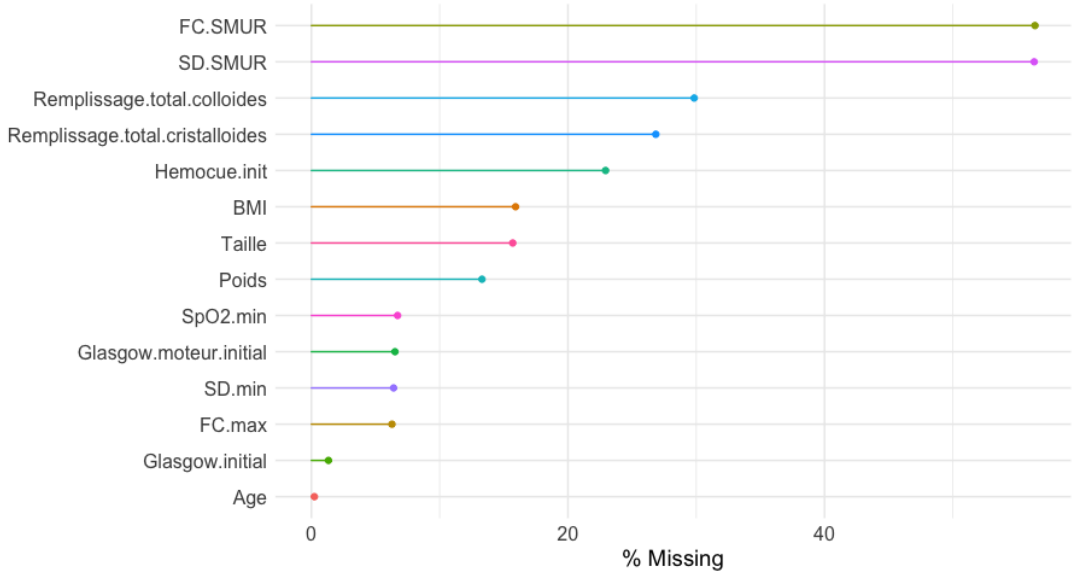

Figure 7: Percentage of missing values in each variable.

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

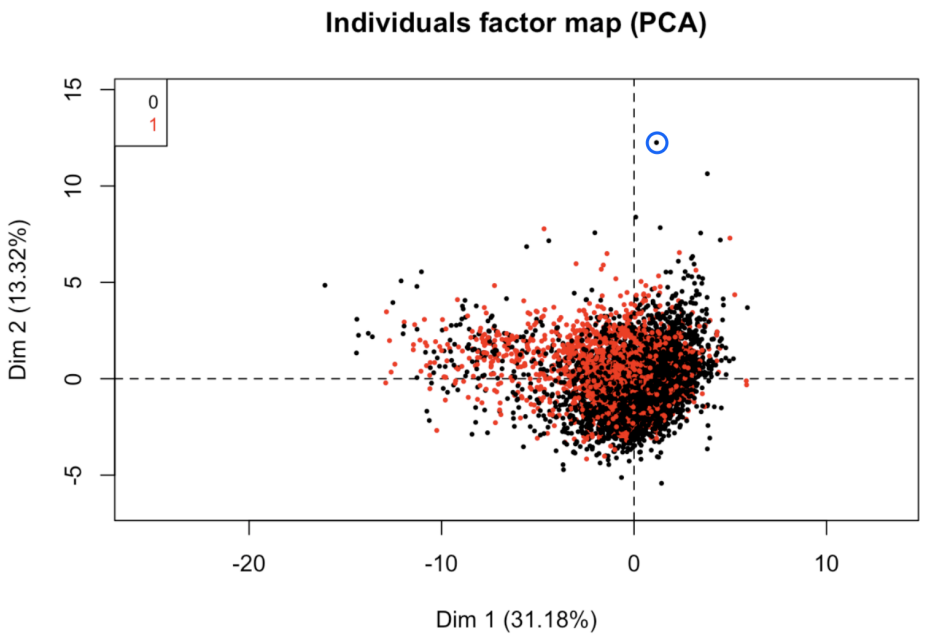

Figure 8: Observation's factor map of PCA. The blue circle shows the outlier. Red points are hemorrhagic shock patients, and black points are patients who did not have hemorrhagic shock. Patient number 3302 (circled in blue) has wrong calculation of BMI.

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 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 20%. In the training set, we selected a model with the suggested BIC with missing values, and used

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.

forward selection. Using the BIC, we selected a model with 8 variables. The estimates of parameters and their standard errors are shown in Table 4.

The TraumaBase medical team indicated 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)* And 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 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. 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 0.8865.

		Predicted outcome	
		1	0
Observed value	**1**	True Positive (109)	False Negative (14)
	0	False Positive (293)	True Negative (859)

Table 5: Confusion matrix for prediction on test set.

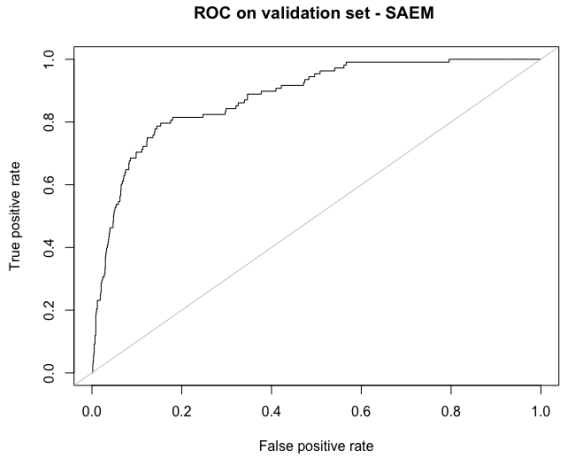

Figure 9: ROC curve of the test set predictions.

7.3 Comparison with other approaches

Finally we compared our 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) (Josse and Husson, 2016), imputation by Random Forest (missForest) (Stekhoven and Buehlmann, 2012), 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 (Stekhoven and Buehlmann, 2012). 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. (2008), 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.

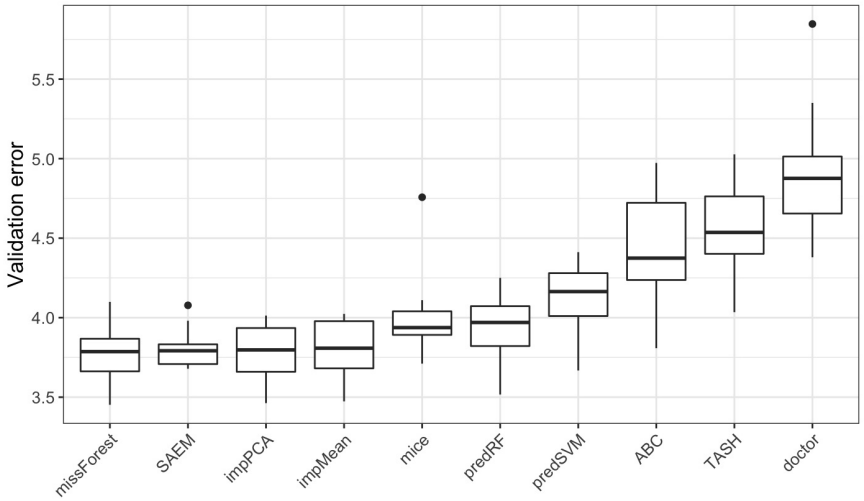

Figure 10: Empirical distribution of prediction errors of different methods over 15 replications for the TraumaBase data.

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. Table 7 in Appendix A.7 gives the details with classical measures (AUC, sensitivity, specificity, accuracy and precision) to compare the predictive performance of the methods. Our 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.

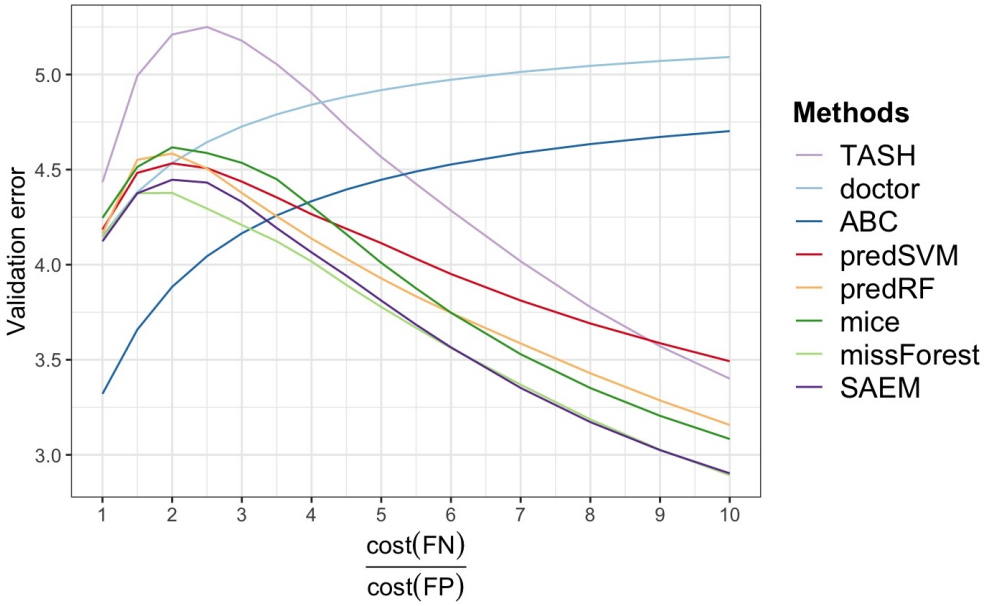

Figure 11: Average prediction errors of different methods, as function of the cost importance $\{\frac{w_0}{w_1} \mid \frac{w_0}{w_1} > 1\}$, over 15 replications for the TraumaBase data.

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.

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 our approach has the ability to outperform existing popular methods dealing with missing data.

8 Discussion

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. In view of the excellent results on the TraumaBase, emergency doctors want to implement our methodology in real time to make a prospective study with missing data.

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 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 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*.

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 (Little and Rubin, 2002). 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 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.

Input: An initial samples $x_{i,\text{mis}}^{(0)} \sim g(x_{i,\text{mis}})$;

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

Output: $(x_{i,\text{mis}}^{(m)}, 1 \leq i \leq n, 1 \leq m \leq M)$.

A.3 Calculation of observed information matrix

Procedure 2 shows how we calculate the observed information matrix.

A.4 Behavior of SAEM: convergence plots for all betas

Figure 12 shows the convergence plot for all the β in one simulation.

Procedure 2 Calculation of observed information matrix.

Input: After drawing MH samples $(x_{i,\text{mis}}^{(m)}, 1 \le i \le n, 1 \le m \le M)$ for unobserved data $(x_{i,\text{mis}}, 1 \le i \le n)$, we have imputed observations, noted as $(z_i^{(m)}, 1 \le i \le n, 1 \le m \le M)$, where $z_{ij}^{(m)} = \begin{cases} x_{i,\text{obs}}, & \text{if } x_{ij} \text{ is observed;} \\ x_{i,\text{mis}}^{(m)}, & \text{if } x_{ij} \text{ is missing.} \end{cases}$.

for $n = 1, 2, \cdots, n$ **do**

for $m = 1, 2, \cdots, M$ **do**

Calculate the gradient:

$\nabla f_{im} = \frac{\partial \mathcal{LL}(\theta; x_{i,\text{obs}}, x_{i,\text{mis}}^{(m)}, y_i)}{\partial \beta} = z_i^{(m)} \left(y_i - \frac{\exp(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j z_{ij}^{(m)})}{1 + \exp(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j z_{ij}^{(m)})} \right)$;

Calculate the Hessian matrix:

$H_{im} = \frac{\partial^2 \mathcal{LL}(\theta; x_{i,\text{obs}}, x_{i,\text{mis}}^{(m)}, y_i)}{\partial \beta \partial \beta^T} = -z_i^{(m)} z_i^{(m)T} \frac{\exp(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j z_{ij}^{(m)})}{\left(1 + \exp(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j z_{ij}^{(m)})\right)^2}$;

$\Delta_i \leftarrow \frac{1}{m}[(m-1)\Delta_i + \nabla f_{im}]$;

$D_i \leftarrow \frac{1}{m}[(m-1)D_i + H_{im}]$;

$G_i \leftarrow \frac{1}{m}[(m-1)G_i + \nabla f_{im} \nabla f_{im}^T]$;

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: $\hat{\mathcal{I}}_M(\hat{\beta})$.

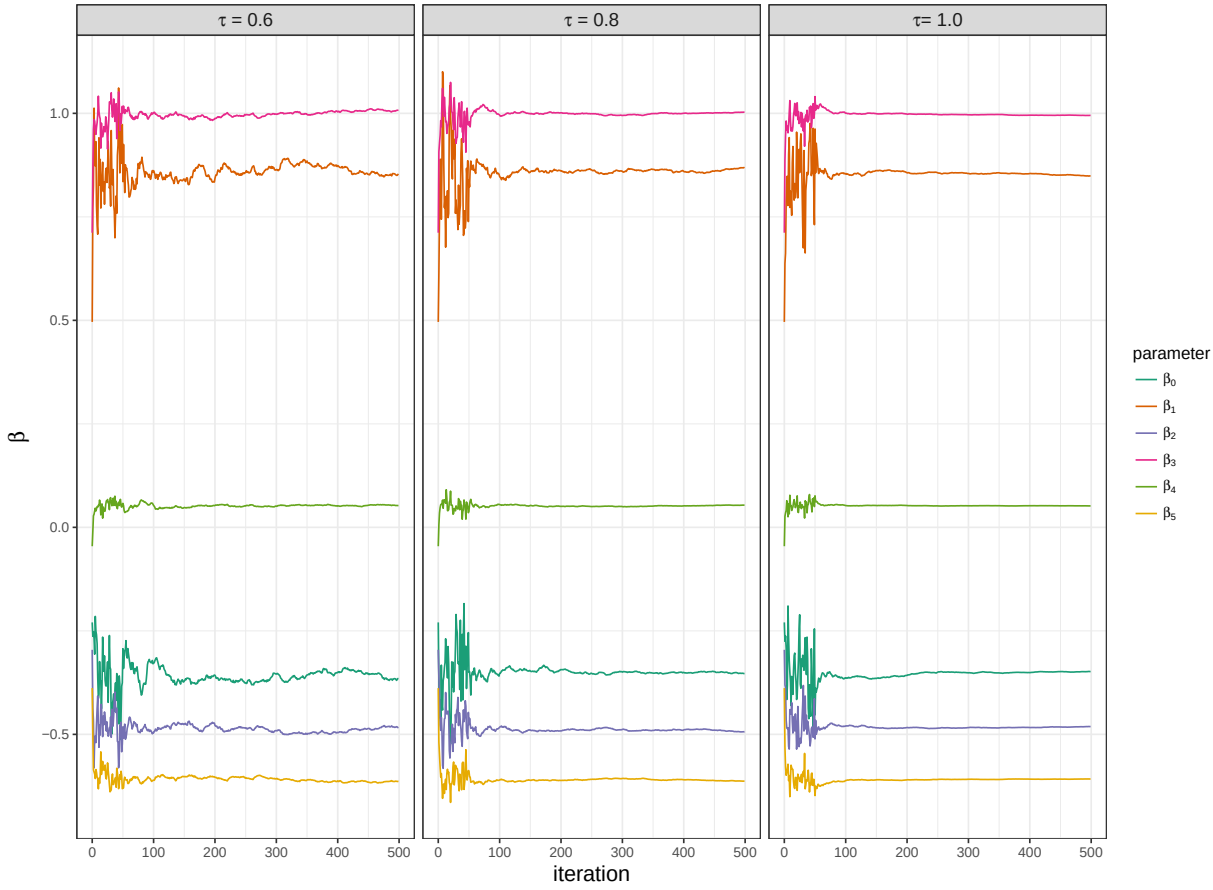

Figure 12: Convergence plots for all β obtained with three different values of τ (0.6, 0.8, 1.0). Each color represents one parameter.

A.5 Simulation results of comparison with MCEM

We generated a small sample with $n = 200$ in order to illustrate the performance of MCEM, which is computationally intensive. The bias and standard error of estimates over 100 simulations are shown in Figure 13.

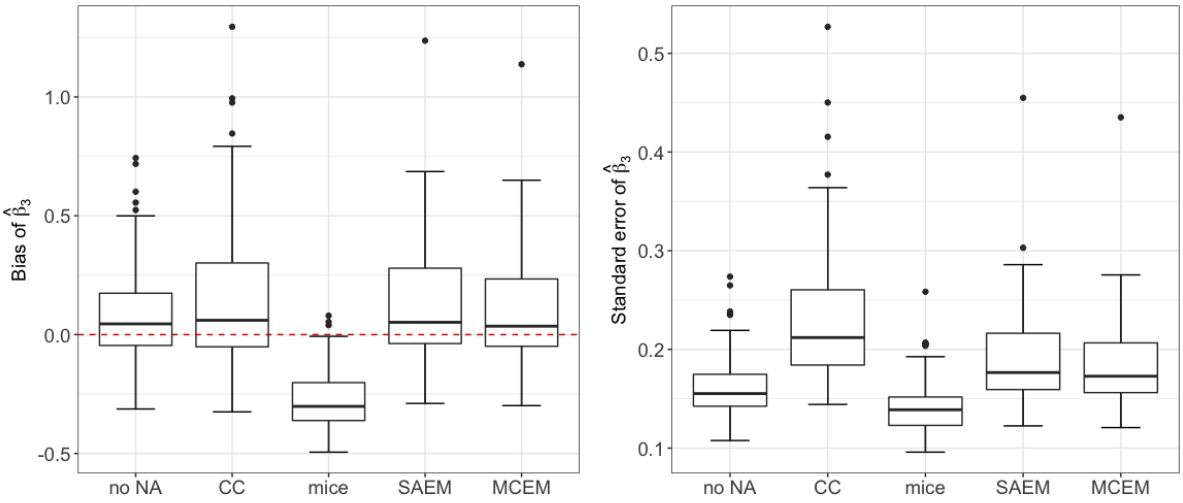

Figure 13: Empirical distribution of the bias and standard error of $\hat{\beta}_3$ obtained over 100 simulations, under MCAR, with $n = 200$ and 10% of missing values, with methods no NA, CC, mice, SAEM and MCEM.

Table 6: Coverage (%) for $n = 200$, correlation C and 10% MCAR, calculated over 100 simulations. Bold indicates under coverage. Inside the parentheses is the average length of corresponding confidence interval over 100 simulations.

parameter	no NA	CC	mice	SAEM	MCEM
β_0	96 (1.61)	96 (2.20)	97 (1.50)	96 (1.73)	96 (1.71)
β_1	98 (1.44)	95 (1.98)	97 (1.40)	97 (1.70)	99 (1.67)
β_2	97 (0.72)	96 (0.98)	96 (0.69)	97 (0.84)	96 (0.82)
β_3	92 (0.63)	90 (0.90)	**46** (0.56)	89 (0.74)	89 (0.72)
β_4	92 (0.30)	96 (0.41)	95 (0.30)	93 (0.34)	92 (0.34)
β_5	94 (0.43)	94 (0.60)	**54** (0.38)	92 (0.50)	92 (0.49)

Table 6 presents the coverage if the confidence interval for all parameters over 100 simulations and inside the parentheses is the average length of corresponding confidence interval over 100 simulations.

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*: 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.

A.7 Details of predictive performance for TraumaBase data

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

Table 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**
Sensitivity	41.1	41.6	42.2	41.0	37.9	41.6	**44.0**
Specificity	**74.6**	74.3	73.2	**75.0**	**75.2**	71.5	66.0
Precision	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:
`https://CRAN.R-project.org/package=misaem`

Codes: Code to reproduce the experiments are provided in:
`https://github.com/wjiang94/miSAEM_logReg`.

References

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Claeskens, G. and Consentino, F. (2008). Variable selection with incomplete covariate data. *Biometrics*, 64:1062–9.

Consentino, F. and Claeskens, G. (2011). Missing covariates in logistic regression, estimation and distribution selection. *Statistical Modelling*, 11(2):159–183.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the em algorithm. *The Annals of Statistics*, 27(1):94–128.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

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

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 107–114.

Hamada, S. R., Gauss, T., Duchateau, F.-X., Truchot, J., Harrois, A., Raux, M., Duranteau, J., Mantz, J., and Paugam-Burtz, C. (2014). 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.

Hamada, S. R., Gauss, T., Pann, J., Dünser, M. W., Léone, M., and Duranteau, J. (2015). European trauma guideline compliance assessment: the etrauss study. *Critical care*, 19:423.

Hamada, S. R., Rosa, A., Gauss, T., Desclefs, J.-P., Raux, M., Harrois, A., Follin, A., Cook, F., Boutonnet, M., Attias, A., Ausset, S., Dhonneur, G., Langeron, O., Paugam-Burtz, C., Pirracchio, R., Riou, B., de St Maurice, G., Vigué, B., Rouquette, A., and

Duranteau, J. (2018). Development and validation of a pre-hospital "red flag" alert for activation of intra-hospital haemorrhage control response in blunt trauma. *Critical Care*, 22(1):113.

Hay, S. I. et al. (2017). 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.

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (1999). Monte carlo em for missing covariates in parametric regression models. *BIOMETRICS*, 55:591–596.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469):332–346.

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

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

Lavielle, M. (2014). *Mixed Effects Models for the Population Approach: Models, Tasks, Methods and Tools*. Chapman and Hall/CRC.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc.

Liu, Y., Wang, Y., Feng, Y., and Wall, M. M. (2016). Variable selection and prediction with incomplete high-dimensional data. *Ann. Appl. Stat.*, 10(1):418–450.

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

McLachlan, G. and Krishnan, T. (2008). *The EM algorithm and extensions*. Wiley series in probability and statistics. Wiley, Hoboken, NJ, 2. ed edition.

Meng, X.-L. and Rubin, D. B. (1991). Using em to obtain asymptotic variance-covariance matrices: The sem algorithm. *Journal of the American Statistical Association*, 86(416):899–909.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rubin, D. B. (2009). *Multiple Imputation for Nonresponse in Surveys*, volume 307. John Wiley & Sons.

Seaman, S., Galati, J., Jackson, D., and Carlin, J. (2013). What is meant by "missing at random"? *Statist. Sci.*, 28(2):257–268.

Stekhoven, D. J. and Buehlmann, P. (2012). Missforest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

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

Wei, G. C. G. and Tanner, M. A. (1990). 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.

Wood, A. M., White, I. R., and Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, 27(17):3227–3246.